Application for *Venia Legendi* in:

**Applied Microeconomics and Econometrics**

Submitted to:

**The Department of Management, Technology and Economics**
**ETH Zurich (Swiss Federal Institute of Technology)**

Mehdi Farsi, PhD
Senior Scientist
Centre for Energy Policy and Economics (CEPE)
ETH Zurich
Zürichbergstr. 18, CH-8032 Zurich

Zurich, October 2008
Introduction

The studies related to this habilitation have been conducted over an extended period of time starting from September 2002, when I began my research at ETH Zurich. My focus has been on empirical research using methodologies drawn from Microeconomic theory. The data used for virtually all the research included in this habilitation are in panel data form. The repeated observations from the individuals allow a better control for omitted variables that are, or can be approximated, as time-invariant characteristics. Econometric modeling of such unobserved heterogeneity and the analysis of those effects on the empirical results and their policy implications are the center of attention throughout these studies.

From the application point of view, the studies cover two distinctive areas of applied Microeconomics, namely:

1. Production theory particularly the firms’ performance and efficiency and the consequences of different organization and ownership types

2. Consumers’ choices: in particular those that entail a discrete (discontinuous) choice

The validity of empirical research in both areas is closely related to the modeling of unobserved or omitted variables. In the case of production especially in strongly heterogeneous sectors such as network industries and health services, previous literature suggests that simplified models can provide misleading results. For instance, in the estimation of cost-efficiency for individual companies, a model that does not include the unobserved external factors might confound the external complexity of the production environment with the productive inefficiency. Noting that the external factors are often likely to be time-invariant, including individual effects in panel data models can improve the estimations. Such models can be readily extended to include richer structures for the stochastic terms. For instance, random coefficients can be used to specify various technologies. In fact, companies operating within the same industry might have different production technologies not only because they might use different methods of production, but also to accommodate their specific networks and environments. Such differences are hardly if at all measurable. For instance the complexity of an electricity network cannot be easily reduced to a single variable. Another origin for unobserved heterogeneity across companies is the fact that our output and input measures are all aggregate. The aggregation of various outputs in a single measure implies that certain potentially varying characteristics are accounted in the same way, thus unobserved variations.

Similarly, considering the heterogeneity is crucial in the analysis of consumer choices, especially because individuals show considerable differences in their preferences. The attitude to risks and tastes therefore the utility functions might vary from one individual to the other. Making policy decisions based on simplistic assumption of uniformity of preferences, based on the estimates for a ‘typical consumer’ could cause substantial errors. This is especially important for issue entailing discrete choice problems. In these cases, because of the strong non-linearity of the model, the parameter estimates might be strongly biased if they are mis-specified. The literature on discrete choice models applied to panel data provides a rich variety of
models with mixed effects and latent classes to study the heterogeneity among the consumers. In particular, these models can provide a basis for tailoring policy programs to specific types of consumers. The data requirement for these models is in general high. Namely, in order to enrich the model regarding the stochastic structure, one might need to use either large data sets if available, otherwise more parsimonious models with regard to explanatory variables. Considering this trade-off is crucial for adequate econometric modeling.

In the remaining part of this report, I provide a list of the refereed articles included in this habilitation, separated according to the two research area mentioned above. Next, I discuss the importance of these studies and describe my contribution in each case. Then, focusing on the main research objectives that I followed during the past six years, I provide a brief description of each paper. I will conclude with comments on further research and my future interests.

**Included articles**

**Articles related to Production Theory:**

Solo papers:


Joint papers:


**Articles related to Consumers’ Choices:**

Solo paper:


Joint papers:


Contributions related to the production theory

Free markets and competition are often considered as a solution to policy concerns about firms’ productive efficiency. However, the economists recognize numerous cases of public services, in which a competitive market not only fail to bring efficiency, but could bear undesired consequences and additional costs for the society. Such failures in the market are associated with a variety of reasons. In particular, promoting competition might increase the total production costs and lower the investments in natural monopolies such as network industries, and/or increase long-term and external costs due to lower quality of service in the sectors that are characterized by strong asymmetry of information such as the health sector.

Lacking the high-powered market incentives these sectors are often scrutinized in public policy debates and occasionally subject to various re-organization and privatization reforms. An often cited theoretical result is that private and for-profit firms have a better financial performance than those in the public and non-profit sectors. However the potential trade-off between quality and efficiency is a main concern especially in the health care sector. A common perception based on economic theory is that private ownership induces a better cost-efficiency while the non-profit sector has a better quality of service. The empirical evidence in both cases is rather mixed. The main problem in detecting quality or efficiency differences between organization types is that those differences are likely to be masked by the unobserved heterogeneity across firms: Namely, health providers face various patient mixes or electricity utilities operate in strongly heterogeneous networks. Using elaborate panel data models part of these unobserved factors can be modeled through firm-specific stochastic effects.

The econometric modeling of such unobserved factors in order to distinguish performance differences across various types of organizations is the main focus of my research in the production theory. In addition, the measurement of quality and cost-efficiency are both contentious issues, which requires elaborate econometric models. My two solo papers in production section (papers 1 and 2) deal with the measurement of cost-efficiency and quality respectively. In the first paper, I develop a mixed effects model to estimate the temporal variations in efficiency in Swiss general hospitals. Considering temporal variations allowed me to reach a better distinction between unobserved external factors and efficiency changes. In the second paper, I use the patient-level discharge data from California hospitals to study the effect of ownership changes on several outcome measures of quality. Ownership conversions are events that can be used to detect the pure quality differences due to ownership from external factors that do not change along with the conversion. The main quality measure that I used was based on inpatient mortality outcomes. In another studied (paper 10) I extended this measure to post-discharge events using a survival model with four states and three transition rates developed by my co-author.
In several other papers, I addressed the issue of efficiency differences across organization types in other contexts and using different methodologies. In papers 4 and 9, I analyzed the cost structure of Swiss hospitals in order to estimate the efficiency differences and the effect of subsidization and teaching activities on operating costs. In paper 12, my coauthors and I studied the impact of hospital take-overs by for-profit chains on nurse employment contracts. We showed that the effects of ownership changes on quality of care could be indirect and complex. In paper 14 and 17, we studied the cost structure of nursing homes in Switzerland in order to assess their efficiency and economies of scale, and also to detect the potential differences between public and private nursing homes.

Another related topic that I covered in my research (papers 3, 5, and 16) was the estimation of the economies of scope and scale in public transport and energy distribution sectors. The actual trends in regulatory reforms in these sectors are toward privatization which is generally facilitated by division and unbundling. The assessment of potential economies of scale and scope is crucial for an optimal organization of these industries. Here again, the econometric problem is the unobserved heterogeneity across different networks and technologies.

I had a major contribution in several studies (papers 6, 8, 10, 13 and 15) aimed at refining the conventional methodologies used for the estimation of productive efficiency, commonly referred to as ‘benchmarking.’ Our focus was on stochastic frontier models. In these studies we applied the recently developed frontier models to several industries especially those that are characterized by strong heterogeneity. I also developed a new method for reconciling the solution for heterogeneity bias in the estimated coefficients of the cost function (due to correlation between individual effects and explanatory variables) with that for avoiding biases in efficiency estimates due to conventional fixed-effects models in stochastic frontier analysis. I have elaborated this method in paper 10.

**Contributions related to the consumers’ choices**

My main contribution in this section, particularly in the discrete choice modeling of stated preferences is the development of an econometric approach that extends the linear random utility model to account for the consumer’s risk-aversion. Such aversion to risk and uncertainty is especially important in the case of new commodities and innovations whose benefits are unknown to the consumers. In my single paper (paper 18) I present the model and apply it to the case of energy efficiency in residential buildings. Using stated preference data from Switzerland I propose a model to quantify the extent to which the consumers discount the benefits of energy efficiency because of risks and illustrate the crucial impacts of risk-aversion in explaining the so-called ‘energy efficiency gap.’

In two other papers (19 and 20), I used several discrete choice models to analyze the data from two experiments in Switzerland. In these papers the unobserved differences across individuals have been considered in the adopted econometric models. In another study (paper 21), we analyze a relevant consumer problem in India namely, choice of energy source for household’s cooking consumption. I modeled the consumers’ decision making process as a discrete choice model with ordered response.
In the paper we assess the effects of different characteristics on the fuel choice and deduce policy conclusions about the promotion of less polluting alternatives. The last paper in this section (paper 22), my co-author and I analyze the data from a survey of post-surgical subjective pain assessment. Though not exactly a consumer choice problem, the methodology is closely related to discrete choice models. In that paper, we showed that accounting for the unobserved characteristics of the patients, could change the estimation results. We the proposed model we estimated a temporal function for analgesic pain relief and showed the effectiveness of preventive relief.

Future research

Based on the set of studies that I presented here, I am convinced that there is much needed research for the following areas:

- Identification of cost efficiency in stochastic frontier models, that is, disentangling external unobserved factors using econometric techniques as well as qualified assessment about the patterns and types. I contend that we could gain substantially if we focus on temporal variations instead of levels. I argue that, similar to the literature on productivity that focuses on growth, the growth in cost efficiency has a great policy importance. Growth has also a practical advantage in that it is relatively easier to identify. As a first step I plan to use my proposed approach (paper 1) in other data.

- Another dimension that I plan to study is the possibility of adapting benchmarking techniques to classifying firms into distinctive groups based on their performance (e.g.: efficient, average, …). Latent class models can be used in this regard. I am using these models in two other ongoing studies (not listed here). In the future I plan to modify the latent class frontier model to incorporate a single technology with several classes of efficiency.

- In the estimation of the economies of scale and optimal scale, I think the approximation point in models with translog form could have an impact on the results. I plan to study this effect using data, preferably from transport or energy sector, and show that a sensible estimation of firm-specific economies of scale might give different results from conventional models. My contention is that in many cases conventional models with a unique approximation point are not appropriate for estimating the optimal scale.

- In the estimation of the economies of scope with quadratic functional forms, the theoretical condition of zero fixed costs (in long-run cost functions) is usually ignored. I believe that in certain cases (especially network industries) this problem can create substantial biases in the estimated economies of scope. In a future paper I plan to propose an econometric model with random coefficients that fully satisfy this theoretical condition and re-estimate the economies of scope in Swiss multi-utilities.

- On the effect of risk aversion on consumers’ and/or investors’ decisions about adopting new technologies (e.g. energy efficient systems), my research (paper 18) provides a good starting. The econometric model proposed in that
paper should be extended to include random effects for the unobserved heterogeneity across individuals. Like the basic model, this would require extensive software programming. I plan to undertake this research as soon as possible.

- The problem of energy demand and alternative fuel choices is a challenging econometric problem that needs much more empirical analysis. I believe that models that account for unobserved heterogeneity such as latent class models can be very useful. Studying the revealed preference data (e.g. Indian household survey data) with this approach that is an open mind to recognize heterogeneity of preferences, could bring some insight to the policy debates about clean energy and the so called ‘energy ladder’ theory.

Acknowledgements

The studies reported in this habilitation draw from several research projects sponsored by various organizations, which has been acknowledged in the corresponding papers. However, I wish to repeat my acknowledgement for the support of the Swiss National Science Foundation (Division I research grant 100012-108288) that favorably responded to my research proposal entitled ‘Measurement of Cost Efficiency in the Presence of Unobserved Heterogeneity.’ That support was crucial for many studies included in this habilitation. Many people have helped me with their advice and suggestions. I particularly wish to thank Martin Jakob, Aurelio Fetz and Michael Künzle for many insightful discussions during their PhD research. I am also grateful to Silvia Banfi, Janet Currie, Bentley MacLeod, Geert Ridder, William Greene, Lester Hunt, Subal Kumbhakar, Robin Sickles, Ricardo Scarpa, Anna Alberini and Leopold Simar for sharing their insights with me in several occasions. This habilitation would not have been possible without Massimo Filippini not only because of his invaluable support and constant encouragements, but also for numerous opportunities for discussion and scientific exchange. Finally I wish to thank all the CEPE staff including part-time student assistants and doctoral students, for the friendly environment they have made possible.
Mehdi Farsi
Poststrasse 10, Wetzikon 8620
Phone: +41 43 333 0361
E-mail: mfarsi@ethz.ch

Address: CEPE, D-MTEC, ETH Zurich
Zürichbergstrasse 18, Zurich, CH-8032
E-mail: mfarsi@ethz.ch
Phone: +41 44 632 0656 (Office)
+41 77 250 5499 (Mobile)
Fax: +41 44 632 1050
http://elbanet.ethz.ch/wikifarm/mfarsi

Present Position: Senior Scientist and Lecturer (permanent position)
Centre for Energy Policy and Economics (CEPE),
Department of Management Technology and Economics (D-MTEC),
ETH Zurich (Swiss Federal Institute of Technology)
Zurich, Switzerland

RESEARCH INTERESTS
Applied Microeconomics; Applied Econometrics;
Industrial Economics; Health Economics; Energy Economics

TEACHING INTERESTS
Econometrics; Statistics; Quantitative Methods;
Microeconomics; Energy Economics

EDUCATION

**PhD**, Economics, University of Southern California, Los Angeles, 2002
*GPA*: 3.96 (4.00 basis)
*Dissertation*: “Essays on Organizational Forms and Performance in California Hospitals”
*Committee*: Janet Currie, Bentley MacLeod, Geert Ridder, Elizabeth Graddy

**MA**, Economics, University of Southern California, Los Angeles, 1999
*GPA*: 3.90 (4.00 basis)

Doctoral Program, Civil Engineering, Ecole Nationale des Ponts et Chaussées,
**DEA**, Civil Engineering, Ecole Centrale de Lyon, France, 1994
With distinction: *Très Bien
GPA*: 16.36  (20.00 basis); *Rank*: 1st among 31)

**MSc**, Soil Mechanics, University of Tehran, Iran, 1991
*GPA*: 3.73  (4.00 basis)

**BSc**, Civil Engineering, University of Tehran, Iran, 1988
*GPA*: 3.51  (4.00 basis)

**WORK EXPERIENCE**

Since 9/2002  
Senior Scientist and Lecturer, D-MTEC, ETH Zurich; and Department of Economics, University of Lugano, Switzerland.

9/1997 - 6/2002  
Research and Teaching Assistant, Department of Economics, University of Southern California, Los Angeles.

Research Assistant, Department of Economics, University of California, Los Angeles.

1/1997 - 8/1997  
Research Assistant, Department of Civil Engineering, University of Southern California, Los Angeles.

Research Engineer, Laboratoire Régional des Ponts et Chaussées, Aix-en-Provence, France.

Structural Engineer, Tehran Metro Company, Tehran, Iran.

**PUBLICATIONS**

**Forthcoming:**


**2008:**


Filippini, Massimo and Mehdi Farsi (2008): Cost Efficiency and Scope Economies in Multi-output Utilities in Switzerland, Strukturberichterstattung Nr. 39, State Secretariat of Economic Affairs (SECO), Bern, Switzerland.

2007:


2006:


2005:


2004:


Prior to 2003:


**WORKING PAPERS**

Farsi, Mehdi: Risk-Aversion and Willingness to Pay for Energy Efficient Systems in Rental Apartments, July 2008; last version submitted for review and publication to *Environmental and Resource Economics*.


**TEACHING EXPERIENCE**

Fall Semesters (regularly, since 2005): Statistics for Business and Economics, Master program in Management Technology and Economics, ETH Zurich.


Fall Semester 2008: Doctoral Seminar in Empirical Methods in Environmental Economics, ETH Zurich, joint with Massimo Filippini and Anna Alberini.


Spring Semester 2006: Managerial Economics, Master Program in Management and Economics, Economics Department, University of Lugano, joint with Massimo Filippini.


Spring Semester 2003 and 2004: Doctoral Seminar in Industrial Economics, ETH Zurich, joint with Massimo Filippini.

1997- 2002, Teaching Assistant, University of Southern California:

**Graduate Courses:**
- Practice of Econometrics, Spring 2002, Instructor: Isabelle Perrigne
- Economic and Financial Time Series, Fall 2001, Instructor: Cheng Hsiao
- Economic and Financial Time Series, Fall 2000, Instructor: Roger Moon
- Microeconomic Theory II, Spring 2000, Instructor: Michael Magill
- Microeconomic Theory I, Fall 1999, Instructor: Bentley MacLeod
- Microeconomic Analysis and Policy, Fall 1998, Instructor: Harrison Cheng

**Undergraduate Courses:**
- International Finance, Spring 1999, Instructor: Caroline Betts
- Principles of Macroeconomics, Spring 1998, Instructor: Pablo Neumeyer
- Introduction to Statistics, Spring 1998, Instructor: Isabelle Perrigne
- Introduction to Statistics, Fall 1997, Instructor: Quang Vuong
- Soil Mechanics Laboratory, Spring 1997, Instructor: Jean-Pierre Bardet
Journal Referee Experience

- Energy Journal
- Energy Economics
- Applied Economics
- Journal of Economic Studies
- Journal of Health Economics
- Health Economics
- Journal of Productivity Analysis
- Southern Economic Journal
- Annals of Public and Cooperative Economics
- Journal of Business Economics and Statistics
- Journal of Policy Analysis and Management
- Journal of Transport Economics and Policy
- Swiss Journal of Economics and Statistics
- European Journal of Health Economics
- Health Care Management Science

AWARDS

- 2006 and 2007: Cited in *Who's Who in the World*
- 2001-02, Graduate School PhD Dissertation Grant, University of Southern California
- 2000-01, NBER Dissertation Fellowship for research on the Economics of the Non-Profit Sector, National Bureau of Economic Research

LANGUAGES

- English: Fluent
- French: Fluent
- German: Basic Knowledge
- Persian: Native Language

Computer skills: Stata, SAS, MATLAB, NLOGIT, Gauss, C++

PERSONAL INFORMATION

- Born in Kerman, Iran, April 1965; Citizenship: Iran
- Marital status: steady partnership
- Hobbies: Piano, Reading Philosophy and History, Swimming

REFERENCES

- Massimo Filippini, Professor of Economics, ETH Zurich and University of Lugano
  Via Maderno 24, Lugano, CH-6900
  Phone: (41) 91 912 4640
  E-mail: mfilippini@ethz.ch

- W. Bentley MacLeod, Professor of Economics, Columbia University
  420 West 118th Street, Room 1038 IAB, New York, NY 10027,
  Phone: (1) 212-854-4520
  E-mail: wbm2103@columbia.edu

- Janet Currie, Professor of Economics, Columbia University
  420 West 118th Street, Room 1038 IAB, New York, NY 10027,
  Phone: (1) 212-854-4520
  E-mail: jc2663@columbia.edu

- Geert Ridder, Professor of Economics, University of Southern California
  Los Angeles, CA 90089-0253
  Phone: (1) 213-740-3511
  E-mail: ridder@usc.edu
Appendix 1

Articles related to Production Theory


The Temporal Variation of Cost-efficiency in Switzerland’s Hospitals: An Application of Mixed Models

Mehdi Farsi

Department of Management, Technology and Economics,
ETH Zurich, Zürichbergstr. 18, Zurich 8032, Switzerland
and
Department of Economics, University of Lugano, 6900 Lugano, Switzerland

April 2008

ABSTRACT

This paper uses a mixed effects model to examine the temporal variation of cost efficiency in Switzerland’s general hospitals. The variations in total costs, the number of empty beds and the length of hospital stays are analyzed using financial data from a sample of 168 hospitals operating from 1998 to 2003, as well as hospitalization records disaggregated to Diagnosis Related Groups. Individual intercepts and random coefficients are used to account for the unobserved time-invariant heterogeneity and the differences in temporal patterns across hospitals and DRG categories. The analysis illustrates the usefulness of mixed models to account for unobserved factors such as quality, with a relatively weak assumption that their temporal variations, rather than their initial levels, be uncorrelated with efficiency changes. The results indicate that hospitals have adopted measures to curtail hospitalizations and reduce empty beds. The extent and effectiveness of these measures vary significantly across individual hospitals. However, there is no evidence in favor of a particular ownership type or subsidization regime. While the link between reduction rates of empty beds and gains in cost-efficiency is statistically significant, the expected association between shortening hospital stays and cost-efficiency cannot be clearly established in the data.

Keywords: general hospitals, stochastic frontier, cost efficiency, mixed models, random coefficients
1. Introduction

The increasing growth of health care costs in Switzerland has raised public concern for containing the hospitalization costs. Starting from 1994, together with the introduction of the mandatory federal insurance law and its implementation in 1996, the Swiss legislators have provided the cantonal authorities with several discretionary measures to control hospitals’ operating costs. Among these measures was the gradual implementation of a prospective reimbursement system based on Diagnosis Related Groups (DRG).1

Thus far, the implementation of DRG-based payment system has been mainly limited to specific services such as ambulatory visits and over-night hospitalizations. Aware of the ongoing reforms, hospital managers are increasingly engaged in the economical planning of their hospitalizations. In particular, the mandatory DRG coding requirement for all hospitalizations introduced in 1998 can be considered as a preface to cost saving pressures. Policy debates reflect a common perception that certain types of hospitals do not have strong incentive for a substantial improvement in their efficiency. Small local hospitals, non-profit providers and university hospitals have often been singled out as inefficient providers.

Several studies tried to detect the efficiency differences across different ownership and organization types (cf. Farsi and Filippini, 2006, 2007; Steinmann and Zweifel, 2003). The main difficulty of such analyses is that the efficiency differences among hospital types might be biased by the potential cost effects of unobserved exogenous factors. However, the required simplifying assumption that the unobserved

1 DRG is a system of classification based on the required hospital resources. DRG codes are assigned by patented computer programs using information on diagnoses, complications/comorbidities and procedures as well as patient’s age and gender. DRGs have been first used by Medicare (the US health insurance program for the elderly) in its case-based reimbursement rules, known as Prospective Payment System. Hospitalization costs of each DRG are usually estimated by statistical analysis of large samples of similar cases. In Switzerland, this information is provided by ‘APDRG Suisse,’ a non-profit association comprising of DRG users throughout the country.
heterogeneity is uncorrelated with efficiency differences, has received little attention. This assumption is particularly debatable if important factors such as quality and/or case mix are not completely observed.

In this paper focusing on growth rates instead of levels of efficiency, I get around the problem of unobserved heterogeneity to the extent that the temporal variation of omitted variables is uncorrelated with efficiency changes. This is a relatively weak assumption in that it allows correlation between heterogeneity and initial values of efficiency. Assuming that the hospitals have undertaken cost-saving measures, I use a mixed effects model to estimate the evolution of cost-efficiency over the “reform period” starting from 1998. Rather than searching for a reliable estimate of a specific hospital’s efficiency at a given period, the focus is upon hospital-specific rates of change in cost-efficiency and their differences across hospital types. Moreover, I analyze the relationship of efficiency changes with observed decreases in empty beds and length of stays. Such analyses can provide some insight on the overall effectiveness of the cost-saving measures adopted by the hospitals and their eventual impact on quality of service.

The data are based on a relatively rich panel of 168 general hospitals operating from 1998 to 2003 and about 108,000 records of the average length of hospitalization of patients with similar DRG’s. The econometric specification is based on a special version of the general parametric framework proposed by Sickles (2005), or the mixed effects model proposed by Kneip et al. (2003), combining individual hospital and DRG fixed effects with random coefficients of the time variables. The adopted model can also be considered as an extension of the random effects model proposed by Cornwell et al. (1990).
While pointing to significant efficiency differences among hospitals regarding their cost-reduction efforts, the results do not provide any evidence in favor of a particular hospital ownership type or subsidization status. The analysis in general indicates that hospitals with relatively important cuts in their empty beds are likely to have relatively high efficiency gains. The evidence regarding the hospitalization length is not conclusive. In most cases, the cost reductions often expected from shortening hospital stays do not appear to be significant.

The rest of the paper is organized as follows. Section 2 provides a critical discussion of the methods of efficiency estimation and justifies the adopted methodology used in this paper. The econometric specification and the explanatory variables are described in Section 3. Section 4 summarizes the data and provides the descriptive statistics of the main variables included in the models. Section 5 presents and analyzes the estimation results and Section 6 concludes the paper with summary of main results and policy implications.

2. Methods

The estimation of firm-specific efficiency is a contentious topic that has been subject of a great body of literature with a variety of econometric models commonly referred to as Stochastic Frontier Analysis. The application of these models to hospitals has been questioned by several authors (Newhouse, 1994; Skinner, 1994; Street, 2003; Folland and Hofler, 2001). The main criticism lies on the aggregation of a myriad of services provided by a hospital into a few output measures, required by any practically manageable multi-output cost function.

Despite these general criticisms the efficiency analysis in health care sector remains commonplace (Hollingsworth and Street, 2006; Worthington, 2004; Rowenta
While admitting the limitations of their approach many authors have adopted various measures for accounting for output characteristics such as case mix severity indexes and other distinctive hospital characteristics (Zuckerman et al., 1994; Linna, 1998; Rosko, 2001; Deily and McKay, 2006; Brown, 2003). Other studies have used econometric modeling strategies that have proved more robust in presence of such heterogeneities (Liu et al., 2007; Bradford et al., 2001) or panel data models that account for unobserved factors through hospital-specific stochastic terms (Farsi and Filippini, 2007).

The frontier literature is especially rich in panel data modeling approaches with a variety of underlying assumptions about temporal variations of efficiency. While models such as Kumbhakar (1990) and Battese and Coelli (1992) assume a uniform variation for all the firms, others such as Kumbhakar (1991), Polachek and Yoon (1996) and more recently Greene (2005) allow for stochastic variation without any correlation over time. The latter models include three stochastic components respectively for efficiency, random noise and time-invariant heterogeneity.

While recognizing that the firm’s efficiency can considerably change over time, a fully stochastic variation over time implies an idiosyncratic nature for the temporal changes. This is probably a too flexible assumption that ignores the fact that efficiency changes are driven by an underlying learning process specific to the firm’s management and their efforts. As Alvarez and Schmidt (2006) point out, even though the randomness appears to be quite important, ‘over longer periods of time, skill persists while luck averages away.’ Even assuming that firms constantly face new technology shocks and market developments that make their resulting productive efficiency look like a stochastic variable, an independent identical time distribution is unrealistic.
As Sickles (2005) points out, in many cases the parametric assumptions help to have a better interpretation of the results. Therefore, a reasonable assumption would be to assign a deterministic functional form for the temporal variation of firm’s efficiency while allowing for changes in the values of the parameters across individual companies. This is the approach adopted by Cornwell et al. (1990) through a quadratic function and Lee and Schmidt (1993) with a linear function both with random coefficients that vary across firms. The functional form and the variation of the individual effects have been later extended to mixed effects models and semi-parametric models respectively by Kneip et al. (2003) and Sickles (2005). These models reconcile the idea of heterogeneity with the need for imposing a time structure upon efficiency changes.

Sickles (2005) provides a general framework for the treatment of time-varying efficiency. He recognizes the vulnerability of efficiency and productivity measures as estimation residuals and ‘reduced form’ concepts that are inevitably based on ad hoc econometric specifications. With a series of Monte Carlo simulations and applying several alternative specifications, the author highlights the difficulties in identifying firm-specific and time-varying efficiency. Sickles (2005) asserts that the robustness, flexibility and precision are the most ‘important distinguishing features’ that should be considered in model specification strategies.

Lack of robustness can be due the reliance of many frontier models on non-testable distribution assumptions often required to distinguish random noise from the efficiency term. For instance, the original frontier approach in cross-sectional data (Aigner et al., 1977; Meeusen et al., 1977) assigns a half-normal distribution to efficiency and a normal distribution to random noise. This model as well as many later extensions relies upon skewed residuals to produce any meaningful values for
efficiency estimates. In many cases, one of the stochastic components might easily
degenerate to zero because of a misspecification of the explanatory variables. This
sensitivity can be exacerbated in panel data models that decompose the residuals into
three components instead of two.

Robustness can be achieved by relaxing the assumptions on the distribution
and correlation structures, usually at a loss of precision or identification. For instance,
considering freely distributed fixed effects instead of random effects allows more
realistic assumptions about the potential correlation between the individual effects and
the explanatory variables. However the fixed effects capture the unobserved time-
invariant factors, which if correlated with efficiency, distort the pattern of efficiency
differences among the companies. In these cases, the potential estimation bias in the
overall efficiency can be anticipated depending on the model. However, assessing the
resulting biases for individual companies is a matter of pure speculation. Therefore,
using fixed effects requires an assumption about the correlation of individual effects
not with explanatory variables, but with efficiency differences.

In this paper, recognizing that time-invariant differences in efficiency are
captured by the fixed effects, thus indistinguishable from the remaining unobserved
heterogeneity, the fixed effects are used to ensure an unbiased estimation of temporal
changes in efficiency to the extent that they are uncorrelated with temporal changes in
other unobserved factors. Therefore, the proposed model combines a fixed effects
approach for intercepts with random effects for time variables representing various

---

2 The overall inefficiency is usually over-stated should the fixed effects be interpreted as inefficiency as
in Schmidt and Sickles (1984), and understated if they are considered as external factors unrelated to
how the efficiency differences could reach implausible levels in the former case. As for the latter cases,
where inefficiency is identified as an additional skewed stochastic term, this author’s experience
suggests that the available algorithms have a high risk of producing unreliable estimates of the fixed
effects. Farsi et al. (2005) propose a solution around the incidental parameters problem by combining
Mundlak’s (1978) specification to Greene’s (2005) random effects model.
temporal patterns across individuals. A formal description of this specification will be presented in the next section.

Another important issue in the estimation of productive efficiency is the study of the sources of inefficiency. The reduced form of the frontier model does not allow in itself an understanding of the inefficiency sources. As Sickles (2005) elegantly points out, a ‘strong institutional understanding of the industry under study’ is required to choose an adequate estimator among the available alternatives that satisfy the generic properties. Given the existing discrepancy and sensitivity issues in the frontier methodology (as discussed earlier), most studies face a recurrent question regarding the validity and reliability of efficiency estimates, namely, whether these estimates are artifact of sampling variations. A common approach is to explore the statistical association between efficiency estimates and the potential sources of inefficiency or to directly integrate such relationships into the frontier model. This approach is however plagued by the possible correlations with third-party unobserved factors such as quality that could bias the results. I argue that the effects of such correlations are attenuated when the relationships are explored between the growth/reduction rates instead of levels. In fact, focusing on temporal changes allows us to reduce the heterogeneity bias due to correlation with time-invariant factors.

For instance, unnecessarily long hospitalizations might be a source of excess costs. This is surely a debatable issue that has been subject of a number of papers. For instance Carey (2000) provides evidence that the US hospitals, facing the policy concerns about rising costs, have reduced the lengths of hospitalizations. Her findings suggest however, that the extent of cost savings has been commonly overstated. Other

---

3 Sickles’ (2005) general framework can be applied with fixed effects for temporal changes as well, however at a considerable loss of the model’s degrees of freedom. For instance in a quadratic form for temporal variations would require 3 fixed parameters for each hospital, which might create a plausibility problem for short and medium panels.
studies suggested that curtailing the hospital stays has led to a deterioration of quality of care and might have a counter-productive effect in the long run. In Switzerland, there is a considerable variation in the average length of stay (LOS) among hospitals with the small local hospitals having significantly longer hospitalizations, suggesting possible inefficiencies (Farsi and Filippini, 2005). Another potential source of inefficiency in hospitals could be related to excess capacity. For instance Gaynor and Anderson (1995) estimate that in the US, the costs of empty hospital beds could amount to 9.5% of the total costs.

Partial efficiency measures based on changes in length of hospital stays and the number of empty beds could be helpful in understanding how the hospitals have dealt with those possible sources of inefficiency. In particular, measures based on LOS are less affected by aggregation bias for, unlike cost data, the records of hospital stays are generally available for individual patients. In this paper, in addition to hospital costs, I use the average LOS at the DRG level and average number of hospital’s empty beds. The statistical relationships between these measures are used to assess the differences in cost-cutting strategies across various hospital types.

3. Model specification

The measure of hospital’s cost efficiency is based on a total cost function with a Cobb-Douglas functional form. The two complementary measures are the excess capacity defined by the hospital’s average number of empty beds and a measure of hospital’s excessive LOS based on the average length of hospitalization. The working

---

4 The adopted cost function is similar to the specification used in Farsi and Filippini (2007), with the difference that here because of the presence of individual fixed effects, a number of variables that are time-invariant or practically stable over time are excluded. Similarly, the choice of Cobb-Douglas from as opposed to flexible forms such as translog is motivated by the trade-off between flexibility and the model’s degrees of freedom, especially restricted here because of fixed effects. Moreover, we do not impose the restriction of linear homogeneity in input prices, because as we see later in the data section, the included input prices do not cover all the input factors.
hypothesis is that the hospitals have adopted measures to contain their operating costs by improving their overall productive efficiency, by reducing their excess capacity, or by curtailing the hospital stays. Including individual fixed effects allows a straightforward identification of the temporal variation of each of the three variables without worrying about the unobserved hospital’s time-invariant characteristics\textsuperscript{5} and their potential correlation with the observed explanatory variables.

The downside is that those efficiency differences across hospitals that are stable over time are entirely captured by the fixed effects, thus inseparable from other time-invariant external heterogeneities. Therefore, any assessment of the hospitals’ efficiency in a given year (relative to other hospitals), is valid only to the extent that the hospitals do not differ significantly with respect to their initial efficiency before the reforms say in 1998. However, it should be noted that the analysis in this paper and the policy conclusions reported here, are strictly based on the temporal changes of efficiency, thus do not require any assumption on efficiency levels. Rather, the required assumption here is that the efficiency gains or losses be uncorrelated with the temporal changes of other unobserved factors such as hospital quality.

The explanatory variables for the cost function include two outputs namely, a DRG-adjusted number of hospitalizations, a measure of ambulatory services offered by the hospital, and three input factor prices \textit{i.e.}, labor price in two categories, non-physician employees and employed physicians, and capital price. The average LOS and the number of medical training positions (interns and medical students) have also been included as output characteristics. For the excess capacity the explanatory variables are specified as follows: the number of hospitalizations and the share of

\textsuperscript{5} These omitted characteristics could include hospital’s specialization level, quality of service, and also case mix severity to the extent that these factors depend on hospital location and long-term factors such as medical staff and reputation. Concerning the case mix it should be noted that the DRG adjustment (used in the specification) is only an imperfect measure of severity, thus the within-DRG variations across hospitals remain unobserved.
patients with private health insurance. The idea here is that hospitals should adjust the number of available beds according to the fluctuations in the demand and also to accommodate the patients entitled to private rooms.

The analysis of hospitalization lengths has been conducted at the DRG level. Namely, the dependent variable is the average LOS for the patients within a given DRG hospitalized in a given hospital-year. Individual fixed effects are considered for each hospital-DRG group. In addition to time variables, the total number of training positions has been included as explanatory variable. The findings in previous studies such as Rogowski and Newhouse (1992) and Simmer et al. (1991) suggest that hospitals with more teaching activities are likely to have longer hospitalizations. As shown by Martin and Smith (1996), the length of stay could depend on several patient characteristics that cannot be summarized in the DRG categories, thus remain among unobserved variables in the present analysis. Part of such variations should be captured by the hospital-DRG fixed effects.

The definition and the summary statistics of all the variables included in the models will be provided in the data section. Now we turn to the econometric specification: The cost model is based on a mixed effects model written as:

$$\ln C_{it} = \beta \ln X_{it} + \gamma Z_i + \rho t + \phi t^2 + \alpha_i + u_{it} + \epsilon_{it},$$

(1)

where \(i\) and \(t\) represent the hospital and year respectively with \(t=0\) representing the first year covered in the sample; \(C\) is the total costs; \(\ln X_{it}\) is the vector of time-varying explanatory variables expressed in logarithm; \(Z_i\) is a vector including all the hospital-specific characteristics that do not vary with time; and \([\beta, \gamma, \rho, \phi]\) is the vector of regression coefficients. The stochastic terms \(\alpha_i\) and \(\epsilon_{it}\) respectively represent the

---

6 In line with the Rogowski and Newhouse, we assume that this effect is a result of ‘indirect’ costs of training medical students rather than hospital’s inefficiency suggested by Simmer et al.
hospital’s individual effect and the random noise. Finally, $u_{it}$ is the inefficiency (here excess costs) of hospital $i$ at year $t$, specified as a quadratic function of time:

$$u_{it} = u_{i0} + \delta_i t + \theta_i t^2,$$

with $u_{i0}$ representing hospital $i$’s initial inefficiency at year $t=0$, $\delta_i$ and $\theta_i$ are random coefficients with a multivariate normal distribution, specified as:

$$\begin{pmatrix} \delta_i \\ \theta_i \end{pmatrix} \sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

(3)

where $(\sigma_1, \sigma_2, \sigma_{12})$ are the parameters to be estimated. The residual term $\epsilon_{it}$ is assumed to be normally distributed with zero mean: $N(0, \sigma_{\epsilon}^2)$ and the individual effects $\alpha_i$ are assumed to be constant fixed effects.

The mean values of the random coefficients $(\delta_i, \theta_i)$ have been set to zero. This is a simplifying assumption that allows the parameters $(\rho, \phi)$ to be identified, while recognizing that the hospital costs might follow a growth pattern that is not related to hospitals’ efficiency, but due to external factors, such as the general progress in medical treatments and pharmaceuticals that are increasingly more costly. Such temporal variations that are not captured by the explanatory variables included in the model, are assumed to be more or less similar for all hospitals, thus represented by the average growth in costs captured by parameter pair $(\rho, \phi)$.

Noting that because of the presence of the fixed effects the coefficient vector $\gamma$ cannot be identified, the model in Equations (1) and (2) can be easily transformed to a random-coefficient model on the deviations from hospital mean, written as:

$$\Delta_i \ln C_{it} = \beta \Delta_i \ln X_{it} + \rho \Delta_i t + \varphi \Delta_i t^2 + \delta_i \Delta_i t + \theta_i \Delta_i t^2 + \epsilon_{it},$$

(4)
where $\Delta x_{it}$ for a generic variable $x_{it}$ is defined as the deviation of the variable from its mean value ($\bar{x}_i$) within hospital $i$:

$$
\Delta x_{it} = x_{it} - \bar{x}_i; \quad \text{with} \quad \bar{x}_i = \frac{1}{T_i} \sum_{t=0}^{T_i} x_{it},
$$

and $T_i$ is the number of periods for hospital $i$.

As it can be seen the above specification does not allow a separate identification of the unobserved heterogeneity represented by fixed effects $\alpha_i$ and the initial inefficiencies denoted by $u_{it0}$. Both of these terms along with the time-invariant variables $Z_i$ (including the intercept) are canceled out in the within transformation. It is important to highlight that while being useful for an effective estimation of the temporal variations free from time-invariant heterogeneity, the fixed effects capture all the ‘between’ variation, namely the long-term and persistent differences across hospitals. Therefore, the marginal effects and elasticities estimated from this model are strictly driven from within-hospital variations that are generally of a transient short-term nature. The implication is that the estimated results can only be used to predict quantities or behaviors that entail a limited range of variation comparable to the within-hospital variations in the sample. This caveat is particularly important for technological characteristics of the production function such as returns to scale, that are best identified through long-term differences between hospitals with different scales of production.

The model used for the analysis of the hospitals’ excess capacity is similar to that described in Equation (4) with the difference that the dependent variable is the number of hospital’s empty beds (instead of total costs) and includes its own the explanatory variables $X$. Another difference is that unlike costs, there is no reason other than efficiency improvement that the excess capacity should uniformly grow or
decrease among all hospitals. Therefore it is reasonable to relax the zero-mean assumption for the individual random coefficients. Thus, the resulting specification can be written as:

\[
\Delta_i \ln E_{it} = \beta^e X_{it} + \delta^e \Delta_i t + \theta^e \Delta_j t^2 + \epsilon^e_{it},
\]

where \( E \) is the number of empty beds and superscript \( e \) denotes parameters related to excess capacity; and the random coefficients are specified as:

\[
\begin{pmatrix}
\delta^e_i \\
\theta^e_i
\end{pmatrix}
\sim N(\mu^e, \Sigma^e); \quad \text{with} \quad \mu^e = \begin{pmatrix}
\rho^e \\
\phi^e
\end{pmatrix}, \quad \Sigma^e = \begin{pmatrix}
(\sigma^e_{11})^2 & \sigma^e_{12} \\
\sigma^e_{12} & (\sigma^e_{22})^2
\end{pmatrix}.
\]

(7)

The analysis of hospitals’ average length of stay (LOS) has been conducted at DRG level observations. Denoting DRG group by subscript \( j \), the model specification for this analysis can be formulated as:

\[
\Delta_{ij} \ln L_{ijt} = \beta^l X_{ijt} + \delta^l_{ij} \Delta_j t + \theta^l_{ij} \Delta_j t^2 + \epsilon^l_{ijt},
\]

where \( L_{ijt} \) is the average LOS for DRG group \( j \) hospitalized in hospital \( i \) during period \( t \); superscript \( l \) denotes parameters related to LOS equation; and the random coefficients and the within operator are respectively defined as:

\[
\begin{pmatrix}
\delta^l_{ij} \\
\theta^l_{ij}
\end{pmatrix}
\sim N(\mu^l, \Sigma^l); \quad \text{with} \quad \mu^l = \begin{pmatrix}
\rho^l \\
\phi^l
\end{pmatrix}, \quad \Sigma^l = \begin{pmatrix}
(\sigma^l_{11})^2 & \sigma^l_{12} \\
\sigma^l_{12} & (\sigma^l_{22})^2
\end{pmatrix},
\]

(9)

\[
\Delta_{ij} x_{ijt} = x_{ijt} - \bar{x}_{ij}; \quad \text{with} \quad \bar{x}_{ij} = \frac{1}{T_{ij}} \sum_{t=0}^{T_{ij}} x_{ijt},
\]

(10)

where \( T_{ij} \) is the number of periods for patients with DRG \( j \) treated in hospital \( i \).

The random-coefficient models described in Equations (4), (6) and (8) will be estimated using the EM algorithm. Based on the estimated parameters and the obtained residuals for each hospital, the hospital specific parameters are calculated.
using a conditional Bayesian predictor denoted hereafter, by a superposed symbol \(^\wedge\). The changes in excess\(^7\) costs, capacity and LOS for a given hospital \(i\) as well as the sector’s growth in total costs due to technological progress can therefore be identified compared to the beginning of the sample period (1998). These temporal changes are respectively specified as:

\[
\begin{align*}
\text{Excess Costs:} & \quad \Delta u_i = \delta_i t + \theta_i t^2 \\
\text{Excess Capacity:} & \quad \Delta u'_i = \delta'_i t + \theta'_i t^2 \\
\text{Excess LOS:} & \quad \Delta u''_i = \delta''_i t + \theta''_i t^2 \\
\text{Sector's Cost Trend:} & \quad \Delta c = \delta_t t + \theta_t t^2
\end{align*}
\]

(11)

In order to test the statistical significance of efficiency differences across different ownership/subsidy types, I apply the Kruskal-Wallis (1952) rank test to the predicted random coefficients \(\delta_i, \theta_i, \delta'_i, \theta'_i, \delta''_i\) and \(\theta''_i\), as well as the estimated total changes realized over the sample period.\(^8\)

4. Data

The data used in this paper consist of two data sets covering 214 general hospitals operating in Switzerland from 1998 to 2003.\(^9\) These data include a “hospital-level data set” based on hospitals’ financial and administrative data (SFSO, 1997a), and a “DRG-level data set” extracted from medical records of individual hospitalizations,\(^10\) including the average LOS and the number of cases by hospital, year and DRG categories (SFSO, 1997b). While the hospital-level data set is used for the analysis of cost-efficiency and excess capacity, the analysis of hospital stays is

---

\(^7\) I use the word excess in a narrow sense, to denote the temporal changes that cannot be explained by the changes in variables included in the model.

\(^8\) Kruskal-Wallis test is a non-parametric test that has been often used in frontier analysis (Singh and Coelli, 2001; Grosskopf \textit{et al.}, 2001). An alternative approach would be to include type indicators as interaction terms in the regression models and test their significance. I preferred the non-parametric test because of its robustness to distribution assumptions.

\(^9\) Specialized clinics, rehabilitation centers and other long-term facilities are excluded.

\(^10\) The original data base includes about a million records by DRG and admission categories.
based on the DRG-level data. The latter data are also used to calculate an average DRG cost weight for each hospital-year that is merged into the hospital-level data. This average cost weight represents a measure of the severity of the patient mix, used for adjusting the number of admissions (more on this later).

The available data contain a number of missing values and invalid observations. In order to have a sufficient number of sample points over time, about 36 hospitals that have been covered for less than three years are excluded from the sample: In fact, the quadratic form of temporal changes requires at least three values for a reasonable identification of individual parameters. Moreover, an adequate efficiency analysis requires a sample of comparable hospitals that satisfy the basic assumptions of the model. Therefore, ten hospitals that have changed ownership status over the sample period are also excluded. Given that the ownership changes are probably related to efficiency reasons, assuming constant parameters for efficiency changes \((\delta, \theta)\) is not realistic for these hospitals.\(^{11}\) In addition a few extremely small hospitals (with less than ten beds) were excluded. Although officially classified as general hospitals, these hospitals appear to belong to a special category of local hospitals whose services might deviate from the short-term treatments commonly provided in general hospitals. The final hospital-level sample includes 863 observations from 168 general hospitals.\(^{12}\)

As for the DRG-level data, the adopted sample has been restricted to the hospitals that have been included in the hospital-level data set and the observations

\(^{11}\) Such an assumption might bias the estimated differences across different ownership types. Assuming a sudden structural change in parameters after the conversion year is also unrealistic, because the ownership changes are usually long processes and the converting hospitals might undergo gradual changes prior to conversion. See Farsi (2004) for some evidence on this issue.

\(^{12}\) A series of probit analyses and t-tests indicated that the excluded observations are not related to an obvious selection of hospitals regarding size (number of beds) or ownership/subsidy types. In any case, given the presence of fixed effects in the model sample selection is not expected to affect the results.
that are based on three or more inpatient cases with hospitalizations longer than 24 hours, from the DRG categories that have a clear definition\textsuperscript{13} according to the Swiss AP-DRG classification version 4.0 (APDRG Suisse, 2003). The final sample, after excluding severe outliers\textsuperscript{14}, consists of 108,227 observations from 162 general hospitals, 492 AP-DRG categories, 826 hospital-year groups and 23,281 hospital-DRG groups. From the 492 DRG’s included in the sample, 223 are classified as surgical procedures. In terms of the number of hospitals and the composition of hospital types regarding ownership, university hospitals and also the distribution across different regions, this sample is very similar to the hospital-level sample used for the analyses of cost and excess capacity. A descriptive summary of the main variables included in the models is provided in Table 1. In the rest of this section these variables will be described.

**Insert Table 1**

The main measure of hospital output is taken as a DRG adjusted number of hospitalizations (cf. Linna, 1998; Rosko, 2001; Heshmati, 2002), obtained by multiplying total admissions by an average DRG cost weight calculated for each hospital-year\textsuperscript{15}. Since the number of outpatient cases is not available in the data, the ambulatory output is approximated by the corresponding revenues adjusted for inflation. This approximation is based on the assumption that the average unit price of ambulatory care is similar across hospitals.

Three input factors are considered: capital, physicians’ input and all other employees’ labor. Similar to Wagstaff and Lopez (1995) and Rosko (2001), capital

\textsuperscript{13} The DRGs described by ‘other’ or ‘non-specified’ were not considered.

\textsuperscript{14} About 1600 severe outliers with average LOS greater than 36.4 days (three times the inter-quartile range) were excluded.

\textsuperscript{15} The average cost weight for any given hospital-year is calculated form the medical data, by dividing the weighted sum of the number of admissions (with weights being the DRG cost weights according to Swiss AP-DRG version 4.0), by the total number of cases. This provides a single measure of inpatient services in contrast with Brown’s (2003) approach with multiple groups with similar DRG weights.
prices, are approximated by the hospital’s total capital expenditure divided by the number of available beds in the hospital. Labor prices are calculated by dividing total salaries by the number of remunerated days. Physicians and non-physicians are considered as two separate labor inputs similar (cf. Folland and Hofler, 2001; Scuffham et al., 1996). The physicians’ labor price represents the average salary of those employed by the hospital and exclude honoraries and fees, accounting on average for about 5% of the hospital’s total costs, usually paid to both employed and unemployed physicians. Both labor prices are proportionally adjusted for social benefits, accounting on average, for about 9% of total costs with the proportions being the respective shares of each group’s salaries. This adjustment captures the potential variation in social benefits due to differences in pension funds as well as the age and seniority of the employees mix.

In line with most hospital cost studies in the literature (with a very few exceptions such as Rosko, 2001), the input prices are assumed to be exogenous. This simplifying assumption usually reflects the difficulty of finding reasonable instrumental variables to account for such endogeneities. Here, the hospital fixed effects alleviate the problem, to the extent that the price endogeneity is time-invariant sources. Moreover, I argue that the problem is less severe in Switzerland, where given the strong restrictions in the labor market, the relative uniformity of capital markets, and the strong monitoring system for quality and maintenance, the hospitals’ ability in affecting the prices are relatively limited.16

---

16 In any case the focus of this study is on efficiency estimates and the endogeneity bias in the price coefficients is of secondary importance. The possible impact of endogeneity on efficiency estimates is an open question that depends on whether a company’s intentions in changing their inputs are interpreted as a quality-neutral effort to improve efficiency or as an intentional change in the quality of inputs. In the latter case, by including the input prices we can provide more realistic values of efficiency adjusted for quality differences, even though the price coefficients are obviously biased.
The three input factor prices considered in the model correspond to about 70 percent of total costs in a typical hospital included in the sample. The available data do not allow an appropriate calculation of the prices of remaining inputs such as medical materials, food, water and power as well as physicians’ fees and other personnel charges. The excluded prices might vary over time and across hospitals. The time-invariant differences are captured by the hospital fixed effects, thus cannot bias the results. As for the temporal variations in the excluded prices, they are partly captured by the time variables included in the cost model, otherwise are assumed to be uncorrelated with temporal variations of efficiency.

The average length of hospitalization has been included in the model (Vita, 1990; Scuffham et al., 1996; Carey, 1997). In addition to representing hospital’s ‘hotel services’ like nursing care and accommodation (Breyer, 1987), this variable provides a measure of severity of the case mix within each DRG. In fact, there is a considerable variation among patients within a DRG, as indicated by the wide range of acceptable hospitalization length provided by the Swiss DRG Association (APDRG Suisse, 2003).

Hospitals’ costs can also be affected by the number of specializations and services offered in a hospital. Here we assume that these factors are time-invariant, thus captured by the fixed effects. The shortcoming of the analysis is mainly related to the quality of care. In fact, it is reasonable to consider that by improving cost-efficiency, certain quality aspects of health care might be compromised. We do not have any reliable data on any measure of quality in Swiss hospitals that show a reasonable variation over the sample period. It should be however noted that the evidence on the effect of quality measures on hospital costs is not conclusive. Zuckermann et al (1994), Rosko (2001) and Vitaliano and Toren (1996) conclude that
quality indicators do not have significant cost effects, whereas others such as Folland and Hofler (2001) suggest a significant effect for structural quality measures such as bed availability and the share of board-certified physicians.

The measure of excess capacity is based on the average number of empty beds in a given hospital-year. This is obtained by subtracting the number of available beds by the total number of patient days divided by 365. The semi-hospitalizations (inpatient stays shorter than 24 hours) are considered as one-day hospitalization. The data show some discrepancy in this measure particularly several negative values. These values have been re-calculated using an alternative measure of hospital’s available beds namely, the number of hospital’s bed-days. The ownership status and subsidization form have been considered in four categories as described in Table 2:

Insert Table 2

5. Results

Table 3 provides the regression results of the hospital-level analysis based on Equations (4), (6) and (8) respectively for total costs, excess capacity and LOS. The results of the cost model point to considerable effects of hospital stays on costs. The variation of other factors such as ambulatory services and the training positions though being statistically significant are practically limited to a few percentage points in terms of elasticity. The estimated coefficients are mostly significant and generally have the expected signs. As discussed earlier, considering that the between-hospital variations are entirely suppressed in the hospitals’ individual fixed effects, the estimated coefficients here might be inadequate for any inference about the
technological characteristics such as returns to scale.\textsuperscript{17} Therefore, in the following discussion we focus on the efficiency estimates and their variations.

\textbf{Insert Table 3}

The estimation results of the cost analysis (Table 3) point to a pattern of increasing growth in hospital’ operating costs, as suggested by the positive coefficients of the time variables with an average growth rate of about 1.6 percent per year. The results also suggest that the temporal changes are significantly different from one hospital to another, as shown by the statistically significant values for the variance of the random effects. The negative covariance between the two random coefficients is consistent with the fact that any growth (decline) is likely to slow down with time. The negative correlation implies for instance, that hospitals that start to cut the costs earlier and more aggressively, will have a relatively lower success later.

The estimation results from the analysis of excess capacity (Table 3, the middle column) indicate that hospitals have decreased their empty beds with a substantial average rate of about 8.6% per year. The negative effect of number of admissions suggests that hospitals with greater outputs have been downsizing more, perhaps because of their greater margins for demand fluctuations.\textsuperscript{18} As expected the share of private-insurance patients shows a positive effect on excess capacity, however, the coefficient is not statistically significant. Similarly the results indicate significant variation across hospitals regarding the empty beds.

Finally the last column (Table 3) provides the results of the DRG-level analysis of the length of hospitalizations. As seen in the table, the estimated annual

\textsuperscript{17} The presence of fixed effects can also explain the lack of statistical significance for some of the variables. Compare for instance with the estimation results reported in Farsi and Filippini (2007).

\textsuperscript{18} This can also be explained by the mechanical negative relationship between admissions and the number of empty beds. Such a relationship might create endogeneity bias in the hospital-specific estimates of growth in excess capacity. However, a preliminary analysis showed that excluding the number of admissions from the model does not cause much difference.
rate of decrease in LOS is about 2% on average. The number of training positions has a positive but statistically insignificant on LOS. The fixed effects at the hospital-DRG groups are expected to capture the differences among DRGs regarding the potentials for reducing LOS, thus decreasing the possible aggregation biases due to different distributions of DRGs across hospitals. These results also indicate a significant variation LOS’s temporal variations, across the included hospitals and also among the DRG groups.

The considerable variation of temporal patterns across individual hospitals suggests that the study of the variations between hospital types could be used to test hypotheses regarding the efficiency patterns in the hospital sector. Before turning to the results of these statistical tests, it is worthwhile to summarize the overall efficiency trends. The average estimated time effects from Table 3 are illustrated in Figure 1. These variations are obtained from Equations (11) averaged over hospitals. As can be seen in the figure, over the five-year span in the sample period (1998-2003) a typical hospital’s costs have grown about 14 percent. This is while the length of hospital stays and the number of empty beds have decreased by about 10 and 18 percent respectively. The substantial rate of decline in LOS and hospital empty beds shown in the figure is indicative of hospitals’ considerable efforts to contain costs.

Insert Figure 1

The considerable growth in the sector’s costs is consistent with the growth of hospital costs in many countries, reported in previous literature. This growth has been often associated with new medical procedures and pharmaceuticals as well as the extension of life expectancy. These are obviously external factors that are modeled by

---

19 An additional analysis of LOS aggregated at the hospital level (available upon request), indicates an average decrease of about 3.3% per year in the length of hospitalizations, suggesting an upward aggregation bias. All other coefficients are very similar to those reported in Table 3.
average trends in the model specification used in this paper. The hospital-specific inefficiency is defined as the hospital’s excess costs as compared to the average increasing trend shown in Figure 1. A useful way of investigation the relationships between costs and other measures, is by dividing the sample into two groups namely hospitals that improved on cost-efficiency and those who showed an efficiency loss. These two groups correspond respectively to negative and positive values for $\Delta u_i$ at the end of the sample period ($t = 5$) obtained from Equation (11). The average temporal variations of excess costs, capacity and LOS in these two groups are depicted in Figure 2 and Figure 3 respectively.

**Insert Figure 2**

**Insert Figure 3**

Figure 2 shows that the 81 hospitals that had an efficiency gain (in costs) have also considerably cut their hospital stays and empty beds. Compared to the overall patterns in Figure 1, these hospitals, while having a relatively high reduction in excess capacity, are not much different from average in terms of LOS. Similarly, the average changes in excess LOS the 87 hospitals with declining cost efficiency over the sample period (Figure 3) show a change of LOS that is totally comparable to the overall average trends (Figure 1). However, the excess capacity takes a somewhat milder reduction here. The trends in both groups of hospitals show an average change of about 8 percentage points in cost-efficiency over the sample period. This might suggest a reasonable targeting benchmark that is comparable to the 2 to 3 percent annual efficiency gain targets set by the UK health care authorities (Jacobs and Dawson, 2003).

A comparison between Figure 2 and Figure 3 points to a distinctive difference in excess capacity changes between the two groups, suggesting that empty beds have
a crucial impact on cost-efficiency. However, it should be noted that the variation among individual hospitals are ignored in the average trends illustrated in these figures. A statistical analysis of the correlations between these efficiency measures can be used to assess the relative importance of hospital stays and excess capacity.

The differences between ownership/subsidy groups listed in Table 2 are analyzed with a series of Kruskal-Wallis and t-tests with unequal variances. The results generally suggest that the differences across hospital types are due to sampling error, not a systematic difference in the underlying distribution. The observed significance level was generally higher than 10% and the results were confirmed using only the linear trends or the resulting change over the five-year span of the sample. Similar results were obtained for all three measures namely, changes in cost-efficiency, excess capacity and LOS.

Noting that the variation among individual hospitals often dominates the variations between hospital types, an important question is whether improvement in cost-efficiency are positively correlated with other measures like limiting the empty beds and shortening hospitalizations, presumably aimed at cost reductions. In order to see an overall picture, the correlation matrix between these measures is provided in Table 4. The listed coefficients are based on Spearman’s rank correlation between the estimated rises for each hospital over the five-year span, obtained by substituting $t = 5$

20 Interpreting the average trends without statistical correlations, could be misleading. For instance, as we see later the excess costs and excess LOS show a positive correlation in the hospital group with efficiency loss, which might seem contradictory to the opposing trends in costs and LOS in Figure 3. However, a positive correlation does not necessarily imply similar average trends. Rather, it implies that the hospitals that are located above the LOS curve are likely to be above the excess capacity curve as well. Another interesting point is the close coincidence of LOS and cost curves in Figure 2, but the lack of statistically significant correlation between the two measures in that group (as we see later).

21 In addition to ownership types, we studied the differences among five typologies based on size and specialization (SFSO, 2001), and five geographical regions (details available upon request). Virtually in all cases, the differences across groups were statistically insignificant. The only exception was canton Ticino (southern region) with greater gains in cost efficiency compared to four other regions. Nevertheless, further tests suggested no significant difference between Ticino and six other cantons (BS, BL, FR, GE, NE and VS), out of 26 Swiss cantons.
in Equations (11). The correlation coefficients are also provided for the two sub-
samples with a gain or loss in cost-efficiency, corresponding to $\Delta u_i(t = 5)$ smaller or
greater than zero respectively.

**Insert Table 4**

The numbers estimated on the entire sample (first two columns) indicate a
positive and significant correlation between efficiency measures related to excess cost
and excess capacity suggesting that hospitals that have been able to decrease the
empty beds are also successful in cutting costs. This statement does not apply to
hospital stays: The null hypothesis of independence between excess costs and excess
LOS cannot be rejected at any reasonable significance level.\(^\text{22}\) However, the
correlation patterns within the two sub-samples (Table 4) point to certain differences
between hospitals showing efficiency gains and losses: In particular, there is a
borderline significant and positive correlation between excess costs and excess LOS
in hospitals that have shown a decline in cost-efficiency. This finding suggests that
among hospitals with relatively poor performance regarding efficiency gains, thus
perhaps with certain slackness in excess LOS, curtailing hospital stays could be an
effective means for improving cost efficiency. On the other hand, in these hospitals
there is no significant correlation between excess capacity and excess costs,
suggesting that lowering excess capacity does not necessarily bring about any cost
savings. This can be explained by the fact that in these hospitals the apparently low
excess capacity might be a result of excessively long hospitalizations.

Finally, the numbers in Table 4 suggest a negative correlation between
temporal changes in excess capacity and excess LOS. However, this correlation is not

\(^{22}\) This result was also confirmed by a series of correlation analyses within various types of hospitals by
ownership/subsidy. Excepting a few cases with significant correlations in levels, the rank correlation
remained statistically insignificant across all sub-samples.
statistically significant within each one of the two sub-samples. Considering that in the short-run, empty beds increase as a result of reduction of hospital stays, this finding suggests that at least in some hospitals, shortening LOS is not completed by sufficient follow-up measures to reduce the resulting excess capacity.

6. Conclusions

Using a panel data mixed effects model we proposed an econometric specification inspired by Sickles’ (2005) general models, for the analysis of temporal variations in Swiss hospitals’ productive efficiency. The model includes fixed effects for unobserved time-invariant factors related to individual hospitals and DRG categories, and random coefficients representing the effects of time variables. The measures of interest are the hospitals’ gains in cost-efficiency, the realized cuts in empty beds and the shortening of hospitalizations over the period starting from 1998 which coincides approximately with the outset of health policy reforms particularly the gradual implementation of prospective payment system in Switzerland.

The results indicate that on average the length of hospitalization and the number of empty beds in a hospital have decreased by about 10 and 18 percent respectively. The results also suggest that after adjusting for the changes in outputs, labor prices and other characteristics such as teaching activities, hospital costs have risen considerably and increasingly over the six-year period from 1998 to 2003, amounting to an average increase of 14 percent in total costs for a typical hospital. It is assumed that this overall increase reflects the external factors such as progress in medical treatments and extension of life expectancy, and the remaining hospital-specific changes in costs are associated with efficiency gains or losses.
There is a considerable variation among individual hospitals concerning cost efficiency gains and also the efforts in cutting the excess capacity and curtailing hospitalizations. In general hospitals that showed a relatively important decrease in excess capacity are likely to show a relative gain in cost-efficiency and *vice versa*. However, the results do not provide any conclusive evidence that gains in cost efficiency be associated with shortening hospital stays. Interestingly, only among hospitals that experienced an efficiency loss over the sample period, relatively low cuts in hospitalization length are likely to be associated with the hospitals with low efficiency gains, suggesting that the length of stay could be an important parameter in these hospitals. This result can also be interpreted as suggestive evidence that hospitals that have a good performance in containing costs do not have much slackness in their hospitalization lengths. While confirming the strong heterogeneity across hospitals regarding efficiency gains, the findings do not provide any evidence in favor of a particular ownership/subsidization type.

The adopted methodology is readily applicable to other industries and the assumptions are easy to understand and interpret. In addition, in line with several models in this field (probably starting from Cornwell, Schmidt and Sickles, 1990) the efficiency estimates do not rely on the skewness of the residuals. The combination of fixed effects with random effects, allows a complete abstraction from the unobserved time-invariant variables whose effects are not primordial for the analysis (fixed effects) while at the same time providing a ‘statistically’ efficient estimation basis for the parameters of interest (random effects).

Given that in presence of strong unobserved heterogeneity, the time-invariant component of efficiency is difficult if at all possible to identify, reliable measures of efficiency gains over time can be helpful in many regulation and policy applications.
This paper illustrates that with certain assumptions, panel data mixed effects can be used for this purpose. However, it is important to consider the implications of the model’s assumptions in each specific application and the resulting policy limitations. In the case studied here, the results are based on the assumption that the potential changes in the unobserved quality of hospital services in response to the reforms and financial pressures, are either uniform across the sector or uncorrelated with the adopted measures of efficiency improvement. Therefore, any possibility of deviation from this assumption should be considered before drawing relevant policy conclusions.

**Acknowledgements**

This research has been partly financed by the Swiss National Science Foundation through research grant 100012-108288, which is gratefully acknowledged. I am also grateful to the Swiss Federal Statistical Office for providing the data and to Massimo Filippini, William Greene, Martin Jakob and André Meister for their support and many helpful comments. I benefited from helpful suggestions of two anonymous reviewers and this journal’s editor which is greatly appreciated. I am solely responsible for the expressed views as well as any remaining errors and omissions.

**References**


Figure 1: Temporal variation of costs, excess capacity and LOS (168 hospitals)

![Figure 1](image1.png)

Figure 2: Variations in hospitals that improved in cost-efficiency (81 hospitals)

![Figure 2](image2.png)

Figure 3: Variations in hospitals that declined in cost-efficiency (87 hospitals)

![Figure 3](image3.png)
Table 1: Descriptive statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital's total costs (CHF '000)</td>
<td>69'655</td>
<td>12'286</td>
<td>924</td>
<td>15'657</td>
<td>32'592</td>
<td>65'129</td>
<td>884'764</td>
</tr>
<tr>
<td>Number of hospitalizations</td>
<td>6'306</td>
<td>7'128</td>
<td>116</td>
<td>1'845</td>
<td>4'096</td>
<td>7'871</td>
<td>50'774</td>
</tr>
<tr>
<td>Number of hospitalizations (AP-DRG adjusted)</td>
<td>5'400</td>
<td>7'065</td>
<td>76</td>
<td>1'370</td>
<td>3'123</td>
<td>6'568</td>
<td>49'251</td>
</tr>
<tr>
<td>Average total cost per hospitalization (CHF '000)</td>
<td>10.02</td>
<td>6.38</td>
<td>1.76</td>
<td>7.04</td>
<td>8.74</td>
<td>11.21</td>
<td>90.13</td>
</tr>
<tr>
<td>Number of patient-days</td>
<td>51'619</td>
<td>58'348</td>
<td>1'068</td>
<td>19'570</td>
<td>32'470</td>
<td>57'419</td>
<td>410'140</td>
</tr>
<tr>
<td>Average length of hospitalizations (CHS)</td>
<td>10.4</td>
<td>6.6</td>
<td>2.0</td>
<td>6.6</td>
<td>8.4</td>
<td>11.5</td>
<td>57.6</td>
</tr>
<tr>
<td>Hospital's outpatient revenues (CHF '000)</td>
<td>10'752</td>
<td>20'458</td>
<td>0</td>
<td>1'301</td>
<td>4'118</td>
<td>10'281</td>
<td>144'802</td>
</tr>
<tr>
<td>Hospital capacity (number of beds)</td>
<td>175.2</td>
<td>202.0</td>
<td>12</td>
<td>63</td>
<td>104</td>
<td>210</td>
<td>1277</td>
</tr>
<tr>
<td>Excess capacity (average % of empty beds)</td>
<td>35.1</td>
<td>52.3</td>
<td>1</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>523</td>
</tr>
<tr>
<td>$P_x$ (capital price) CHF '000 per bed</td>
<td>28.04</td>
<td>26.68</td>
<td>1.46</td>
<td>11.05</td>
<td>17.19</td>
<td>36.28</td>
<td>242.57</td>
</tr>
<tr>
<td>$P_L$ - physicians $^b$ (CHF per day)</td>
<td>334.51</td>
<td>114.22</td>
<td>66.80</td>
<td>263.03</td>
<td>321.15</td>
<td>393.43</td>
<td>781.63</td>
</tr>
<tr>
<td>$P_L$ - other employees $^c$ (CHF per day)</td>
<td>178.11</td>
<td>33.09</td>
<td>69.43</td>
<td>158.91</td>
<td>176.98</td>
<td>196.85</td>
<td>302.01</td>
</tr>
<tr>
<td>Number of medical training position</td>
<td>41.6</td>
<td>91.3</td>
<td>1</td>
<td>6</td>
<td>14</td>
<td>31</td>
<td>583</td>
</tr>
<tr>
<td>Share of private-insurance admissions $^d$</td>
<td>0.28</td>
<td>0.22</td>
<td>0.00</td>
<td>0.15</td>
<td>0.22</td>
<td>0.31</td>
<td>1</td>
</tr>
</tbody>
</table>

Average length of full hospitalizations excluding semi-hospitalizations (days):

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital-level sample</td>
<td>11.3</td>
<td>6.4</td>
<td>3.7</td>
<td>7.8</td>
<td>9.0</td>
<td>12.3</td>
<td>57.6</td>
</tr>
<tr>
<td>DRG-level sample</td>
<td>9.7</td>
<td>6.1</td>
<td>1.0</td>
<td>5.2</td>
<td>8.0</td>
<td>12.6</td>
<td>36.3</td>
</tr>
</tbody>
</table>

Average AP-DRG cost weight:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital-level sample</td>
<td>0.806</td>
<td>0.110</td>
<td>0.520</td>
<td>0.740</td>
<td>0.789</td>
<td>0.854</td>
<td>1.334</td>
</tr>
<tr>
<td>DRG-level sample</td>
<td>1.008</td>
<td>0.783</td>
<td>0.112</td>
<td>0.582</td>
<td>0.795</td>
<td>1.161</td>
<td>21.597</td>
</tr>
</tbody>
</table>

- Unless stated otherwise, the numbers are based on the hospital-level sample.
- The hospital-level sample includes 863 observations from 168 hospitals (1998-2003).
- The DRG-level sample includes 108,227 observations from 492 AP-DRG categories.
- All monetary values are adjusted by the global consumer price index relative to 2003 prices.
- Semi-hospitalizations (shorter than 24 hours) are considered as one-day hospitalizations.
- Employed physicians' average salary, adjusted for social benefits and excludes fees.
- Average salary (adjusted for social benefits) of all hospital employees except physicians.
- Based on hospital discharges; includes cases with private and semi-private insurance.
Table 2: Number of hospitals by category (1998-2003)

<table>
<thead>
<tr>
<th>Ownership/Subsidy status</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-subsidized For-Profit (FP)</td>
<td>27</td>
<td>16.07</td>
</tr>
<tr>
<td>Non-subsidized Non-Profit (NP)</td>
<td>16</td>
<td>9.52</td>
</tr>
<tr>
<td>Public (PUB)</td>
<td>81</td>
<td>48.21</td>
</tr>
<tr>
<td>Private subsidized (SUB)</td>
<td>44</td>
<td>26.19</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>168</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
Table 3: Estimation results

<table>
<thead>
<tr>
<th></th>
<th>Total Costs</th>
<th>Excess Capacity</th>
<th>Length-of-Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of hospitalizations (AP-DRG adjusted)</strong></td>
<td>0.300*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Outpatient revenues</strong></td>
<td>0.025*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average length of hospitalizations</strong></td>
<td>0.228*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.022)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>$P_K$ (capital price)</strong></td>
<td>0.124*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>$P_L$ - physicians</strong></td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>$P_L$ - others</strong></td>
<td>0.050*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of training positions</strong></td>
<td>0.021*</td>
<td>0.0046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.0046)</td>
<td></td>
</tr>
<tr>
<td><strong>Time (linear trend)</strong></td>
<td>0.016*</td>
<td>-0.086*</td>
<td>-0.019*</td>
</tr>
<tr>
<td></td>
<td>(.006)</td>
<td>(.035)</td>
<td>(.0022)</td>
</tr>
<tr>
<td><strong>Time (squared)</strong></td>
<td>0.002*</td>
<td>0.010</td>
<td>-0.00034</td>
</tr>
<tr>
<td></td>
<td>(.001)</td>
<td>(.006)</td>
<td>(.00039)</td>
</tr>
<tr>
<td><strong>Number of hospitalizations</strong></td>
<td></td>
<td>-0.447*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.112)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Share of private-insurance admissions</strong></td>
<td></td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.208)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.062*</td>
<td>0.317*</td>
<td>0.187*</td>
</tr>
<tr>
<td></td>
<td>(.006)</td>
<td>(.036)</td>
<td>(.0027)</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>0.011*</td>
<td>0.053*</td>
<td>0.032*</td>
</tr>
<tr>
<td></td>
<td>(.001)</td>
<td>(.007)</td>
<td>(.00051)</td>
</tr>
<tr>
<td>$\sigma_{12}$</td>
<td>-0.894*</td>
<td>-0.915*</td>
<td>-0.936*</td>
</tr>
<tr>
<td></td>
<td>(.027)</td>
<td>(.024)</td>
<td>(.0023)</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>0.040*</td>
<td>0.287*</td>
<td>0.220*</td>
</tr>
<tr>
<td></td>
<td>(.001)</td>
<td>(.009)</td>
<td>(.00061)</td>
</tr>
<tr>
<td>Log Likelihood (restricted)</td>
<td>1288.16</td>
<td>-305.90</td>
<td>-1958.14</td>
</tr>
<tr>
<td>Number of observations</td>
<td>863</td>
<td>863</td>
<td>108,227</td>
</tr>
</tbody>
</table>

* means significant at 5%; Standard errors are given in parentheses; all variables except share of private insurance admissions are in logarithms; the hospital-level sample includes 863 records from 168 hospitals; the DRG-level sample includes 108,227 observations from 492 AP-DRG’s treated in 162 hospitals; the sample period covers from 1998 through 2003.
Table 4: Spearman’s rank correlation between three measures of growth between 1998 and 2003

<table>
<thead>
<tr>
<th></th>
<th>Overall (N=168 hospitals)</th>
<th>Hospitals showing an improvement in cost-efficiency (N=81)</th>
<th>Hospitals showing a decline in cost-efficiency (N=87)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Excess Capacity</td>
<td>Excess Length-of-Stay</td>
<td>Excess Capacity</td>
</tr>
<tr>
<td>Excess Costs</td>
<td>0.200**</td>
<td>-0.054</td>
<td>0.244**</td>
</tr>
<tr>
<td>Ex. Capacity</td>
<td>1</td>
<td>-0.184**</td>
<td>1</td>
</tr>
<tr>
<td>Average decrease (%)</td>
<td>17.7</td>
<td>9.7</td>
<td>27.0</td>
</tr>
</tbody>
</table>

** Significant at 5%; * Significant at 10%.
Changes in Hospital Quality

After Conversion in Ownership Status

Mehdi Farsi*

Department of Economics, University of Lugano

and

Swiss Federal Institute of Technology,

ETH Zentrum, WEC, 8092 Zurich, Switzerland

March 2004

Abstract

This paper examines the effects of conversions between For-Profit and Not-For-Profit forms on quality of medical care in California hospitals. The sample includes elderly patients treated in California’s private hospitals from 1990 to 1998 for Acute Myocardial Infarction and Congestive Heart Failure. The results suggest that converted hospitals have experienced quality changes before conversion and that ignoring these changes may bias the estimates of conversion effects. Both conversions are found to have some adverse consequences: Hospitals that converted to FP form show an increase in AMI mortality rates, while those converted to NFP status indicate an increase in CHF mortality outcomes.

JEL classification: L31; I18; I11

Keywords: Ownership conversions; Hospital quality; Mortality; Medicare beneficiaries

* Tel.: +41-1-632-0656; Fax: +41-1-632-1050

E-mail address: farsi@cepe.mavt.ethz.ch
1. Introduction

Recent conversions of not-for-profit (NFP) hospitals to for-profit (FP) status have raised public concerns about possible detrimental effects on quality of care (cf. Goddeeris and Weisbrod (1998), Kuttner (1996a-b) and Ho and Hamilton (2000)). A common perception is that NFP institutions are committed to providing quality care regardless of costs. In fact, following Arrow (1963) theoretical models often assume that providers choose the NFP form of organization in order to signal this high commitment to quality (Frank and Sulkever (1994) and Glaeser and Schleifer (1998)).

There is a growing literature on the effects of ownership status on the quality of care. However, only a limited number of papers have studied the effect of conversion from one form to the other. All these studies have assumed that the conversion effects on quality appear after the conversion. There is however, some evidence that conversions are usually preceded by financial difficulties (Sloan (2002) and Mark (1999)). Converted hospitals may therefore be subject to some gradual changes that potentially affect the quality of care prior to conversion. For instance, hospitals in financial distress may have started to deteriorate in quality long before conversion. In this case even if the conversion does improve the situation, failure to control for the pre-conversion changes may lead to the conclusion that conversion resulted in lower quality.

---

1 NFP hospitals are private hospitals that are owned by non-for-profit foundations, charity organizations or churches. Here, public hospitals are not considered as NFP.

2 Sloan et al. (2003) is a recent example. See Sloan (2000) and Baker et al. (2000) for extensive surveys of this literature.
This paper examines the effects of conversions between FP and NFP forms in the California hospital market over the nine-year period between 1990 and 1998 using models with hospital fixed effects. It differs from the previous literature in several aspects. First, the time-variations of quality before conversion are taken into account by controlling for a linear trend before each type of conversion. Secondly, in contrast with Sloan (2002) and Shen (2002), there is no restrictive assumption on the conversion effects on quality. Finally, while all the above studies used a national sample of hospitals (Medicare data), this paper considers a relatively uniform sample including California’s private short-term hospitals. The US states are quite different regarding FP sector’s share in hospital markets, ranging from states with virtually no FP sector to markets dominated by FP hospitals. Since the presence of FP hospitals may affect the behavior of the NFP hospitals in the same area, the NFP hospitals in different states may be significantly different from each other. California hospital market is characterized by a relatively large FP sector that has remained more or less constant over the study period.

One of the important difficulties in studying the ownership effects on hospital quality is the selection bias. Patients with acute diseases are likely to choose the closest hospital. For instance, paramedics are instructed to take heart attack patients to the nearest hospital. In this paper the patient mix characteristics that are related to the hospital’s location are taken into account through hospital fixed effects. Patient selection into a

---

3 Sloan (2002) assumes that conversion effects are symmetric. Shen (2002) on the other hand, assumes that the effect of conversion does not depend on the status prior to conversion. She assumes for instance that a conversion from public to NFP form has the same effect as a conversion from FP to NFP status.

4 See Kessler and McClellan (2001) for some evidence.
hospital may also be different before and after its conversion. For instance, an institution that changes from NFP to FP status may step up efforts to discourage the admission of unprofitable patients. Moreover, as suggested by Geweke et al. (2003), patients with higher unobserved severity are more likely to choose higher-quality hospitals. Therefore, in the presence of unobserved severity factors, conversion effects may be biased (selection bias). In order to identify the direction and importance of such biases, I exploit the fact that patients admitted through the Emergency Room are less affected by systematic selection. Generally, these patients do not have time to plan their hospitalizations and are likely to go to the closest hospital. Moreover, California hospitals are required to treat patients in emergency situations regardless of their insurance coverage. These considerations suggest that the measured effects of conversion should be less biased among the ER patients.

The results of this paper generally indicate that conversions to both FP and NFP forms may have adverse effects on quality. While conversion to FP status is found to increase the in-hospital mortality of AMI patients, conversion to NFP form has increased the mortality probability in the CHF sample. These results suggest that health outcomes in different diagnoses may represent different dimensions of hospital quality. This paper’s findings also suggest that hospitals that convert from one status to another may be subject to certain changes prior to conversion and neglecting such variations may lead to a considerable bias in the estimation of conversion effects.

---

5 Although many hospitals violate these requirements, there is no significant difference in propensity to violate between FP and NFP hospitals in California (Blalock and Wolfe, 2001).
The rest of the paper proceeds as follows: Section 2 reviews some of the previous literature. A description of the data and the adopted measures of quality is given in section 3. Section 4 explains the econometric methodology and discusses potential sources of bias. Section 5 provides the results and section 6 concludes the paper.

2. Background

Between 1970 and 1995, 330 of 5,000 NFP hospitals (about 7%) converted to FP type (Cutler and Horwitz, 2000). These conversions accelerated in the mid-90s. For example, 58 conversions occurred in 1995, up from 34 in 1994 (Kuttner, 1996a). These developments have spurred a large literature on the effects of FP and NFP status on quality of care, but the results are far from conclusive largely because of the difficulty of controlling adequately for patient selection. As Kessler and McClellan (2001) suggest, more productive hospitals may attract sicker patients. Geweke et al. (2003) provide some evidence that patients with higher unobserved severity are more likely to be hospitalized in high quality institutions.

Studies such as Gowrisankaran and Town (1999), Ettner and Hermann (2001) and McClellan et al. (1994) suggest that many patients choose the closest hospital, but this does not mean that FP status can be treated as exogenous determinant of mortality because FP hospitals are more likely to locate in areas with better insured patients, for example in areas with high proportion of Medicare patients (Norton and Staiger, 1994). McClellan and Staiger (2000) found that NFP hospitals have slightly lower mortality rates in a sample of elderly AMI patients, but reported that the estimated effects fell by almost
half when county fixed effects were included in the model. Sloan (2002) cites these results and concludes that FP hospitals tend to be located in areas with higher mortality rates.

The evidence in general points to the importance of hospital location in patient selection. Given that the hospital location does not change with conversion, the study of conversion effects can shed some light on the issue. Namely, using panel data models with fixed-effects one can account for the unobserved, location-related severity factors. The existing literature on the effects of conversions on quality of care is mainly limited to three papers: Sloan (2002), Picone et al. (2002) and Shen (2002).

All these papers used the Medicare data. Sloan examined the effects of conversions on patients admitted for stroke, hip fracture, coronary heart disease, congestive heart failure and pneumonia. He finds that conversions have no effect on the in-hospital mortality or on the proportion of uninsured patients. His results indicate however, that pneumonia patients treated in hospitals that converted to FP status experienced an increased rate of complications. Sloan argues that the failure to find a significant effect on the in-hospital mortality may reflect shorter hospitalizations after conversion to FP status. Using the same samples, Picone et al. (2002) found that one to two years after conversion to FP status, patients’ mortality increases suggesting a decline in quality. Shen (2002) studied the effect of conversions on the health outcomes of Medicare heart attack patients. Her findings suggest that conversion to FP ownership has resulted in a significant increase in mortality probability.
All the three papers have concluded that conversion to FP status has caused a decline in quality, while conversion to NFP status has not shown any significant changes in quality. However, all these studies have assumed that the quality of a converting hospital has remained constant prior to conversion. Given that conversions, especially those hospitals that have been taken over by hospital chains, are usually caused by financial problems, the assumption of constant quality of service does not appear to be realistic. It is important to note that financial difficulties are not always together with a decline in quality. It might as well be the case that in an increasingly competitive market, financial problems are due to a high quality of care such as a large nursing staff or expensive materials. It is also reasonable to assume that multi-hospital FP firms are more interested in hospitals that have maintained a high level of quality, which implies a good reputation, a large number of potential clients and better possibilities of savings by eventually lowering the quality.

3. Data

The data used in this paper consist of two main data sets prepared by the California’s Office of Statewide Health Planning and Development (OSHPD). The first set is the Patient Discharge Data that includes all the discharge abstracts for all patients discharged from a Californian hospital from 1990 to 1998. Patients younger than 65 are excluded from the sample to obtain a relatively uniform sample of Medicare beneficiaries. The variables include patients’ basic characteristics like age, gender and race, length-of-

---

6 It is interesting to note that in California, a large part of conversions are the result of a takeover by a FP or NFP hospital chain. See Spetz et al. (2000) for more details.
stay, severity of the disease, the diagnosed conditions and procedures used for treatment. Severity of illness is defined in four levels (extreme, major, moderate, and minor) according to APR-DRG (All Patient Refined Diagnosis-Related Group) classification. This severity measure and its validity are discussed later in this section.

The second data set is the California’s Hospitals Disclosure Data (CADD) from 1989 to 1998. This data set consists of the information obtained from the hospital financial reports submitted annually to the Department of Health Services. All non-federal hospitals are required to report. Hospital characteristics like ownership status, size (number of beds) and type of the hospital are extracted from this data set. Spetz et al (1999, 2000) and Mitchell et al. (2001) report that the ownership information reported in CADD has a lot of reporting errors. The main problems are non-standard reporting periods, multiple reports in a single year and late reporting or failure to report ownership changes. This data set is corrected using the information reported in the appendix of Spetz et al (1999) along with other corrections using Internet sources and a few direct contacts with hospitals.7 The status changes in private hospitals are completely checked and corrected. In order to avoid the potential errors related to the conversions of public hospitals, I excluded all hospitals that were public at least for one year within the sample period. In the analyses reported in this paper the unit of observation is a hospital-year

7 See Currie et al. (2002) for more details about the corrections in the data.
with the year being the fiscal year beginning June 30th.\textsuperscript{8} The patient-level data are merged with hospital characteristics using the admission month and year for each patient.

\textit{Conversions}

The changes in California’s acute-care\textsuperscript{9} hospital market share in private FP, private NFP, and public sectors between 1989-90 and 1997-98 are given in table 1. The changes in the number of hospitals in each sector and the average hospital size in terms of available beds are also given. These numbers suggest that during this period, public and FP hospitals became fewer but larger. While the number of NFP hospitals increased, their size remained practically unchanged.\textsuperscript{10} The total number of acute-care hospitals decreased from 462 hospitals in 1989-90 to 410 hospitals in 1997-98. The NFP sector is however less affected by this consolidation trend. As we will see later, conversions are responsible for part of these asymmetrical changes across different sectors.

The data show three major types of conversion: FP to NFP, NFP to FP, and public to NFP. There are only a few conversions in other directions. Between 1990 and 1998, 11

\textsuperscript{8} This is the OSHPD’s fiscal year and the most frequent reporting period in the data. For the cases where the reporting period does not match with the fiscal year, the data are arranged such that every report is considered in the fiscal year that covered the largest part of the reporting period. In the hospital-years with multiple reports, a single observation is created using the weighted average over the reports (weights being the lengths of the reporting periods). In cases where conversion in status is in a multiple-report year, the new (old) status is considered if the conversion occurred in the first (second) half of the fiscal year.

\textsuperscript{9} By acute-care hospitals I mean all non-psychiatric hospitals that reported inpatient care. Rehabilitation centers are also excluded.

\textsuperscript{10} According to the data the average capacity of the NFP hospitals increased gradually from 217 beds in 1989-90 to 227 in 1993-94 and then decreased to 217 beds in 1997-98.
hospitals converted from FP to NFP form, while 15 NFP hospitals became FP. At the same period 14 hospitals converted from public to NFP status. Among more than 500 acute-care hospitals that have operated in California, about 56 hospitals had at least one conversion during this period. These hospitals on average, account for about 13% of hospital beds in California. Table 2 gives the distribution of conversions between FP and NFP status over time. The number of hospitals and hospital beds are both given. As suggested by these numbers, the conversions are spread over the nine years and do not show any clear temporal pattern.

The variations in the size of the converted hospitals (given in table 2) suggest that among both FP and NFP hospitals, the relatively large institutions have been more likely to convert in ownership status. However, given that hospital capacity is an endogenous parameter that can change with conversions, it is not included in the model.\footnote{My regressions (not included in the paper) indicate that converted hospitals may change their capacity. However, including the number of beds in the regressions does not change the results of the paper.} Other hospital characteristics that do not change with conversion, are captured by hospital fixed effects.\footnote{For instance, Keeler et al (1992) found that among all hospital characteristics, the involvement in teaching activities has the most significant effect on their quality measures. However, there is no association between conversions and teaching status in our sample. In fact there are only two FP hospitals that have teaching status, one of which is a non-converting hospital and the other has converted from NFP status, but kept its teaching affiliations after conversion.} Market-specific characteristics are not considered in the analysis. First, because the time-invariant and general time trends in market characteristics are respectively captured by the hospital fixed effects and year dummies, and the immediate effect of
conversions on market shares does not seem to be significant.\textsuperscript{13} Secondly, the market share of FP hospitals in California remained almost constant over the study period.

\textit{Patient-level data}

Hospitalizations of California’s elderly patients for the following two diagnostic categories have been chosen: acute myocardial infarction (AMI), and congestive heart failure (CHF).\textsuperscript{14} In each case the sample contains all the patients of 65 years of age and older\textsuperscript{15}, hospitalized with the corresponding condition as principal diagnosis.\textsuperscript{16} Elderly patients provide relatively more homogeneous samples not only regarding age-related risk factors, but also because of a single insurance coverage by Medicare. Moreover, the relatively high risk among these patients helps avoid the small-sample problem caused by rare outcomes such as mortality.

The choice of diagnoses is also based on the variety of treatment methods available. One can expect a higher variation in hospital quality for cardiac diseases whose

\textsuperscript{13} My preliminary analyses (not included in the paper) show that an approximately constructed Herfindahl index based on county borders has no significant effect in mortality regressions as long as hospital fixed effects and year dummies are included in the model.

\textsuperscript{14} Four other diagnostic groups, malignant lung cancer, hypertensive heart disease, diabetes mellitus, and hospitalizations due to motor vehicle traffic accidents, were also studied using a similar methodology. However, these samples did not show any significant ownership effects, and are excluded from the paper to avoid unnecessary repetition.

\textsuperscript{15} I excluded a few patients older than 99 years.

\textsuperscript{16} The corresponding codes according to the International Classification of Diseases, 9\textsuperscript{th} version, Clinical Modification (U.S. Department of Health and Human Services) are as follows: AMI: 410.xx, CHF: 428.0, 402.x1, 398.91, 404.x1, and 404.x3.
treatment is chosen from a relatively wide range of procedures. There has been a great amount of innovation in the treatment of cardiac diseases in general and CHF in particular (Braunwald and Bristow, 2000). Since the main measure of quality is based on the in-hospital mortality, the diagnoses are chosen from the most important causes of death. According to the California mortality data, AMI and CHF are ranked among the most deadly diseases in California and throughout the US.\(^\text{17}\)

Table 3 gives the distribution of the patients and a descriptive summary of some of the features of hospitalizations by sector. The size of the samples varies from about 252,000 for the AMI group to 486,000 for CHF patients. NFP hospitals have the largest share (about 80 percent) of hospitalizations in private hospitals. FP hospitals have the highest mortality rate and the most severely ill cases (according to reported APR-DRG classification) in AMI sample and the lowest mortality and least severe case-mix in CHF group. These numbers also indicate that FP hospitals attract older patients. NFP hospitals have the highest rate of ER admissions and the longest hospitalizations.

The selection patterns observed in table 3, suggest that an unbiased estimation of ownership effects requires controlling for severity variations across hospitals. The risk factors considered in this paper include demographic covariates like age, gender, race (black/non-black), and ethnicity (Asian and Hispanic groups). Age is considered as five age groups: 65 years to 69, 70 to 74, 75 to 79, 80 to 84, and 85 and older. I also control for the interaction terms of race and gender with age groups. Moreover, a severity index is

constructed for every patient based on the APR-DRG classification. Finally, in the case of CHF sample where the diagnosis consists of four main categories, these categories are identified according to the first three digits of the principal diagnosis ICD-9-CM code and are taken into account using three binary indicators.

The APR-DRG measure of severity has been shown to be a powerful predictor of mortality. However, this measure is not directly used as a risk-adjustment factor. First, since it includes all the relevant diagnoses reported at discharge, regardless of whether they are developed before or after admission, it may include some “preventable” complications as well as “natural” comorbidities. Secondly, given that the Medicare reimbursement system is based on the patient’s diagnosis group, hospitals have an incentive to over-report complications. This problem, known as upcoding or DRG creep, may occur differently among hospitals with different ownership status.

---

18 APR-DRG is a system of classification of diseases with severity categories, patented by 3M Health Information Systems. This severity measure is not available for most of the discharges that occurred in 1990 and 1991. APR-DRG system defines the severity as the "extent of physiologic decompensation or organ system loss of function". Using information like principal diagnoses, procedures, multiple comorbidities, and age, it provides four severity-of-illness and risk-of-mortality subclasses within each DRG (Diagnosis-Related Group). See www.3Mhis.com and the 3M’s APR-DRG Software’s brochure.

19 See Romano and Chan (2000) for evidence regarding AMI patients.

20 For instance, Medicare reimbursements increase about 40% if an AMI patient has CHF complication (Psaty et al., 1999). Silverman and Skinner (2001) and Psaty et al. (1999) provide some evidence of over-reporting the severity of illness by hospitals. See also Foundation for Health Care Quality (1997) section 2.

21 Studying Medicare inpatient claims between 1989 and 1997 for pneumonia patients, Silverman and Skinner (2001) provide evidence suggesting that upcoding is more common among FP hospitals and also those NFP hospitals that converted to FP type.
paper, a severity index based on the APR-DRG classification is used. This index measures the difference between the APR-DRG severity measure of the patient and the average severity of patients within the same hospital-year-diagnosis group. Since this measure only represents the variation within hospital-year, differential upcoding cannot create any bias in the estimation of conversion effects.

*Measures of quality*

One of the most commonly used outcome measures of quality is the risk-adjusted in-hospital mortality. There are several validation studies suggesting that adjusted mortality rates can be used as a measure of hospital quality. Thomas et al. (1993) studied the in-hospital mortality rates for ten diagnostic groups of patients separately. For many but not all of these groups, the results showed a significant relationship between risk-adjusted in-hospital mortality and the hospital's quality as evaluated by peer reviews based on explicit and implicit process criteria. The strongest evidence of validity was obtained for cardiac diseases, which may suggest less selection for this kind of patients. Kahn et al. (1990) found similar results using mortality rates 30 days after admission. Significant relationship of risk-adjusted 30-day mortality and several process measures of quality was found in four out of five examined conditions.

Based upon these studies, the risk-adjusted in-hospital mortality probability is adopted as the main measure of quality in this paper. Like most other health outcomes that potentially have some information about hospital quality, mortality is a rare outcome and sometimes takes a long time to manifest, making its measurement difficult. Especially since the hospitals have some discretion on discharging the patients, the differences in
hospitals’ discharge/transfer policies may distort the in-hospital mortality from the “real” mortality risk. However, this issue seems to be relatively insignificant for cardiac diseases, which generally show a high correlation between in-hospital and long-term mortality. For instance, Rosenthal et al. (2000) find a strong correlation between 30-day (post-admission) mortality rates and in-hospital death rates for a sample of 13,800 CHF patients. They also provide evidence that the small differences in hospitals ranking caused by replacing in-hospital death rates by 30-day mortality rates are not resulted from the differences in hospital discharge practices.

As the numbers in table 3 indicate, the selected diagnoses have relatively high in-hospital death rates. Moreover in both groups, a relatively large part of deaths occur in acute-care hospitals. For instance during 1998 in California, 29.1% of 17,422 deaths caused by AMI and more than half of deaths caused by CHF occurred in short-term hospitals. However, these arguments are not perfectly satisfying. I therefore study the robustness of the results to potential differences in discharge and transfer practices across hospitals. This issue is discussed in more detail later.

Another outcome measure used in this paper is the risk-adjusted probability of early re-admission of AMI patients following discharge from a hospital. Usually re-admission within a short period (typically one month) after an initial discharge is considered as an undesired outcome that could be avoided by the original provider (Thomas and Holloway, (1991) and Carey and Burgess (1999)). In some cases re-admission within longer periods of time was used as an indicator of poor quality (Cutler, 1995). However, most of preventable readmissions occur within 10 days of a previous
discharge (Frankl et al., 1991). Several authors have found that the variations in re-admission probability are related to patient’s clinical conditions rather than hospital quality (cf. Thomas and Holloway (1991), Thomas (1996), and Ludke et al. (1993)). However, a re-admission for an AMI patient may imply another heart attack, thus a significant increase in patient’s mortality risk. The re-admission measure used in this paper is based on unscheduled re-hospitalizations with AMI as the principal diagnosis within one, two and three months after an initial discharge.22

4. Methods

The empirical model used in this paper can be formulated as follows:

\[ m_{ijt} = \beta X_{ijt} + \gamma Z_{jt} + \tau Y_t + \lambda_j + \epsilon_{ijt} \]  \hspace{1cm} (1)

where \( m_{ijt} \) is the quality indicator of patient \( i \) hospitalized in hospital \( j \) in year \( t \). The quality indicators are binary variables representing the patient’s mortality outcome or whether the patient was readmitted after discharge.

\( X_{ijt} \) is the vector of patient’s characteristics including five age groups, gender, race (black/non-black), two dummies for ethnicity (Asian, Hispanic) and the pair-wise interactions of age groups with gender and with race. This vector also includes a constructed severity index as defined in the previous section, as well as three additional dummies for CHF sample, which represent its main diagnostic sub-categories. \( Y_t \) is the vector of year dummies and \( \lambda_j \) is the hospital-specific fixed effect. Finally \( \epsilon_{ijt} \) represents

22 I also considered re-admissions within six months. The results were very similar and therefore not reported.
an \( i.i.d. \) random error that represents the unobserved heterogeneities among patients, hospitals, and years.

\( Z_{jt} \) is the vector of hospital status. This vector includes four conversion indicators that represent the state of hospital with respect to the conversion year. The coefficients of these variables measure the effect of conversions on quality. The first two of these variables are post-conversion dummy variables, NFP-to-FP and FP-to-NFP. Each of these dummies is set to one if the hospital has gone through the specified conversion in a previous year, and zero otherwise. The other two variables are linear trends (\( t_{\text{NFP-to-FP}} \) and \( t_{\text{FP-to-NFP}} \)) that measure the number of years before the conversion. These variables are set to zero for the conversion year and all the following years and take negative values equal to the number of years before conversion. The trend variables are also set to zero for the hospitals whose ownership status was stable throughout the sample period (no conversion).

For instance if a hospital has converted from NFP to FP status in 1993, the NFP-to-FP dummy for that hospital is one in 1993 and all the following years and is zero for all the years prior to 1993 and the trend variable \( t_{\text{NFP-to-FP}} \) for the same hospital takes 0, -1, -2 and -3 respectively in 1993, 92, 91 and 90 and zero for all other years. The conversion year is the omitted year or the baseline. The negative values are chosen to ensure that the positive (negative) values of the trend coefficients represent the annual growth (decrease) in the dependent variable prior to conversion.

Since the effect of risk factors differs across different health conditions, equation (1) is estimated separately in the two diagnostic groups. The standard errors are corrected
for the correlation of errors within hospital-year groups.\textsuperscript{23} The least squares method is used to estimate the model.\textsuperscript{24} This method may seem inconsistent with the dichotomous dependent variable. However, it should be noted that insofar as hospital-level effects are concerned, the model in equation (1) can be integrated to an equivalent aggregate model with hospital-years as its observation units. The dependent variable can thus be considered as an aggregate mortality probability that is specific to hospital-year groups.\textsuperscript{25}

\textit{Patient selection}

Patient level data can be used to estimate hospital-specific measures of quality. However, these measures are affected by a variety of confounding factors such as caseload characteristics. Patients with different severity may favor hospitals in one sector over another. Hospitals may also have different incentives in targeting certain groups of patients or avoiding “costly” patients to make more profits. An unbiased estimation of

\textsuperscript{23} This correction is done by clustering the sample in hospital-year groups using cluster command in Stata. In this method the errors are only required to be independent across groups and can be correlated within groups. Consequently, the variations within groups contribute little to the estimation precision. The standard errors are therefore more realistic than those obtained with the independence assumption, which may be under-estimated. See Moulton (1990) for an illustration of the downward bias in standard errors in grouped data and Rogers (1993) for more details on clustering technique. Our estimations show however that clustering does not change the estimated standard errors much. This result is consistent with Moulton’s contention that the problem does not arise in a fixed effects model (see Moulton (1986)).

\textsuperscript{24} The advantage of the least square method (compared to Logit or Probit) is in that no distribution assumption is imposed on the error term.

\textsuperscript{25} Note that the usual heteroscedasticity of OLS estimators with dichotomous dependent variables does not arise here because the errors across hospital-year groups do not have a dichotomous nature and the errors can be correlated within groups.
ownership effects on hospital performance requires a sufficient adjustment for the unobserved risk factors that potentially vary across different sectors.

To the extent that patients go to the closest hospital and hospital location does not vary with ownership changes, hospital fixed effects ($\lambda_j$) can capture the selection effects. The emergency nature of heart diseases especially diagnoses like AMI can help in this regard. Similarly, patients admitted through ER are less affected by selection. Comparing the results between such patients and the whole sample can help identify the direction of selection biases.

There is a possibility that certain types of hospitals get rid of their sickest patients by premature discharges or transfers to other hospitals. In this case the mortality rates of such hospitals will be biased downward. But controlling for the length of hospital stays reduces this bias. Comparing specifications with and without controlling for the hospital stays allows to understand if the mortality differences are due to different lengths of hospitalization. For instance, suppose that high-quality hospitals, say NFP ones, have systematically longer hospitalizations for risky patients. In this case these hospitals’ mortality will be biased upward if the length-of-stay is not taken into account. On the other hand, since the length-of-stay is endogenous, controlling for it will result in an endogeneity bias resulting an underestimation of the mortality in NFP hospitals. The two specifications can therefore provide upper and lower bounds of the potential bias associated with differences in hospital stays.
Pre- and post-conversion effects

The first hypothesis studied in this paper posits that NFP status is associated with a relatively high quality of service. In this case the quality rises after conversion from FP to NFP and deteriorates by a similar amount because of conversion from NFP to FP form. This hypothesis can be tested by comparing the coefficients of the two binary indicators representing post-conversion states (NFP-to-FP and FP-to-NFP). Under the symmetry hypothesis these coefficients must be opposite but similar in absolute value.

Another important question concerns the quality changes prior to conversion. Conversions are mostly a consequence of the sale of the hospital to a new owner. Such decisions are usually made a few years before the actual transactions occur. One can therefore expect that the converted hospitals have gone through some changes before the conversion. These pre-conversion changes may affect the hospital quality in different ways. For instance, a hospital that is subject to financial problems and perhaps to a deteriorating quality is more likely to be taken over by other firms and eventually convert to another status. In this case if the pre-conversion effects are not taken into account, the estimations may suggest that the quality has actually fallen after the conversion even though the hospital may have actually improved.\textsuperscript{26} Moreover, one may expect that NFP and FP hospitals differ in the way they cope with financial difficulties: While a FP hospital may lower the quality facing such problems, a NFP one may want to maintain a high quality of care. Financial problems may also arise if a NFP hospital decides to raise

\textsuperscript{26} Note that if the pre-conversion effects are not taken into account, the estimated effect of conversion is based on the difference between average quality measures before and after conversion.
its quality of service. After conversion the new managers/owners may decide to lower the quality. In such a scenario, quality of the NFP hospital may have substantially increased before conversion, and conversion has a declining effect on quality. However, if the pre-conversion changes are neglected the estimations may show an improvement in quality because of conversion.

Potential quality changes prior to conversion are captured by the two linear trends ($t_{NFP-to-FP}$ and $t_{FP-to-NFP}$) as defined earlier. These trends approximate the pre-conversion change in quality with a constant annual growth rate starting long before conversion. However, one may argue that the pre-conversion changes may be limited to only a few years before conversion. It is however difficult to specify the outset of these changes. In order to avoid an arbitrary starting year, I used a linear approximation throughout. Given the relatively small number of conversions in the sample, the linear approximation is the best possible approximation.\footnote{One can argue that these changes may have a non-linear form. I added the squares of these trends to the model. None of the second-order terms showed any significant effect. I also tried an alternative with 8 year dummies for any specific year before conversion. However, because of the large number of variables and relatively small number of observations in each group, virtually all the coefficients were statistically insignificant.} Moreover, for the following reasons this approximation does not affect the main effects of conversions. First, as we go back to the starting year (1990) the number of observation points decreases quite rapidly, thus a relatively low weight in the estimations. For instance, only for a fifth of converted hospitals the sample covers eight years before conversion. Secondly, although a linear approximation might underestimate the potentially large effects occurring right before conversion, it can give
an overall picture of the pre-conversion changes and does not bias the post-conversion effects represented by the two post-conversion indicators.

5. Results

Mortality outcomes

The estimation results for AMI and CHF mortality outcomes are respectively given in tables 4 and 5. In each group, the results are shown in two panels one for the entire sample (columns I, II and III) and one for the subsample of patients admitted through ER (columns IV, V and VI). In columns I and IV the changes before conversion are neglected and in columns III and VI the average length of hospital stay is included in the model. The first observation is that the symmetry hypothesis is strongly rejected in all specifications. That is, conversions in opposite directions do not have opposite effect on quality. Secondly, as expected the effect of severity indexes is positive and highly significant.28

The first columns in both tables suggest that a conversion from NFP to FP does not have any significant effect on mortality whereas a FP to NFP conversion results in a significant increase in mortality incidence. These results may reflect at least partially, the pre-conversion changes. The results listed in column II (table 4) show that after controlling for a linear trend in mortality before conversion, the results are reversed suggesting that conversion to FP status raises the mortality of AMI patients. This result is

28 A further analysis (not reported here) shows that excluding the severity deviations does not affect the results but as expected decreases the model’s explanatory power, reflected in a significantly lower R².
consistent with Shen (2002)’s findings that the conversions to FP status resulted in an increase in the mortality incidence of AMI patients.29 The results in column II also suggest the hospitals that converted from FP to NFP form have experienced a gradual increase in AMI mortality before conversion. This result is consistent with the scenario that the FP hospitals lower the quality of service faced with financial difficulties, but the NFP hospitals tend to maintain the quality until they are taken over by a FP firm, which can lower the quality after the conversion. The results of CHF sample (table 5) indicate however, that controlling for pre-conversion changes does not change the initial estimation results (compare columns I and II). These results suggest that the conversions from FP to NFP had a negative effect on quality.

The estimation results with control for the length-of-stay (LOS) are listed in column III (tables 4 and 5). These results indicate that including the LOS does not change the results significantly, suggesting that the results are not driven by systematic differences in hospitals’ discharge practices. Notice that if the estimated mortality differences were due to differences in LOS, they should have decreased after controlling for hospital stays and this is not the case. The results show that the LOS has a negative effect on mortality among AMI patients (table 4), but a positive effect on the CHF mortality (table 5). It should be noted that there is obviously a negative correlation between LOS and the in-hospital death probability.30 On the other hand, the hospital stays may represent part of the unobserved severity of cases, thus have a positive effect on

---

29 Shen (2002) used a national sample including 300 hospitals that changed ownership between 1985 and 1996.

30 Patients who die do not stay in the hospital, resulting in a mechanical negative relation.
mortality. This difference between AMI and CHF sample may be related to the fact that the in-hospital mortality in the AMI case is more than twice as that of CHF (see table 3). The high frequency of death outcomes may cause a relatively high negative correlation with the LOS.

Comparing the results between the entire sample and the sub-sample of ER patients can help identify the direction of potential biases due to patient selection. The results in both tables 4 and 5 (compare columns II and V) indicate that focusing on ER patients does not significantly change the estimation effects of conversion. If the higher (or lower) mortality rates were only because of the differences in case-mix severity, one can expect that the estimated differences are lower in cases that are less affected by selection (like ER patients). However, the results indicate that the estimated effects rather increase by focusing on ER admissions, shown by the slightly higher absolute values in column V compared to column II, suggesting that the unobserved severity factors may lead to an underestimation of quality differences. This result is consistent with the suggestive evidence in the previous literature that more severe patients are more likely to choose higher-quality hospitals.

Table 4 also shows that if we focus on the ER sample, the NFP hospitals that converted to FP form shows a gradual decrease in AMI mortality rates before conversion (column V). This result may suggest that the FP firms are attracted in buying high-quality NFP hospitals in order to make profits by reducing the quality after the conversion.
Re-admission rates

The probability of readmission among AMI patients is analyzed using re-hospitalization with AMI as a principal diagnosis within 1, 2, and 3 months after an initial discharge. The estimation results are given in table 6. As expected the effect of severity measure is positive. These results suggest that patients admitted to hospitals that converted to NFP status are more likely to be re-hospitalized after an initial treatment. Hospitals that converted to FP status show a similar change but the effects are statistically insignificant in most cases. However, the extremely low values of R-square in these regressions indicate that the re-admission probabilities are influenced by a relatively large number of unobserved factors. Due to the low explanatory power of this model, these results do not appear to be conclusive. This result is consistent with the evidence provided by Thomas (1996) and Ludke et al. (1993) who conclude that re-hospitalization probability is affected by individual patients’ clinical conditions and cannot adequately represent the hospital quality.

6. Conclusions

Between 1990 and 1998 California has witnessed 11 conversions from FP to NFP status and 15 conversions from NFP to FP form. The quality changes before and after conversions have been analyzed. The estimations are based on models that control for possible selection biases by including hospital fixed effects and controlling for the relative severity of patients. The samples are restricted to elderly patients who are automatically eligible for Medicare benefits.
The first measure of quality is the in-hospital mortality of AMI patients. The results suggest that the converted hospitals have gone through a gradual change in quality before conversion. The NFP hospitals that converted to FP status show a slightly falling AMI mortality rate before conversion, followed by a significant increase after conversion. Conversely, the FP hospitals that have converted to NFP status show a gradual growth in AMI mortality outcomes before conversion, which slows down after conversion. The observed changes in AMI mortality before and after conversions are consistent with the contention that the NFP hospitals have a commitment to quality. Assuming that the conversions are driven by financial difficulties, these results indicate the FP and NFP hospitals react differently to such problems: While FP firms are willing to lower the quality the NFP hospitals are likely to maintain their high quality of service. These results also provide suggestive evidence that the FP firms are more interested in taking over NFP hospitals that have a more than average quality of service, thus more profitable through possible reduction of quality.

The above results are not confirmed by the other two measures of quality, that is the in-hospital mortality rate of CHF patients and the re-admission probability of patients treated for AMI. The CHF mortality has been found to increase after conversion from FP to NFP status. These results suggest that conversions may have different effects on various aspects of hospital quality. In the case of AMI re-admission rates, the model’s explanatory power is too low for any conclusive statement.

The estimations are not sensitive to whether or not the average length of hospital stay is included in the model, which suggests that the estimated differences in mortality
are not driven by potential differences between hospital types in hospital discharge/transfer policy. Comparing the results between the entire sample and the sample consisting of the ER admissions indicates that the mortality differences may be slightly underestimated by patient selection. However, the relatively small differences suggest that as long as the hospital fixed effects are included, the selection bias does not considerably affect the results.

This study provides some evidence that the converted hospitals may be subject to quality changes before conversion. Controlling for these changes is crucial for an unbiased estimation of conversion effects. While the results suggest that some of the public concern over conversions to FP form may be warranted, this paper’s general conclusion is that hospital quality is a complex multi-dimensional concept, and is unlikely to be uniformly affected by hospital ownership status.

Acknowledgements

I am grateful to Janet Currie, Bentley MacLeod, Geert Ridder and Massimo Filippini for valuable comments and many discussions. I also wish to thank the editor and two anonymous referees for their helpful suggestions. The data were provided by the California Office of Health Planning and Development and the UCLA Institute for Social Science Research, which is gratefully acknowledged. I am indebted to the National Bureau of Economic Research for their financial support through dissertation fellowship 25-2154-02-0-43-003. I am solely responsible for all the remaining errors and the views expressed in the paper.
References

Arrow, Kenneth (1963) “Uncertainty and the welfare economics of medical care”


Table 1 - Share of California acute-care hospitals by sector in 1990 and 1998

<table>
<thead>
<tr>
<th>Year</th>
<th>FP</th>
<th>NFP</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989-90</td>
<td>20.3</td>
<td>63.3</td>
<td>16.4</td>
</tr>
<tr>
<td>1997-98</td>
<td>20.5</td>
<td>64.8</td>
<td>14.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of hospitals</th>
<th>1989-90</th>
<th>1997-98</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989-90</td>
<td>134</td>
<td>238</td>
<td>90</td>
</tr>
<tr>
<td>1997-98</td>
<td>107</td>
<td>233</td>
<td>70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Average size (number of beds)</th>
<th>1989-90</th>
<th>1997-98</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989-90</td>
<td>124</td>
<td>217</td>
<td>148</td>
</tr>
<tr>
<td>1997-98</td>
<td>150</td>
<td>217</td>
<td>163</td>
</tr>
</tbody>
</table>

Table 2 - Conversions in California acute-care hospitals between FP and NFP forms

<table>
<thead>
<tr>
<th>Fiscal year starting from the end of June</th>
<th>FP to NFP</th>
<th>NFP to FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>90-91</td>
<td>11 (1672)</td>
<td>15 (3480)</td>
</tr>
<tr>
<td>91-92</td>
<td>152</td>
<td>232</td>
</tr>
<tr>
<td>92-93</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>93-94</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>94-95</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>95-96</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>96-97</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>97-98</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

- Total number of beds is given in brackets. Hospital size is considered as the average number of beds.

Table 3 - Descriptive summary of hospitalizations in private acute-care hospitals

<table>
<thead>
<tr>
<th>Diagnostic group:</th>
<th>AMI</th>
<th>CHF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of admissions:</td>
<td>249,332</td>
<td>482,235</td>
</tr>
<tr>
<td>Distribution of admissions (%)</td>
<td>For-Profit: 17.8</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 82.2</td>
<td>79.0</td>
</tr>
<tr>
<td>Average in-hospital death rate (%) by status and year:</td>
<td>For-Profit: 14.5</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 13.1</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>1990: 15.9</td>
<td>7.34</td>
</tr>
<tr>
<td></td>
<td>1998: 11.4</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>Overall: 13.3</td>
<td>5.79</td>
</tr>
<tr>
<td>Average age (years):</td>
<td>For-Profit: 76.2</td>
<td>78.7</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 76.0</td>
<td>78.4</td>
</tr>
<tr>
<td></td>
<td>Overall: 76.0</td>
<td>78.5</td>
</tr>
<tr>
<td>Percent of patients with extreme or major severity categories:</td>
<td>For-Profit: 46.7</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 45.2</td>
<td>38.0</td>
</tr>
<tr>
<td></td>
<td>Overall: 45.5</td>
<td>37.8</td>
</tr>
<tr>
<td>Percent of admissions through ER:</td>
<td>For-Profit: 69.0</td>
<td>60.3</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 70.5</td>
<td>67.9</td>
</tr>
<tr>
<td></td>
<td>Overall: 70.2</td>
<td>66.3</td>
</tr>
<tr>
<td>Average length-of-stay (days):</td>
<td>For-Profit: 5.97</td>
<td>5.52</td>
</tr>
<tr>
<td></td>
<td>Not-For-Profit: 6.39</td>
<td>5.62</td>
</tr>
<tr>
<td></td>
<td>Overall: 6.31</td>
<td>5.60</td>
</tr>
</tbody>
</table>
Table 4- Mortality regressions (Acute Myocardial Infarction patients)

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>Patients Admitted through ER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Converted from NFP to FP</td>
<td>.0088 (.0074)</td>
<td>.019* (.010)</td>
</tr>
<tr>
<td>Converted from FP to NFP</td>
<td>.025** (.012)</td>
<td>.0074 (.011)</td>
</tr>
<tr>
<td>Time trend before NFP to FP conversion</td>
<td>- .0028 (.0018)</td>
<td>- .0027 (.0018)</td>
</tr>
<tr>
<td>Time trend before FP to NFP conversion</td>
<td>.0064** (.0023)</td>
<td>.0083** (.0030)</td>
</tr>
<tr>
<td>Severity deviation</td>
<td>.13** (.0012)</td>
<td>.13** (.0012)</td>
</tr>
<tr>
<td>Length-of-stay (days)</td>
<td>-</td>
<td>- .012** (.0004)</td>
</tr>
<tr>
<td>R-square</td>
<td>.128</td>
<td>.128</td>
</tr>
<tr>
<td>Sample size</td>
<td>249,332 hospitalizations</td>
<td>1,522 hospital-years</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets.
- ** indicates significant at 5% level.
- * indicates significant at 10% level.
- Standard errors are clustered in hospital-year groups.
- The time trends are compared to the conversion year for each hospital.
- Hospital fixed effects, year dummies and patients demographics (5 age groups, gender, race and interaction of age groups with race and gender dummies) are included in the model but not shown in the table.
Table 5- Mortality regressions (Congestive Heart Failure patients)

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>Patients Admitted through ER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Converted from</td>
<td>-.0004</td>
<td>-.005</td>
</tr>
<tr>
<td>NFP to FP</td>
<td>(.0040)</td>
<td>(.0049)</td>
</tr>
<tr>
<td>Converted from</td>
<td>.015**</td>
<td>.018**</td>
</tr>
<tr>
<td>FP to NFP</td>
<td>(.0047)</td>
<td>(.0055)</td>
</tr>
<tr>
<td>Time trend before</td>
<td>-.000006</td>
<td>-.0008</td>
</tr>
<tr>
<td>NFP to FP conversion</td>
<td>- (         )</td>
<td>(.0008)</td>
</tr>
<tr>
<td>Time trend before</td>
<td>-.0012</td>
<td>-.0013</td>
</tr>
<tr>
<td>FP to NFP conversion</td>
<td>- (         )</td>
<td>(.0012)</td>
</tr>
<tr>
<td>Severity deviation</td>
<td>.074**</td>
<td>.074**</td>
</tr>
<tr>
<td></td>
<td>(.0009)</td>
<td>(.0009)</td>
</tr>
<tr>
<td>Length-of-stay (days)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R-square</td>
<td>.0560</td>
<td>.0560</td>
</tr>
<tr>
<td>Sample size</td>
<td>482,235</td>
<td>1,537 hospitalizations</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets.
- ** indicates significant at 5% level.
- * indicates significant at 10% level.
- Standard errors are clustered in hospital-year groups.
- The time trends are compared to the conversion year for each hospital.
- Hospital fixed effects, year dummies and patients demographics (5 age groups, gender, race and interaction of age groups with race and gender dummies) are included in the model but not shown in the table.
### Table 6- Readmission of AMI patients

<table>
<thead>
<tr>
<th>Readmission within:</th>
<th>1 month</th>
<th>2 months</th>
<th>3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>Converted from NFP to FP</td>
<td>.0046</td>
<td>.0035</td>
<td>.0066</td>
</tr>
<tr>
<td></td>
<td>(.0039)</td>
<td>(.0059)</td>
<td>(.0055)</td>
</tr>
<tr>
<td>Converted from FP to NFP</td>
<td>.014**</td>
<td>.014*</td>
<td>.014*</td>
</tr>
<tr>
<td></td>
<td>(.0065)</td>
<td>(.0076)</td>
<td>(.0065)</td>
</tr>
<tr>
<td>Time trend before NFP to FP conversion</td>
<td>-</td>
<td>.0003</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0012)</td>
<td></td>
</tr>
<tr>
<td>Time trend before FP to NFP conversion</td>
<td>-</td>
<td>-.00003</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0009)</td>
<td></td>
</tr>
<tr>
<td>Severity deviation</td>
<td>.0013**</td>
<td>.0012**</td>
<td>.0026**</td>
</tr>
<tr>
<td></td>
<td>(.0005)</td>
<td>(.0005)</td>
<td>(.0006)</td>
</tr>
<tr>
<td>R-square</td>
<td>.0100</td>
<td>.0100</td>
<td>.0097</td>
</tr>
</tbody>
</table>

Average re-admission rates:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FP hospitals</td>
<td>4.5%</td>
<td>5.7%</td>
<td>6.4%</td>
</tr>
<tr>
<td>NFP hospitals</td>
<td>3.4%</td>
<td>4.6%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Overall</td>
<td>3.6%</td>
<td>4.8%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

- The sample includes 202,864 observations (1,503 hospital-years) consisting of AMI elderly patients with valid ID who were discharged alive after an initial hospitalization.
- Standard errors are given in brackets.
- ** indicates significant at 5% level.
- * indicates significant at 10% level.
- Standard errors are clustered in hospital-year groups.
- The time trends are compared to the conversion year for each hospital.
- Hospital fixed effects, year dummies and patients demographics (5 age groups, gender, race and interaction of age groups with race and gender dummies) are included in the model but not shown in the table.
ECONOMIES OF SCALE AND SCOPE IN MULTI-UTILITIES *

Mehdi Farsi, Aurelio Fetz, Massimo Filippini
Department of Management, Technology and Economics, ETH Zurich
Zurichbergstr. 18, CH-8032 Zurich, Switzerland
Tel. +41-44-632 06 50, Fax. +41-44-632 10 50
and
Department of Economics, University of Lugano
mfarsi@ethz.ch, afetz@ethz.ch, mfilippini@ethz.ch

Abstract

This paper explores the economies of scale and scope in the electricity, gas and water utilities. These issues have a crucial importance in the actual policy debates about unbundling the integrated utilities into separate entities, a policy which has often been supported by the ongoing reforms in the deregulation of network industries. This paper argues that the potential improvements in efficiency through unbundling should be assessed against the loss of scope economies. Several econometric specifications including a random-coefficient model are used to estimate a cost function for a sample of utilities distributing electricity, gas and/or water to the Swiss population. The estimates of scale and scope economies are compared across different models and the effect of heterogeneity among companies are explored. While indicating considerable scope and scale economies overall, the results suggest a significant variation in scope economies across companies due to unobserved heterogeneity.

JEL Classification: C33, D24, L11, L25, L94, L95

* This study has benefited from the financial support of the Swiss National Science Foundation through research grant 100012-108288 and also that of the State Secretariat for Economic Affairs (SECO), which is gratefully acknowledged. The authors also wish to thank Adonis Yatchew and two anonymous reviewers for their very helpful suggestions.
1. Introduction

In Switzerland’s energy sector, there is a certain tendency that local utility companies operate in both electricity and gas distribution as well as in the provision of water. Generally, this horizontal integration strategy allows the local multi-utility companies to save on costs by exploiting the economies of scope and to provide customers with an integrated set of services. As pointed out by Baumol, Panzar et al. (1982), economies of scope can result from sharing or joint utilization of inputs such as labor and capital. The distribution companies use similar equipment such as wires, overhead line and similar skills such as those required for network operation and maintenance. Synergies also exist in advertising and billing activities. Another source of cost savings is due to economies of ‘massed reserves’ (Waldman and Jensen (2001)). Multi-utility companies can make use of the same reserve capacity for maintenance and emergency repair activities.

During the last two decades the introduction of high levels of competition in the electricity and gas sectors of several EU-member countries has raised the general question of the necessity of unbundling services of utility companies. The regulatory reforms have been toward a separation of activities in the form of functional, legal or ownership unbundling, which are often believed to lower entry barriers and boost competition. However, the importance of the potential synergies through ‘horizontal’ integration has been recognized in the recent European regulatory recommendations (cf. DG Energy & Transport (2004)). An effective policy for unbundling multi-utilities, requires a reliable assessment of the scope economies and their variation with the company’s size and other characteristics.

Despite its policy importance, there are only a few studies that have studied the issue of scope economies in multi-utilities. In general, these studies suggest that
the scope economies are considerable at least for relatively small companies. However, the evidence as to the extent and statistical significance of the scope economies is rather mixed. A major difficulty in estimating scope and scale economies is the fact that utilities operate with different networks with various environmental and technical characteristics, which might induce various levels of synergies across different services. Many of these characteristics are not observed or difficult to measure. Such omitted variables could bias the estimation results. Moreover, the differences among companies could be beyond their variation in output and size. In fact, the strong heterogeneity among utilities operating in such different environments, suggests that a cost function with constant coefficients might be inadequate for a reliable analysis of scope economies.

Given that such network characteristics can be considered more or less constant over time, panel data can be used to account at least partially, for such heterogeneity and perhaps assess the potential biases. However, to our knowledge none of the previous studies in this field has used the advantages of panel data models to account for heterogeneity among companies.

Benefiting from a data set from 87 companies over a nine-year period, this paper applies two panel data models, a GLS model with random intercept and a random coefficient model, to estimate the scope and scale economies for individual firms. The variation across individual companies has been studied regarding both observed and unobserved heterogeneity. The results suggest significant scope and scale economies at most output levels and regardless of the variation in observed characteristics. The analysis also highlights the effect of unobserved heterogeneity across companies, suggesting that sophisticated econometric specifications such as random coefficients may be superior for analyzing the potential variation in scope and scale economies beyond the observed characteristics such as output patterns and customer density.
The rest of the paper is organized as follows. Section 2 presents the background along with a brief review of previous literature. The model specification and methods are presented in Section 3. Section 4 describes the data and Section 5 presents the regression results. The definition of scale and scope economies and their estimates are discussed in Section 6. The paper ends with a summary of main results and policy conclusions.

2. Background

The ongoing regulatory reforms in the energy sector in many countries have adopted measures toward unbundling public utilities into separate operations. The traditional models based on vertical integration in single sectors are often rejected. Especially, in the electricity sector the vertically integrated companies are generally required to unbundle the production, transmission and distribution functions. For instance, the directive 2003/54/EC of the European Parliament and of the EU Council of 26 June 2003 requires a legal and functional unbundling of the utilities operating in a single sector.

As opposed to ‘vertical’ unbundling that is generally being promoted by the ongoing reforms, the ‘horizontal’ unbundling of multi-utilities has remained an open question with less clear-cut recommendations. The unbinding guidelines released by the EU Directorate-General of Energy and Transport (DG Energy & Transport (2004)) state that the extent of management separation between activities related to different sectors “can only be decided on a case by case basis”. Further it is highlighted that a clear answer to this unbundling question requires a “balanced assessment of, on the one hand, the need for independence and, on the other hand, the interest of multi-utility operators to look for possible synergies.” While allowing certain flexibility in unbundling multi-utilities, this note
requires the policy makers to assess the extent of the economies of scope before taking policy decisions.

According the EU policy directive all the utilities with fewer than 100,000 customers can be exempt from any functional unbundling requirement. The distinction of small and large companies is based on the relative insignificance of scope economies in large companies that exploit scale economies. Such discriminative policies allow small companies to benefit from other synergies than scale economies. Since Switzerland is among the European countries with a large number of small companies in its energy sector, it provides a policy-relevant context for exploring the economies of scope. Moreover, although Switzerland does not belong to the European Union, the Swiss unbundling requirements upcoming in the near future, will probably reflect those discussed in the European directives. This study can provide the Swiss policy-makers with some insight concerning the effectiveness of similar regulatory measures in Switzerland.

Unbundling the services into separate functions allows a greater efficiency through stronger and more transparent competition that can be separately introduced in electricity, gas and water sectors. However, the implementation of the unbundling requirements will reduce the possibility of exploiting the economies of scope. The analysis of scope economies and its assessment across different companies can have important policy implications for the actual policy debates on the regulatory reforms in the Swiss gas and electricity sectors. Therefore, it is relevant for the Swiss federal authorities to identify if and to what extent multi-utility companies are able to use the scope and scale economies to reduce their costs in comparison to a group of single-utility companies. This question is in line with the issue of multiproduct natural monopoly raised by Baumol, Panzar et al. (1982), which has been applied to local public services. In the presence of
economies of scope a multiproduct firm is more economical than separate spe-
cialized firms. As first identified by Mayo (1984a), such economies are espe-
cially significant in relatively small companies. Therefore, the choice to exempt
small and medium-size companies from the unbundling requirements could be
sustained by economic arguments.

In the literature there are only a few studies on the economies of scope in multi-
utilities: Mayo (1984a), Chappell and Wilder (1986) and Sing (1987) in elec-
tricity and gas distribution, and Fraquelli, Piacenza et al. (2004) and Piacenza and
Vannoni (2004) in electricity, gas and water sectors. Mayo (1984a) and
Chappell and Wilder (1986) estimate a quadratic cost function for two cross
sectional data sets from the US electricity and gas distribution sectors. Mayo
(1984a) reports scope economies only for small companies, whereas Chappell
and Wilder (1986) conclude significant scope economies over most of output
ranges. Sing (1987), also using a cross-sectional data set including electricity
and gas distributors, estimates a generalized translog cost function with a Box-
Cox transformation for outputs. In addition to the factor prices of labor, capital
and fuel, he includes the customer density as an output characteristic. While
reporting diseconomies of scope for the sample mean Sing (1987) finds scope
synergies for certain output combinations, without any clear pattern with respect
to the outputs magnitude.

The relatively recent papers by Fraquelli, Piacenza et al. (2004) and Piacenza
and Vannoni (2004) use data from 90 Italian electricity, gas and water distribu-
tors over 3 years. However the data is pooled across the years and no panel data
models are applied. They compare different functional forms such as the trans-
log cost function with a small value transformation, the generalized translog, the
separable quadratic and the composite cost function introduced by Pulley and
Braunstein (1992). They conclude that economies of scope exist but their statistical significance can only be asserted over small outputs.

A summary of the above studies and their main results is presented in Table 1.

As we can see, panel data has hardly been utilized to date. The short panels used in the recent studies by Fraquelli, Piacenza et al. (2004) and Piacenza and Vannoni (2004) probably have not allowed the authors to account for unobserved heterogeneity and correlation in the error terms. Another interesting study is Yatchew (2000) who applied a semi-parametric model to a 3-year panel data set of Canadian electricity distributors. Focusing on scale economies that author uses an additional dummy variable to account for the economies of scope gained by joint distribution of water and electricity.

Table 1: Summary of previous empirical studies of multi-utilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional form</td>
<td>Quadratic and flexible fixed costs, quadratic</td>
<td>Quadratic</td>
<td>Generalized translog</td>
<td>Translog, generalized translog, separable quadratic and composite</td>
<td>Translog, generalized translog, separable quadratic, composite and general form (Pulley and Braunstein (1992))</td>
</tr>
<tr>
<td>Estimation method</td>
<td>OLS</td>
<td>OLS</td>
<td>SUR</td>
<td>NLSUR</td>
<td>NLSUR</td>
</tr>
<tr>
<td>Outputs</td>
<td>Electricity and gas distribution</td>
<td>Electricity and gas distribution</td>
<td>Electricity and gas distribution</td>
<td>Electricity, gas and water distribution</td>
<td>Electricity, gas and water distribution</td>
</tr>
<tr>
<td>Factor prices</td>
<td>Labor, fuel</td>
<td>-</td>
<td>Labor, capital, fuel</td>
<td>Labor, other inputs</td>
<td>Labor, other inputs</td>
</tr>
<tr>
<td>Other factors</td>
<td>-</td>
<td>-</td>
<td>Customer density</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Economies of scope</td>
<td>Exist only for small companies (+0.77%), for large companies diseconomies (up to -11.7%)</td>
<td>Exist over most of the output ranges, +12% for small, -10% for largest companies</td>
<td>Output combinations of both scope economies and diseconomies, no economies of scope for the mean output (-7.2%)</td>
<td>Exist, but significant only for companies producing less than the median output</td>
<td>Exist with all the models except with the translog cost function. For the median output between 16 and 64%</td>
</tr>
<tr>
<td>Economies of scale</td>
<td>Product-specific economies of scale for gas over all outputs, for electricity only for small companies</td>
<td>Global and product-specific economies of scale exist</td>
<td>Product-specific economies of scale for electricity, diseconomies for gas</td>
<td>All the models show economies of scale except the translog model</td>
<td>All the models show economies of scale except the translog model</td>
</tr>
</tbody>
</table>
Given that the energy distribution companies operate in strongly heterogeneous environments, accounting for firm-specific unobserved factors might change the estimates of scope and scale economies. The moderately long panel data set used in this study allows the use of panel data models that can account for such heterogeneity and assess their effects on the estimations.

Before turning to the model it is worth noting that the Swiss energy sector is a fragmented market characterized by a strong heterogeneity across the 3,023 communities. With a total of 940 electricity utilities, 124 gas companies and 2,995 water distributors Switzerland’s energy sector is characterized by its staggeringly large number of distributors with a prevalence of small and medium size companies throughout the 3,023 Swiss communities (cf. Dymek and Glaubitz (2003), VSG (2007) and Föllmi and Meister (2005)). Multi-utilities play an important role in all three sectors: The share of multiproduct utilities in the electricity and gas sectors is respectively about 35 and 75 percent of the total national consumption. With a roughly estimated share of 80 percent of the total national consumption, multi-utilities are also dominant in the water sector.1 In general multi-utilities tend to be active in all three sectors. The share of double-output utilities is relatively low (limited to a few percentage points), especially in the gas sector.

3. Model specification and estimation method

The model specification is based on a cost function with three outputs (electricity, gas and water). The model also includes a measure of the characteristic of

1 The numbers for electricity and gas are based on the data from 127 electricity distributors and 80 gas companies that respectively provide about 90% of electricity and gas consumption in Switzerland. The share in water distribution is estimated based on the available data from 95 companies that provide about 41 percent of the national water consumption.
the service area and three sector-specific linear time trends capturing technological changes. Moreover, four input prices are also included in the model. As in Sing (1987) customer density is introduced as a service area characteristic. This variable should capture, at least partially, the impact on costs of the heterogeneity of the service area of the companies. In fact, differences in networks and environments influence the production process and, therefore, the costs. Of course, we are aware that the heterogeneity of the service area cannot be summarized into one single variable. Unfortunately, the information is not available for all network and environmental characteristics. Thus, many of these characteristics are omitted from the cost function specifications. As we see later these omitted factors are represented by firm-specific stochastic components in the adopted panel data econometric models.

If it is assumed that the firm minimizes cost and that the technology is convex, a total cost function can be written as:

$$ C = C(q^{(1)}, q^{(2)}, q^{(3)}, w^{(1)}, w^{(2)}, w^{(3)}, w^{(4)}, r, \tau^{(1)}, \tau^{(2)}, \tau^{(3)}) $$  

(1),

where $C$ represents total costs; $q^{(1)}$, $q^{(2)}$ and $q^{(3)}$ are respectively the distributed electricity, gas and water during the year, $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ and $w^{(4)}$ are respectively the input factor prices for labor and capital services and the purchased electricity and gas; $r$ is the customer density measured by the number of customers divided by the size of the service area measured in square kilometers; and the sector-specific linear trends are represented by $\tau^{(1)}$, $\tau^{(2)}$ and $\tau^{(3)}$ respectively for electricity, gas and water sectors.

Following Baumol, Panzar et al. (1982) and Mayo (1984a) we use a quadratic cost function. This form has been considered as one of the most relevant options for estimating scope economies (Tovar, Jara-Diaz et al. (2007)). Unlike loga-
rithmic forms, this functional form accommodates zero values for outputs thus, allows a straightforward identification of scope economies. Although logarithmic functions could be used with an arbitrary small value transformation for zero values, it has been shown that this approach could result in large errors in the estimation of scope economies (Pulley and Humphrey (1993)). As in our case, many output values for electricity, gas and water distribution are zero, such estimation errors may lead to misleading conclusions about scope economies.

The choice of the quadratic functional form has been also in close relationship with the econometric specification possibilities for the available panel data that will be described later. In fact, unlike other functional forms, the quadratic functional form can be easily estimated with panel data models. For instance, the application of panel data models (especially the random effects models) in non-linear models such as Box-Cox or the composite model (Fraquelli, Piacenza et al. (2004), Mayo (1984b)) is not straightforward. Given the potential importance of the unobserved heterogeneity in the data we focused on the quadratic functional form that is readily adaptable to panel data models. Especially as the utilities operate in environments characterized by strong heterogeneity and given the fact that the integrated companies as well as specialized utilities are included in the data, the omitted variables could have an important effect that can be better accounted for in panel data models. By a similar argument we excluded the equation system approach with factor share equations as this approach cannot easily accommodate random effects specification.

One disadvantage of the quadratic form is that the linear homogeneity of the cost function in input prices cannot be imposed by parametric restrictions without compromising the flexibility of the functional form (Caves, Christensen et al. (1980)). A fairly common approach around this issue is the normalization of all monetary variables by one of the common factor prices referred to as nu-
meraire price (see Farsi, Fetz et al. (2007), Featherstone and Moss (1994) and Jara-Diaz, Martinez-Budria et al. (2003)). However, depending on which input factor is chosen as the numeraire, the normalized model has non-unique solutions that might result in certain discrepancy across the estimates.\(^2\) Considering this drawback, we favored the non-normalized version of the model that has a greater flexibility as well as a better robustness. Especially in the context of this paper, in which the main focus is on the output coefficients that determine the economies of scope and scale, imposing the linear homogeneity restriction does not appear to provide any added value into the analysis.\(^3\)

The adopted quadratic cost function using a random effects specification can be written as follows:

\[
C_{it} = \alpha_0 + \sum_m^M \alpha_m^w q_{it}^{(m)} + \frac{1}{2} \sum_m^M \sum_n^M \alpha_m^w q_{it}^{(m)} q_{it}^{(n)} + \sum_p^P \beta_p^w w_{it}^{(p)} D_i^{(p)} + \alpha^r r_{it} + \sum_m^M \gamma_m^{(m)} D_i^{(m)} + u_i + \varepsilon_{it}
\]

(2)

where superscripts \(m\) and \(p\) denote respectively, the number of products (1, 2, 3) and the number of input factors (1, 2, 3, 4), and subscripts \(i\) and \(t\) denote respectively the company and year. The stochastic terms \(u_i\) and \(\varepsilon_{it}\) represent respectively the firm-specific individual effects and the error term. The factor prices \(w\) and the density variable \(r\) are introduced in a linear way (following Mayo (1984a)). The dummy variables \(D_i^{(p)}\) take one if the corresponding input factor

\(^2\) Because of its additive form the obtained quadratic models are not equivalent. This is in contrast with multiplicative models such as translog in which normalization is perfectly invariant to the choice of the numeraire and equivalent to a single parametric restriction.
has been used in the production. These dummies, relevant only for electricity and gas prices, allow to exclude the corresponding term if the company does not distribute electricity or gas (see Isaacs (2006) for this approach). The linear trends $\tau_i^{(m)}$ are specific to the sector as each one of the sector might be subject to a different technological progress. Similarly, dummy variables $D_i^{(m)}$ represent the cases in which the company distributes the corresponding product (electricity, gas and water). Finally $\alpha_0$ is the intercept. The alternative specification would be a flexible fixed cost model as in Mayo (1984a) and Panzar (1989), which includes several intercepts depending on the sector or the utility’s output combination. We explored this possibility, but given that the estimated intercepts are not significantly different from each other, we favored the simpler model with a single intercept.

The quadratic form is a flexible functional form that can be considered as a second-order Taylor approximation of any arbitrary function around a local approximation point. In this paper following the commonly used approach in the literature (e.g. Jara-Diaz, Martinez-Budria et al. (2003)), the sample mean has been used as the approximation point. This normalization has been obtained by demeaning all the included explanatory variables (subtracting from their mean values). Therefore the intercept $\alpha_0$ captures the total costs of production at the sample mean.

---

3 This has been confirmed by a supplementary analysis (available upon request) in which we have considered normalizing the costs and input prices by the labor price. The results suggest no significant change as far as the scope and scale economies are concerned.
The above cost function has been specified as a random effect GLS model with:

\[ u_i \sim iid(0, \sigma_u^2) \]  

This model has a clear advantage over an alternative cross-sectional model that pools the data across companies, thus simplifies the firm-specific effects in a constant intercept. Using individual effects \( u_i \), the GLS model allows for certain variation among companies regarding the model’s intercept, that as pointed out by Jara-Diaz, Martinez-Budria et al. (2003), has an important effect on the estimates of economies of scope. The main assumption is that the random effects \( u_i \) are uncorrelated with the explanatory variables, a restriction that could be relaxed in a fixed-effects specification.

However, the reliability of fixed-effects estimators depends on the extent of within-company variations that is, the variation of costs and outputs of given companies over time. As Cameron and Trivedi (2005) pointed out, the fixed-effects approach has an important weakness in that the coefficients of explanatory variables are “very imprecise” if the variable’s variation over time is dominated by that across companies (between variation). The data used in this study show a relatively low within variation (variation over time) in some of the variables, especially, the ratios between the three outputs remain more or less constant within a given company. The extremely low variation in some of the vari-

---

4 We have also estimated an alternative random effects model with AR1 serial correlation. The results (available upon request) do not show any significant difference between the corresponding coefficients.

5 Such correlation might create ‘heterogeneity bias’ in the estimates (more on this later). The term ‘heterogeneity bias’ probably coined by Chamberlain (1982), has also been used for the bias due to ignoring variation of regression coefficients across individuals (e.g. Asteriou and Hall (2007)).

6 Johnston and DiNardo (1997) also show that the ‘attenuation’ bias due to measurement errors is exacerbated in the fixed-effects models depending on the fraction of the within variation due to ‘mismeasurement’ especially when the explanatory variables are correlated across time. In our case it is plausible that the reporting errors have a contribution in the observed within variations.
ables coupled with the presence of the second-order terms in the quadratic functional form also exacerbate the risk of multicollinearity, thus unreliable results.\(^7\)

Moreover, the fixed-effects estimators are strongly conditioned upon the companies included in the sample, thus not convenient for boundary predictions at output bundles with zero values that are required for the estimation of scope economies.\(^8\) In fact the definition of the economies of scope relies on a comparison of the company’s costs of producing all outputs with those of the same company with zero production in certain outputs. However, changes from positive output to zero output usually do not occur within a specific company. Therefore, the economies of scope can only be identified through the variations between a given company and other companies that are similar in all aspects but have little or zero production in those outputs. In the fixed-effect model such between variations are entirely captured by the company’s individual effect, thus excluded from the cost function. Considering the above discussion, we excluded the fixed-effect model and focused on the random effects framework. We recognize however, the limitation of the adopted models concerning the assumption that omitted factors are uncorrelated with the explanatory variables.

In the random effects model the unobserved firm-specific heterogeneity is accounted for by individual effects. These factors might be correlated with the explanatory variables, in which case the estimations might be affected by ‘het-

\(^7\) Following a referee’s suggestions we estimated several fixed-effects models. The results (available upon request) indicate that the estimates of the main output coefficients are quite sensitive to the included variables and occasionally counter-intuitive, suggesting that the within variation is not sufficient in order for the fixed-effects model to provide sensible results.

\(^8\) As pointed out by Hsiao (2003), while the fixed-effects model is more appropriate for conditional predictions for individuals, the random effects is a better specification for unconditional (population-
ergogeneity bias. One improvement over the GLS model in this respect could be obtained by including random coefficients for those explanatory variables. The variation of these coefficients should capture part of the correlation of the random intercept with the corresponding variables. Moreover, the unobserved firm-specific heterogeneity could also apply to marginal costs represented by the coefficients of the cost function. Therefore, we also estimate the cost function using a random coefficient (RC) model. In this model the three output coefficients, the intercept and the output characteristics are assumed to be random variables with a normal distribution across companies.

The quadratic cost function with the adopted random coefficient specification can be written as follows:

\[
C_i = \alpha_i^0 + \sum_{m}^{M} \alpha_i^m q_i^{(m)} + \frac{1}{2} \sum_{m}^{M} \sum_{n}^{M} \alpha_i^{mn} q_i^{(m)} q_i^{(n)} + \sum_{p}^{P} \beta_i^{p} y_i^{(p)} D_i^{(p)} + \alpha_i^r r_i + \sum_{m}^{M} \gamma_i^m \tau_i^{(m)} D_i^{(m)} + \varepsilon_i 
\]

(3),

where \( \alpha_i^m \sim N(\mu_{\alpha_i^m}, \sigma_{\alpha_i^m}^2) \), for \( m=0,1,2,3 \), and \( \alpha_i^r \sim N(\mu_{\alpha_i^r}, \sigma_{\alpha_i^r}^2) \). Similar to the GLS model, all the explanatory variables are normalized to their sample means. The above random coefficient model has been estimated using a simulated maximum likelihood method. The firm-specific parameters are estimated for individual companies as their conditional expectation.

See also Cameron and Trivedi (2005) and Verbeek (2004) for a discussion of this issue.

\(^{9}\) For a presentation of this model see Cameron and Trivedi (2005). See also Biørn, Lindquist et al. (2002) for an application of this model in the estimation the returns to scale among heterogeneous technologies.
The random coefficient model described above provides a relatively rich specification that allows for interaction of unobserved factors such as network characteristics with outputs and customer density. However, it has a shortcoming in that it imposes the normality assumption on the random intercept. Therefore, the choice of the best model between the two depends on the trade-off between refining the econometric specification against the distribution restrictions. As we see later, as far as the estimates of the economies scale and scope are concerned, the results are not sensitive to this choice. Another important issue is that the specification of random coefficients can be extended to other variables. The benefits of such extensions should be assessed against the entailed numerical difficulties as well as the interpretation problems.\footnote{Following the suggestion of a referee we estimated several alternatives in which the input prices especially capital price have also random coefficients. The results (available upon request) indicate that adding random coefficients to the model can cause convergence problems and numerical instability, otherwise, counterintuitive results that are difficult to interpret. These problems could be explained by the relatively large number of explanatory variables in the model and the fairly limited number of companies included in the data.}

4. Data

The unbalanced panel data set used for this analysis contains financial and technical information from 87 companies observed during the nine-year period between 1997 and 2005. The companies in the sample cover about 42% of total electricity, 67% of total gas and 22% of total water distribution in Switzerland. Among these companies, 33 are fully integrated and offer electricity, gas and water. 11 companies offer electricity and water, 3 companies distribute gas and water and 2 companies electricity and gas. The remaining companies are specialized companies from which 23 are active only in electricity distribution, 12 only in gas distribution and 3 only in water distribution. The presence of just a
few number of specialized water distributors could be considered as a drawback for the estimation of economies of scope. However, this limitation should be considered together with the fact that in a fair number of companies in the sample, the distribution of gas and electricity constitutes a small fraction of the total output.\textsuperscript{11}

The data were collected from the companies’ annual reports containing financial and technical information.\textsuperscript{12} As pointed out by Kaserman and Mayo (1991), the degree of vertical integration can have an important impact on costs, thus affecting the estimates of economies of scope. The problem does not arise in gas and water sectors, in which companies have a uniform level of integration with the generation section (fully integrated in the case of water and completely separate in gas companies). In order to abstract from the effect of vertical integration in electricity distribution, companies with more than 10% self-generation of total electricity distribution were excluded.

The variables for the cost function specification were constructed as follows. Total costs ($C$) are calculated as the total expenditures of the energy and water distribution firms in a given year. The outputs $q^{(m)}$ are measured by the total

\textsuperscript{11} The number of these companies depends on the units used for measuring the various outputs. For instance, if we choose the units such that the sample median values will have the same order of magnitude (GWh for gas and electricity $10^4$ cubic meters for water output) there are 14 companies whose water output is more than two third of their total output.

\textsuperscript{12} Information on the size of the firm’s distribution area is from the “Arealstatistik 2002” from the Federal Statistical Office and from the “Preisüberwacher”.
quantity delivered to the customers. The measurement units are GWh for electricity and gas and million cubic meters for water.\textsuperscript{13}

Input prices are defined as factor expenditures per factor unit. Labor price ($w^{(1)}$) is defined as the ratio of annual labor costs to the total number of employees as full time equivalent. As data on full time equivalent was not available for 40 companies and taking the number of employees including part time workers would underestimate the labor price, a correction was done by taking the mean with the labor price of the companies within the same canton. Following Friedlaender and Chiang (1983), the capital price ($w^{(2)}$) is calculated as residual cost (where residual cost is total cost minus labor and electricity and gas purchases) divided by the network length.\textsuperscript{14} For the multi-utilities, the prices were weighted by the share of the residual costs in each sector to the total residual costs in all sectors (see also Fraquelli, Piacenza et al. (2004) for this approach). The electricity and gas price is defined as the expenditures of purchasing the input factors divided by the amount purchased (in MWh).

Table 2 provides the sample’s descriptive statistics. All the costs and prices are adjusted for inflation using consumer price index and are measured in year 2000 Swiss Francs (CHF). As can be seen in the table, the sample shows a considerable variation in all three outputs.

\textsuperscript{13} The distributed gas is generally reported in energy units rather than volume units. Given that the gas distributors in Switzerland mainly use the same source of imported natural gas with a uniform quality, we do not expect that the change of measurement unit has any effect on the results.

\textsuperscript{14} More precise measures of capital stock and expenditures can be obtained from a perpetual inventory approach. Unfortunately such inventory data was not available.
Table 2: Descriptive statistics (622 observations)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit</th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>Total cost</td>
<td>CHF Million</td>
<td>1.52</td>
<td>35.7</td>
<td>79.0</td>
</tr>
<tr>
<td>$q^{(1)}$</td>
<td>Electricity distribution</td>
<td>GWh</td>
<td>0</td>
<td>115.4</td>
<td>405.9</td>
</tr>
<tr>
<td>$q^{(2)}$</td>
<td>Gas distribution</td>
<td>GWh</td>
<td>0</td>
<td>78.8</td>
<td>363.8</td>
</tr>
<tr>
<td>$q^{(3)}$</td>
<td>Water distribution</td>
<td>Million m$^3$</td>
<td>0</td>
<td>1.1</td>
<td>3.4</td>
</tr>
<tr>
<td>$w^{(1)}$</td>
<td>Labor price</td>
<td>CHF/ employee</td>
<td>75'575</td>
<td>103'610</td>
<td>104'863</td>
</tr>
<tr>
<td>$w^{(2)}$</td>
<td>Capital price</td>
<td>CHF/ km</td>
<td>8'165</td>
<td>26'421</td>
<td>34'018</td>
</tr>
<tr>
<td>$w^{(3)}$</td>
<td>Electricity price</td>
<td>CHF/ MWh</td>
<td>43.5</td>
<td>103.2</td>
<td>101.0</td>
</tr>
<tr>
<td>$w^{(4)}$</td>
<td>Gas price</td>
<td>CHF/ MWh</td>
<td>16.3</td>
<td>29.1</td>
<td>30.3</td>
</tr>
<tr>
<td>$r$</td>
<td>Density</td>
<td>Customers/ km$^2$</td>
<td>2.3</td>
<td>230.1</td>
<td>348.0</td>
</tr>
</tbody>
</table>

5. Results

The estimation results obtained from the GLS model are given in Table 3. These results show that the output and input price coefficients are highly significant and have the expected positive sign.

As expected, the effect of customer density (coefficient $\alpha^r$), is negative, showing that an increase in the customer density decreases costs. The coefficients of the linear trends suggest different technological progress across the three sectors. The results, while suggesting a cost decrease in the electricity networks, indicate a growth in operating costs in both gas and water sectors. These differences might also be related to the differences in the regulation systems for these sectors. It is interesting to note that although almost all public utilities are undergoing regulatory reforms, the electricity distributors have been subject to a relatively more advanced de-regulation process. However, the relative growth

---

15 The first official attempt for the de-regulation of the Swiss electricity market dates back to 2002.
of costs in gas and water networks might be related to the relatively higher age of these networks, thus a more accentuated need for new investments.

Another interesting observation is the considerable variation of the random intercept as reflected in the estimate of $\sigma_u$. The significant variation of the fixed costs across companies might be considered as a support for models with flexible fixed costs suggested by Mayo (1984a) and Panzar (1989). However, our additional estimations with a similar model but a varying intercept across different sectors suggest no statistically significant differences across sectors. This result combined with a relatively important variation in the random effects indicates that the variation in the fixed costs across companies might be mainly due to unobserved variations across companies. However, as we will see later from the random coefficient model’s results the GLS model could overstate the variation of intercept because it assumes constant slopes for all companies.
### Table 3: Regression results (GLS model)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha^1$ (Electricity output)</td>
<td>152'698 **</td>
<td>(3'318)</td>
</tr>
<tr>
<td>$\alpha^2$ (Gas output)</td>
<td>42'659 **</td>
<td>(4'210)</td>
</tr>
<tr>
<td>$\alpha^3$ (Water output)</td>
<td>2'266'445 **</td>
<td>(504'478)</td>
</tr>
<tr>
<td>$\alpha^{11}$</td>
<td>-22.33 **</td>
<td>(1.33)</td>
</tr>
<tr>
<td>$\alpha^{22}$</td>
<td>0.18</td>
<td>(1.54)</td>
</tr>
<tr>
<td>$\alpha^{33}$</td>
<td>-43'314 *</td>
<td>(22'532)</td>
</tr>
<tr>
<td>$\beta^1$ (Labor price)</td>
<td>132.75 *</td>
<td>(75.77)</td>
</tr>
<tr>
<td>$\beta^2$ (Capital price)</td>
<td>139.85 **</td>
<td>(33.78)</td>
</tr>
<tr>
<td>$\beta^3$ (Electricity price)</td>
<td>127'777 **</td>
<td>(52'794)</td>
</tr>
<tr>
<td>$\beta^4$ (Gas price)</td>
<td>562'209 **</td>
<td>(111'478)</td>
</tr>
<tr>
<td>$\gamma^1$ (Electricity sector)</td>
<td>-7'207.54 **</td>
<td>(2'973.91)</td>
</tr>
<tr>
<td>$\gamma^2$ (Gas sector)</td>
<td>-2'639'987 **</td>
<td>(331'928)</td>
</tr>
<tr>
<td>$\gamma^3$ (Water sector)</td>
<td>945'850 **</td>
<td>(390'922)</td>
</tr>
<tr>
<td>$\gamma^4$</td>
<td>1'544'447 **</td>
<td>(461'136)</td>
</tr>
<tr>
<td>$\alpha^0$</td>
<td>90'140'600 **</td>
<td>(1'926'850)</td>
</tr>
<tr>
<td>$\hat{\sigma}_u$</td>
<td>10'586'724</td>
<td></td>
</tr>
<tr>
<td>$\hat{\sigma}_e$</td>
<td>9'411'338</td>
<td></td>
</tr>
</tbody>
</table>

** and * indicate 5% and 10% significance level respectively. Standard errors are given in parentheses.

In the random coefficient model, it is assumed that the intercept and the first-order coefficients of output and customer density vary across companies. The random coefficient model was estimated with the simulated likelihood method using quasi-random Halton draws.\textsuperscript{16}

---

\textsuperscript{16} The number of draws has been fixed to 1000. The model was also estimated with several numbers of draws between 100 and 1'000. The results indicate that after 500 draws, the estimations become stable.
Table 4 lists the regression results obtained from the random coefficient model. The first important observation is that the estimated coefficients are slightly (but mostly not significantly) different from those obtained from the GLS model. However, the estimated standard deviations of the random coefficients are all statistically significant for electricity and gas output as well as the customer density. This suggests that there is a significant variation in the output and density coefficients across companies. As for the intercept the standard deviation shows a considerably lower value than that obtained from the GLS model ($\hat{\sigma}_u$ in Table 3), suggesting that ignoring the heterogeneity in slopes can result in an overestimation of the variations of the fixed costs across companies.

The random coefficient estimators can be used to estimate the conditional expectation of firm-specific coefficients. These estimates show that for the intercept and the output coefficients, all the coefficients are positive, while for the customer density coefficient, 13 companies (out of 87) have positive coefficients. This can be explained by the fact that as customer density increases, certain companies incur extra costs through congestion effects or some unobserved network characteristics. The estimates of the variances of the random effects in both models (Table 3 and Table 4) show that there is a considerable unobserved firm-specific heterogeneity. We will see later if and how ignoring this heterogeneity could affect the estimates of scale and scope economies.

---

17 We explored the possibility that the congestion effect might be related to some observed variables by including a square term for customer density and accounting for the network location in rural/urban areas. The results do not show statistical significant effect which could lead to any conclusive evidence in this regard.
Table 4: Regression results (Random coefficient model)

<table>
<thead>
<tr>
<th></th>
<th>Quadratic cost function (RC)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Electricity output)</td>
<td>162'889 ** (1'533)</td>
<td>15'652 ** (524)</td>
<td></td>
</tr>
<tr>
<td>(Gas output)</td>
<td>50'132 ** (1'639)</td>
<td>14'921 ** (882)</td>
<td></td>
</tr>
<tr>
<td>(Water output)</td>
<td>1'562'760 ** (184'820)</td>
<td>9'289 (48'903)</td>
<td></td>
</tr>
<tr>
<td>( \alpha_{11} )</td>
<td>-32.21 ** (0.95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{22} )</td>
<td>-0.63 (0.539)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{23} )</td>
<td>-12'262 (8'684)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{12} )</td>
<td>25.78 ** (2.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{13} )</td>
<td>-1'704 ** (154)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_{23} )</td>
<td>-399.68 ** (150.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Labor price)</td>
<td>126.56 ** (30.94)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Capital price)</td>
<td>128.91 ** (15.97)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Electricity price)</td>
<td>91'957 ** (23'154)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Gas price)</td>
<td>522'290 ** (64'763)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha' )</td>
<td>-3'829.95 ** (983.31)</td>
<td>14'200.7 ** (1'157.4)</td>
<td></td>
</tr>
<tr>
<td>(Customer density)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Electricity sector)</td>
<td>-2'488'370 ** (198'250)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Gas sector)</td>
<td>916'995 ** (424'060)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Water sector)</td>
<td>1'323'520 ** (432'471)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha^0 )</td>
<td>93'564'700 ** (682'477)</td>
<td>108'524 (355'020)</td>
<td></td>
</tr>
</tbody>
</table>

** and * indicate 5% and 10% significance level respectively. Standard errors are given in parentheses.

The estimation results presented in Table 3 and Table 4 can be used to compute the estimated of the economies of scale and scope. These results along with a formal description of the concepts will be presented in the following section.
6. Scale and Scope Economies

Following Baumol, Panzar et al. (1982) the global economies of scale in a multi-output setting are defined as:

\[
SL = \frac{\sum_{m} q^{(m)} \cdot \partial C}{\partial q^{(m)}}, \quad (4)
\]

where \( q=(q^{(1)}, q^{(2)}, q^{(3)}) \) for \( m=1 \) (electricity), 2 (gas) and 3 (water). Global economies of scale describe the cost behavior due to proportional changes in the entire production. The returns to scale are increasing, constant or decreasing if the corresponding ratio (SL) is greater, equal or less than one.

Economies of scope are present when costs can be reduced by joint production of multiple outputs. Following Baumol, Panzar et al. (1982) the degree of global economies of scope across three products is defined as the ratio of excess costs of separate production to the costs of joint production of all outputs:

\[
SC = \frac{C(q^{(1)}, 0, 0) + C(0, q^{(2)}, 0) + C(0, 0, q^{(3)}) - C(q)}{C(q)}, \quad (5)
\]

A positive (negative) value for the above expression implies the existence of global economies (diseconomies) of scope.

Scope and scale economies are usually estimated using the deterministic part of the cost function at some representative outputs. In previous studies these representative outputs are generally obtained by setting the outputs at different points of their sample distribution such as median and other quartiles. As seen in Equation (5), a correct estimation of economies of scope relies on adequately predicting of costs at certain points that are at the sample boundary or completely out
of the sample. The precision of such predictions depends on the econometric specification. As discussed earlier, a GLS model provides a relatively accurate out-of-sample prediction. The random-coefficient model has an additional advantage with respect to heterogeneity bias in the coefficients. The predictions required for estimating scope and scale economies in Equations (4) and (5), can also be conducted at the individual company level, using the individual estimates of company-specific random effects and coefficients. The individual company-level estimates can better represent the actual output patterns. The company-level cost predictions might however entail relatively large estimation errors. In this paper, we have used both approaches.

Using Equations (4) and (5) and the regression results, the values of scope and scale economies have been estimated for five hypothetical companies with representative output combinations. These companies are characterized by the 1st, 2nd, 3rd and 4th quintiles and the sample median of the non-zero output values and customer density. A summary of these results is provided in Table 5. These results suggest the presence of scope and scale economies at most output levels. The estimates also show a well-behaved variation: as outputs increase (decrease) both scale and scope economies fall (rise).

### Table 5: Point estimates of global economies of scope and scale

<table>
<thead>
<tr>
<th>Representative firm</th>
<th>Economies of Scope</th>
<th>Economies of Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLS</td>
<td>RC</td>
</tr>
<tr>
<td>1. Quintile</td>
<td>0.37</td>
<td>0.27</td>
</tr>
<tr>
<td>2. Quintile</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>Median</td>
<td>0.17</td>
<td>0.12</td>
</tr>
<tr>
<td>3. Quintile</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>4. Quintile</td>
<td>0.03</td>
<td>-0.003</td>
</tr>
</tbody>
</table>

The representative points are based on positive values of the three outputs as well as the customer density. Input prices and time trends are kept constant at their sample mean values. The random effects (and coefficients) are assumed to be at their mean values.
Representative sample points such as output quintiles correspond to hypothetical productions that vary in overall scale and density as they represent a more or less similar ratio between all outputs. In this case the firms with “non-typical” mixtures of outputs and customer density would not be represented. In order to study the variation of scale and scope economies in the sample, based on the actual levels of production rather than hypothetical values, we computed the economies of scope and scale for each individual company. Note that the definitions of global economies of scope and scale as defined in equations (4) and (5) is directly applicable only to all-positive-output combinations. In order to extend the estimates to other companies we have chosen a hypothetical all-positive output for each one of these companies. While keeping the positive observed values, we replaced the zero values by a positive value constructed based on the company’s overall scale relative to all the companies in the sample. For any given company the “overall scale factor” is defined as that company’s maximum output standardized by the mean value and standard deviation of that output observed in the sample. For any given company the hypothetical output of a given zero output is constructed by multiplying the company’s overall scale factor by the sample mean value of that output.

An alternative method would be to limit the estimates to the companies with all-positive outputs. However, the fact that the fully integrated companies might be a selection of companies in that they exploit the economies of scope and might have a lower fixed costs, could distort the estimates of scope economies.18

Table 6 and Table 7 respectively provide a summary descriptive of the distribution of the estimates of the global economies of scope and scale across the com-

18 We have also estimated these values for the 33 fully integrated companies. The results do not show much difference.
panies included in the sample. The results obtained from both GLS and RC models are listed. The first and third columns provide the estimates obtained by ignoring the random effects, namely the means of the random coefficients are considered. In the second and fourth columns, the firm-specific random effects are included in the calculation of scale and scope economies. The input prices and the time trends have been set equal to their mean values over the entire sample. Both GLS and RC estimates suggest the existence of scope and scale economies across a major part of the sample. Looking across the numbers from both models indicate that more than 60 percent of the companies can exhibit economies of scope and at least 80 percent can benefit from economies of scale.

Table 6: Distribution of global economies of scope estimated for individual companies

<table>
<thead>
<tr>
<th>Quintile</th>
<th>GLS a</th>
<th>GLS b</th>
<th>RC a</th>
<th>RC b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.05</td>
<td>-0.11</td>
<td>-0.02</td>
<td>-0.18</td>
</tr>
<tr>
<td>2.</td>
<td>0.09</td>
<td>0.05</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Median</td>
<td>0.14</td>
<td>0.15</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>3.</td>
<td>0.17</td>
<td>0.19</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>4.</td>
<td>0.25</td>
<td>0.33</td>
<td>0.18</td>
<td>0.29</td>
</tr>
</tbody>
</table>

a) Individual random effects are not taken into account. b) Individual firm-specific random effects are included in the computations. The values are estimated for all individual observations. Input prices and time trends are kept constant at their sample mean values.

Table 7: Distribution of global economies of scale estimated for individual companies

<table>
<thead>
<tr>
<th>Quintile</th>
<th>GLS a</th>
<th>GLS b</th>
<th>RC a</th>
<th>RC b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1.08</td>
<td>0.97</td>
<td>1.04</td>
<td>1.00</td>
</tr>
<tr>
<td>2.</td>
<td>1.11</td>
<td>1.09</td>
<td>1.06</td>
<td>1.05</td>
</tr>
<tr>
<td>Median</td>
<td>1.12</td>
<td>1.15</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>3.</td>
<td>1.13</td>
<td>1.19</td>
<td>1.08</td>
<td>1.09</td>
</tr>
<tr>
<td>4.</td>
<td>1.22</td>
<td>1.28</td>
<td>1.13</td>
<td>1.24</td>
</tr>
</tbody>
</table>

a) Individual random effects are not taken into account. b) Individual firm-specific random effects are included in the computations. The values are estimated for all individual observations. Input prices and time trends are kept constant at their sample mean values.

Assuming that the larger companies have a lower potential of scale and scope economies (as suggested by Table 5), these results indicate that all small and
moderate-sized utilities can benefit from significant savings through scale and scope economies. However, as seen in Table 6 and Table 7 the extent of these economies can vary depending on the adopted model and the approach used for accounting the estimated effects of unobserved factors. The first and third columns in both tables indicate that if the random effects are not considered in the computations, GLS and RC models provide a quite similar distribution of scale and scope economies across companies. However, a comparison of the first and third columns with the second and fourth ones respectively, suggests that including the individual random effects results in a wider range of variation in scale and scope economies. These results indicate that the economies of scope and scale could be influenced by unobserved factors beyond output and density. We could not find any conclusive pattern suggesting a one-sided bias because of ignoring such unobserved heterogeneity. The results suggest however that compared to GLS model, the RC model provides a lower overall estimate of both economies, as seen in slightly lower median values. This could be explained by the fact that the RC model gives a relatively lower weight to differences regarding fixed costs because part of these costs might be captured by random coefficients.

However, it should be noted that some of the observed variation in the above tables might be related to the relatively large estimation errors of the fixed costs across all models. Considering that the reliability of the individual estimates remains a contentious issue, we contend that the extreme values especially those of scope economies should be considered with caution. Overall these results suggest that a great majority of the companies can benefit from significant economies of scope and scale. Considering the median values these savings vary depending on the model, from 4 to 15 percent for scope economies and 7 to 15 percent for scale economies. Especially the small multi-utilities benefit from considerable scope economies that could reach 20 to 30 percent of total costs.
7. Conclusions

Using a panel data set from the distribution utilities operating in water, gas and electricity sectors this paper has studied the economies of scope and scale in multi-output utilities. A random effect panel data (GLS) model and a random-coefficient (RC) model have been used to explore the effect of unobserved heterogeneity across different networks. While the GLS model considers the unobserved heterogeneity as various cost shifts across companies, the RC model includes variations in marginal effects of outputs and customer density. Compared to cross-sectional model, the GLS specification provides a better control for omitted variables. The RC model provides an additional improvement regarding the potential heterogeneity bias in the coefficients’ estimates.

This paper also shows that the computation of the economies of scope and scale can be extended to include the estimates of firm-specific individual effects, namely the conditional expectation of the random intercept and random coefficients. While admitting that such company-level estimates may entail relatively large estimation errors at the individual level, we assert that the overall results could represent a better picture of scope and scale economies based on actual levels of outputs and network characteristics rather than simplified hypothetical values.

From the results three general observations can be pointed out. First, the results confirm the existence of significant scope and scale economies in a majority of multi-utilities, which can be considered as suggestive evidence of natural monopoly in multi-utilities. This conclusion is confirmed across the two models and regardless of whether the individual firm-specific stochastic terms are included in the estimations. Secondly, considerable variation of the estimated values among individual companies suggests that the economies of scope and scale can depend on unobserved network characteristics as well as output patterns and
customer density. Finally, the variations across the models indicate that the overall point estimates are not very sensitive to the specification of unobserved firm-specific factors.

The results of this paper show that even after accounting for unobserved heterogeneity, the scope economies exist in a majority of the multi-utilities, suggesting that additional costs could result from unbundling the multi-utility companies. In the actual situation many companies avoid these additional costs through scope economies. Especially for small companies the savings associated with scope economies are considerable.

In this study it is assumed that there is no functional separation between distribution and supply functions. While being possibly unrealistic in some EU countries, this assumption closely reflects Switzerland’s actual situation and most probably, its future development. In fact, under the EU policy directive the utilities with fewer than 100,000 customers can be exempt from any functional unbundling requirement. As most of the distribution companies in Switzerland are relatively small with only a few companies having more than 100,000 customers, with a likely adoption of policies similar to those of EU, the distribution and supply are likely to remain integrated in the future. Therefore, the results of this study are especially relevant for the context of Switzerland as well as in many similar cases in other countries.

References


EFFECTS OF OWNERSHIP, SUBSIDIZATION
AND TEACHING ACTIVITIES ON
HOSPITAL COSTS IN SWITZERLAND

Mehdi Farsi †,‡ Massimo Filippini †,‡

† Department of Management, Technology and Economics,
ETH Zurich, Zurichbergstr. 18, Zurich 8032, Switzerland
‡ Department of Economics, University of Lugano, 6900 Lugano, Switzerland

April 2007

ABSTRACT
This paper explores the cost structure of Swiss hospitals, focusing on differences due to teaching activities and those related to ownership and subsidization types. A stochastic total cost frontier with a Cobb-Douglas functional form has been estimated for a panel of 148 general hospitals over the six-year period from 1998 to 2003. Inpatient cases adjusted by DRG cost weights and ambulatory revenues are considered as two separate outputs. The adopted econometric specification allows for unobserved heterogeneity across hospitals. The results suggest that teaching activities are an important cost driving factor and hospitals that have a broader range of specialization are relatively more costly. The excess costs of university hospitals can be explained by more extensive teaching activities as well as the relative complexity of the offered medical treatments from a teaching point of view. However, even after controlling for such differences university hospitals have shown a relatively low cost-efficiency especially in the first two or three years of the sample period. The analysis does not provide any evidence of significant efficiency differences across ownership/subsidy categories.

Keywords: general hospitals, teaching hospitals, stochastic frontier, cost efficiency
1. Introduction

The increasing growth of health care costs in Switzerland has raised the public interest in identifying the possibilities of improvement in productive efficiency. General hospitals (specialized clinics excluded) that account for about a quarter of national health expenditures have been subject of much debate but few studies. Farsi and Filippini (2006) and Steinmann and Zweifel (2003) have found significant differences in productivity and cost-efficiency among hospitals. Identifying the sources of such differences is an important policy issue that has not been explored sufficiently.

Ownership and subsidization as well as research and teaching activities have been considered as important cost-driving factors among Swiss hospitals. University hospitals have been often criticized for being excessively costly. Many policy-makers believe that public and subsidized hospitals are not as efficient as private facilities. However these policy debates remain qualitative and lack sufficient empirical evidence.

The present study addresses the above issues using data form a national sample of 148 general hospitals operating from 1998 to 2003. Compared to the previous research on Swiss hospitals this paper benefits from a larger data set and several additional variables especially those related to teaching activities. Moreover, the adopted methodology is based on some of the recent developments in stochastic frontier panel data models.

The analysis indicates that teaching activities can explain part of cost differences among hospitals. The results suggest that university hospitals while showing relatively high inefficiency, have improved over the sample period. There is no evidence of statistically significant efficiency differences among various
ownership/subsidy types. The estimation results also point to unexploited economies of scale in a majority of the studied hospitals.

The rest of the paper is organized as follows. Section 2 provides a general description of the adopted methodology with discussions of the functional form and econometric models. Section 3 describes the model specification. The data and descriptive statistics are given in Section 4. The estimation results are presented and discussed in Section 5. Section 6 concludes the paper.

2. Methodology

Though many authors (Zuckerman et al., 1994; Linna, 1998; Rosko, 2001) have used cost frontier models to evaluate hospitals' efficiency, the application of such models in the health-care sector has been criticized (Newhouse, 1994; Skinner, 1994). The main arguments against these models are related to the unobserved heterogeneity due to differences in case-mix and quality of care as well as the errors incurred by aggregation of outputs. Horrace and Schmidt (1996), Jensen (2000) and Street (2003) have highlighted the sensitivity issues in the efficiency ranking of individual firms, which has been considered the frontier models’ main objective.

Virtually all frontier models rely on an assumption that the inefficiencies can be represented by certain asymmetric component of the stochastic error term. Such assumptions provide a convenient practical basis for separating the random noise from the inefficiency term. Moreover, they are often based on a sensible distribution assumption that assigns relatively high likelihood to full efficiency, thus providing a basis for comparison of individual firms with the bulk of the sample. However, it should be noted that these assumptions are in principle non-testable, and as pointed
out by Street (2003), might create sensitivity problems for they link the identification of inefficiencies to the skewness of the residuals to a certain direction.

Admitting the difficulties involved in identifying the individual firm’s relative efficiency, other studies (Folland and Hofler, 2001; Hadley and Zuckerman, 1994; Farsi et al., 2005) show the practical use of stochastic frontier analysis for comparing the performance across groups of providers. In addition, the new developments in stochastic frontier models for panel data proposed by Greene (2005) provide a better account of the hospital-specific unobserved heterogeneity. Following this thread of literature, we adopt a stochastic cost frontier approach to explore the efficiency differences across hospital types. The null hypothesis posits a similar distribution of inefficiency across different types, while the alternative suggests that hospital types differ in cost-efficiency perhaps due to various incentive mechanisms e.g. ownership/subsidy status, or differences in objective functions such as teaching and research purposes in university hospitals.

Functional form

Griffin et al. (1987) provide a comprehensive list of alternative functional forms and propose a series of criteria for model selection in cost and production analyses. The most important restrictions are related to the sample size and the estimation method. As the number of variables increase, most functional forms require a geometrically increasing number of parameters, thus necessitate much larger samples. The optimal choice is therefore a functional form that can be estimated with available estimation procedures and limits the number of parameters while using as many relevant variables as possible. One of the most commonly used functional forms is the Cobb-Douglas (log-linear) model (cf. Greene, 2004, 2003; Linna, 1998). Thanks to its
limited number of variables this form has a practical advantage in estimation and interpretation, over more complicated forms. The main shortcoming of this model is the assumption of constant rate of scale economies which is considered as restrictive because by using the same proportional increase in output, small companies usually gain more than large firms.

The potential changes in scale elasticity with output can be analyzed using flexible functional forms such as translog. However, a translog model requires the estimation of a large number of parameters. Furthermore, the included second-order terms could cause multicollinearity, which can affect the model’s statistical performance. Especially with the multiple error component model used in this study and the available sample size, such problems could induce numerical problems resulting in degenerate stochastic terms. In fact, our preliminary analyses showed that a numerically feasible estimation of a translog cost frontier with non-degenerate stochastic components was only possible with simplified specifications that excluded several important output characteristics.

Using a parsimonious translog model with a homothetic cost function and its corresponding Cobb-Douglas model, we performed an exploratory analysis to identify the effect of functional form on the results. The results indicate that: first, the main estimated coefficients do not change much across the two functional forms. Particularly, the main output coefficients used for estimating the scale economies are quite similar to those of the complete model used here. Secondly, the efficiency estimates obtained from the translog model are highly correlated (higher than 90%) with those of the corresponding parsimonious Cobb-Douglas model. The main differences in efficiency estimates appeared when we included the deleted variables and more importantly with a change in econometric specification. These results
suggest that in our case the choice of explanatory variables and the econometric specification of unobserved heterogeneity have a greater importance than the functional form. We therefore decided to focus on the Cobb-Douglas form that allows a larger number of explanatory variables.

Resulting from a minimization problem given input prices and outputs, cost functions must be non-decreasing in outputs and concave and linearly homogeneous in input prices (Cornes, 1992). In particular, the latter condition is usually imposed by dividing the input prices by a numeraire price, thus ensuring the input shares add up to 1. In this paper, this condition is not imposed mainly because as we will see later, the available data does not allow a complete account of all input factors.

**Econometric models**

There are a number of econometric approaches to estimate stochastic cost frontier models (Kumbhakar and Lovell, 2000). The original cost frontier model (Aigner et al., 1977) applied to panel data can be written as:

\[
\ln T C_{it} = f (Y_{1it}, \ldots, Y_{mit}; P_{1it}, \ldots, P_{nit}; Z_{1it}, \ldots, Z_{kit}) + u_{it} + v_{it} \tag{1}
\]

where subscripts \(i\) and \(t\) represent the firm and year respectively; \(TC\) is the total costs; \(Y_{m}(m=1, \ldots, M)\) are the outputs; \(P_{n}(n=1, \ldots, N)\) are the input factor prices; \(Z_{k}(k=1, \ldots, K)\) are output characteristics and other exogenous factors that may affect costs; \(v_{it}\) is the random noise or unobserved heterogeneity; and \(u_{it}\) is a positive stochastic term representing inefficiency, typically with a normal-half-normal distribution:

\[
u_{it} \sim \left|N(0, \sigma_{u}^{2})\right|, \quad v_{it} \sim N(0, \sigma_{v}^{2}).\]

The firm’s inefficiency is estimated using the
conditional mean of the inefficiency term as proposed by Jondrow et al. (1982), that is:
\[ E[u_t | \varepsilon_t], \]
where \( \varepsilon_t = u_t + v_t. \)

Assuming a time-invariant inefficiency term \( u_{it} = u_i \), this term can be identified by panels’ individual fixed or random effects. The resulting specifications (Pitt and Lee, 1981; Schmidt and Sickles, 1984) relax the distribution assumptions on stochastic terms, in particular in the fixed effect specification the individual firm effects \( (u_i) \) do not need to be uncorrelated with explanatory variables. Several authors (Battese and Coelli, 1992; Cornwell et al., 1990; Sickles, 2005) have extended the above panel data models to include time-variant inefficiency. Others (Greene, 2004, 2005; Kumbhakar and Hjalmarsson, 1995; Kumbhakar; 1991; Polachek and Yoon, 1996) have adopted another approach in which a stochastic firm-specific term (fixed or random effect) is added into the original stochastic frontier model presented in Equation (1). This approach allows a distinction of unobserved time-invariant heterogeneity across firms, which is particularly important in hospitals characterized by strong unobserved heterogeneity associated with case mix and quality differences.

In particular the random intercept frontier model (‘true’ random effects frontier model) proposed by Greene (2004, 2005) has been successfully used in other sectors (Farsi et al., 2005). This model can be obtained by adding a firm-specific stochastic term \( \alpha_i \sim N(0, \sigma^2) \), on the right-hand-side of Equation (1). As opposed to alternatives with fixed effects, this model does not have the incidental parameters problem. The main difficulty of this model is in its numerically cumbersome estimation method. As the likelihood function does not have a closed form, this model is estimated using Simulated Maximum Likelihood (SML) method, in which \( \alpha_i \)’s are simulated by random draws. Because of non-linearity of errors in the number of
simulations, the SML estimators require a large number of simulations or might show sensitivity to the draws (Gouriéroux and Monfort, 1996).

In this paper, we use the Greene’s true random effect frontier model, labeled here as TRE. We use pseudo-random Halton draws to minimize the potential sensitivity of the results to simulations. Number of draws has been fixed to 1000. Our sensitivity analysis using several options suggested that the estimations are not sensitive when the number of draws is higher than a few hundred. The inefficiency is estimated using the (simulated) conditional mean of the inefficiency term ($u_i$) given by $E[u_i | \hat{\omega}_i]$, where $\hat{\omega}_i = \alpha_i + u_i + v_i$. In addition to the TRE model, we estimated the original pooled frontier model as shown in Equation (1). The contrasting difference between the two models is that unlike the pooled model, in the TRE specification, the persistent cost differences are excluded from inefficiency estimates. In this sense the two models can be used to provide complementary estimates of persistent and transient inefficiencies.

Differences across ownership/subsidization types

Although, economic theory predicts lower costs for organizations with relatively high-powered financial incentives such as for-profit and non-subsidized firms, the empirical evidence is rather mixed. While some studies (Eakin, 1991; Steinmann and Zweifel, 2003) conclude no significant differences, a few others (Li and Rosenman, 2001; Carey, 1997) report slightly lower costs in for-profit hospitals compared to non-profit ones. In this paper, the effects of ownership/subsidization status on efficiency are studied using a two-stage method. This method is based on testing the significance of differences across hospital groups. We use the Kruskal-
Wallis (1952) rank test (KW) as well as the t-test with unequal variances. The KW test is a non-parametric test that has been often used in frontier analysis (Singh and Coelli, 2001). Given that the hospital types are more or less constant over the sample period (no change in subsidy status and only 9 cases of ownership change), the tests have also been performed on the hospital average values over the sample period but have not shown much difference in the outcome.

The two-stage approach has a disadvantage in that the first-stage estimation errors may affect the results of the test in the second-stage. These errors may lead to an under-rejection of the null hypothesis postulating similar cost-efficiencies across different categories (Farsi and Filippini, 2004). On the other hand, the two-stage approach allows the use of non-parametric statistical tests based on efficiency ranks rather than efficiency values that are subject to relatively large estimation errors and sensitive to outliers. An alternative approach is to include type indicators in the regressions and test the significance of the corresponding coefficients. We performed a GLS estimation of this alternative specification to confirm the results of the two-stage procedures. Our data show that the subsidization status has not changed over the sample period and only 9 hospitals have changed ownership status from one year to another.

3. Model specification

The specification used in this study is based on two main outputs: hospitalizations and ambulatory care. In line with Linna (1998), Rosko (2001) and Heshmati (2002) the main measure of hospitalization output is taken as a DRG weighted number of hospitalizations (denoted by $Y$). This approach was preferred over the alternative based on multiple output categories based on DRG weights (Brown,
2003), mainly because such categories might be arbitrary as the DRG weights define the cost intensity of the cases rather than different outputs.

Since the number of outpatient cases is not available in the data, the ambulatory output is approximated by the corresponding revenues adjusted for inflation (AMB). This approximation is based on the assumption that the average unit price of ambulatory care is similar across hospitals. Three input factors are considered: capital, physicians’ input and all other employees’ labor. Similar to Wagstaff and Lopez (1995) and Rosko (2001), capital prices (PK), are approximated by the hospital’s total capital expenditure divided by the number of available beds in the hospital.

Labor prices ($PL_1$ and $PL_2$) are calculated by dividing total salaries by the number of remunerated days. In line with Folland and Hofler (2001) and Scuffham et al. (1996) among others, physicians and non-physicians are considered as two separate labor inputs. The physicians’ labor price represents the average salary of those employed by the hospital and exclude honoraries and fees, accounting on average for about 5% of the hospital’s total costs, usually paid to both employed and unemployed physicians. Both labor prices are proportionally adjusted for social benefits, accounting on average, for about 9% of total costs. These charges are proportionally distributed to physician and non-physician groups, the proportions being the respective shares of each group’s salaries. This adjustment captures the potential variation in social benefits across hospitals due to differences in pension funds as well as the age and seniority of the employees mix.

The three input factor prices considered in the model correspond to about 70 percent of a hospital’s total cost on average. The available data do not allow an appropriate calculation of the prices of remaining inputs such as medical materials,
food, water and power as well as physicians’ fees and other personnel charges. The excluded prices are obviously not constant and neglecting their variation could affect the estimation results. However, some of these variations are probably captured by the three included factor prices. For instance, physicians’ fees are likely to be correlated with physicians’ salaries. Another concern is the accuracy of the price data that may create bias in the price coefficients. However, other coefficients will not be affected if these measurement errors and the unobserved factor prices are uncorrelated with explanatory variables.

Similar to Vita (1990), Scuffham et al. (1996) and Carey (1997) the average length of hospitalization (LOS) has been included in the model. In addition to representing hospital’s ‘hotel services’ like nursing care and accommodation (Breyer, 1987), this variable provides a measure of severity of the case mix within each DRG. In fact, there is a considerable variation among patients within a DRG, as indicated by the wide range of acceptable hospital stays provided by the Swiss DRG Association (APDRG Suisse, 2003). The number of hospital departments (medical units MU) is also included to represent the range of specializations offered in the hospitals. Each one of these centers provides a single specialization. Hospitals with a wider variety of medical specializations are expected to be more costly than those with similar output but from fewer specializations. Another cost-driver is the number of non-medical units (TU) including medico-technical, therapeutic and infrastructure units. The operation costs of these units are also included in the observed total costs.

The share of outpatient clinics over total medical units (AMBC) operated by the hospital is also included as a complementary measure of ambulatory output in the model. The level of hospital’s teaching activities is measured by the total number of internship positions (NP) offered in the hospital. The internships might have different
levels of complexity and specialization. In order to account for such differences we used the Swiss Medical Association’s classification that applies to internship positions and hospital departments. We included the percentage of the internship positions recognized as the two top categories (AB) and that of the hospital’s departments accredited for specialized medical training (FMH). These two variables are expected to represent the complexity level of the hospital’s medical care.

Hospitals’ costs can also be affected by the quality of care. The evidence on the effect of quality measures on hospital costs is not conclusive. Zuckermann et al (1994), Rosko (2001) and Vitaliano and Toren (1996) conclude that quality indicators do not have significant cost effects, whereas others such as Folland and Hofler (2001) suggest a significant effect for structural quality measures such as bed availability and the share of board-certified physicians. This may be explained by the fact that unlike outcome or process measures the structural quality is usually easier to observe and has a more directly measurable effect on costs. As we do not have access to patient-level data or any reliable outcome measure quality from Swiss hospitals, in this paper we focus on structural measures of quality. In addition to the share of accredited medical units and training positions, we included the hospital’s nurse per bed ratio (NB) to represent the quality of nursing care.

We also included two binary indicators for emergency room (ER) and geriatrics department (GER). While emergency services are usually involved with relatively severe cases, geriatrics cases are less intensive in medical care thus less costly. Year dummies (Y99 through Y03) are included to capture the overall technological progress and the potential temporal variations in unobserved variables such as reporting procedures.

The specification of the true random effects model can therefore be written as:
\[
\ln TC_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 \ln AMB_{it} + \gamma_1 \ln PK_{it} + \gamma_2 \ln PL_{it} + \gamma_3 \ln PL_{2it} \\
+ \omega_1 \ln LOS_{it} + \omega_2 \ln NB_{it} + \omega_3 \ln MU_{it} + \omega_4 \ln TU_{it} + \omega_5 \ln NP_{it} \\
+ \rho_1 AMBC_{it} + \rho_2 FMH_{it} + \rho_3 AB_{it} + \delta_1 ER_{it} + \delta_2 GER_{it} \\
+ \delta_0 Y99_{it} + \delta_{00} Y00_{it} + \delta_{01} Y01_{it} + \delta_{02} Y02_{it} + \delta_{03} Y03_{it} + \alpha_i + u_{it} + v_{it}
\]

The stochastic components \(\alpha_i, u_{it}\) and \(v_{it}\) respectively represent the hospital-specific random effect, inefficiency term and random noise with a normal-half-normal distribution. The pooled model is a special case, in which the stochastic component \(\alpha_i\) is set equal to zero.

The above specification leaves out several important factors. Namely the available data do not provide any measure of severity variation for outpatients and those within each DRG or a better measure of hospital’s ambulatory output especially the ER visits. More importantly, the model lacks an outcome measure of less observable quality differences across hospitals. Assuming that these unobserved factors are independent of the included explanatory variables, their omission does not bias the results. However, this might be a restrictive assumption as in many cases, the correlation between those differences and hospital size or type is rather plausible. For instance, assuming a positive relation between costs and quality, higher unobservable quality for non-profit hospitals would imply an overestimation of their inefficiency compared to other hospital types. The results of this paper should therefore be considered within the limits of the available data and the related simplifying assumptions.

4. Data

The data used in this paper are extracted from the annual financial and administrative data reported by general hospitals to the Federal Statistical Office (SFSO, 1997a) from 1998 to 2003. These data have been merged with another data set.
consisting of an aggregate extraction of the medical data of the Swiss hospitals with records for individual hospitalizations (SFSO, 1997b). The extracted medical data consist of the number of cases by AP-DRG in each hospital-year, including about a million observations. Using the cost weights from Swiss AP-DRG version 4.0 (APDRG Suisse, 2003), we calculated an average cost weight for each hospital-year. The adjusted number of admissions is then calculated by multiplying these average cost weights by the number of admissions recorded in the administrative data.

After excluding the observations with missing and invalid values from an unbalanced panel with 1082 observations from 221 general hospitals, the final sample was created with 623 observations from 148 hospitals operating from 1998 through 2003. The excluded observations are mainly those with missing DRG data or erroneous values for outpatient revenues. We also excluded three hospitals with fewer than 20 beds. In general, the excluded observations with missing or suspicious values include higher proportion of small-size hospitals. T-tests suggest that the excluded observations are from hospitals with significantly lower number of beds (an average of 110 beds). However, similar tests indicate that there is no significant difference in average cost per hospitalization across the two groups. A descriptive summary of the sample listed in Table 1, shows a considerable variation among hospitals in most variables. Particularly while the average cost of a hospitalization varies from 4,500 to 54,000 Francs, the average DRG cost weight ranges from .52 to 1.47 and an average hospitalization lasts from 4 to about 50 days.

**Insert Table 1**

The sample also includes all the five university hospitals in Switzerland. These hospitals stand out from the rest of the sample in several ways. With an average size of 1030 beds, DRG cost weight of 1.07, LOS of 8.3 days and an average cost of
19,700 Francs per hospitalization these hospitals are on average larger and treat a relatively severe patient mix with relatively short but expensive hospitalizations. These statistically significant differences (shown by t-test) might suggest the possibility of a different technology, hence a different cost function in university hospitals. However our preliminary regressions on a sample excluding the university hospitals indicate that the results do not change significantly, suggesting that these hospitals can be pooled with the rest of the sample. Moreover, the correlation between the efficiency estimates within each econometric model is higher than 98% between the samples with and without university hospitals.

The number of general hospitals in the sample and their average capacity by ownership/subsidy types are listed in Table 2. All public hospitals and most private non-profit hospitals are subsidized, whereas in the private for-profit sector, a large fraction of hospitals are not. Table 2 also lists the average hospital size measured by the number of beds for each ownership/subsidy type. Public hospitals with an average of 262 beds are by far the largest providers of health care in the sample. Subsidized hospitals are also considerably larger than non-subsidized ones.

5. Results

Table 3 lists the regression results of the cost frontier analysis as in Equation (2). The estimated coefficients are mostly significant and generally have the expected signs. Overall, the differences across the two models, while being statistically significant in many cases, are not considerable for practical purposes. The results especially those of the TRE model are also comparable to a similar model estimated by GLS given in the appendix (Table A.1). According to the TRE model a ceteris
paribus increase in the adjusted number of hospitalization by 1% will result in about 0.7% increase in total costs. As expected, the effect of ambulatory output is much smaller, suggesting a marginal cost of about five times.

Insert Table 3

The regression results indicate that LOS is an important predictor of hospital costs. Given that hospital stays are on average about 10 days, this implies that a difference of one day in the hospital’s average LOS is approximately equivalent to 4% of total costs. This could be considered as an important policy implication in the context of Switzerland, where local hospitals have been criticized for their excessively long hospitalizations. However, the apparently considerable savings by curtailing hospital stays should be considered with caution. First, the costs of medical treatment are not evenly distributed over the course of a hospitalization and the final days are usually less costly. Secondly, as confirmed by the smaller effect in the TRE model that has a better control for unobserved heterogeneity, LOS variable also captures part of the unobserved differences in case mix severity that are beyond the hospital management’s control.

As expected, the price coefficients are positive and significant. However, these estimates significantly differ from the average actual share of the corresponding input factors (about 7, 11 and 53 percent for capital, physician services and other employees). This result can be related to the fact that because of labor contracts and other institutional and practical restrictions hospitals are not fully responsive to changes in input prices. This might imply that hospitals do not completely minimize their total costs. It should also be noted that hospitals might have other objectives in addition to cost minimization, in which case functions based on cost optimization can still be used as a “behavioral” cost functions and can be helpful in studying the firms’
behavior rather than their production technological characteristics (Breyer, 1987; Bös, 1986).

As seen in Table 3, the number of hospital units has a significant effect on total costs, suggesting that hospitals with a wider range of specialization and also those with more non-medical services are relatively more costly. However the estimated coefficients suggest that such cost differences are relatively small. The marginal cost of internship positions is also low but statistically significant. The results predict an average increase of 0.9% in total costs for 10% increase in the number of positions. The teaching quality regarding medical specialization has also a statistically significant effect on hospital costs, but the marginal effects remain quite low.

The share of ambulatory clinics has a negative and significant effect, consistent with the fact that ambulatory visits are usually less costly than inpatient care. The TRE model suggests that for instance, an increase of 10 percentage points in the share of ambulatory clinics thus 10 points decrease in the share of inpatient units, results in a decrease of about 3% in the hospital’s total costs. The nurse per bed ratio has a relatively high and significant effect, indicating that the cost of nursing care is quite considerable. As expected, the ER dummy has a positive coefficient and the geriatrics dummy has a negative effect. The coefficients of the year dummies suggest a positive growth in hospital costs starting from 2000, with an average annual rate of 1 to 3 percent.

Regarding scale economies, the results listed in Table 3 indicate that the returns to scale (inverse of the main output elasticity) are on average significantly higher than 1 (1.4 or 1.6 depending on the model). This suggests that the majority of general hospitals in Switzerland do not fully exploit the potential scale economies.
However, it should be noted that these economies are likely to be marginal for large hospitals with more than 130 to 200 beds (Vita, 1990; Crivelli et al., 2001; Aletras, 1999; Dranove, 1998).

*Overall cost-efficiency*

Table 4 provides a descriptive summary of the inefficiency scores estimated by the two models. The inefficiency scores obtained from the two models are significantly correlated with a correlation coefficient of 0.56 and a Spearman rank correlation of .53. According to the pooled model, the inefficiency is less than 8 percent for half of the sample but a quarter of the studied hospitals show 11% or more excess costs. As expected, the pooled model’s estimates are generally higher than those of the TRE model that separates part of the hospital-specific heterogeneity. The latter model suggests that on average, about 6 percent of the hospital costs cannot be explained by the included explanatory variables or by a symmetric stochastic term.

Noting that the “true” inefficiencies cannot be exactly identified, the estimated inefficiencies can be interpreted as the excess costs compared to the best observed practice provided that the omitted variables are captured by symmetric stochastic components. While we cannot favor one model over the other, we assert that given the important unobserved factors related to quality and case-mix severity the TRE model is likely to give a better picture of excess costs. On the other hand in the context of Swiss hospitals, because of strong regulation and institutional restrictions managers might be unable to adapt with changing conditions thus persistent inefficiencies might be relatively important. In this case the TRE model might understate the sector’s overall inefficiency.
These inefficiency estimates are in general lower than those reported in previous studies for the Swiss hospitals (Farsi and Filippini, 2006; Steinmann and Zweifel, 2003; Steinmann et al., 2004). However, the differences can be explained by several additional characteristics included in this paper, such as teaching and specialization variables as well as a different methodology in separating heterogeneity from inefficiency differences. The results are comparable to similar estimates reported in the literature for the US hospitals, ranging from 5 to 15 percent (Zuckerman et al., 1994; Folland and Hofler, 2001; Eakin, 1991), but differ from other studies particularly those on European samples, which estimate generally higher levels of inefficiency amounting to 20 to 30 percent (Linna, 1998; Wagstaff, 1989; Wagstaff and Lopez, 1995; Steinmann et al., 2004; Bruning and Register, 1989). It should be noted that even the seemingly low values estimated from the TRE model are equivalent to considerable excess costs amounting, for instance in 2003, to about 590 million Francs out of the actual total costs of 10.7 billion Francs for the hospitals in the sample. The 6 percent average inefficiency is also equivalent to 2 or 3 years efficiency lag according to the efficiency targets set by the UK health care authorities (Jacobs and Dawson, 2003).

Cost-efficiency in university hospitals

The estimation results suggest that university hospitals are on average less efficient than other hospitals. However, this difference is not statistically significant in the TRE model. Excepting the university hospitals the average efficiency estimates do not show any significant changes over time. University hospitals however show a
different pattern with a relatively high inefficiency in the first years (1998 to 2000) and a decreasing trend over the sample period (Figure 1).

**Insert Figure 1**

Several t-tests on the university hospitals’ efficiency scores across different years suggest that the efficiency improvement in university hospitals is statistically significant. According to these estimates, from 1998 to 2003, university hospitals have considerably reduced their excessive costs. Part of these changes could be explained by the variation of case mix severity. In fact, the trends in AP-DRG cost weights suggest that the severity of the patient mix has grown relatively more in university hospitals (Figure 2). Over the sample period the average cost weight for university hospitals has increased from 0.99 to 1.17 whereas the corresponding change in other hospitals is from 0.78 to 0.84. Given that in Switzerland, DRG coding has been introduced in 1998, some of such increases might be related to changes in the quality of DRG coding especially in university hospitals that, having relatively severe cases, require a more elaborate coding practice. In this case the observed changes in efficiency of university hospitals could be an artifact of a different DRG coding.

**Insert Figure 2**

In order to explore the relationship between changes in severity and inefficiency, we estimated another model similar to Equation (2), with the only difference that the number of admissions is not adjusted for AP-DRG cost weights. The inefficiency estimates of this analysis still show a slight but still statistically significant improvement in university hospitals over the sample period. These results indicate that part of efficiency gains in university hospitals could be related to the fact
that these hospitals increasingly treat more severe cases. However, even if we assume that the observed severity trends are entirely related to gradual effect of better coding practices, the results still indicate that on average university hospitals have improved.

**Effects of ownership/subsidy types**

The average inefficiency estimates are listed by ownership and subsidization categories in Table 5 and Table 6 respectively for the pooled and TRE models. These results point to some differences among various hospital types. We explored the significance of these differences with several Kruskal-Wallis and t-tests. Several possible groupings have been considered. In summary, the TRE model’s inefficiency estimates do not show any statistically significant difference across hospital types. The estimates obtained from the pooled model are significant at 10% level only for a single case, suggesting a higher efficiency in subsidized versus non-subsidized hospitals. Overall, consistent with the results reported in previous studies (Farsi and Filippini, 2006; Steinmann and Zweifel, 2003), this analysis suggests that after controlling for other factors, subsidization and ownership do not have any significant effect on hospital costs. This is also confirmed by the GLS model in which the three type indicators remain statistically insignificant (Table A.1).

Insert Table 5

Insert Table 6

6. Conclusions

Using a stochastic cost frontier model we explored the cost-efficiency differences across various hospital types. Consistent with the previous studies, the
results point to considerable unexploited scale economies in a majority of the studied hospitals. The results also suggest that hospitals with a wider range of specializations are relatively more costly than those specializing in fewer categories of medical services. However, the cost differences resulting from specialization are limited to a few percentage points for a relatively large change in the number of services.

The richer data compared to the previous studies on Swiss hospitals were used to identify the effect of teaching activities on hospital costs through the number of internships and measures of teaching accreditation. The results suggest that the considerable excess costs of university hospitals, reported in previous studies, can be explained by more extensive teaching activities in those hospitals as well as the relative complexity of the offered medical care as assessed for training purposes. However, our analysis indicates that even after controlling for such differences university hospitals have shown a relatively poor cost-efficiency in the first two or three years of the sample period. The results also point to a statistically significant improvement of efficiency of university hospitals over the sample period.

Finally, the statistical tests do not provide any evidence of statistically significant efficiency differences across ownership and subsidization categories. This result has been confirmed by a panel data model that integrates the ownership/subsidy indicators. However, lack of evidence for significant efficiency advantage of one type over another might be restricted to the available data, thus should be considered with caution. In fact, the potential correlation between hospital types and other cost driving factors might mask the actual ownership/subsidy effects.

The present analysis has two main shortcomings that call for further study. First, the available data do not allow for a sufficient account of differences regarding the quality of care and case-mix severity. Although the adopted econometric
specification accounts for unobserved hospital-specific heterogeneity through individual random effects, the potential correlation of omitted variables with hospital types is neglected. Secondly, the cost frontier model while having a practical convenience in comparing the individual hospitals with the “best” observed practice, prove to be sensitive to the model specification.

Acknowledgements

The authors are grateful to the Swiss Federal Statistical Office for providing the data and their financial support. They also wish to thank this journal’s editor and two anonymous referees for their helpful suggestions and André Meister, Luca Stäger and Luca Crivelli for their help in understanding the data. Part of this research has been financed by the Swiss National Science Foundation through research grant 100012-108288, which is gratefully acknowledged. The views expressed in this paper are those of the authors and do not necessarily reflect the position of any institution.

Appendix

Insert Table A.1

References


Figure 1: Efficiency trend in university hospitals

Figure 2: Average AP-DRG cost weight
Table 1: Descriptive statistics

<table>
<thead>
<tr>
<th>Continuous variables</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital's total costs (CHF '000)</td>
<td>86'697</td>
<td>139'667</td>
<td>5'036</td>
<td>884'764</td>
</tr>
<tr>
<td>Number of hospitalizations</td>
<td>7'542</td>
<td>7'814</td>
<td>367</td>
<td>50'774</td>
</tr>
<tr>
<td>Number of hospitalizations (AP-DRG adjusted)</td>
<td>6'555</td>
<td>7'825</td>
<td>208</td>
<td>49'251</td>
</tr>
<tr>
<td>Average total cost per hospitalization (CHF '000)</td>
<td>10.17</td>
<td>4.53</td>
<td>4.39</td>
<td>53.78</td>
</tr>
<tr>
<td>Average AP-DRG cost weight</td>
<td>0.8224</td>
<td>0.1127</td>
<td>0.52</td>
<td>1.4735</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dummy variables</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergency Room</td>
<td>0.9165</td>
</tr>
<tr>
<td>Geriatrics</td>
<td>0.5217</td>
</tr>
<tr>
<td>Year 1998</td>
<td>0.1108</td>
</tr>
<tr>
<td>Year 1999</td>
<td>0.1573</td>
</tr>
<tr>
<td>Year 2000</td>
<td>0.1846</td>
</tr>
<tr>
<td>Year 2001</td>
<td>0.1878</td>
</tr>
<tr>
<td>Year 2002</td>
<td>0.1878</td>
</tr>
<tr>
<td>Year 2003</td>
<td>0.1717</td>
</tr>
<tr>
<td>Private for-profit Hospital</td>
<td>0.0979</td>
</tr>
<tr>
<td>Private non-profit hospital</td>
<td>0.3355</td>
</tr>
<tr>
<td>Public hospital</td>
<td>0.5666</td>
</tr>
<tr>
<td>Subsidized hospital</td>
<td>0.9085</td>
</tr>
<tr>
<td>University hospital</td>
<td>0.0385</td>
</tr>
</tbody>
</table>

- All monetary values are adjusted by the global consumer price index relative to 2003 prices.
- Semi-hospitalizations (shorter than 24 hours) are considered as one-day hospitalizations.
- a Excludes semi-hospitalizations (over-night hospital stays shorter than 24 hours).
- b Employed physicians' average salary, adjusted for social benefits and excludes fees.
- c Average salary (adjusted for social benefits) of all hospital employees except physicians.
- d Includes medicotechnical, therapeutic and infrastructure units.
# Table 2: Number of hospitals and average size by ownership/subsidy (1998-2003)

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC</th>
<th>PRIVATE NON-PROFIT</th>
<th>PRIVATE FOR PROFIT</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBSIDIZED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td>353</td>
<td>175</td>
<td>38</td>
<td>566</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td>262</td>
<td>155</td>
<td>194</td>
<td>224</td>
</tr>
<tr>
<td><strong>NON SUBSIDIZED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td>34</td>
<td>34</td>
<td>23</td>
<td>57</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td>80</td>
<td>80</td>
<td>133</td>
<td>101</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>353</td>
<td>209</td>
<td>61</td>
<td>623</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td>262</td>
<td>143</td>
<td>171</td>
<td>213</td>
</tr>
</tbody>
</table>
Table 3: Estimation results

<table>
<thead>
<tr>
<th></th>
<th>Pooled Model</th>
<th>True RE Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hospitalizations (AP-DRG adjusted)</td>
<td>0.6914* (0.0152)</td>
<td>0.6315* (0.0069)</td>
</tr>
<tr>
<td>Average length of hospitalizations</td>
<td>0.5100* (0.0194)</td>
<td>0.4023* (0.0088)</td>
</tr>
<tr>
<td>Outpatient revenues</td>
<td>0.1498* (0.0093)</td>
<td>0.1456* (0.0037)</td>
</tr>
<tr>
<td>$P_k$ (capital price)</td>
<td>0.1211* (0.0096)</td>
<td>0.1225* (0.0045)</td>
</tr>
<tr>
<td>$P_L$ - physicians</td>
<td>0.0744* (0.0179)</td>
<td>0.0311* (0.0085)</td>
</tr>
<tr>
<td>$P_L$ - others</td>
<td>0.2333* (0.0406)</td>
<td>0.1533* (0.0183)</td>
</tr>
<tr>
<td>Nurse per bed</td>
<td>0.2406* (0.0262)</td>
<td>0.1423* (0.0118)</td>
</tr>
<tr>
<td>Number of medical units</td>
<td>0.0380* (0.0149)</td>
<td>0.0573* (0.0070)</td>
</tr>
<tr>
<td>Number of non-medical units</td>
<td>-0.0321 (0.0358)</td>
<td>0.0743* (0.0179)</td>
</tr>
<tr>
<td>Number of training positions</td>
<td>0.1003* (0.0108)</td>
<td>0.0882* (0.0049)</td>
</tr>
<tr>
<td>Fraction of ambulatory clinics</td>
<td>-0.1747* (0.0705)</td>
<td>-0.3037* (0.0329)</td>
</tr>
<tr>
<td>Fraction of medical units recognized by FMH</td>
<td>0.0150 (0.0369)</td>
<td>0.0476* (0.0184)</td>
</tr>
<tr>
<td>Fraction of training positions A and B</td>
<td>0.0471* (0.0237)</td>
<td>0.0191 (0.0110)</td>
</tr>
<tr>
<td>Emergency Room</td>
<td>-0.0727* (0.0224)</td>
<td>0.0279* (0.0114)</td>
</tr>
<tr>
<td>Geriatrics</td>
<td>-0.0345* (0.0126)</td>
<td>-0.0281* (0.0057)</td>
</tr>
<tr>
<td>Year 1999</td>
<td>-0.0168 (0.0204)</td>
<td>-0.008 (0.0112)</td>
</tr>
<tr>
<td>Year 2000</td>
<td>0.0111 (0.0204)</td>
<td>0.0128 (0.0096)</td>
</tr>
<tr>
<td>Year 2001</td>
<td>0.0313 (0.0200)</td>
<td>0.0402* (0.0100)</td>
</tr>
<tr>
<td>Year 2002</td>
<td>0.0525* (0.0202)</td>
<td>0.0604* (0.0097)</td>
</tr>
<tr>
<td>Year 2003</td>
<td>0.0511* (0.0207)</td>
<td>0.0618* (0.0099)</td>
</tr>
<tr>
<td>Constant ($\alpha$)</td>
<td>0.2081 (0.2418)</td>
<td>1.2165* (0.1097)</td>
</tr>
<tr>
<td>$\sigma_u$</td>
<td>-</td>
<td>0.1378* (0.0031)</td>
</tr>
<tr>
<td>$\sigma$ [where: $\sigma^2 = \sigma_u^2 + \sigma_v^2$]</td>
<td>0.1592* (0.0002)</td>
<td>0.0872* (0.0036)</td>
</tr>
<tr>
<td>$\lambda = \sigma_u / \sigma_v$</td>
<td>1.0698* (0.1094)</td>
<td>1.5521* (0.2124)</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>391.59</td>
<td>610.19</td>
</tr>
</tbody>
</table>

* Significant at 5%; Standard errors are given in parentheses; Dependent variable is hospital's total costs in logs; All variables except dummies and the three fractions are in logarithms.
Table 4: Descriptive summary of inefficiency estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pooled</em></td>
<td>0.0926</td>
<td>0.0401</td>
<td>0.0299</td>
<td>0.0656</td>
<td>0.0859</td>
<td>0.1084</td>
<td>0.3249</td>
</tr>
<tr>
<td><em>True RE</em></td>
<td>0.0568</td>
<td>0.0291</td>
<td>0.0114</td>
<td>0.0387</td>
<td>0.0502</td>
<td>0.0673</td>
<td>0.2622</td>
</tr>
</tbody>
</table>

Table 5: Average inefficiency by ownership/subsidy type (pooled model)

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC</th>
<th>PRIVATE NON-PROFIT</th>
<th>PRIVATE FOR-PROFIT</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBSIDIZED</strong></td>
<td>0.0896</td>
<td>0.0941</td>
<td>0.093</td>
<td>0.0916</td>
</tr>
<tr>
<td><strong>NON SUBSIDIZED</strong></td>
<td>-</td>
<td>0.1099</td>
<td>0.0906</td>
<td>0.1021</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>0.0896</td>
<td>0.0967</td>
<td>0.0954</td>
<td>0.0926</td>
</tr>
</tbody>
</table>

Table 6: Average inefficiency by ownership/subsidy type (TRE model)

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC</th>
<th>PRIVATE NON-PROFIT</th>
<th>PRIVATE FOR-PROFIT</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBSIDIZED</strong></td>
<td>0.0551</td>
<td>0.0586</td>
<td>0.0594</td>
<td>0.0565</td>
</tr>
<tr>
<td><strong>NON SUBSIDIZED</strong></td>
<td>0.0591</td>
<td>0.0605</td>
<td>0.0597</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>0.0551</td>
<td>0.0587</td>
<td>0.0598</td>
<td>0.0568</td>
</tr>
</tbody>
</table>
Table A.1: Total cost function with a GLS model with random effects

<table>
<thead>
<tr>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private For-Profit</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Private Non-Profit</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Subsidized</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of hospitalizations (AP-DRG adjusted)</td>
<td>0.6238*</td>
</tr>
<tr>
<td></td>
<td>(0.0168)</td>
</tr>
<tr>
<td>Average length of hospitalizations</td>
<td>0.3914*</td>
</tr>
<tr>
<td></td>
<td>(0.0195)</td>
</tr>
<tr>
<td>Outpatient revenues</td>
<td>0.1467*</td>
</tr>
<tr>
<td></td>
<td>(0.0098)</td>
</tr>
<tr>
<td>( P_k ) (capital price)</td>
<td>0.1202*</td>
</tr>
<tr>
<td></td>
<td>(0.0091)</td>
</tr>
<tr>
<td>( P_L ) - physicians</td>
<td>0.0300*</td>
</tr>
<tr>
<td></td>
<td>(0.0127)</td>
</tr>
<tr>
<td>( P_L ) - others</td>
<td>0.1719*</td>
</tr>
<tr>
<td></td>
<td>(0.0299)</td>
</tr>
<tr>
<td>Nurse per bed</td>
<td>0.1472*</td>
</tr>
<tr>
<td></td>
<td>(0.0197)</td>
</tr>
<tr>
<td>Number of medical units</td>
<td>0.0600*</td>
</tr>
<tr>
<td></td>
<td>(0.0159)</td>
</tr>
<tr>
<td>Number of non-medical units</td>
<td>0.0839*</td>
</tr>
<tr>
<td></td>
<td>(0.0362)</td>
</tr>
<tr>
<td>Number of training positions</td>
<td>0.0826*</td>
</tr>
<tr>
<td></td>
<td>(0.0102)</td>
</tr>
<tr>
<td>Fraction of ambulatory clinics</td>
<td>-0.2867*</td>
</tr>
<tr>
<td></td>
<td>(0.0700)</td>
</tr>
<tr>
<td>Fraction of medical units recognized by FMH</td>
<td>0.0481</td>
</tr>
<tr>
<td></td>
<td>(0.0294)</td>
</tr>
<tr>
<td>Fraction of training positions A and B</td>
<td>0.0162</td>
</tr>
<tr>
<td></td>
<td>(0.0158)</td>
</tr>
<tr>
<td>Emergency Room</td>
<td>0.0260</td>
</tr>
<tr>
<td></td>
<td>(0.0255)</td>
</tr>
<tr>
<td>Geriatrics</td>
<td>-0.0305*</td>
</tr>
<tr>
<td></td>
<td>(0.0122)</td>
</tr>
<tr>
<td>Year 1999</td>
<td>-0.0121</td>
</tr>
<tr>
<td></td>
<td>(0.0086)</td>
</tr>
<tr>
<td>Year 2000</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>(0.0091)</td>
</tr>
<tr>
<td>Year 2001</td>
<td>0.0356*</td>
</tr>
<tr>
<td></td>
<td>(0.0087)</td>
</tr>
<tr>
<td>Year 2002</td>
<td>0.0566*</td>
</tr>
<tr>
<td></td>
<td>(0.0091)</td>
</tr>
<tr>
<td>Year 2003</td>
<td>0.0581*</td>
</tr>
<tr>
<td></td>
<td>(0.0097)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.2490*</td>
</tr>
<tr>
<td></td>
<td>(0.2103)</td>
</tr>
</tbody>
</table>

* Significant at 5% or less; Standard errors are given in parentheses; Dependent variable is hospital's total costs; All variables except fractions and dummies are in logarithms.
Economies of scale and scope in local public transportation

Mehdi Farsi, Aurelio Fetz, Massimo Filippini

Address for correspondence: Aurelio Fetz, Department of Management, Technology and Economics, ETH Zurich, Zurichbergstr. 18, CH-8032 Zurich, Switzerland (afetz@ethz.ch). Mehdi Farsi and Massimo Filippini are also at D-MTEC, ETH Zurich; Professor Filippini is also at the Department of Economics, University of Lugano. The authors acknowledge the financial support of the Swiss National Science Foundation through research grant 100012-108288. They also wish to thank the editor Steven Morrison and an anonymous referee for their helpful suggestions and Martin Hohmann for his assistance in compiling the data.

Abstract

This paper analyzes the cost structure of the Swiss urban public transport sector in order to assess scale and scope economies. A multi-output cost function has been estimated for a panel data set of companies operating trolley-bus, motor-bus and tramway systems. The results suggest increasing returns to scale and economies of scope. This analysis has important policy implications in view of the ongoing reforms in several European countries in which competitive tendering is occasionally used to assign the provision of transport services to unbundled franchised monopolies. The significant scope economies provide some evidence in favor of integrated multi-mode operation as opposed to unbundling.

Date of final manuscript: January 5, 2007
1.0 Introduction

During the last two decades several EU-member countries have introduced a competitive tendering procedure in the assignment of franchised monopolies in the local transport industry. This process has been initiated by the European Directive 1191/69/EU (modified by 1893/91/EU) that encourages the member countries to use competitive tenders in cases where the providers are not owned by home states. The implementation of tendering procedures in the urban transit industry is, however, not straightforward, because in many cases the incumbents are large multi-mode transit operators that combine different transport systems such as motor-bus, tramway and trolley-bus. In such cases companies specialized in a single transit service face a barrier to market entry because of the comparative advantage of the incumbent multi-mode companies. Therefore, due to fewer potential bidders for the multi-mode transit services, the benefits from competitive tendering would not be exploited completely.1

In order to induce more competition and avoid insufficient number of bids, there is a tendency among the public authorities toward opening separate competitive tenders for different modes of transport. This kind of ‘unbundling’ has been already used in the urban public transport systems in several European countries.2 When transport modes are unbundled,3 bidding can be opened to both single-mode operators and multi-mode companies. On the other hand, a multi-mode transit company serving the entire local market can completely exploit the potential scope and scale economies and reduce the planning costs of urban transport, since the local authority must not

---

1 For a discussion of the problems in the application of competitive tendering processes in the local transport sector see Cambini and Filippini (2003).
2 For instance, in Rome a competitive tendering procedure for some additional lines has been already utilized. See Cambini and Boitani (2006) for a discussion of this issue.
3 By unbundling we refer to legal unbundling (as opposed to ownership unbundling) which means the divestiture into more than one company. In principle the resulting companies can belong to a single holding firm.
coordinate a large set of services provided by different operators in order to have a well integrated network. Moreover, such integrated companies potentially provide a better quality of service with more stable and coordinated timetables to the extent that the disruptions can be minimized by substituting across transport modes.

The choice between a single tendering procedure for the entire transport services and unbundling the modes into separate tenders is a crucial policy question that has extremely important impacts on the organization of the local transport system namely, the operation mode (single or multiple) in different parts of a network as well as the planning of final services such as frequencies, number of lines etc. Therefore, it is relevant for the local authorities to know if and to what extent multi-mode suppliers could use the scope and scale economies to reduce their costs in comparison to a group of single-mode operators. This question is in line with the important issue of natural monopoly raised by Baumol, Panzar et al. (1982), applied to the local transport sector.

In the presence of economies of scope a multi-output firm is more economical than separate specialized firms. Following Baumol, Panzar et al. (1982) and Bailey and Friedlaender (1982) the scope economies can result from sharing or the joint utilization of inputs. In the case of local public transportation such sharable inputs are labor, capital and energy. Local public transport companies which combine several transport modes use similar equipment such as wires, overhead line and similar skills such as driving, management and network maintenance. Such synergies also apply to activities like advertising, scheduling and ticketing. Another source of cost savings is due to economies of massed reserves (Waldman and Jensen, 2001). Multi-output transportation companies can make use of the same reserve capacity for maintenance and buildings.
The purpose of this study is to make a contribution to the above debate on the introduction of competitive tendering procedures in the urban bus transportation sector. This paper explores the empirical evidence of scale and scope economies in 16 multi-mode transport companies operating in Switzerland from 1985 to 2003. A normalized total cost function with quadratic form has been estimated. The results suggest that scope economies exist for at most of the output levels observed in the data. This study provides some evidence in favor of the status quo regarding multi-mode transport companies. The potential competition benefits of unbundling remain to be explored.

The rest of the paper proceeds as follows: Section 2 provides a brief review of the relevant literature and presents the adopted specification. The concepts of scope and scale economies are defined in section 3. The data are described in section 4. Section 5 presents the estimation results and discusses their implications. The main conclusions are summarized at the end.

2.0 Model specification and econometric methods

There is a great body of literature on the cost structure of single output bus companies. Filippini and Prioni (1994), Fraquelli, Piacenza et al. (2004a) and Shaw-Er, Chiang et al. (2005) are among the recent empirical examples. However, only a few studies have addressed the issue of scope economies in urban transit systems. Authors such as Gillen and Oum (1984) studied companies operating with a single transport mode but in a multi-product set-up. In these cases the multiple outputs are defined on the basis of service type namely, urban, intercity etc. Previous studies on the economies of scope across different modes of transport (such as motor-bus, tramway, and trolley-bus) are rare and mostly outdated. The most relevant ones in this category are Viton (1992), Viton (1993) and Colburn and Talley (1992), both of which ana-
lyzed the long run cost structure of urban multi-mode transit system in the U.S.

Viton (1992) studied the cost structure of a sample of 289 urban transit companies operating in the U.S. between 1984 and 1986. Six modes are distinguished: motor-bus, rapid-rail, streetcar, trolley-bus, demand responsive mode and a last mode including all other modes. Viton uses a quadratic total cost function with the following explanatory variables: six outputs, measured in vehicles-miles, price of labor and the average speed in each one of the six modes. Empirical results highlight the presence of economies of scale and scope. However, the extent of the economies of scope depends on the post-consolidation level of the wage: If wages remain unchanged after consolidation, economies of scope exist for certain transportation modes. If, on the other hand, wages rise due to consolidation, economies of scope are smaller or even negative. Colburn and Talley (1992) analyze the economies of scale and scope of a single urban multi-service company using quarterly data from 1979 to 1988. Four modes are distinguished: motor-bus, dial-a-ride, elderly service, and van pool service. Colburn and Talley used a translog total cost function with the following explanatory variables: four outputs, measured in vehicles-miles, and three factor prices (labor, fuel and capital). The empirical results reported in that study indicate unexploited scale economies. However, the evidence of cost complementarity is limited to certain combinations involving motor-bus and the three para-transit services (elderly service, and van pool service).

In this paper we consider three modes that are typically used in most European urban transit systems namely, motor-bus, trolley-bus and tramway. We will employ a panel data econometric approach. To our knowledge this paper is the first empirical study of a European urban transit system that provides evidence about the economies of scale and scope across transport modes.
The model specification is based on a cost function with three outputs namely, transport services in three modes and two inputs, labor and capital. The model also includes a measure of network size and a time trend. If it is assumed that the firm minimizes cost and that the technology is convex, a total cost function can be written as:

\[ C = C(y^{(1)}, y^{(2)}, y^{(3)}, w^{(1)}, w^{(2)}, n, t), \]  

where \( C \) represents total costs; \( y^{(1)}, y^{(2)} \) and \( y^{(3)} \) are the numbers of seat-kilometers provided by trolley-bus, motor-bus and tramway systems respectively; and \( w^{(1)} \) and \( w^{(2)} \) are the factor prices for labor and capital respectively.

The network size (\( n \)) is measured by the number of stops\(^4\) and \( t \) is the linear trend which captures the shift in technology representing technical change.

Following Baumol, Panzar et al. (1982) and Mayo (1984) we use a quadratic cost function.\(^5\) Unlike logarithmic forms, this functional form accommodates zero values for outputs thus, allows a straightforward identification of scope economies. Although logarithmic functions could be used with an arbitrary small value transformation for zero values, it has been shown that this approach could result in large errors in the estimation of scope economies (Pulley and Humphrey, 1993). As in our case, many output values for trolley-buses and tramways are zero, such estimation errors may lead to misleading conclusions about scope economies. However, one disadvantage of the quadratic form is that the linear homogeneity of the cost function in input prices cannot be imposed by parametric restrictions without sacrificing the flexibility of the functional form (Caves, Christensen et al., 1980).

\(^4\) In two alternative specifications we respectively used area size and network length (sum of the three modes) instead of number of stops. Neither variable has shown any statistically significant effect at 5 per cent significance level. This can be partly explained by relatively high density variation within a service area and also variation of shape and complexity across different networks.

\(^5\) A quadratic function requires an approximation of the underlying cost function at a local point, which in our case is taken at the sample mean. Thus, all independent variables are normalized by their sample mean values.
The quadratic cost function specification can therefore be written as follows:

\[
C_{it} = \alpha_0 + \sum_{m}^{M} \alpha^m_{m(m)} y_{it}^{(m)} + \frac{1}{2} \sum_{m}^{M} \sum_{n}^{M} \alpha^m_{mn} y_{it}^{(m)} y_{it}^{(n)} + \sum_{p}^{P} \beta^p w_{it}^{(p)} + \alpha^n n_{it} + \alpha^t t_{i} + \varepsilon_{it},
\]

(2)

where superscripts \(m\) and \(p\) denote respectively, the number of products (1, 2, 3) and the number of input factors (1, 2), subscripts \(i\) and \(t\) denote respectively company and year. Variable \(y\) is a product quantity, \(w\) a factor price, \(t\) a time trend and \(n\) is a network characteristic. The factor prices and the network variable are introduced in a linear way (following Mayo (1984) and Viton (1992), respectively).

The linear homogeneity in input prices can be imposed by normalization of prices namely, by dividing the costs and all factor prices by one common factor price (see Featherstone and Moss (1994) and Jara-Diaz, Martinez-Budria et al. (2003)).

The normalized quadratic cost function takes the following form:

\[
C'_{it} = \alpha_0 + \sum_{m}^{M} \alpha^m_{m(m)} y_{it}^{(m)} + \frac{1}{2} \sum_{m}^{M} \sum_{n}^{M} \alpha^m_{mn} y_{it}^{(m)} y_{it}^{(n)} + \sum_{p}^{P-1} \beta^p w_{it}^{(p)} + \alpha^n n_{it} + \alpha^t t_{i} + \varepsilon_{it},
\]

(3)

where \(C'_{it}\) is the normalized cost and \(w_{it}^{(p)}\) is the normalized input prices. In this case, given that the model does not have any second-order term for input prices, the linear homogeneity restriction is equivalent to excluding one price coefficient.

The econometric model (3) is estimated for an unbalanced panel data set consisting of 16 companies over 19 years (300 observations). The repeated observations of a same company allow the use of panel data models that can account for unobserved heterogeneity across companies. However, as the number of companies is smaller than the number of periods \((N<T)\), this data set is an unusual case for widely used panel data specifications such as fixed effects and random effects models, in

---

6 We have also estimated the model without linear homogeneity restriction (as in Equation 2). The results (available upon request) are more or less similar to those of the normalized model. However, the statistical significance for scope and scale economies is slightly different across the two models with the normalized model having generally more significant values.
which $T$ is small relative to $N$.\(^7\) When sample period is relatively short, one can assume that the individual effects remain constant. In long panel data on the other hand these effects might change over time, resulting in serial correlation of errors.\(^8\) Both fixed and random effects models can be extended to include serial correlations with an autoregressive model of order 1 as in Cochrane-Orcutt approach (Cochrane and Orcutt, 1949). However, in fixed/random effects models a great part of the between variations (variations among companies) can be suppressed into the firm-specific effects. Given the small size of the sample and the relative importance of between variations in identifying the scope and scale economies, a pooled model seems to be adequate for our study.

Therefore, we decided to use a heteroscedastic model with autoregressive errors, as proposed by Kmenta (1986).\(^9\) The Kmenta approach, also known as the cross-sectionally heteroscedastic and time-wise autoregressive model, is attractive when $N$, the number of units, is lower than $T$, the number of periods, or when the within variation of many explanatory variables is low. In this model, the cross-sectional heteroscedasticity captures the unobserved heterogeneity across companies,\(^10\) while the serial correlation is modeled through the autoregressive error structure as follows:

\[
\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad \text{(autoregressive errors)}
\]

\[
E(u_t^2) = \sigma_i^2 \quad \text{(heteroscedasticity)}
\]

where $\rho_i$ is a coefficient of first-order autocorrelation. It is assumed that the correlation parameter varies across the firms. Therefore, the unobserved heterogeneity across

---

\(^7\) For a detailed presentation of panel data models, see Greene (2003) and Baltagi (2001).
\(^8\) The significant test statistics from autocorrelation test in panel data (Wooldridge, 2002) indicates the presence of serial correlation in the data.
\(^9\) The model has been also estimated using the fixed and random effects approaches. The results (available upon request) show that the estimated coefficients are generally similar to those reported in the paper. However, in the fixed effects model the coefficient of the network size variable is negative. This counterintuitive result could be due to extremely low within variation of that variable.
\(^10\) A modified Wald test on an OLS model shows the existence of heteroscedasticity in our data.
firms is accounted by firm-specific variances and serial correlation coefficients. The Kmenta method consists of two sequential feasible generalized least squares (FGLS) transformations to remove autocorrelation and cross-sectional heteroscedasticity respectively (Baltagi, 2001; Kmenta, 1986).

3.0 Scale and scope economies

Following Baumol, Panzar et al. (1982) global economies of scale\(^{11}\) in a multi-output setting are defined as:

\[
SL = \frac{C(y)}{\sum_{m} y^{(m)} * \frac{\partial C}{\partial y^{(m)}}}, \quad (6)
\]

where \(y=(y^{(1)}, y^{(2)}, y^{(3)})\) for \(m=1\) (trolley-bus), 2 (motor-bus) and 3 (tramway). Global economies of scale describe the cost behavior due to proportional changes in the entire production.

In addition, product-specific economies of scale are based on changes of one output or an output pair, while all other outputs are held constant. Product-specific economies of scale to the product \(m\) are defined as:

\[
SL_m = \frac{C(y) - C(y^{(-m)})}{y^{(m)} * \frac{\partial C}{\partial y^{(m)}}}, \quad (7)
\]

\(^{11}\) In the definition of economies of scale we do not follow Caves, Christensen et al. (1984) by distinguishing between economies of scale and economies of density due to the complexity of the weighting of different network elements.
where $C(y) - C(y^{(m)})$ represents the incremental cost resulting from output $m$ and $C(y^{(-m)})$ is the costs of producing all the outputs jointly except output $m$.

Similarly, in the case of joint production of outputs $m$ and $n$ the product-specific economies of scale can be written as:

$$SL_{mn} = \frac{C(y) - C(y^{(-mn)})}{y^{(m)} \ast \frac{\partial C}{\partial y^{(m)}} + y^{(n)} \ast \frac{\partial C}{\partial y^{(n)}}}.$$

where $C(y^{(-mn)})$ is the costs of producing all the outputs except outputs $m$ and $n$.

All the above definitions represent the ratio of the expansion rate in all or certain output(s) to the rate of the resulting increase in costs. For any one of these cases (global, output $m$, or output pair $m$ and $n$) the returns to scale are increasing, constant or decreasing if the corresponding ratio ($SL$, $SL_m$ or $SL_{mn}$) is greater, equal or less than one.

Economies of scope are present when costs can be reduced by joint production of multiple outputs. Following Baumol, Panzar et al. (1982) the degree of global economies of scope in the production of three products is defined as the ratio of excess costs of separate production to the costs of joint production of all outputs:

$$SC = \frac{C(y^{(1)}, 0, 0) + C(0, y^{(2)}, 0) + C(0, 0, y^{(3)}) - C(y)}{C(y)}.$$

A positive (negative) value for the above expression implies the existence of global economies (diseconomies) of scope.
In addition to the above measure, product-specific measures can be defined for any given output or output pair. The product-specific economies of scope for output $m$ ($SC_m$) measure the relative increase in costs resulting from separating the production of output $m$ from all other outputs:

$$SC_m = \frac{C(y^{(m)}) + C(y^{(-m)}) - C(y)}{C(y)}. \quad (10)$$

In line with Fraquelli, Piacenza et al. (2004b) the degree of product-specific economies of scope for output pair $m$ and $n$ with the remaining output being zero, is defined as:

$$SC_{mn} = \frac{C(y^{(m)}) + C(y^{(n)}) - C(y^{(m)}, y^{(n)})}{C(y^{(m)}, y^{(n)})}. \quad (11)$$

Product-specific economies (diseconomies) of scope exist if $SC_m$ or $SC_{mn}$ is greater (smaller) than zero.

4.0 Data

The sample consists of sixteen public transport companies, which cover all the local public transit services within the urban centers in Switzerland. Most of these companies participate in a transport association that ensures coordination with the regional public transport system. There is no overlap between the offered transport services across companies. Therefore, all these companies can be considered as independent local monopolies.

---

12 Swiss Federal Statistical Office classifies local public transport into urban and rural categories (BFS, 1985-97). The sample used in this paper excludes rural transport companies. These companies are generally bus operators that cover relatively long distances with low frequencies in rural areas, which is considered as a different transport operation compared to urban transport.

13 These associations are created for setting the prices and organizing the ticketing. They also ensure that passengers can use a single time-table and travel throughout the entire associated network with only one ticket. However, the participating companies operate independently.
For the years between 1985 through 1997 the data have been extracted from the annual statistics on public transport reported by the Swiss Federal Statistical Office (BFS, 1985-97). The data for the following years (1998–2003) have been collected from companies’ annual reports.\textsuperscript{14} Because of a merger with a regional transportation company in 1999, one company was excluded from the sample after the merger. The available information in the dataset includes costs, total number of employees, network length, total numbers of trolley-buses, motor-buses and tramways, vehicle-kilometers, delivered passengers and total number of seats in each transportation mode.

The variables for the cost function specification were calculated as follows. Total costs ($TC$) are calculated as the total expenditures of the local public transit firms in a given year. The output $y$ is measured by the number of seat-kilometers provided by motor-buses, trolley-buses and tramways, respectively. This is a pure supply output measure that has been used in previous studies for bus companies, such as Fazioli, Filippini et al. (1993), Farsi, Filippini et al. (2006). Filippini and Prioni (2003) compared a model with bus-kilometers with one with seat-kilometers as output. The bus-kilometer output variable has the disadvantage that the size of the bus is not taken into account. Alternatively some authors have used passenger revenue (as in Button and O'Donnell, 1985) or passenger trips (Berechman, 1987; Bhattacharyya, Kumbhakar et al., 1995; Windle, 1988).

Labor price ($w_1$) is defined as the ratio of annual labor costs to the total number of full-time equivalent employees. The largest fraction of total costs is for labor costs (61 per cent on average). Following Friedlaender and Chiang (1983), the capital price ($w_2$) is calculated as residual cost (where residual cost is total cost minus labor).

\textsuperscript{14} The Federal Statistical Office does not provide data on individual companies after 1997.
divided by the total number of seats in the operator’s fleet.\textsuperscript{15} Capital price is therefore a measure of all non-labor inputs including materials and energy input.\textsuperscript{16} Unfortunately no data were available which would allow us to calculate the capital stock using the capital inventory method. The use of a simple indicator for capital stock can be justified by the fact that the bus companies do not possess a significant stock of capital apart from the rolling stock.

Table 1 provides the sample’s descriptive statistics. All the costs and prices are adjusted for inflation using consumer price index and are measured in year 2000 Swiss Francs (CHF). As can be seen in the table, the sample shows a considerable variation in all three outputs. All the companies in the sample provide motor-bus transport. As there were only seven companies (out of 16) with non-zero tramway services, we see from the table that the median output of the tramways is zero.

Table 1: Descriptive statistics (300 observations)

Among the sixteen companies in the sample, six offer all three modal transit services; nine provide motor-bus and trolley-bus services; and one firm offers motor-bus and tramway services. All the companies in the sample except one are multi-output operators.\textsuperscript{17} Therefore the single-output cases can be considered as exceptions in Switzerland’s urban transit system. This issue might create problems in estimating some of the product-specific scale and scope economies, which are based on extrapolation of costs for specific output combinations with zero values. However, this

\textsuperscript{15} For an application of this approach in the bus industry see Filippini and Prioni (2003) and Farsi, Filippini et al. (2006).

\textsuperscript{16} The energy price is not included directly because energy costs are only a small fraction of total costs (3.4 per cent on average and less than 6.3 per cent for 95 per cent of observation) and also because it does not vary much over time.

\textsuperscript{17} The exception is a bus company that has taken all its trolley-buses out of service in the last three years of the sample period.
should be considered in view of the fact that these cost values are second-order approximations at boundary points. These points being far from the sample mean might incur relatively high approximation errors in any case.

Moreover, a careful analysis of the sample shows that a considerable number of the companies are highly specialized in a single transport mode. For instance, there exist four two-output companies (about a quarter of the entire sample) whose motor-bus or tramway output consists of more than 80 percent of their total output in terms of seat-kilometers. Therefore, the estimated scale and scope economies can be reasonably extended to such cases provided that the single-output and multi-output operators use a more or less similar production technology.

For the estimation of product-specific scope and scale economies we will focus on output combinations that are realistic in comparison to the observed cases in the sample. Therefore, we exclude the single-output case with trolley-bus and the two-output case with motor-bus/tram combination, because there is no case with strong specialization that can be considered closely similar to these two cases.

5.0 Results

The estimation results are given in Table 2. As expected, the first order output coefficients ($\alpha^1$, $\alpha^2$ and $\alpha^3$) and that of the input price ($\beta^2$) are positive and highly significant. The first-order coefficients suggest that tramway system has the highest marginal cost followed by trolley-bus and motor-bus. This order can be explained by the relatively high costs of tramway and trolley-bus systems that require an electricity grid and an additional railway network in the case of tramway. Both these elements have relatively high capital and labor costs that accrue the corresponding marginal costs. The quadratic output coefficients terms ($\alpha^{11}$, $\alpha^{22}$ and $\alpha^{33}$) are negative (except
\(\alpha^{11}\) which is statistically insignificant), suggesting that the marginal cost of any given output is decreasing in that output.

As labor price is used for normalization, the coefficient \(\beta^1\) is excluded from the normalized quadratic cost function. As expected, the sign of the coefficient \(\alpha_n\) is positive, showing that a higher number of stops increase costs. The negative coefficient \(\alpha_t\) show that companies have reduced their operating costs in the sample time period. The autocorrelation coefficient has also been estimated for all companies. The considerable variation of these values confirms the assumption that the correlation structure varies across firms. These coefficients are greater than .7 for half of the companies, suggesting the importance of serial correlation in our sample.\(^{18}\)

**Table 2: Regression results**

The results presented in Table 2 can be used to estimate the economies of scale and scope.

In order to study the variation of scale and scope economies in the sample, we considered several representative sample points regarding outputs. In particular, we estimated the scale economies respectively for output values at the sample mean, median, 1\(^{st}\) and 3\(^{rd}\) quartiles of non-zero output values. For all non-output variables that enter in the equations, we considered the sample mean values. For instance, the median point consists of the medians of outputs after excluding the zero values with all other variables kept at their mean values. As discussed in the data section, for the

\(^{18}\) The estimated correlation coefficients range from .03 to .99 with an average value of .64 and with a value greater than .5 for 12 companies.
product-specific economies, only the realistic output combinations are estimated. Namely, the single-output trolley bus and the two-output tram/motor-bus are excluded.

Table 3 shows the point estimates of scale economies along with their standard errors. The estimated global scale economies are significantly greater than one, suggesting increasing returns to scale at all the considered output levels. This implies that the scale economies are not fully exploited in a great majority of the companies. The results also indicate that most of the estimated product-specific coefficients are significantly different from one. Only for a joint production of trolley-bus and motor-bus the hypothesis of constant product-specific returns to scale cannot be rejected.

Table 3: Estimates of the economies of scale

Table 4 shows the point estimates for economies of scope. Similar to the previous table, only the realistic output combinations are included. Most of the numbers in Table 4 are significantly greater than zero, indicating scope economies at all output levels and across all three output types. For instance the global scope economies of 0.25 at the sample median means that at the median output levels total costs are on average 25 percent lower when offering all the three outputs by one company than produced by three specialized firms. In general economies of scope are decreasing with an increase in outputs. The computed values are statistically different from zero at the 1st quartile and median output levels as well as some of the mean output levels.

While global scope economies remain significant for all output levels, the product-specific economies of scope seem to be exhausted after a certain level of output, as indicated by mostly insignificant values at the 3rd quartile output level. The
product-specific economies of scope in the single-output case, that is when one output is produced by a company and the other two by another company, are in a similar range for both tram and motor-bus at the 1st quartile and median output levels (9 – 13 per cent for median and 19 – 21 per cent for 1st quartile). The joint production of trolley-bus and tramway also yields economies of scope. Even higher are the cost savings from a joint production of trolley-bus und motor-bus (37 per cent at the 1st quartile output level).

Table 4: Estimates of the economies of scope

It should be noted that the value of the global economies of scale depends upon both product specific economies of scale and economies of scope. The results listed in Table 3 and Table 4 suggest that the global scale economies are driven by strong economies of scope and to a lower extent by product-specific economies of scale. As pointed out by Baumol, Panzar et al. (1982) decreasing average incremental costs of each product along with global scope economies imply subadditivity in costs. Therefore, the results of this paper provide suggestive evidence that the industry can be characterized as a natural monopoly.

6.0 Discussion

Several European countries have introduced a competitive tendering procedure in the assignment of franchised monopoly in the local transport industry. In the case of multi-mode systems the regulator has to decide to open the competitive tendering procedure for supplying the entire transport services or to unbundle the multi-mode
systems and open separate tenders for different modes of transport. In order to make the decision the regulator should have information on the economies of scope. Few studies have addressed the issue of scope economies in local transport systems.

The goal of this paper is to make a contribution to the ongoing debate about tendering local transport services. It is argued that the unbundling of transport modes in competitive tendering brings about efficiency benefits since the level of potential bidders is high and thus competition is more intense. On the other hand, integrated companies benefit from the potential scope economies and reduce the total cost of service planning, since the local authority has not to integrate the entire range of services provided by the different operators. Of course, in this case the potential number of bidders would be relatively low, since it is difficult for a small operator to provide services in a big city. Therefore, the potential benefits from competition for the market would be lower.

The tradeoff between these two efficiency gains lies at the core of the policy debate about unbundling. Using an empirical analysis of the cost structure of the urban transit companies in Switzerland this paper provides an assessment of scope and scale economies that would be compromised as a result of unbundling. The considered transport modes are motor-bus, trolley-bus and tramway systems.

The estimation results indicate considerable economies of scope, suggesting that unbundling a multi-mode company into single-output companies might lead to higher costs as the synergies in the joint production can no longer be exploited. Moreover, the results indicate increasing returns to scale in almost all outputs which combined with cost complementarity, can be considered as a suggestive evidence for natural monopoly.
The results of this study provide some insight to the efficiency trade-off of unbundling between the loss of economies of scope and the gain of higher cost efficiency from the introduction of competition for market entry. The assessment of the efficiency gains of unbundling through greater competition remains an open question that needs further research. In particular, it is not clear that the unbundling can be effective in lowering barriers to market entry for some of the transport modes such as tramways that require relatively high infrastructure costs.

An alternative to competitive tendering procedures for the multi-mode transport system could be the introduction of incentive regulation schemes such as ‘yardstick competition’ in which cost-efficiency is induced by controlling each local monopolist based on average costs of similar firms (Shleifer, 1985). The advantage of such regulatory instruments is that they allow a complete exploitation of the economies of scale and scope while avoiding the implementation problems related to competitive tendering policies for urban transit systems.

Finally, it should be noted that unbundling might also have undesired consequences for quality of service. Namely, the integrated multi-mode operators have a better flexibility in minimizing the disruptions and lower the instability of their timetables by reallocating across different modes. Moreover, with lower transaction costs related to information and communication, such companies have a greater possibility in creating trust among the consumers, thus reaching higher levels of quality as perceived by the society.
References


Abstract

This paper studies the cost structure of gas distribution utilities in Switzerland. Three stochastic frontier models are applied to a panel of 26 companies operating from 1996 to 2000. Efficiency is assumed to be constant over time. The analysis highlights the importance of output characteristics such as customer density and network size. The results suggest that the utilities could slightly reduce their operating costs by improving efficiency. There is no evidence of significant unexploited scale economies. However, our analysis indicates that the estimates of scale economies could be sensitive to the assumptions regarding the variation of output with output characteristics.

* Corresponding author: Mehdi Farsi, D-MTEC, ETH Zurich, ZUE, Zurichbergstr. 18, CH-8032, Switzerland. Tel.: +41-44-632-0656; Fax: +41-44-632-1050; E-mail: m.farsi@ethz.ch
The authors would like to thank the editor and two anonymous referees for their helpful suggestions and Jorg Wild for his support and collaboration.
1. Introduction

In the last decade, many countries started to liberalize their energy markets. Since 2000, the European Union has been gradually opening its gas networks to third party access and the consumers are allowed to choose their optimal contracts. Switzerland is expected to follow its neighboring countries and open its gas market to competition. The general idea is to introduce competition in the wholesale and retail markets, and to have a regulated natural monopoly in the transmission and distribution sectors. Therefore, the network access prices need to be regulated. Generally, the regulation can follow a traditional approach such as rate-of-return regulation or an incentive-based mechanism like price-cap or yardstick regulation.\footnote{See Joskow and Schmalensee (1986), Littlechild (1983), Shleifer (1985) and Laffont and Tirole (1993) for a discussion of these approaches.} The application of these three types of regulation approaches calls for a better understanding of the cost structure and the efficiency of the distribution companies. Especially in the case of incentive-based approaches, the regulator could use this information to induce efficient performance. For instance, predicted costs can be used in setting the yardstick competition targets or the efficiency scores can be used as the X-factor in price cap formulas to reward/punish companies according to their performance.\footnote{For a discussion on the use of inefficiency indicators in applied regulation see Rossi and Ruzzier (2000) and Farsi and Filippini (2004).} Moreover, information about the potential scale economies can help the regulators to evaluate the possibilities of cost saving by encouraging mergers and joint ventures among small companies.

Inefficiency in production may come from two different sources: deficiency in applying the technology (technical inefficiency) and suboptimal allocation of resources (allocative inefficiency). Productive inefficiency subsumes these two concepts and can be
measured by input or output oriented measures.³ One of the commonly used measures of productive inefficiency is the deviation from minimum costs to produce a given level of output with given inputs prices. This measure, although usually referred to as cost inefficiency, does not include the inefficiencies due to suboptimal scale of production.

Generally, there are two main approaches for estimating cost inefficiency: the non-parametric approach originated from operations research, and the econometric approach.⁴ The parametric methods use econometric theory to estimate a cost function with a specified form, where the inefficiency is modeled as additional stochastic term. The non-parametric methods like Data Envelopment Analysis (DEA) calculate an efficient deterministic frontier by linear programming and do not require a pre-specified functional form. Murillo-Zamorano (2004) provides an account of advantages and shortcomings of each one of these methods. In this paper we focus on parametric methods. The main advantage of these methods over non-parametric approaches is the separation of the inefficiency effect from the statistical noise due to data errors and omitted variables.

It should be noted that the results in term of inefficiency obtained using different approaches can be quite different. Thus, it is important to consider certain consistency conditions. Bauer et al. (1998) and Rossi and Ruzzier (2000) propose a series of criteria to evaluate if the inefficiency estimates obtained from different models are “mutually consistent”, that is, lead to comparable inefficiency scores and ranks. Farsi and Filippini (2004) recommend using the results of a benchmarking analysis as a complementary instrument in incentive regulation and not in a mechanical way.

³ See Russel (1998) for a discussion of different measures of productive efficiency.
The main goal of this paper is to study the sensitivity of inefficiency estimates using panel data models. We apply different parametric models to a five-year panel of 26 Swiss gas distributors. The inefficiency scores estimated from four different models are compared and the consistency of the estimates across different models is discussed. The economies of scale and density are also estimated. A slightly modified definition of scale economies is used to study the sensitivity of the results to the assumption of equi-proportional changes in output characteristics.

The rest of the paper is structured as follows: Section 2 provides a brief review of the empirical literature in gas distribution sector. Section 3 discusses the different cost frontier models used in the paper. The data are described in section 4. Section 5 presents the estimation results and discusses their implications regarding inefficiencies. Section 6 discusses the results on economies of scale and density. At the end, the main conclusions are summarized.

2. Review of the literature

The literature on econometric estimation of cost or production functions in gas distribution companies is scarce. Table 1 lists a selection of these papers. Hollas and Stansell (1988)\textsuperscript{5} are probably the first authors who analyzed this industry by modeling technical and allocative inefficiency. Their model is a behavioral translog profit function that includes output and price “shifters” for four types of companies. Their method allows identifying the relative efficiency of different types with respect to each other, but it does not provide any firm-specific inefficiency estimate. In their specification, they include the price of fuel, labor price, customer density and the fixed capital input measured in daily throughput capacity of the distribution system.

\textsuperscript{5} See Hollas and Stansell (1994) for a similar application to estimate the economic efficiency of public and private gas distribution utilities.
Kim and Lee (1996) highlight the importance of accounting for output characteristics in estimating a translog cost function for gas distributors. In addition to the labor price and the unit price of pipeline, these authors include the customer density, the average “customer size” measured as average consumption and the “supply rate” measured as the number of total customers relative to the number of total potential customers. Bernard et al. (1998) consider the load factor and the network length as major cost drivers that should be included as output characteristics.

Granderson and Linvill (1999) and Granderson (2000) used an eleven-year panel of 20 U.S. interstate natural gas transmission companies to produce a benchmark for regulation. As inputs, they specify labor, fuel, the weight of the transmission pipelines, and the capacity of compressor station and estimate a translog cost frontier by a random effects GLS model (Schmidt and Sickles, 1984). They also used DEA to get non-parametric estimates of inefficiency and compare the results. Although the results show that when using the non-parametric approach, the inefficiency estimates are lower, the inefficiency ranking stays more or less the same.\(^6\)

Fabbri et al. (2000) estimated a total distribution translog cost function for 31 Italian companies observed during two years. They use the yearly average cost per employee as labor price, the book-value of equipment divided by the length of the distribution network as capital price and the price of material and services is calculated as the residual expenses divided by network length. Output is measured as the volume of gas delivered and the number of customers. Their specification also includes the ratio of network length to the number of customers, share of urban population, the average altitude of the service area, and dummy variables for ownership differences and time effects. Their results suggest a more cost

---

\(^6\) This result is generally consistent with those reported by Carrington et al. (2002) who used the DEA approach and a translog input distance function to estimate efficiency of gas distribution companies.
efficient production in private firms. These authors also found that the economies of scale are not significant at the output levels in the data. On the other hand, economies of density appear to be considerable. These results are in line with most of the findings reported in other studies.

### Table 1: An overview of previous studies

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Database</th>
<th>Functional Form</th>
<th>Measures of output and characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sing (1987)</td>
<td>Cross section of 108 privately owned US gas and electricity distribution utilities, 1981</td>
<td>Hybrid translog cost function with Box-Cox transformation on outputs</td>
<td>Volume of gas delivered / customers per service area</td>
</tr>
<tr>
<td>Bernard et al. (1998)</td>
<td>Cross section of 131 Canadian gas extension projects</td>
<td>Box-Cox cost function</td>
<td>Max. daily demand / pipe length</td>
</tr>
</tbody>
</table>

Rossi (2001) estimated a stochastic frontier production function using the approach suggested by Battese and Coelli (1992). Rossi used the network length as a proxy for capital input, and the number of employees as the labor input. In addition, the concession area, the ratio of residential sales to total sales and the maximum demand are considered as the environmental variables. The number of customers is used as a single output. In one of his specifications the results suggest significant diseconomies of scale, but another model’s results do not reject the hypothesis of optimal scale.
3. Methodology

A frontier cost function defines minimum costs given output level, input prices and the existing production technology. Failure to attain the cost frontier implies the existence of technical and/or allocative inefficiency. This section provides a description of the cost frontier models and the specification used in this paper. The adopted methodology is based on a comparison of different models with respect to the estimated cost function parameters and estimated inefficiency scores. The main goal is to study the limitations of different models in benchmarking and the sensitivity of inefficiency scores to econometric modeling.

3.1 Cost Frontier Models

In this paper, we consider the estimation of a stochastic frontier cost function using different panel data models. The theoretical development of stochastic frontier models in panel data has been subject of a great body of literature. Recent developments such as random parameter frontier models usually require relatively large samples with sufficient variation. Considering the small size of our sample and the limited number of periods these methods do not appear to be effective. Considering the characteristics of our data set and the purpose of the paper, we decided to use three classical frontier models for panel data and a variant of one of these models.

The first model is the stochastic cost frontier approach proposed by Aigner et al. (1977). This model is a cross sectional model but could be applied to a panel data set by

---

7 See Kumbhakar and Lovell (2000) for a review and Greene (2005) and Tsionas (2002) for some recent developments.

8 In a preliminary analysis we applied some of these models especially the true random effects proposed by Greene (2005, 2004). The results generally indicate that due to the insufficient variation in the data, some of the random terms degenerate to zero. This suggests that these types of specifications are too rich for our data. For an application of these models in other network industries, see Farsi et al. (2006, 2005a).
pooling the data across different years. This model, which we refer to as the “pooled model”, can be written as:

\[
\ln C_{it} = \ln C \left( y_{it}, w_{it} \right) + u_i + v_{it} \geq 0, \quad i=1, 2, ..., N \quad \text{and} \quad t=1, 2, ..., T. \quad (1)
\]

In this specification the error term is composed of two independent parts: The first part \( u_{it} \), is a one-sided non-negative disturbance reflecting the effect of inefficiency, and the second component \( v_{it} \), is a symmetric disturbance capturing the effect of noise. The statistical noise is assumed to be normally distributed, while the inefficiency term \( u_i \) is assumed to follow a half-normal distribution. This model can be estimated using Maximum Likelihood Estimation (MLE) method. Individual inefficiencies can be estimated by the conditional expectation of the inefficiency term, \( E \left( u_i \left| u_i + v_i \right. \right) \), proposed by Jondrow et al. (1982).

In the pooled model, a given firm’s inefficiency is supposed to vary with time, but could take a value completely independent of its level in the previous years. An extension of this model applied to panel data has been proposed by Pitt and Lee (1981). This model can be written as follows:

\[
\ln C_{it} = \ln C \left( y_{it}, w_{it} \right) + u_i + v_{it} \geq 0, \quad i=1, 2, ..., N \quad \text{and} \quad t=1, 2, ..., T. \quad (2)
\]

In this specification it is assumed that the inefficiency term \( u_{it} \) follows a half-normal distribution across firms but stays constant over time within a given firm. Although this assumption might appear unrealistic given the fairly long period covered in the sample, our preliminary analyses indicate that the temporal variations of inefficiency are not significant in our data.\(^9\) Similarly, this model is estimated using MLE and the individual inefficiencies

\(^9\) We first explored the time variation by including a linear trend and year dummies in explanatory variables. None of these time variables showed any significant effect. We also applied an extension of Pitt and Lee (1981)’s model proposed by Battese and Coelli (1992) in which the inefficiency is assumed to follow a time-
are estimated by the conditional expectation of the inefficiency term,\(^{10}\) given by 
\[ E(u_i | \varepsilon_{i1}, ..., \varepsilon_{iT}) \],
where \( \varepsilon_i = u_i + v_i \) with \( i = 1,2,...,T \).

The assumptions about the distribution of the random terms \( u_i \) and \( v_{it} \) can be relaxed by rewriting equation 2 as:

\[ \ln C_u = \ln C(y, w) + \alpha + u_i^* + v_i \]
with \( u_i = u_i^* - \min \{ u_i^* \} \) and \( \alpha = \min \{ u_i^* \} \), (3)
where \( u_i^* \) is the firm-specific random effect. The resulting model, proposed by Schmidt and Sickles (1984), can be estimated using the feasible Generalized Least Squares method.

The remaining restrictive assumption is that the random effects are uncorrelated with the observed characteristics included in the cost function. Schmidt and Sickles (1984) propose a solution to relax this assumption by using the fixed effects specification. The main shortcoming of this model is that the inefficiency measures may be confounded with time-invariant factors, which could not be included in the model.\(^{11}\) To avoid this problem and at the same time controlling for the potential correlation of firm-specific effects with explanatory variables, we applied Mundlak (1978)'s formulation to the GLS model.\(^{12}\) The correlation of firm-specific unobserved effects with explanatory variables are captured in an auxiliary equation given by:

\[ \alpha_i = \gamma X_i + \delta_i, \text{ where } X_i = \frac{1}{T_i} \sum_{t=1}^{T_i} X_{it} \text{ and } \delta_i \sim iid(0, \sigma_\delta^2). \] (4)

varying exponential decay function. The results indicate that the coefficient of time is statistically insignificant.

Moreover, the estimated efficiencies are very close to those obtained from the Pitt and Lee model.

\(^{10}\) Jondrow et al. (1982) provide the expression for conditional mean in cross sectional data. The extension for panel data is given in Kumbhakar and Lovell (2000), page 111 and Greene (2002), section 24.

\(^{11}\) See Farsi and Filippini (2004) for a discussion of advantages and drawbacks of fixed and random effects models.

\(^{12}\) See also Farsi et al. (2005b) for a similar specification.
$X_{it}$ is the vector of all explanatory variables and $\gamma$ is the corresponding vector of coefficients. This formulation divides the firm-specific term into two components: The first part can be explained by the exogenous variables and is interpreted as heterogeneity, whereas the remaining component $\delta_i$ is orthogonal to the explanatory variables and is assumed to reflect the company’s inefficiency. Equation (4) is incorporated in the main regression equation:

$$\ln C_{it} = \ln C(y_{it}, w_{it}) + \gamma \bar{X}_i + \delta_i + v_{it} \quad \text{with} \quad u_i = \delta_i - \min_i \delta_i . \quad (5)$$

This model is estimated by the GLS method. Mundlak (1978) showed that this formulation of the GLS model results in a coefficient vector of $\ln C(y_{it}, w_{it})$ that is equivalent to the within estimator (fixed effects model) and thus would be unbiased even in presence of firm effects that are correlated with explanatory variables. It should be noted that the fixed effects model (FE) can also be used to estimate the inefficiencies (as proposed by Schmidt and Sickles, 1984). However, we do not use the FE model here, because the estimated coefficients in this model do not bring any additional value to our study as they are exactly equal to those obtained from the GLS model with Mundlak’s adjustment. Moreover, the inefficiency estimates from the FE model are overestimated because they include some of the factors (such as area size and customer density) that are almost time-invariant. This issue has been highlighted in an example reported by Farsi and Filippini (2004).

### 3.2 Specification of the Cost Function

Gas distribution companies operate in networks with different shapes and environmental characteristics, which directly affect costs. The output is measured as total volume of natural gas delivered. Input factors consist mainly of the gas purchased from a transmission company, labor and capital. Therefore, there are in principle two alternatives for

---

13 For a proof of this statement, see also Hsiao (2003), Section 3.4.2a.
measuring inefficiency: an integrated total cost approach and a network operating cost approach, where expenditures for gas purchases are excluded. The network costs approach has a practical advantage in that the estimated average costs can be directly used in a price-cap formula. However, this approach neglects the potential inefficiencies in the choice of the gas delivery contract. Therefore, in this paper we adopt the total cost approach. The cost function is specified as:

\[ TC = f(Y, P_C, P_L, P_E, LF, TB, CUD, ASIZE), \]  

where \( TC \) represent total costs; \( Y \) is the energy value of the delivered gas measured in MWh; and \( P_C, P_L \) and \( P_E \) are respectively the prices of capital, labor and purchase price of natural gas. In addition to these variables, we also include the load factor \( LF \), the number of terminal blocks \( TB \), the customer density \( CUD \) and the area size \( ASIZE \) as output characteristics.

The load factor is defined as the ratio of annual average flow of gas to the annual peak flow per hour. It is a measure of how constant the network capacity is used throughout the year. A higher value of load factor implies a lower variation in consumption. The load factor is a demand characteristic and cannot be directly influenced by the company. A network with a low load factor needs more capacity. We therefore expect the coefficient of the load factor to be negative, implying lower costs for companies with more evenly distributed network use.

The number of terminal blocks (\( TB \)) is another output characteristic included in the model. Terminal blocks are usually located at the entrance of the buildings and serve several end-users. It is also possible that two or more buildings share a terminal block when they are internally connected. Typically, a terminal block is owned and maintained by the gas distribution company, whereas the pipeline following the block belongs to the customer. The number of blocks is expected to have a positive effect on costs through higher hookup, maintenance and billing costs. Often, the customer density, measured as number of clients per
kilo-meter of network length (CUD), is also considered as an important indicator of costs.\textsuperscript{14} The average maintenance cost per customer is lower in networks with higher density, suggesting a negative sign for CUD. The area size (ASIZE) of each utility is included to measure the company’s size.\textsuperscript{15} Larger service areas generally require larger and more spread networks, thus more operating and maintenance costs.

The regularity conditions require that the cost function in equation (6) be non-decreasing in input prices and output, and linearly homogeneous and concave in input prices.\textsuperscript{16} In the empirical literature, two main functional forms have been used: the translog and the Cobb-Douglas form. In general, the translog form provides a more flexible framework especially regarding the scale economies, which can vary with the output. However in this paper, given the small size of the sample and the large number of parameters\textsuperscript{17} in the translog model, we use the Cobb-Douglas form.

The Cobb-Douglas specification of the cost function in (6) can be written as:

\[
\ln \left( \frac{TC}{P_L} \right) = \beta_0 + \beta_Y \ln Y + \beta_C \ln \left( \frac{P_C}{P_L} \right) + \beta_L \ln \left( \frac{P_L}{P_L} \right) + \beta_{LF} LF + \beta_{TB} \ln TB + \beta_{CUD} \ln CUD + \beta_{ASIZE} \ln ASIZE \tag{7}
\]

\textsuperscript{14} Sing (1987) uses customers per square mile of service area as density.

\textsuperscript{15} An alternative measure would be the network length. Given that the latter variable is highly correlated with the area size (correlation coefficient of 0.94), we decided to use the actual length of the network in calculating the capital price and customer density and the area size as a measure of the utility’s size.

\textsuperscript{16} See Cornes (1992) for a discussion of the properties of cost functions.

\textsuperscript{17} In our specification the number of coefficients in translog model would be 36, which results in a relatively small number of degrees of freedom in a sample of 26 companies with 129 observations.
The linear homogeneity condition is imposed by normalizing costs and prices to the labor price. All variables except the load factor are in logarithms. The load factor is a positive value lower than one, thus is not transformed into logarithm.\textsuperscript{18}

4. Data

Based on the year 2000 statistics, there are about 128 gas distribution companies in Switzerland. Generally, the distribution companies own the distribution network in which they operate and there is no overlap between the service areas of individual companies. The transmission network is owned and operated by Swiss Gas, a company mainly held by four major public regional gas distributors. Swiss Gas is in charge of transmitting about three fourths of the total national gas consumption. The remaining part is delivered by the transmission pipelines owned by neighboring countries. Distributors purchase the gas from transmission companies and deliver it through their own distribution networks to the end-use consumers.

The data used in this paper is based on a mail survey from 26 distribution companies accounting for about 57 percent of the total gas consumption in Switzerland. These companies participated in the survey in a voluntary basis. None of the four regional companies are included. All the participants except one provided the data for five years from 1996 to 2000. The 1996 data is not available for one company. Thus, the sample consists of 129 observations. The data collected consist of financial and technical information.

The companies in the sample represent about a fifth of Swiss gas distributors but own about 40 percent of the total length of the gas distribution network in Switzerland. This implies that many small gas distributors are under-represented in the sample. Moreover, the

\textsuperscript{18} We also estimated the model with logarithm of LF. The estimated inefficiencies (not reported here) do not vary much and show a very high correlation with those estimated from the adopted models (generally higher than 0.99).
average volume of distributed gas per network length in these companies is higher than the national average value. With regard to the service area, our sample covers 42% of all Swiss communities served with gas. Table 2 provides a descriptive summary of the main variables used in the analysis. Total costs $TC$ are the total annual operating costs plus the gas purchases from the transmission sector. Tax expenditures and non-operating costs are excluded. Output $Y$ is measured by the total amount of gas delivered to end-consumers and to downstream distributors.  

Table 2: Descriptive statistics (129 observations)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs ($TC$) Thousand CHF</td>
<td>21411</td>
<td>24477</td>
<td>2592</td>
<td>135382</td>
</tr>
<tr>
<td>Annual output ($Y$) in MWh</td>
<td>548515</td>
<td>729301</td>
<td>58000</td>
<td>4174000</td>
</tr>
<tr>
<td>Firm-average annual labor price ($P_L$) CHF per employee</td>
<td>96161</td>
<td>15963</td>
<td>61830</td>
<td>139460</td>
</tr>
<tr>
<td>Firm-average annual capital price ($P_C$) CHF per meter network length</td>
<td>28.56</td>
<td>11.43</td>
<td>12.53</td>
<td>75.78</td>
</tr>
<tr>
<td>Firm-average annual energy price ($P_E$) $10^{-2}$ x CHF per kWh</td>
<td>2.51</td>
<td>0.47</td>
<td>1.65</td>
<td>3.82</td>
</tr>
<tr>
<td>Load Factor ($LF$) %</td>
<td>34.99</td>
<td>6.68</td>
<td>14.51</td>
<td>57.91</td>
</tr>
<tr>
<td>Number of customers ($NUMB$)</td>
<td>4537</td>
<td>5744</td>
<td>509</td>
<td>29605</td>
</tr>
<tr>
<td>Number of terminal blocks ($TB$)</td>
<td>4423</td>
<td>5691</td>
<td>756</td>
<td>29575</td>
</tr>
<tr>
<td>Service area ($ASIZE$) in hectares</td>
<td>2104</td>
<td>1840</td>
<td>320</td>
<td>8310</td>
</tr>
<tr>
<td>Number of customers per km network length ($CUD$)</td>
<td>20.26</td>
<td>7.19</td>
<td>4.34</td>
<td>32.57</td>
</tr>
<tr>
<td>Network length in km</td>
<td>214.3</td>
<td>213.4</td>
<td>37.7</td>
<td>1122</td>
</tr>
</tbody>
</table>

- All monetary values are in 2000 Swiss Francs (CHF), adjusted by the consumer price index.

Labor price ($P_L$) is defined as the ratio of total annual labor costs, including social security costs, to the number of full time equivalent employees. The price of energy ($P_E$) is

---

19 About one fourth of the companies have gas resale to other distributors. In an alternative specification (not reported here) we controlled for these companies with a dummy variable. Since this dummy is not significantly different from zero we decided to exclude it from the final specification.
the average unit price of the purchased gas. The capital price ($P_C$) is calculated as sum of expenditures other than labor expenses and gas purchases divided by the network length. These expenditures include interest payments and depreciation as well as material costs and other services included in operating costs. Capital stock includes the distribution network as well as other equipment such as monitoring and control systems and the final connections and metering equipment. In fact, lacking any other reliable measure of total capital stock, we assumed that the capital stock is more or less proportional to the network length. Moreover, as the network is the major part of the capital stock of a gas distributor, network length has been used as a proxy physical measure of capital in the calculation of capital prices. All costs and prices are adjusted for inflation using the Switzerland’s consumer price index and are measured in year 2000 Swiss Francs.

5. Estimation Results

Estimation results for the four models are given in table 3. Results show that the output and input price coefficients are positive and highly significant across all models. Furthermore, the coefficients are not significantly different from one model to another, suggesting the results for the parameters do not depend on distributional assumptions of the error and inefficiency term. All output characteristics show the expected signs, although the coefficient of the load factor is only significantly different from zero in the GLS model.20

The results indicate that all the group mean coefficients in the GLS-Mundlak model are insignificant at 5%. This finding is supported by the Hausman test, which does not reject the hypothesis of similar coefficients between the GLS and fixed-effects models (P-value of 0.39). This in turn suggests that the firm effects are not correlated with the explanatory variables. Compared to other models, the model with Mundlak adjustment show generally higher standard errors. In particular, the coefficient of the service area is not significantly different from zero (at 5% significance level). This result can be explained by the fewer

---

20 Similar results were obtained using the logarithm of the load factor.
degrees of freedom in this model. Another important observation is that the estimated coefficients from the pooled model are more or less similar to those obtained from the three panel data models. This result also suggests that the firm-specific unobserved factors do not create a significant bias in the coefficients.

Table 3: Regression results

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Pooled</th>
<th>Pitt &amp; Lee</th>
<th>GLS</th>
<th>GLS &amp; Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_Y$</td>
<td>0.718*</td>
<td>0.691*</td>
<td>0.696*</td>
<td>0.638*</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.034)</td>
<td>(0.021)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>$\alpha_{PC}$</td>
<td>0.259*</td>
<td>0.262*</td>
<td>0.248*</td>
<td>0.250*</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.021)</td>
<td>(0.016)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>$\alpha_{PE}$</td>
<td>0.624*</td>
<td>0.589*</td>
<td>0.596*</td>
<td>0.588*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.030)</td>
<td>(0.020)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>$\alpha_{LF}$</td>
<td>-0.175</td>
<td>-0.243</td>
<td>-0.283*</td>
<td>-0.158</td>
</tr>
<tr>
<td></td>
<td>(0.075)</td>
<td>(0.193)</td>
<td>(0.098)</td>
<td>(0.146)</td>
</tr>
<tr>
<td>$\alpha_{TB}$</td>
<td>0.198*</td>
<td>0.225*</td>
<td>0.202*</td>
<td>0.259*</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.047)</td>
<td>(0.030)</td>
<td>(0.090)</td>
</tr>
<tr>
<td>$\alpha_{ASIZE}$</td>
<td>0.089*</td>
<td>0.091*</td>
<td>0.117*</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.032)</td>
<td>(0.033)</td>
<td>(0.100)</td>
</tr>
<tr>
<td>$\alpha_{CUD}$</td>
<td>-0.155*</td>
<td>-0.194*</td>
<td>-0.163*</td>
<td>-0.278*</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.075)</td>
<td>(0.026)</td>
<td>(0.074)</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>2.798*</td>
<td>2.712*</td>
<td>2.565*</td>
<td>2.926*</td>
</tr>
<tr>
<td></td>
<td>(0.202)</td>
<td>(0.477)</td>
<td>(0.276)</td>
<td>(0.593)</td>
</tr>
<tr>
<td>$\gamma_Y$</td>
<td>-</td>
<td>-</td>
<td>0.082</td>
<td>(0.064)</td>
</tr>
<tr>
<td>$\gamma_{PC}$</td>
<td>-</td>
<td>-</td>
<td>0.012</td>
<td>(0.038)</td>
</tr>
<tr>
<td>$\gamma_{PE}$</td>
<td>-</td>
<td>-</td>
<td>0.048</td>
<td>(0.069)</td>
</tr>
<tr>
<td>$\gamma_{LF}$</td>
<td>-</td>
<td>-</td>
<td>0.0656</td>
<td>(0.251)</td>
</tr>
<tr>
<td>$\gamma_{TB}$</td>
<td>-</td>
<td>-</td>
<td>-0.064</td>
<td>(0.097)</td>
</tr>
<tr>
<td>$\gamma_{ASIZE}$</td>
<td>-</td>
<td>-</td>
<td>-0.073</td>
<td>(0.108)</td>
</tr>
<tr>
<td>$\gamma_{CUD}$</td>
<td>-</td>
<td>-</td>
<td>0.137</td>
<td>(0.080)</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets. * means significantly different from zero at least at 99%.
- $s_\nu$ and $s_u$ in the GLS models are respectively the standard deviation of residuals associated with $v_i$ and $u_i$. In Pitt and Lee’s model, $\sigma_v$ and $\sigma_u$ are the model parameters and are statistically significant at $p=0.05$.

Since total costs and the relevant explanatory variables are in logarithms, the estimated coefficients can be interpreted as cost elasticities. For instance, the output coefficient suggests
that on average, a one percent increase in the amount of gas delivered will increase the costs by about 0.7 percent. Similarly, a one percent increase in the number of terminal blocks will increase the costs by about 0.2 percent. The cost elasticities with respect to factor prices are positive and of similar magnitude in all models. The estimated coefficient for capital price ($\alpha_{PC}$) represents the share of costs attributed to capital, which is about 25 percent. This result is very close to the average capital share in the actual data of about 29 percent. Also, the value of the coefficient of energy price ($\alpha_{PE}$) reflects its share in the sample of about 59 percent.

As expected, the negative sign of the coefficient of the load factor suggests that networks that are more evenly utilized throughout the year, are relatively less costly. The results also indicate that the customer density has a negative effect on total costs (negative $\alpha_{CUD}$). The coefficient of area size ($\alpha_{ASIZE}$) is consistent with the contention that ceteris paribus larger areas imply longer distances, thus higher organization and maintenance costs. Moreover, larger networks are more likely to be complex.

Table 4 provides a descriptive summary of the inefficiency estimates. These estimates denoted by $u_i$ in our models, represent the relative excess cost of a given firm compared to a minimum level that would have been achieved if the firm had operated as efficiently as the “best practice” observed in the sample. The results suggest an average inefficiency of 6 to 7.5 percent (median values of 5 to 7 percent), which is quite stable across models. As it can be seen in the table the inefficiency estimates obtained from the pooled model and Pitt and Lee specification are quite similar. The fact that the assumption regarding the temporal variation of inefficiency does not affect the results, suggests that the inefficiencies could be considered as more or less constant over time. Constant efficiency estimates are not surprising, because although one might expect to observe changes in efficiency over the five-year sample period, the studied companies operate in a heavily regulated system that has not changed significantly. The results also show that the extreme inefficiency estimates slightly decrease
when the correlation of unobserved firm-specific variables is taken into account (Mundlak formulation).

<table>
<thead>
<tr>
<th>Table 4: Inefficiency measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>95th Percentile</td>
</tr>
<tr>
<td>Minimum</td>
</tr>
</tbody>
</table>

The pairwise Pearson correlation matrix of the inefficiency estimates presented in table 5, shows a generally high correlation. This suggests that the estimates are reasonably robust to specification. The results of Spearman rank correlation (omitted from the paper) are similar to those in table 5. However, our analysis of efficiency ranks indicates that although the most and least efficient firms remain stable across different models, the companies in the first and last quintiles change from one model to another. These results indicate that Bauer et al. (1998)’s mutual consistency criteria are not fully satisfied. Therefore, the estimated individual inefficiency scores could incur considerable estimation errors that could also affect ranks. The results are consistent with Horrace and Schmidt (1996) who show that even a panel with 6 periods cannot provide reasonable estimates of individual efficiency scores. The results are however more reliable concerning the average inefficiency in the sector or in a group of companies. As shown in table 4, on average inefficiency is responsible for about 5 to 7 percent of the operating costs in the Swiss gas distribution utilities.
Table 5: Pairwise Pearson correlation between inefficiency estimates

<table>
<thead>
<tr>
<th></th>
<th>Pooled</th>
<th>Pitt &amp; Lee</th>
<th>GLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitt &amp; Lee</td>
<td>0.837</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GLS</td>
<td>0.850</td>
<td>0.956</td>
<td>1</td>
</tr>
<tr>
<td>GLS &amp; Mundlak</td>
<td>0.857</td>
<td>0.873</td>
<td>0.935</td>
</tr>
</tbody>
</table>

6. Economies of scale and density

In the cost function framework, returns to scale can be defined in terms of the relative changes in costs due to an increase in output, namely the inverse of output elasticity in a cost function.\(^{21}\) In network industries however, the output variation is generally together with a change in output characteristics such as network size. In such cases, the concept of density economies is used to describe the effect of changes in output with the network characteristics being fixed (cf. Caves, Christensen and Tretheway, 1984 and Caves et al., 1985). As for the scale economies it is generally assumed that, in line with Caves, Christensen and Swanson (1981), as the production scale increases, all outputs and output characteristics vary at the same proportion. In this case the economies of scale can be defined as the inverse of the sum of the elasticities with respect to outputs and output characteristics, whereas the economies of density are defined as the inverse of output elasticity.\(^{22}\) Based on these definitions, when scale or density economies are greater than one, the production operates at increasing returns to scale.

---

\(^{21}\) In general, this definition differs from the concept of returns to scale in production, which is defined as the change in output resulting from an equi-proportional increase in inputs. Chambers (1988) makes the distinction by referring to the cost-based measure as “returns to size”. However, in the case of homothetic production functions (as assumed in this paper), these measures coincide.

\(^{22}\) See also Farsi et al. (2006) for the mathematical expression and Panzar and Willig (1977) for a formal definition.
scale, suggesting potential savings by increasing output. Conversely, values lower than one indicates diseconomies of scale.

Table 6 lists the values of scale and density economies based on the above definitions estimated from the GLS-Mundlak specification. Since the coefficients of the estimated models do not differ much across specifications, in this section we focus on the GLS-Mundlak model. The confidence intervals have also been listed in the table. As scale and density economies are non-linear functions of the estimated coefficients, we used the delta method\textsuperscript{23} to calculate the standard errors and the confidence intervals. These results show that the value of economies of density is larger than one, suggesting that distributors could lower their average costs by increasing the output if the service area and the number of customers remain constant. This is reflected by the value of the output coefficient \( \alpha_Y \), which implies that a one percent increase in output would increase costs by about 0.64 percent.

### Table 6: Economies of scale and density

<table>
<thead>
<tr>
<th>Measure</th>
<th>Term</th>
<th>Value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economies of density</td>
<td>( 1/\alpha_Y )</td>
<td>1.57</td>
<td>1.28 1.86</td>
</tr>
<tr>
<td>Economies of scale</td>
<td>( 1/(\alpha_Y + \alpha_{TB} + \alpha_{ASIZE}) )</td>
<td>0.94</td>
<td>0.80 1.09</td>
</tr>
</tbody>
</table>

- 95% confidence intervals are calculated by the delta-method.

If in the process of increasing output, both area size and number of customers increase with the same proportion, the economies of scale falls to 0.94, which is not significantly different from one (see table 6). This result, more or less consistent with the previous literature\textsuperscript{24}, suggests that on average extending the network does not help to save operating costs. However, these results are based on the conventional definition of multiproduct

\textsuperscript{23} The delta method is an intuitive method to estimate the standard errors of any function of random variables, such as coefficient estimates in a regression model. See Oehlert (1992) for more details.

\textsuperscript{24} For a summary on economies of scale in gas distribution networks, see Fabbri et al. (2000).
economies of scale, which assumes that all outputs and output characteristics increase by the same proportion. This assumption could be unrealistic in practice. For instance a ten percent extension of the network might result in a lower increase in output, because the newly connected customers could have a lower density and consumption. Especially, this relationship might also depend on the network and environmental characteristics of the company. In fact a simple analysis of changes over the sample period shows that despite relatively strong inter-correlation, the output characteristics do not vary at the same proportions.

The data indicate that while the company’s gas output has shown an average growth of about 9 percent over the sample period, the number of terminal blocks have grown by 11 percent on average and the covered area size by about 4 percent. Moreover, the ratio of change differs across companies. Focusing on the ten companies with a non-zero change in their network over the sample period, we calculated the ratio of relative change in an output characteristic to the relative change in the amount of delivered gas (output Y) for each company. This ratio is on average about 1.7, 1.3 and 0.3 respectively for the number of terminal blocks (TB), area size (AS) and customer density (CUD). These values suggest that for the companies that had any extension over the sample period, one percent increase in output has been on average, associated with 1.7% extension in area size, 1.3% increase in the number of terminal blocks and a small increase of 0.3% in customer density. This implies that a given increase in output may require a larger extension in the network.

The assumption of equal proportions can be relaxed by weighting the elasticities of each output by its corresponding variation ratio with respect to output. Table 7 gives two variants of scale economies considering non-uniform proportions. The variation ratios ($\rho_{TB}$, $\rho_{ASIZE}$ and $\rho_{CUD}$) can be specified according to the case at hand. Here, we set these values to their

---

25 In 16 out of 26 companies the area size has remained constant.
corresponding estimates from the data. The first variant measures the economies of scale associated with an increase in output, when the number of terminal blocks and the area size and therefore the network are extended but the customer density remains constant, while the second variant includes the additional effect of customer density. Economies of scale 1 (table 7) can be compared with its equi-proportional counterpart in table 6. The results suggest that the scale economies may be overestimated if all changes are considered with the same proportion. As table 7 shows, in both variants, the economies of scale are lower than one, suggesting diseconomies of scale. However, given that the estimates are sensitive to the adopted values for the proportions, these results cannot be generalized and should be considered with caution.

Table 7: Economies of scale redefined

<table>
<thead>
<tr>
<th>Measure</th>
<th>Terms</th>
<th>Value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economies of scale 1</td>
<td>$1/(\alpha_Y + \rho_{TB} \alpha_{TB} + \rho_{ASIZE} \alpha_{ASIZE})$</td>
<td>0.78</td>
<td>0.64 0.93</td>
</tr>
<tr>
<td>Economies of scale 2</td>
<td>$1/(\alpha_Y + \rho_{TB} \alpha_{TB} + \rho_{ASIZE} \alpha_{ASIZE} + \rho_{CUD} \alpha_{CUD})$</td>
<td>0.83</td>
<td>0.66 0.99</td>
</tr>
</tbody>
</table>

* $\rho_{TB}$, $\rho_{ASIZE}$ and $\rho_{CUD}$ are respectively the ratios of relative change in TB, ASIZE and CUD to the relative change in output ($Y$) over the sample period. These values averaged over the ten companies with non-zero change are $\rho_{TB}=1.66$, $\rho_{ASIZE}=1.30$ and $\rho_{CUD}=0.26$. Confidence intervals are at 95% confidence level and are calculated by the delta-method. The estimation errors of the slopes are neglected in calculating the confidence interval.

7. Conclusions

The application of three cost frontier models to a five-year panel of 26 gas distribution companies in Switzerland suggests an average inefficiency of about seven percent in the sector. This result is robust across all the models. The individual efficiency scores and ranks estimated from different models show a strong correlation. However, the companies identified as “best” and “worst” practices change across models. These results show that the mutual “consistency” requirements (Bauer et al., 1998) are not fully satisfied. Therefore, the individual efficiency estimates cannot be directly used as X-factors in price cap formulas.
However, the robustness of average efficiency estimates and also the cost function coefficients across different specifications suggest that the results can be used in setting target efficiency scores and cost prediction. The regulators can use such instruments to improve efficiency through yardstick competition.

The results also highlight the importance of environmental and output characteristics. Especially, the customer density, measured as number of customers per kilometer of network, has a decreasing effect on costs, while the area size has a positive effect. As for the scale and density economies the results are more or less consistent with the findings of studies performed in other countries, in that they provide evidence of considerable density economies but insignificant or weak scale economies. This implies that distributors could decrease their average costs by increasing the output as long as they use the same network but the extension of networks does not result in any significant economies. Our analysis of variations over the sample period shows that the output characteristics do not vary at the same proportion as assumed in the definition of the scale economies. An alternative definition that accounts for the proportions estimated from the sample, suggests that the estimates of scale economies might be sensitive to the assumptions on the proportions between outputs.

References


APPLICATION OF PANEL DATA MODELS IN
BENCHMARKING ANALYSIS OF THE ELECTRIVITY
DISTRIBUTION SECTOR*

Mehdi Farsi† Massimo Filippini† William Greene‡

† Department of Management, Technology and Economics
Swiss Federal Institute of Technology
Zürichbergstr. 18, CH-8032 Zurich, Switzerland
and
Department of Economics, University of Lugano
Via Maderno 24, 6900 Lugano, Switzerland

‡ Department of Economics, Stern School of Business
New York University
44 West 4th St., New York, NY 10012, USA
† Center for Energy Policy and Economics

June 2005

Correspondence: Massimo Filippini, CEPE, ETH, Zürichbergstr. 18, CH-8032, Zurich, Switzerland
Phone: +41-1-632-0649, Fax: +41-1-632-1050, E-mail: mfilippini@ethz.ch

* The authors are grateful to the Swiss Federal Office for Education and Science for their financial support.
They also wish to thank Aurelio Fetz for his assistance.
ABSTRACT

This paper explores the application of several panel data models in measuring productive efficiency of the electricity distribution sector. Stochastic Frontier Analysis has been used to estimate the cost-efficiency of 59 distribution utilities operating over a nine-year period in Switzerland. The estimated coefficients and inefficiency scores are compared across three different panel data models. The results indicate that individual efficiency estimates are sensitive to the econometric specification of unobserved firm-specific heterogeneity. When these factors are considered as a separate stochastic term, the efficiency estimates are substantially higher indicating that conventional models could confound efficiency differences with other unobserved variations among companies. The results suggest that alternative panel models such as the “true” random effects model proposed by Greene (2005) could be used to evaluate the possible impacts of unobserved factors such as network effects on efficiency estimates.
1. INTRODUCTION

Transmission and distribution of electricity have been considered as natural monopolies, thus less affected by the recent waves of deregulation in power industry. However, as competition is being introduced into generation sector, regulatory reform and incentive regulation of distribution utilities have become more common. In traditional cost-of-service regulation systems companies recover their costs with a risk-free fixed rate of return and therefore have little incentive to minimize costs. The incentive-based schemes on the other hand, are designed to provide incentive for productive efficiency by compensating the company with its savings. A variety of methods have been proposed in the literature. Main categories of incentive-based schemes used for electricity utilities are: price or revenue cap regulation schemes, sliding-scale rate of return, partial cost adjustment, menu of contracts, and yardstick regulation.\footnote{See Joskow and Schmalensee (1986) for a review of regulation models.} Jamasb and Pollitt (2001) provide an extensive survey of different regulation practices in electricity markets around the world. Virtually most of the models used in practice, are based on ‘benchmarking’ that is, measuring a company’s productive efficiency, i.e. technical and/or cost efficiency, against a reference performance.\footnote{Other measures of performance such as quality and productivity are not considered here. This paper focuses on productive (in)efficiency, which can be decomposed into technical and allocative (in)efficiencies (cf. Kumbhakar and Lovell, 2000). Another source of inefficiency is related to the size (scale) of the production unit (cf. Chambers, 1988). However, scale inefficiency is usually beyond the firm’s control, thus generally not considered in benchmarking.}

There exist a variety of methods for efficiency measurement.\footnote{See Kumbhakar and Lovell (2000) and Coelli et al. (1998) for extensive discussion of these methods.} As pointed out in Jamasb and Pollit (2003), Estache et al. (2004) and Farsi and Filippini (2005), different methods could lead to significantly different individual efficiency estimates.\footnote{Jamasb and Pollit (2003) report substantial variations in estimated efficiency scores and rank orders across different approaches (parametric and non-parametric) and among different econometric models applied to a cross sectional sample of European power distribution utilities. Similar results are reported by Farsi and Filippini (2004, 2005) in a sample from Switzerland. Estache et al. (2004) provide more or less similar discrepancies between parametric and non-parametric methods applied to a sample of power distributors from South America. Other authors like Horrace and Schmidt (1996), Street (2003) and Jensen (2000) reported substantial errors and}
especially important for in most cases, there is no clearly defined criterion for choosing a unique method among several legitimate models. Moreover, the inefficiency estimates can have great financial consequences for the firms and therefore, their reliability is crucial for an effective regulation system. In particular, if the estimated inefficiency scores are sensitive to the benchmarking methods, a more detailed analysis to justify the adopted approach is required. For instance, Bauer et al. (1998) have proposed a series of criteria that can be used to evaluate if the results in terms of inefficiency level obtained from different approaches and models are mutually “consistent”, that is, lead to comparable inefficiency scores and ranks. However, in many cases because of a considerable discrepancy, these criteria are not satisfied. This can be considered as an improvement over the benchmarking models commonly used in electricity networks, which have been frequently criticized.5

In the literature we can distinguish two main approaches to measure efficiency – the econometric (parametric) approach and the linear programming (non-parametric) method.6 Although the latter category, particularly Data Envelopment Analysis, has become popular among electricity regulators, both approaches have advocates in the scientific community. The purpose of this paper is not to stress the advantages and disadvantages of these two different approaches, but to present how some limitations of frontier models can be overcome if panel data are available.7 This paper focuses on econometric methods as they can be relatively easily adapted to panel data. Productive efficiency can also be estimated using production or cost frontiers. In this paper we focus on the latter category that can be readily used to estimate cost-efficiency.

---

5 For instance see Shuttlewoth (2003) and Irastorza (2003) for criticisms of benchmarking approaches in electricity industry.
6 See Murillo-Zamorano (2004) for a general presentation of the different methodologies.
7 In contrast with cross-sectional data, panels provide information on same companies over several periods.
As opposed to cross-sectional data, panels provide information on same companies over several periods. Repeated observation of the same company over time allows an estimation of unobserved firm-specific factors, which might affect costs but are not under the firm’s control. Individual companies operate in different regions with various environmental and network characteristics that are only partially observed, it is crucial for the regulator to distinguish between inefficiency and such exogenous heterogeneity. Several recently developed models such as Greene (2005, 2004), Alvarez, Arias and Greene (2004) and Tsionas (2002) have addressed this issue using alternative panel data models. Some of these models have proved a certain success in other applications such as public transportation networks in that they give more plausible efficiency estimates. These results raise an important question as to whether (or to what extent) the sensitivity problems can be solved by using panel data and the adapted frontier models. This question is especially important in the electricity sector, in which the application of benchmarking has been frequently criticized based on reliability of efficiency estimates. Moreover, given that in many countries the regulatory reforms have been in effect for several years, an increasing number of regulators have access to panel data. However, the number of empirical studies is still insufficient to provide a general answer to this question. This paper studies the performance of an alternative panel data econometric frontier model to distinguish unobserved firm-specific heterogeneity from inefficiency in the context of electricity distribution networks.

The results of this paper suggest that the alternative panel data models can separate part of the unobserved heterogeneity from inefficiency estimates, thus can be considered as a promising complement to other regulatory instruments such as cost prediction (as proposed in Farsi and Filippini, 2004) and case-by-case analyses. The rest of the paper proceeds as

---

8 See Farsi, Filippini and Kuenzle (2006) and Farsi, Filippini and Greene (2005) for applications in bus and railway transports respectively.

follows: Section 2 discusses the application of stochastic frontier models in panel data. The model specification and the adopted econometric methods are described in Section 3. Following a brief description of the data, the estimation results are presented and discussed in Section 4. And Section 5 summarizes the main conclusions.

2. PANEL DATA AND STOCHASTIC FRONTIER MODELS

The first use of panel data models in stochastic frontier models goes back to Pitt and Lee (1981) who interpreted the panel data random effects as inefficiency rather than heterogeneity.\textsuperscript{10} This tradition continued with Schmidt and Sickles (1984) who used a similar interpretation applied to a panel data model with fixed effects. Both models have been extensively used in the literature. A main shortcoming of these models is that any unobserved, time-invariant, firm-specific heterogeneity is considered as inefficiency. In more recent papers the random effects model has been extended to include time-variant inefficiency. Cornwell, Schmidt and Sickles (1990) and Battese and Coelli (1992) are two important contributions in this regard. In particular the former paper proposes a flexible function of time with parameters varying among firms. However, in both these models the variation of efficiency with time is considered as a deterministic function that is commonly defined for all firms. We contend that the time variation of inefficiency may be different across firms. Even within a given firm, these variations could depend on unobserved factors thus can be assumed as a stochastic term rather than a deterministic function of time.

As shown by Alvarez, Arias and Greene (2003), even in cases where inefficiency is due to time-invariant factors such as constant managers’ capability, the resulting cost inefficiencies can vary over time. These authors assume that the management skills are one of

\textsuperscript{10} Pitt and Lee (1981)’s model is different from the conventional RE model in that the individual specific effects are assumed to follow a half-normal distribution. Important variations of this model were presented by Schmidt
the inputs that can interact with other time-variant input factors thus, create time-variant cost inefficiency. This result is consistent with the economic theory in that a firm’s inefficiency is a dynamic phenomenon and cannot be constant. Firms constantly face new events and technologies, which they gradually learn how to deal with and apply. As the learning process continues, inefficiency with regards to the existing technologies decrease but other new events and technologies appear. Therefore the overall inefficiency of a firm depends on not only the managers’ efforts but on the effect of new technologies and events on the production process. Based on this argument, the inefficiency can best be modeled as a time-variant stochastic term. On the other hand a major part of the unobserved heterogeneity such as network and location-related factors can be considered as constant over time.

The discrepancy in efficiency estimates in conventional panel data models has been shown in Horrace and Schmidt (1996) and Farsi and Filippini (2004). A common feature of all these models is that they do not fully separate the sources of heterogeneity and inefficiency at the firm level. In fact, the time-variant error term in these models could include a major part of inefficiencies whereas the firm-specific effects that are interpreted as inefficiency could be mainly due to time-invariant heterogeneity.

An alternative approach is to consider an additional stochastic term for cost efficiency. Theoretically, a stochastic frontier model in its original form (Aigner, Lovell and Schmidt, 1977) can be extended to panel data models, by adding a fixed or random effect in the model. There are however few papers that have explored this possibility. The first development can be attributed to Kumbhakar (1991) who proposed a three-stage estimation procedure to solve the model with time- and firm-specific effects. Polachek and Yoon (1996) attempted to

---


11 See also Heshmati and Kumbhakar (1994) and Kumbhakar and Hjalmarsson (1995) for two applications of this model. Note that in the latter paper, it is assumed that both time- and firm-specific effects are part of inefficiency.
estimate a panel data frontier model with firm dummies using a one-step procedure. Greene (2002a) discussed the numerical obstacles that have apparently delayed such a development.

As shown by Greene (2002a), assuming that the inefficiency term follows a distributional form, both models with random and fixed effects can be estimated using maximum likelihood estimation methods. These models are referred to as “true frontier models” in that they are a straightforward extension of original frontier framework (in line with Aigner et al., 1977) to panel data. He proposed numerical solutions for both models, which he respectively refers to as ‘true’ fixed and random effects models (see also Greene, 2005). In particular, Greene’s true random effects model has proved useful in efficiency measurement of network industries (Farsi, Filippini and Greene, 2005).

3. MODEL SPECIFICATION

To illustrate the differences across models, we focus on three panel data models: GLS model in line with Schmidt and Sickles (1984), MLE model as in Pitt and Lee (1981), and the True Random Effects (TRE) model as proposed by Greene (2005, 2004). These methods have been applied to a panel of 59 Swiss distribution utilities. A triple-input single-output production function has been considered. The output is measured as the total number of delivered electricity in kWh, and the three input factors are set as capital, labor and the input power purchased from the generator. Capital price is measured as the ratio of capital expenses (depreciation plus interest) to the total installed capacity of the utility’s transformers in kVA. The capital costs are approximated by the residual costs that is, total costs minus labor and purchased power costs. Labor price is defined as the average annual salary of the

---

12 The sample used in this study is similar to the one used by Farsi and Filippini (2004).
13 Because of the lack of inventory data the capital stock is measured by the capacity of transformers, which are the main device used to transfer electricity in the network.
firm’s employees. For those companies that produce part of their power the average price of input electricity is assumed to be equal to the price of purchased power.

The costs of distribution utilities consist of two main parts: the costs of the purchased power and the network costs including labor and capital costs. There are therefore two alternatives for measuring cost efficiency in power distribution utilities: total costs approach and network costs approach. The network costs approach has a practical advantage in that the estimated average costs can be directly used in a price-cap formula.\(^{14}\) However, this approach neglects the potential inefficiencies in the choice of the generator and also in the possibilities of substitution between capital and input energy. In this paper we use the first approach based on the total costs.

In addition to input prices and output, several output characteristics are included. The resulting specification of the cost function can be written as:

\[
C = C(Y, P_K, P_L, P_P, LF, CU, AS, HGRID, DOT) \tag{1},
\]

where \(C\) represents total cost; \(Y\) is the output in kWh; \(P_K\), \(P_L\) and \(P_P\) are respectively the prices of capital, labor and input power; \(LF\) is the ‘load factor’ defined as the ratio of utility’s average load on its peak load; \(CU\) is the number of customers; and \(AS\) the size of the service area served by the distribution utility. \(HGRID\) is a binary indicator to distinguish the utilities that operate a high-voltage transmission network in addition to their distribution network and \(DOT\) is a dummy variable representing the utilities whose share of auxiliary revenues is more than 25 percent of total revenues.

The specification of the cost frontier used in this analysis is similar to the one used in the previous section. Here, we included two additional variables. A Cobb-Douglas functional form has been adopted. We excluded the flexible forms like translog to avoid the potential risk of multicollinearity among second order terms due to strong correlation between output

\(^{14}\) Notice that the price cap is generally applied to the network access.
characteristics. Moreover, given the purpose of this study, we want to use a simple specification and avoid an excessive number of parameters required in the flexible functional forms.

After imposing the linear homogeneity in input prices the adopted cost function can be written as:

$$\ln \left( \frac{C}{P_r} \right)_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_K \ln \left( \frac{P_K}{P_r} \right)_{it} + \beta_L \ln \left( \frac{P_L}{P_r} \right)_{it} + \gamma_1 \ln LF_{it}$$

$$+ \gamma_2 \ln AS_{it} + \gamma_3 \ln CU_{it} + \delta_1 HGRID_{it} + \delta_2 DOT_{it} + r_{it}$$

$$\text{with } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T_i$$

This specification is similar to that used in Farsi and Filippini (2004) with the only difference that here we excluded two explanatory variables whose effects proved to be statistically insignificant.\(^\text{15}\) Subscripts \(i\) and \(t\) denote the company and year respectively and \(r_{it}\) is the stochastic term.

Quality of service usually measured by the number of interruptions is among the excluded variables. Given that in Switzerland, practically there has been no outage cases, we can assume that all the utilities operate at a sufficient level of quality reinforced by a tight regulation system. Therefore, we contend that the quality differences are not significant. Another excluded variable is the network length. In our model, this variable is proxied by the service area.

All the three models are based on the specification given in (2). The differences are in the specification of the residuals \((r_{it})\). This term is composed of two components, one of which \((\alpha_i)\) being time-invariant (firm-specific) and the other \((\varepsilon_{it})\) varying across observations. Table 2 summarizes the econometric specification of the models used in this study. The table also provides the estimation procedure for the efficiency scores. These scores are relative

\(^{15}\) The excluded variables are the linear trend and the dummy variable representing the forested areas.
efficiencies on a scale of 0 to 1 against the best practice. The conditional expectations are estimated using the procedure proposed by Jondrow et al. (1982).\textsuperscript{16}

\begin{table}[h]
\centering
\caption{Econometric specifications of the stochastic cost frontier}
\begin{tabular}{llll}
\hline
$r_{it} = \alpha_i + \epsilon_{it}$ & Model I & Model II & Model III \\
\hline
Firm-specific component $\alpha_i$ & \textit{GLS} & Half-normal & \textit{True RE} \\
$\sim iid (0, \sigma^2_\alpha)$ & $N^+ (0, \sigma^2_\alpha)$ & $N (0, \sigma^2_\alpha)$ \\
Time-variant component $\epsilon_{it}$ & $iid (0, \sigma^2_\epsilon)$ & $N (0, \sigma^2_\epsilon)$ & $\epsilon_{it} = u_{it} + v_{it}$ \\
& $\sim N^+ (0, \sigma^2_u)$ & $u_{it} \sim N^+ (0, \sigma^2_u)$ & $v_{it} \sim N (0, \sigma^2_v)$ \\
Inefficiency & $\hat{\alpha}_i - \min \{\hat{\alpha}_i\}$ & $E\left[ u_{it} | \hat{r}_{i1}, \hat{r}_{i2}, \ldots, \hat{r}_{iT} \right]$ & $E\left[ u_{it} | \hat{r}_{it} \right]$ \\
Relative efficiency (0-1) & $e^{-(\hat{\alpha}_i - \min \{\hat{\alpha}_i\})}$ & $e^{-\text{u}_i} | \hat{r}_{i1}, \hat{r}_{i2}, \ldots, \hat{r}_{iT} \right]$ & $E\left[ e^{-\text{v}_{it}} | \hat{r}_{it} \right]$ \\
\hline
\end{tabular}
\end{table}

4. DATA AND ESTIMATION RESULTS

The data consist of an unbalanced panel of 59 Switzerland’s distribution utilities over a 9-year period from 1988 to 1996. The sample includes 380 observations with a minimum of four observations per company. From about 900 power distribution companies in Switzerland, the companies included in the sample deliver about a third of Switzerland’s electricity consumption, thus can be considered as representative of relatively large distribution utilities in the country.\textsuperscript{17} The descriptive statistics are given in Table 3.

\textsuperscript{16} See also Greene (2002b) and Battese and Coelli (1992).

\textsuperscript{17} See Farsi and Filippini (2004) for more details on the data set and a general description of the Swiss power distribution sector in Switzerland.
Table 3. Descriptive statistics (380 observations)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs per kWh output (CHF)</td>
<td>.188</td>
<td>.0303</td>
<td>.128</td>
<td>.323</td>
</tr>
<tr>
<td>Annual output ((Y)) in GigaWh</td>
<td>263.51</td>
<td>390.36</td>
<td>17</td>
<td>2301.5</td>
</tr>
<tr>
<td>Number of customers ((CU))</td>
<td>26975.6</td>
<td>36935.8</td>
<td>2461</td>
<td>220060</td>
</tr>
<tr>
<td>Load Factor ((LF))</td>
<td>.5541</td>
<td>.0727</td>
<td>.3219</td>
<td>.9817</td>
</tr>
<tr>
<td>Service Area ((AS)) in km(^2)</td>
<td>15,467</td>
<td>35,376</td>
<td>176</td>
<td>198,946</td>
</tr>
<tr>
<td>Average annual labor price ((P_L)) per employee (CHF 1000)</td>
<td>101.27</td>
<td>32.55</td>
<td>43.36</td>
<td>253.89</td>
</tr>
<tr>
<td>Average capital price ((P_K)) in CHF per kVoltAmpere installed capacity</td>
<td>95.06</td>
<td>39.35</td>
<td>32.08</td>
<td>257.98</td>
</tr>
<tr>
<td>Average price of input power ((P_r)) in CHF/kWh</td>
<td>.105</td>
<td>.0210</td>
<td>.0583</td>
<td>.161</td>
</tr>
<tr>
<td>High-voltage network dummy ((HGRID))</td>
<td>.35</td>
<td>.4776</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Auxiliary revenues more than 25% ((DOT))</td>
<td>.397</td>
<td>.490</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- All monetary values are in 1996 Swiss Francs (CHF), adjusted for inflation by Switzerland’s global consumer price index.

The estimated parameters of the cost frontier are listed in Table 4. This table shows that almost all the coefficients are highly significant and have the expected signs. The results are more or less similar across different models. It should be noted that the three models are similar in the sense that they all have a firm-specific and a time-variant stochastic term, but differ in the distribution of these terms. Moreover, in all the models it is assumed that the firm-specific term is uncorrelated with the time-variant one.\(^{18}\)

\(^{18}\) Potential correlations may bias the coefficients. The assumption of no correlation can be relaxed using a fixed effects model (cf. Farsi and Filippini, 2004). However, given that in this paper the main focus is on the efficiency estimates and the coefficients have only a secondary importance, we decided to focus on random-effects models.
A descriptive summary of the efficiency estimates from different models is given in Table 5. The results indicate quite similar estimates for the GLS and MLE models, with a difference of about .02 in the median and average values. This can be explained by the fact that these models have a similar interpretation of inefficiency as a time-invariant factor. The True RE model predicts on the other hand, a much higher average efficiency rate. According to this model, the companies are on average 96% efficient. Noting that this model assumes a time-variant inefficiency term and a separate stochastic term for firm-specific unobserved heterogeneity, these results suggest that the other models overestimate the inefficiency. This conclusion is valid to the extent that inefficiencies do not remain constant over time.

### Table 4. Cost frontier parameters - Panel data (1988-1996)

<table>
<thead>
<tr>
<th></th>
<th>GLS</th>
<th>MLE</th>
<th>True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnY</td>
<td>.783*</td>
<td>.031</td>
<td>.789*</td>
</tr>
<tr>
<td>lnCU</td>
<td>.150*</td>
<td>.033</td>
<td>.145*</td>
</tr>
<tr>
<td>lnAS</td>
<td>.052*</td>
<td>.009</td>
<td>.046*</td>
</tr>
<tr>
<td>lnLF</td>
<td>-.234*</td>
<td>.038</td>
<td>-.211*</td>
</tr>
<tr>
<td>lnPL</td>
<td>.044*</td>
<td>.013</td>
<td>.044*</td>
</tr>
<tr>
<td>lnPK</td>
<td>.173*</td>
<td>.009</td>
<td>.166*</td>
</tr>
<tr>
<td>HGRID</td>
<td>.074*</td>
<td>.026</td>
<td>.108*</td>
</tr>
<tr>
<td>DOT</td>
<td>.049*</td>
<td>.021</td>
<td>.033</td>
</tr>
<tr>
<td>Constant</td>
<td>-.854*</td>
<td>.360</td>
<td>-.870*</td>
</tr>
<tr>
<td>$\sigma_\alpha$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_u$ (half-normal)</td>
<td>-</td>
<td>-</td>
<td>.146*</td>
</tr>
<tr>
<td>$\sigma_v$ (normal)</td>
<td>-</td>
<td>-</td>
<td>.040*</td>
</tr>
</tbody>
</table>

* significant at p=.05; The sample includes 380 observations from 59 companies.
Table 5. Summary statistics of efficiency scores (1988-96)

<table>
<thead>
<tr>
<th></th>
<th>GLS</th>
<th>MLE</th>
<th>True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>.723</td>
<td>.735</td>
<td>.861</td>
</tr>
<tr>
<td>Maximum</td>
<td>1</td>
<td>.993</td>
<td>.996</td>
</tr>
<tr>
<td>Average</td>
<td>.868</td>
<td>.887</td>
<td>.957</td>
</tr>
<tr>
<td>Median</td>
<td>.857</td>
<td>.877</td>
<td>.966</td>
</tr>
<tr>
<td>95 percentile</td>
<td>.981</td>
<td>.990</td>
<td>.990</td>
</tr>
<tr>
<td>N</td>
<td>380</td>
<td>380</td>
<td>380</td>
</tr>
</tbody>
</table>

The correlation coefficients between the efficiency estimates from different models are listed in Table 6. As expected these results indicate a high correlation between the GLS and MLE estimates. However, the True RE estimates are only weakly correlated with those of the two other models. The correlation between efficiency ranks shows a similar pattern, thus excluded from the paper. These results suggest that the assumption about the inefficiency term is crucial for the estimations. The assumption that inefficiencies are random over time is more realistic than considering constant inefficiency. In fact, the regulated firms cannot sustain a constant level of inefficiency for a long period of time. Not only are they presumably induced to improve their efficiency they constantly face new technological and organizational problems. On the other hand there are a host of parameters such as network characteristics and location related factors that remain more or less constant. Therefore, the assumptions of the True RE model appear to be more consistent with the real world. The results in Table 6 indicate that if the model does not separate unobserved heterogeneity from inefficiency, the efficiency estimates could be misleading.

Table 6. Correlation between efficiency from different models (1988-96)

<table>
<thead>
<tr>
<th></th>
<th>GLS</th>
<th>MLE</th>
<th>True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLS</td>
<td>1</td>
<td>.970</td>
<td>.042</td>
</tr>
<tr>
<td>MLE</td>
<td>.970</td>
<td>1</td>
<td>.055</td>
</tr>
</tbody>
</table>
5. CONCLUSION

The results of frontier analyses of electricity distribution utilities presented in the literature point to sensitivity problems in the benchmarking methods commonly used in the regulation practice. The discrepancy appears to be high when the efficiency scores or ranks are considered for individual companies, whereas the efficiency of the whole sector or large groups of utilities prove to be more or less robust. This general result applies to both parametric and non-parametric methods. A possible explanation of this inconsistency problem can be related to the difficulty of benchmarking models in accounting for unobserved heterogeneity in environmental and network characteristics across companies. Parametric panel data models could be helpful to solve at least partially this heterogeneity problem. In this paper we applied several stochastic frontier models to a panel of Swiss distribution utilities. Consistent with previous research, the results suggest that the panel data models cannot completely solve the problem. However, the alternative models like the ‘true’ random effect model (cf. Greene, 2005) can be helpful to disentangle unobserved heterogeneity from inefficiency estimates. This study along with the previous empirical literature suggests that the estimation errors for individual efficiency scores are rather high. Given these possible errors, the direct use of benchmarking results in regulation could have significant financial consequences for the companies. Therefore, the benchmarking results should not be directly applied to discriminate companies through different individual X-factors. Such differentiations require a complementary study of individual cases. However, the results can be used as an instrument to minimize the information asymmetry between the regulator and the regulated companies. For instance benchmarking can be used as a guide to classify the companies into several efficiency groups.

An interesting feature of parametric methods is that they can be used to predict the costs/revenues for each company within a confidence interval. Therefore, such methods can
be employed to implement a yardstick regulation framework in line with Schleifer (1985). The prediction power of these models can be considerably improved by using panel data. For instance, Farsi and Filippini (2004) show that panel data models can have a reasonably low out-of-sample prediction error.\textsuperscript{19} This method could be used as an alternative to conventional use of benchmarking methods. In practice this regulation approach implies that the regulator predicts a confidence interval of the expected costs of a given utility accounting for its unobserved characteristics and considering a level of efficiency. The utilities are then required to justify any costs in excess of the predicted range.

A similar approach has been used in the regulation of water supply in Italy, where a yardstick competition model has been applied (cf. Antonioli and Filippini, 2001). This regulation method is based on an interactive approach: The company proposes its tariff in the first stage. The regulator estimates a price cap for the firm using a benchmarking analysis and adjusting for observed differences among companies. The proposed tariff is approved if it does not exceed an acceptable range around the estimated price cap. Otherwise, the tariffs can be renegotiated with the requirement that the company justify its excessive tariff before any revision.

References


\textsuperscript{19} For instance that study reports that a GLS model (similar to the one used in this paper) can achieve a one-year ahead prediction error of 3 percent on average while keeping the maximum error at 10 percent level.


COST EFFICIENCY IN REGIONAL BUS COMPANIES: AN APPLICATION OF ALTERNATIVE STOCHASTIC FRONTIER MODELS*

Mehdi Farsi†‡  Massimo Filippini†‡  Michael Kuenzle†

† Center for Energy Policy and Economics
Swiss Federal Institute of Technology
ETH Zentrum, WEC, 8092 Zurich, Switzerland

‡ Department of Economics, University of Lugano
Via Maderno 24, 6900 Lugano, Switzerland

January 29, 2005

* The authors wish to thank Subal Kumbhakar, Antonio Alvarez, Steven Morrison, the editor and an anonymous referee for their helpful suggestions. We would also like to thank participants at the North American Productivity workshop (Toronto, 2004) for their comments on an earlier version of the paper. Any remaining errors are solely the responsibility of the authors.
Abstract

This paper evaluates cost and scale efficiencies of Switzerland's regulated rural bus companies operating in regional networks. The adopted methodology can be used in benchmarking analyses applied to incentive regulation systems. Moreover, the estimations can be used to evaluate the bidding offers for the tendering processes predicted by the ongoing reform policies. Since these companies operate in different regions with various characteristics that are only partially observed, it is crucial for the regulator to distinguish between inefficiency and exogenous heterogeneity that influences the costs. A number of stochastic cost frontier models are applied to a panel of 94 companies over a 12-year period from 1986 to 1997. The main focus lies on the ability of these models to distinguish inefficiency from the unobserved firm-specific heterogeneity in a network industry. The estimation results are compared and the effect of unobserved heterogeneity on inefficiency estimates is analyzed.
1. Introduction

In many European countries the regional public bus services are being reorganized. In line with the EU policy the Swiss government has introduced important regulatory reforms in the public transport system, including regional bus companies. The new policy act predicts a tendering process for the provision of regional bus services. With the implementation of the new system, the applying companies will bid in competitive auctions and the access rights will be granted to the company with the lowest subsidies request. This system is believed to introduce greater incentives for competitive behavior. However, given the limited number of bidding companies in most regions, it is not clear to what extent the new policies lead to efficient production. Moreover, the incumbents, mostly public companies, might have an advantageous position in such auctions. Benchmarking methods can be used to evaluate the requested subsidies and proposed costs by individual companies or to adjust the minimum bidding prices.

Benchmarking analysis is based on comparing the costs of individual companies to the ‘best’ (most cost-efficient) observed practice. These deviations, often labeled as ‘cost-inefficiency’ can also be used to adjust the amount of subsidies paid to individual bus operators. Moreover, predicted costs of the benchmark practice could be used to gain information regarding the future evolution of costs incurred by the companies operating in a service area, and to re-evaluate the claimed subsidies.¹

In order to use the efficiency estimates of individual companies in regulation, it is important to have precise measurement methods. In particular, because of

¹ See Farsi and Filippini (2004) for a discussion on the use of cost prediction in the regulation of public utilities.
considerable cost differences across various networks, it is crucial to distinguish the cost difference due to unobserved heterogeneity in external factors from the excess costs due to the company’s inefficiency. Benchmarking can be conducted using econometric methods such as stochastic frontier models, which have been developed in a variety of forms during the past two decades. All these models in one way or another separate the heterogeneity from cost-inefficiency. Especially, with panel data at hand, the unobserved heterogeneity can be better identified because the time-invariant elements of heterogeneity can be separately specified by firm-specific effects.

The first application of panel data models in stochastic frontier analysis was introduced by Pitt and Lee (1981). These authors formulated the firm-specific error component as a half-normal distribution, which they interpreted as inefficiency. In the following years, several models have been developed to incorporate the observed firm-specific heterogeneity. For instance, Jha and Singh (2001), Piacenza (2002) and Dalen and Gomez-Lobo (2003) use single equation models proposed by Battese and Coelli (1995) to incorporate some exogenous variables to explain the determinants of the inefficiency component in the bus transportation industry. However, most of these models have a shortcoming in that they cannot disentangle firm’s inefficiency from cost differences due to unobserved characteristics of the service area. Especially, transport companies operate in networks with different shapes and structures, which result in different coordination problems and thus lead to different costs. These characteristics are usually given and cannot be controlled by the companies. Some of these exogenous factors are either unavailable or too complex to be measured by

---

2 Kumbhakar and Lovell (2000) provide an extensive survey of this literature.
3 For the advantages of single stage models, see Wang and Schmidt (2002).
single indicators. Unfortunately, when unobserved heterogeneity is present the inefficiency estimates can be biased.

Greene (2004, 2005) proposes alternative panel data models, which can better distinguish between unobserved firm-specific heterogeneity and inefficiency. These models extend the previous models by adding an additional stochastic error component for the heterogeneity.\textsuperscript{4} Such models are particularly useful in transport industries where the network and environmental characteristics are mostly unobserved or hard to measure, but play an important role on the operating costs.

The purpose of this study is to analyze the performance of different panel data frontier models with regard to estimated coefficients, inefficiency scores and estimates of economies of scale and density. Especially, we focus on the ability of different models to distinguish unobserved heterogeneity from inefficiency. Alternative models are applied to a sample of 94 Swiss rural bus companies from 1986 to 1997. It is concluded that in the studied sample, Greene’s “true” random effects model has a considerable advantage over other models in separating heterogeneity from inefficiency.

The rest of the paper is organized as follows: Sections 2 and 3 present the model specification and the methodology respectively. The data are explained in section 4. Section 5 presents the estimation results and discusses their implications, and section 6 provides the conclusions.

\textsuperscript{4} A similar model but with a three-stage estimation procedure has been proposed by Kumbhakar (1991) and Heshmati and Kumbhakar (1994).
2. Model Specification

A bus transit company can be considered as a production unit that operates in a given network and transforms labor and capital services and energy into units of transport services. Since in most cases not only the network but also the schedule of a bus operator is regulated and predetermined, it is common to estimate a cost rather than a production function. Different specifications have been used in the literature. Often, output is measured in terms of either passenger- or seat-kilometers. To capture some of the heterogeneity of different service areas, most specifications include additional output characteristics such as the number of stops, network length or average commercial speed. Most of these studies also include a time trend to capture the potential changes in technology.

The total cost frontier can therefore be written as the following function:

$$TC = f(Y, N, P_L, P_C, t),$$

where $TC$ is the total annual cost and $Y$ is the output represented by the total number of seat-kilometers. $N$ represents the network length. $P_C$ and $P_L$ are respectively the capital and labor prices. We considered an alternative specification including energy prices. The estimated coefficients did not change significantly and the coefficient of the energy price was generally insignificant. Moreover, because of a number of missing values for energy costs a two-input model allows a larger sample. Therefore,

---

5 See Berechman (1993) for an overview of the application of cost functions in public transport.
we consider labor and capital as the main input factors. However, as we see later, the capital price is calculated from all non-labor expenses, thus includes the variations in energy prices.

It is generally assumed that the cost function given in (1) is the result of cost minimization given input prices and output and should therefore satisfy certain properties namely, linear homogeneity and concavity in input prices and monotonicity in input prices and output.\(^7\) Input prices and output are assumed to be exogenous, thus beyond the firm’s control. In the case of Swiss bus transport companies, the municipalities and the cantons specify the output by regulating the frequency of the service. The input prices can also be regarded as given, because these companies have a relatively small share in the labor and capital markets, thus cannot influence the prices through monopsony.

To estimate the cost function (1), a translog functional form is chosen. This flexible functional form is a local, second-order logarithmic approximation to any arbitrary twice-differentiable cost function. It places no \textit{a priori} restrictions on the elasticity of substitution and allows the economies of scale to vary with the output level. The translog approximation to (1) is written as:

\[
\ln \left( \frac{TC_i}{P_{e_i}} \right) = \alpha_0 + \alpha_Y \ln Y_i + \alpha_N \ln N_i + \alpha_K \ln \frac{P_{e_i}}{P_{n_i}} + \frac{1}{2} \alpha_{yy} (\ln Y_i)^2 + \frac{1}{2} \alpha_{NN} (\ln N_i)^2 + \frac{1}{2} \alpha_{KK} (\ln \frac{P_{e_i}}{P_{n_i}})^2 + \alpha_{yk} \ln Y_i \ln \frac{P_{e_i}}{P_{n_i}} + \frac{1}{2} \alpha_{YN} (\ln Y_i \ln N_i) + \frac{1}{2} \alpha_{KY} (\ln \frac{P_{e_i}}{P_{n_i}} \ln N_i) + \alpha_{kY} \ln \frac{P_{e_i}}{P_{n_i}} \ln N_i + \alpha_i t_i
\]

(2)

with \(i = 1, 2, \ldots, N\) and \(t = 1, 2, \ldots, T\).

\(^7\) For more details on the properties of the cost function, see Chambers (1988), p. 52.
where subscripts $i$ and $t$ denote the company and year respectively. The technical change is specified as a linear trend and is assumed to be neutral with respect to cost minimizing input ratios.\(^8\) The translog form requires that the underlying cost function be approximated around a specific point like the sample mean or median. Here, the sample median is chosen because it is less affected by outliers and thus the approximation will have better precision. As can be seen in equation (2), linear homogeneity in input prices is imposed by dividing total costs and input prices by labor price. The other theoretical restrictions are verified after the estimation.

Apart from estimating cost inefficiency, the estimation of a cost function enables us to derive important characteristics of bus supply technology such as economies of density and scale. The distinction between scale and density economies is particularly important in network industries. In such cases, a company’s size is related to both its output level and its network size, which do not necessarily vary with a simple one-to-one relationship. For this reason it is important to distinguish cost changes that occur uniquely because of output changes within a fixed network and cost changes resulting from a proportional change in both network and output.

Economies of density are defined as the inverse of the elasticity of costs with respect to output that is, the relative increase in total cost resulting from an increase in output, holding all input prices and the network size fixed.\(^9\)

\[
ED = \left( \frac{\partial \ln C}{\partial \ln y} \right)^{-1} = \left( \alpha_y + \alpha_{yy} \ln y + \alpha_{yk} \ln \frac{P_K}{P_L} + \alpha_{yn} \ln N \right)^{-1}.
\]  

\(^8\) In other words the technical change does not alter the optimal input bundles.  
\(^9\) See also Caves, Christensen and Tretheway (1984).
The existence of economies of density implies that the average costs of a bus operator decrease as physical output increases. Economies of density exist if the above expression \((ED)\) has a value greater than one. For values of \(ED\) below one, we identify diseconomies of density. In the case of \(ED = 1\), the company’s output minimizes its average costs given the network’s size.

Slightly different is the definition of economies of scale \((ES)\). Here, the increase in total costs is brought about by an increase in company’s scale that is in both output and the network size, holding the factor prices constant. However, since the changes in output and network size are inter-related, the definition of scale economies requires an assumption in this respect. The commonly used definition is the one proposed by Caves, Christensen and Tretheway (1984), which assumes that any increase in size raises the network size and the outputs with the same proportion. Based on this assumption, \(ES\) is defined as:

\[
ES := \left( \frac{\partial \ln C}{\partial \ln y} + \frac{\partial \ln C}{\partial \ln N} \right)^{-1}
\]

\[
= \left( \alpha_y + \alpha_{yy} \ln y + \alpha_{yk} \ln \frac{P_k}{P_L} + \alpha_{yn} \ln N + \alpha_{N} + \alpha_{Ny} \ln \frac{P_k}{P_L} \right)^{-1}.
\]

Similarly, economies of scale exist if \(ES\) is higher than 1.

It should be noted that the above definitions of scale and density economies are in terms of cost elasticity and do not necessarily correspond to the definitions derived from the production function. In fact, only in homothetic production functions, where the optimal input bundles vary proportionately, the two definitions
are equivalent. Here, we do not impose such an assumption. However, as in this paper
we are interested in the cost effects of output, we define the scale and density
economies as the inverse of the corresponding cost elasticities.\textsuperscript{10}

3. Methodology

The effects of unobserved heterogeneity on inefficiency estimates are studied
by a comparative analysis of four econometric models. These models are a pooled
cross section model in line with Aigner, Lovell and Schmidt (1977); a random effects
model as in Pitt and Lee (1981); a fixed effects model as in Schmidt and Sickles
(1984); and a random intercept frontier model (also known as “true” random effects
model) proposed by Greene (2004, 2005). The deterministic part of all models is
based on the specification given in equation (2).

Model I (Aigner, Lovell and Schmidt, 1977) is a pooled frontier model, in
which the error term is divided into two components: a normally distributed error $v_{it}$,
capturing general measurement errors and heterogeneity and a half-normal random
term $u_{it}$, representing the inefficiency as a one-sided non-negative disturbance. This
model can be written as:

$$
\ln TC_i = \alpha_0 + x_{it}' \beta + v_i + u_i,
$$

where $v_i \sim iid \ N(0, \sigma_v^2)$ and $u_i \sim iid \ N^+(0, \sigma_u^2)$. (5)

$\alpha_0 + x_{it}' \beta$ represents the deterministic part of the cost function as in equation (2), and

$N^+(0, \sigma_u^2)$ stands for the positive part of a normal distribution. Both error
components are assumed to be uncorrelated with each other and the regressors. This

\textsuperscript{10} See Chambers (1988) for more details about this issue. To avoid confusion this author refers to the
inverse of cost elasticity as the “economies of size” rather than economies of scale (see page 72).
model is estimated by Maximum Likelihood and the inefficiency component is estimated from the residuals \( \varepsilon_i = v_i + u_i \) by the conditional expectation \( E(u_i | \hat{\varepsilon}_i) \), proposed by Jondrow et al. (1982). In this model, the observations of a same company are considered as independent sample points. Therefore, the panel structure of the data is completely ignored. This issue can be addressed by considering a random effects model (model II) as in Pitt and Lee (1981). Similar to model I, a normal-half-normal composite error term is considered. The difference is that here, the observations for a specific company possess a common error component. This model can be formulated as:

\[
\ln TC_i = \alpha_0 + \mathbf{x}_i^\prime \mathbf{\beta} + v_i + u_i, \\
\text{where } v_i \sim iid \ N(0, \sigma_v^2) \text{ and } u_i \sim iid \ N^+ (0, \sigma_u^2). 
\]  

(6)

The model is estimated by Maximum Likelihood method. The firm-specific inefficiency is estimated using the conditional mean of the inefficiency term \( (u_i) \) proposed by Jondrow et al. (1982),\(^{11}\) that is: \( E[u_i | \hat{\varepsilon}_{i1}, \hat{\varepsilon}_{i2}, \ldots, \hat{\varepsilon}_{iT}] = E[u_i | \bar{\varepsilon}_i] \), where \( \varepsilon_i = u_i + v_i \) and \( \bar{\varepsilon}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \hat{\varepsilon}_{it} \).

A limitation of this model is the assumption that the firm-specific stochastic term is assumed to be uncorrelated with the explanatory variables. In fact, most frontier models assume that inefficiency is uncorrelated with explanatory variables included in the cost function.\(^{12}\) However, the firm-specific term \( (u_i \text{ in this model}) \) may also contain other unobserved environmental factors, which may be correlated with explanatory variables and thus may bias the coefficients. For example, larger networks are likely to be more spread, thus incur higher coordination and

\(^{11}\) See also Greene (2002a).

\(^{12}\) This assumption can be justified based on the fact that the apparent excess costs that are correlated with exogenous variables may be due to factors beyond the firm’s control.
maintenance costs. In this case as the overall spread of the network is not observed, its positive correlation with the network’s size may create an upward bias in the coefficient of the network’s size.

The fixed effects model (model III) can overcome this heterogeneity bias problem, by taking the firm-specific effects as constants. In this model the estimated coefficients are unbiased even in the presence of such correlations. The inefficiency estimates are obtained using the procedure proposed by Schmidt and Sickles (1984). This model is given by:

$$\ln TC_i = \alpha_0 + x_i' \beta + v_i + u_i,$$

(7)

where $v_i \sim iid (0, \sigma_v^2)$, and $u_i = \alpha_i - \alpha_0$. The firm-specific $\alpha_i$’s are the so-called fixed effects.\(^{13}\) The model can therefore be re-written as:

$$\ln TC_i = \alpha_i + x_i' \beta + v_i.$$

(8)

This model is estimated by applying Ordinary Least Squares. The positive inefficiency scores are then calculated as $\hat{u}_i = \hat{\alpha}_i - \min(\hat{\alpha}_i)$. From this expression, it can be seen that the company with the smallest firm-specific component is regarded as fully efficient and defines the common intercept: $\hat{\alpha}_0 = \min(\hat{\alpha}_i)$.

Finally, Greene’s true random effects model (model IV) is an extension of Aigner et al.’s frontier model that includes an additional time-invariant random term to capture the firm-specific heterogeneity effect on cost by a random intercept component. It can be written as:

$$\ln TC_i = \alpha_0 + \alpha_i + x_i' \beta + v_i + u_i,$$

where

$$\alpha_i \sim iid N(0, \sigma_\alpha^2), v_i \sim iid N(0, \sigma_v^2) \text{ and } u_i \sim iid N^+(0, \sigma_u^2).$$

(9)

\(^{13}\) See Hsiao (2003) for details.
As before, all distributions are assumed do be independent from each other and from the regressors. This model is estimated using Simulated Maximum Likelihood method.\textsuperscript{14} The inefficiency is estimated using the conditional mean of the inefficiency term ($u_{it}$) given by $E[u_{it} \mid \theta_{it}]$, where $\omega_{it} = \alpha_{i} + \epsilon_{it}$.\textsuperscript{15}

Aigner et al.’s model (model I) is formulated as a cross sectional model and thus, ignores the panel aspects of the data. This might lead to inaccurate results due to misspecification by ignoring firm-specific unobserved factors. In the true random effects model (model IV), this problem is addressed by including a separate stochastic term for firm-specific heterogeneity. Such heterogeneity is also accounted in the fixed and random effects models (models II and III). However, both these models impose additional restrictions that might affect the inefficiency estimates. In fact in both models, the firm-specific unobserved effects are interpreted as efficiency differences. Moreover, inefficiency is assumed to be constant over time. Both these assumptions might be quite restrictive. Given that in network industries, a considerable part of the unobserved factors are related to the network complexity and are beyond the firm’s control, thus cannot be considered as firm’s inefficiency. As the time-invariant part of unobserved heterogeneity is primarily captured by the firm-specific effects, the inefficiency estimates are likely to be biased in these models. As for the second assumption, both economic theory and empirical evidence suggest that cost-inefficiency varies with time. New technology shocks and learning are among the reasons why inefficiency varies over time and across individuals. Using a translog production function, Alvarez, Arias and Greene (2003) have shown that even in cases

\textsuperscript{14} See Greene (2001, 2005) for a discussion of the estimation method. For the simulation, 100 Halton draws were used. Our estimations with higher numbers of draws showed that the results are not sensitive to the number of draws.

\textsuperscript{15} See Greene (2002a) for more details.
when the management’s efficiency is constant, the technical efficiency could vary with time.\textsuperscript{16} Moreover, the assumption of time-invariant inefficiencies is not realistic in a relatively long panel such as our sample.

The true random effects model does not require any of these assumptions. However, if the firm-specific heterogeneity is correlated with the explanatory variables, the estimated parameters of the cost function might be biased. In another paper (Farsi, Filippini and Kuenzle, 2003), we proposed an adjustment based on Mundlak (1978)’s formulation to reduce the possible biases in this model.\textsuperscript{17} However, in the present study, our analysis (not reported here) indicates that the estimation results are fairly close with or without this adjustment. Thus, we decided to focus on the model without adjustment. With two heterogeneity terms, this model is expected to provide a better distinction between inefficiency and other unexplained variations. This advantage is especially important in network industries, in which a significant part of unobserved differences is due to time-invariant factors.

In our comparative analysis we consider two aspects of the models’ performance. The first dimension is the estimation of the cost function’s coefficients. In cases such as bus companies (or in general network industries), explanatory variables and costs can be influenced by a number of unobserved network characteristics. For instance, increasing density of stops will increase the costs due to higher infrastructure expenditures, or a ramified network will lead to a higher labor and capital demand than a single-line network. But longer networks might be relatively more complex, in which case complexity is an unobserved factor that is correlated with the network length. A Hausman test is used to confirm that the firm-

\textsuperscript{16} In translog form this time variation is due to interaction of time-invariant inefficiency with explanatory variables.

\textsuperscript{17} See also Farsi, Filippini and Greene (2004) for an application of this method in railway companies.
specific effects are correlated with the explanatory variables. In this case a fixed effects estimator would be unbiased and could thus be used as a benchmark. Therefore, the extent of heterogeneity bias across different models can be compared according to the overall distance of their parameter estimates with respect to those of the fixed effects model.

One can argue that models with more general error structures, such as model IV, have lower biases because the residuals can capture a larger part of the correlations between unobserved heterogeneity and explanatory variables, thus leaving the coefficients less affected. However, the residuals are by definition uncorrelated with explanatory variables and the extent to which they may confound such correlations with errors may significantly vary from one sample to another. Especially, since the frontier estimators are non-linear, the prediction of the biases is not straightforward. This theoretical discussion is beyond the scope of this paper. Here we rather focus on the evaluation of the models with respect to our sample.

The second aspect of the models’ performance concerns the estimation of inefficiency scores. The specification of inefficiency in each model relies on certain assumptions on its error components, which do not violate the consistency of estimated parameters. Therefore, an unbiased estimation of the cost function is not a sufficient condition for a reliable estimation of inefficiency. Given that the “real” inefficiency scores are not known, a high correlation between the inefficiency estimates from different models is usually considered as an indication of the validity of individual approaches. However, as we will see, our results show a rather weak

---

18 This test performed on a GLS random effects model (not reported here), rejects the hypothesis of no correlation between regressors and firm effects.

19 See Farsi and Filippini (2004) for an example of overestimation of inefficiency in the fixed-effects model, which in principle gives unbiased estimates of cost function’s coefficients.
correlation. Therefore, our assessment of various models relies on plausibility arguments.

In particular, the purpose of this paper is to study whether the true random effect model can help solve some of the mentioned problems. It should be noted that the validity of the results depends on the study sample and may vary from one case to another. Therefore, our purpose is not to identify a unique all-purpose model. Rather, our comparative analysis highlights in each one of the models, the effect of unobserved heterogeneity on inefficiency estimates.

4. Data

The data used in this paper are extracted from the annual reports of the Swiss Federal Office of Statistics on public transport companies. The companies operating in main urban centers are excluded from the sample. Most of these companies operate both inner-city tramways and buses, whose functioning is quite different from rural bus transport. Our data set includes information on all the 170 rural companies operating in Switzerland during the study period. However, the data is not available for all years. In several cases lack of information is due to closure or merging with other companies. We decided to exclude the companies that have fewer than four observations. That is, all the companies in the final sample have at least four years of non-missing data. Therefore companies that were closed or taken over by other companies after a short period of operation are excluded. Obviously, such companies are not comparable with other companies because their closure may have been related

20 We also dropped one observation that we suspected as erroneous because of extremely low reported total costs compared to the same company’s total costs reported in other years.
to their excessive costs or other peculiar reasons. Moreover, since the panel models used in this study require in one way or another the estimation of firm-specific effects, four observations per firm appears to be a reasonable minimum. We also excluded Swiss Post\(^\text{21}\) and all its sub-contractors from the sample, because a considerable part of these companies’ revenues is related to package transport and other postal services. Therefore, all the companies included in the sample are mainly involved in passenger transport.

The final data set is an unbalanced panel with 985 observations including 94 operators over a 12-year period from 1986 to 1997. The number of periods per firm varies from 4 to 12 with an average of 10.5 years. The available information includes total costs, total number of employees, network length, total numbers of bus-kilometers and passenger-kilometers as well as those of buses and seats. Table 1 provides a descriptive summary of the main variables used in the analysis.

The variables for the cost function specification were calculated as follows. Total costs \(TC\) are calculated as the total expenditures of the bus companies in a given year. The output \(Y\) is measured by the number of seat-kilometers, which is calculated by multiplying the total number of bus-kilometers in any given year by the average number of seats per bus in that year. It should be pointed out that this calculation is based on the assumption that the number of seats in a bus does not vary considerably in a given company’s fleet in a given year. This is a reasonable assumption because a typical bus company in Switzerland possesses a uniform fleet. Generally, in order to reduce their maintenance costs, companies purchase their vehicles in large quantities from the same supplier and the same model. The number of seats includes both sitting and standing places. In other studies such as Windle (1988), Bhattacharyya et al.

\(^{21}\) Swiss Post, a public company funded by the federal government, mainly in charge of mail delivery and financial services, operates public transport in about 60% of Switzerland’s rural bus network.
(1995) and Jha and Singh (2001), the number of passenger-kilometers is used as output. However, since in Switzerland the rural buses are rarely running at full capacity and they have to run according to the frequency set by the regulators, a considerable number of seats are likely to be empty in a typical bus travel during an off-season period. Therefore, the number of passenger-kilometers is not a representative measure of output. Alternatively, several authors like Berechman (1987), Matas and Raymond (1998) and Fazioli et al. (2003), use bus-kilometers as the output measure. Given that the average vehicle size is likely to vary across different bus companies in our sample, this measure can distort the output in favor of companies with smaller thus less costly buses. The number of seat-kilometers measures the kilometers traveled by the fleet capacity, which is not sensitive to occupancy rate and at the same time account for the variation of vehicle size across companies. In Switzerland the minimum required frequency of bus services (set by the communities) does not considerably change with the actual occupancy rates. Especially in remote areas, it is not unusual that buses occasionally run with few passengers. Therefore, we contend that in the context of Switzerland’s rural bus systems, this measure is more relevant for cost estimations.22

---

22 In any case, the three mentioned output measures are highly correlated in our data and our preliminary estimations suggest that the results are similar regardless of the adopted measure.
Table 1: Descriptive statistics based on 985 observations

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>1. Quartile</th>
<th>Median</th>
<th>3. Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs (TC)</td>
<td>3106</td>
<td>4802</td>
<td>425</td>
<td>1270</td>
<td>3410</td>
</tr>
<tr>
<td>Thousand CHF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output (Y)</td>
<td>47986</td>
<td>85556</td>
<td>5715</td>
<td>16403</td>
<td>53127</td>
</tr>
<tr>
<td>Thousand seat-kilometers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network length (N) km</td>
<td>43</td>
<td>70</td>
<td>13</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Capital price (Pc) CHF/Seat</td>
<td>1343</td>
<td>606</td>
<td>927</td>
<td>1225</td>
<td>1612</td>
</tr>
<tr>
<td>Labor price (Pl) CHF per employee per year</td>
<td>80749</td>
<td>27133</td>
<td>66417</td>
<td>80872</td>
<td>91586</td>
</tr>
<tr>
<td>Number of seats</td>
<td>1067</td>
<td>1721</td>
<td>184</td>
<td>439</td>
<td>1215</td>
</tr>
<tr>
<td>Number of employees</td>
<td>22</td>
<td>35</td>
<td>3</td>
<td>9</td>
<td>23</td>
</tr>
</tbody>
</table>

- All monetary values are in 1997 Swiss Francs (CHF), adjusted for inflation using Switzerland’s consumer price index.

Input prices are defined as factor expenditures per factor unit. Labor price ($P_l$) is defined as the ratio of annual labor costs to the total number of employees.\(^{23}\) Following Friedlaender and Chang (1983), the capital price ($P_c$) is calculated as residual cost divided by the total number of seats (both standing and sitting), where residual cost is total cost minus labor cost.\(^{24}\) Unfortunately, we do not have the required data to calculate the capital stock using the capital inventory method. The use of a simple indicator is justified by the fact that the bus companies do not possess a significant stock of capital apart from the rolling stock, which could be considered as a relatively uniform stock. All the costs and prices are adjusted for inflation using the Switzerland’s consumer price index and are measured in 1997 Swiss Francs. The network length is also included in the explanatory variables as an output characteristic. It is expected that due to organization and coordination problems, all

\(^{23}\) Given the range of variation of salaries in the data we can safely assume that a large majority of the employees in our sample are full-time.

\(^{24}\) See also Filippini and Prioni (2003) for a similar approach.
other factors being constant, longer networks are expected to be more costly. Other output characteristics such as the number of stops per kilometer of network were initially considered. However, given that these variables and some of their interactions proved to be highly correlated with other explanatory variables, we decided to exclude them from the equation to avoid the possibility of multicollinearity.  

5. Estimation Results

The estimation results for the four models are given in table 2. These results show that the output and input price coefficients are positive and highly significant across all models. The estimated coefficient of output from the pooled model (I) is particularly different from those of other models. Noting that model I completely ignores the panel structure of the data, its estimates are likely to be biased through omitted firm-specific factors.

Since total costs and all the continuous explanatory variables are in logarithms and normalized by their medians, the estimated first order coefficients can be interpreted as cost elasticities evaluated at the sample median. For instance, the output coefficients suggest that on average a one percent increase in seat-kilometers will increase the costs by about 0.25 to 0.73 percent depending on the adopted specification. The cost elasticity of the network length is as expected positive ($\alpha_N$) and significant. This implies that the increase in network length will increase total

---

25 We only dropped the variables that had extremely high correlation with other variables (the correlation coefficient of about 0.99). The omission of these variables would not significantly bias the results, because their effects are captured by other variables. However, including such variables creates a near-singularity problem, which might cause high estimation errors.
costs. This result is consistent with previous empirical studies such as Filippini and Prioni (1994, 2003) and Windle (1988).26

The median cost elasticity with respect to the factor price is positive and of similar magnitude in all models. The estimated coefficient for capital price \((\alpha_{pc})\) represents the share of costs attributed to capital at the median production unit, which varies from 51 to 54 percent depending on the model. This result is more or less consistent with the actual data that show a capital share of about half for the sample median. Additionally, the estimated cost function is concave\(^27\) in input prices suggesting that the companies have a cost-minimizing behavior in response to changes in prices.

The coefficient of the linear time trend is significant and positive in all models except in model I, which shows an insignificant effect. These results suggest an annual increase of about 1% in total costs. This result can be explained by the fact that the production technology has not much changed in bus transport. The increase in costs may be related to higher quality of service and increased security requirements.

Although the Hausman specification test’s results suggest that the firm specific effects in a GLS model are correlated with the regressors, the results in table 2 indicate that most of the coefficients do not vary considerably across models II, III and IV. In particular the estimated coefficients of model IV are within a reasonable range of the unbiased estimates of the fixed effects model. Several likelihood ratio tests\(^28\) were performed to test whether the cost function coefficients are similar across models. As expected model I, that ignores the panel structure of the data, is

---

26 It should be noted that Filippini and Prioni (2003) studied the Swiss bus systems though using a different sample in a shorter time period and without cost frontier models.
27 In this context the concavity condition reduces to \(\alpha_{pc} \leq 0\).
28 In the case of true random effects model, as the likelihood function is simulated, a Wald test was used instead. Note that the two tests are asymptotically equivalent.
significantly different from all other models. These tests also suggest statistically significant differences across the other three models. However, these differences are mainly limited to the output coefficient ($\alpha_Y$). In fact, a test on the similarity of all other coefficients was not rejected at p= .05, suggesting that apart from the output coefficient the estimated cost function is similar across these three models.

Table 2: Regression results

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Model I Pooled</th>
<th>Model II RE (ML)</th>
<th>Model III FE</th>
<th>Model IV True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_Y$</td>
<td>0.734*</td>
<td>0.326*</td>
<td>0.247*</td>
<td>0.351*</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.040)</td>
<td>(0.024)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$\alpha_N$</td>
<td>0.122*</td>
<td>0.244*</td>
<td>0.240*</td>
<td>0.264*</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.023)</td>
<td>(0.033)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>$\alpha_{PC}$</td>
<td>0.512*</td>
<td>0.535*</td>
<td>0.540*</td>
<td>0.525*</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.008)</td>
<td>(0.015)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>$\alpha_{YY}$</td>
<td>0.079*</td>
<td>-0.027*</td>
<td>-0.010</td>
<td>0.014*</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.09)</td>
<td>(0.018)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>$\alpha_{NN}$</td>
<td>0.083*</td>
<td>0.063</td>
<td>0.027</td>
<td>0.119*</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.054)</td>
<td>(0.064)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>$\alpha_{PCPC}$</td>
<td>-0.162*</td>
<td>-0.262*</td>
<td>-0.264*</td>
<td>-0.278*</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.015)</td>
<td>(0.025)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>$\alpha_{YN}$</td>
<td>0.003</td>
<td>-0.026</td>
<td>-0.026</td>
<td>-0.094*</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.023)</td>
<td>(0.032)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$\alpha_{VPC}$</td>
<td>-0.067*</td>
<td>-0.095*</td>
<td>-0.098*</td>
<td>-0.115*</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.016)</td>
<td>(0.016)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>$\alpha_{NPC}$</td>
<td>0.129*</td>
<td>0.093*</td>
<td>0.102*</td>
<td>0.116*</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.011)</td>
<td>(0.024)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>$\alpha_T$</td>
<td>-0.005</td>
<td>0.011*</td>
<td>0.015*</td>
<td>0.010*</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.001)</td>
<td>(0.002)</td>
<td>(0.001)</td>
</tr>
</tbody>
</table>

$\sigma_{\alpha_t}$ - Standard errors are given in brackets. * means significantly different from zero at least at 5%.
Table 3 provides a descriptive summary of the inefficiency estimates from different models. These estimates represent the relative excess cost of a given firm compared to a minimum level that would have been achieved if the firm had operated as efficiently as the ‘best practice’ observed in the sample, since $u_t = \ln TC_t - \ln \hat{TC}_t$, where $TC_t$ are the predicted costs of the regression model, including the random error term $\nu_t$.\(^{29}\) In comparing different models it should again be stressed that models II and III assume constant inefficiency over time. Moreover, in these models all the unobserved firm-specific differences are interpreted as inefficiency. As expected, model II and III predict rather implausible inefficiency scores averaging about 1.15 and 1.46. At their face values these numbers suggest an excess cost of more than 100 percent for a typical company. These high values indicate that the heterogeneity across companies is an important driver of cost differences and that neglecting it may create a substantial upward bias in inefficiency scores.

In model I the inefficiency estimates are in a more realistic range, with an average of 0.25 and a maximum value of 0.73. These values, though still arguably quite high, are substantially lower than those predicted by models II and III. However, the inefficiency scores obtained from model I are likely to be overestimated, because in fact they might capture some of the network-specific unobserved heterogeneity, which is not accounted for separately. Model IV, which has two separate stochastic terms for inefficiency and firm-specific heterogeneity, has inefficiency estimates of about 0.09 on average, which stands for a cost saving potential of about 9 percent.\(^{30}\) Although the maximum value of 0.47 appears as excessive, this model’s results

\(^{29}\)Note that cost efficiency can be alternatively defined as the optimal costs divided by actual costs that is, $CE = \exp(-u)$, where $u$ is the relative excess cost given in table 3.

\(^{30}\)This result is consistent with the average inefficiency levels reported in Dalen and Gomez-Lobo (2003) for the Norwegian bus industry.
suggest that in 95 percent of the cases the relative excessive cost is below 15 percent. The high value of estimated inefficiency in the remaining 5 percent can be explained by statistical errors. Therefore, compared to the other models the inefficiency estimates from model IV are plausible and remain within a reasonable range of variation.

Table 3: Inefficiency measures

<table>
<thead>
<tr>
<th></th>
<th>Model I Pooled</th>
<th>Model II RE (ML)</th>
<th>Model III FE</th>
<th>Model IV True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.249</td>
<td>1.147</td>
<td>1.457</td>
<td>0.090</td>
</tr>
<tr>
<td>Median</td>
<td>0.228</td>
<td>1.070</td>
<td>1.408</td>
<td>0.082</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.732</td>
<td>2.825</td>
<td>3.383</td>
<td>0.473</td>
</tr>
<tr>
<td>95th Percentile</td>
<td>0.472</td>
<td>2.316</td>
<td>2.854</td>
<td>0.153</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.042</td>
<td>0.031</td>
<td>0</td>
<td>0.019</td>
</tr>
</tbody>
</table>

The pair-wise correlation coefficients between the inefficiency estimates from different models are listed in table 4. In order for the correlation coefficients to be comparable, they are calculated at the firm level using 94 observations (one observation for each firm). Namely, in models with time-variant efficiency, the inefficiency score is calculated as the firm’s average inefficiency score over the sample period. For models with time-variant inefficiency the correlation coefficients are also given over the total of 985 observations. As expected, in most cases the correlation coefficients are rather low, suggesting substantial differences across models. Some of these differences can be explained by large sampling errors incurred for the estimation of inefficiency for individual companies, especially in cases where the inefficiency can vary with time. This problem for cross-sectional data and short panels is documented by Horrace and Schmidt (1996), Street (2003) and Jensen.
(2000). Obviously, to the extent that inefficiencies remain constant over time, a longer panel can help. Nevertheless, the assumption of constant inefficiency can be unrealistic in long panels.

However, the weak correlation between efficiency estimates across different models suggests that these models differ not only with respect to individual companies’ inefficiency scores but also give significantly different efficiency rankings. In particular, model IV shows a negative correlation with models II and III and a weak positive correlation with model I. Such weak correlation implies that the individual companies might get completely different evaluations depending on the adopted model. To the extent that model IV is a legitimate model that separates unobserved heterogeneity from inefficiency, these results suggest that other models might give a misleading assessment of individual companies. For instance, our estimations show that the company regarded as fully efficient in the fixed effects model (model III), is a company that operates in a relatively short network with a single line. However, this company’s relatively low costs might be related to its simple network, rather than high efficiency. This explanation is consistent with the results of the true random effects model (model IV) that ranks the same company as highly inefficient.

As table 4 shows, the high correlation between inefficiency estimates from models II and III (coefficient of .987) is a striking exception to the rule. This result can be explained by the fact that the inefficiency estimates in both models mainly represent the network-specific heterogeneity. Therefore, the high correlation between

---

31 The rank correlations show a pattern similar to table 4. These results are omitted to avoid repetition.
32 All three coefficients are significantly different from zero at 5% significance level.
33 Model IV suggests that with an inefficiency score of about 10%, this company is less efficient than the average company in the sample.
these two models can only suggest that the estimation of unobserved network effects is not much sensitive to whether fixed or random effects specification is used.

Table 4. Pair-wise Pearson correlation between inefficiency estimates

<table>
<thead>
<tr>
<th></th>
<th>Model I Pooled</th>
<th>Model II RE (ML)</th>
<th>Model III FE</th>
<th>Model IV True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model II</td>
<td>0.496</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model III</td>
<td>0.426</td>
<td>0.987</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Model IV</td>
<td>0.083 (0.371)</td>
<td>-0.082</td>
<td>-0.098</td>
<td>1</td>
</tr>
</tbody>
</table>

- The correlation coefficients have been estimated over the firms (94 observations).
- Correlation coefficient based on 985 observations is given in brackets.

To further test whether the inefficiency estimates differ across various models, a series of t-tests have been performed. The results unequivocally reject the hypothesis that the inefficiency estimates across any pair of models are on average identical. Table 5 shows the estimates of scale and density economies as given in equations (3) and (4), obtained from different models. The results are listed for three representative companies at the first quartile, median and the third quartile outputs. We identified the median (1st/3rd quartile) company as the company that produces the sample median (1st/3rd quartile) of the number of seat-kilometers and considered that company’s corresponding network length in the estimation of density and scale economies. Since the factor prices are assumed to be exogenous, they are held constant at their median values for all three cases.

---

34 We considered alternative definitions for representative companies, e.g. median of both output and network. However, the results are mainly the same insofar as the following discussion is concerned.
Table 5. Economies of scale and density estimates

<table>
<thead>
<tr>
<th></th>
<th>Model I Pooled</th>
<th>Model II RE</th>
<th>Model III FE</th>
<th>Model IV True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED at 1st Quartile</td>
<td>1.545</td>
<td>2.592</td>
<td>3.468</td>
<td>2.224</td>
</tr>
<tr>
<td>ED at Median</td>
<td>1.370</td>
<td>2.780</td>
<td>3.571</td>
<td>2.115</td>
</tr>
<tr>
<td>ED at 3rd Quartile</td>
<td>1.208</td>
<td>3.544</td>
<td>4.485</td>
<td>3.086</td>
</tr>
<tr>
<td>ES at 1st Quartile</td>
<td>1.500</td>
<td>1.719</td>
<td>1.910</td>
<td>1.490</td>
</tr>
<tr>
<td>ES at Median</td>
<td>1.343</td>
<td>1.915</td>
<td>2.059</td>
<td>1.713</td>
</tr>
<tr>
<td>ES at 3rd Quartile</td>
<td>1.008</td>
<td>1.907</td>
<td>2.248</td>
<td>1.879</td>
</tr>
</tbody>
</table>

The results listed in table 5 show a considerable amount of variation between different models. As can be seen in this table, both economies of density and scale are greater than one in all three representative cases, suggesting the presence of unexploited economies in most companies in the sample.35 In particular, the relatively high values of density economies indicate that a more intensive use of a given network would considerably lower the average cost per seat-kilometer. However, it should be noted that the intensity of demand in a given network is beyond the company’s control. An increase in output usually requires some extension in the network, which can be represented by scale economies.

The estimated scale economies from all models also suggest the existence of considerable potential for cost saving through extending the networks. As expected, the economies obtained from an increase in output density in a given network (density economies) are relatively higher than those gained by extending a company’s network (scale economies). The presence of unexploited scale economies in all three

---

35 Only the estimated value of economies of scale for the third quartile for the pooled model is not significantly different from one.
representative cases suggests that most companies are smaller than the cost-minimizing size at which such economies are fully utilized. The small size of rural bus companies in Switzerland is related to the development of this industry that has been historically associated with the growth of small and fragmented user communities.

The high variation of scale and density economies across various models (see table 5) can be partially explained by the models’ differences with respect to the unobserved network effects. If these effects are correlated with explanatory variables (such as output and network length) the values obtained from the fixed effect model (model III) are unbiased and those of the other three models are biased. Particularly, the values estimated by the pooled model (model I) are likely to be biased downward. This model’s estimates are comparable to those reported by Filippini and Prioni (2003), who applied a cross-sectional cost-frontier model to a panel data of 34 bus companies in Switzerland. In one of their specifications, the median company’s size (about twice as our sample median) has been found to be very close to the cost-minimizing size. These results suggest that ignoring the unobserved firm-specific effects can bias the estimated coefficients. In fact such biases are driven by possible correlation of unobserved effects with output and network length. For instance it is plausible that larger networks are more complex in terms of unobserved factors, thus more costly. Such correlations are likely to be positive, thus lead to an underestimation of scale economies.

Theoretically, an unbiased estimation of scale and density economies can be obtained from the fixed effect model (III), because it allows with the regressors correlated firm specific effects. If such correlation is present and positive, then the coefficients of model II and IV are biased upwards. However, because of the large
number of parameters in this model (at least as many as the companies in the sample),
the precision of the results depends on the number of periods in the sample and the
companies’ output variation over the sample period. As indicated in table 2, the
coefficients of most of the second order terms are statistically insignificant in this
model. This can be explained by the relatively high standard errors in the fixed effects
model. Therefore, estimated economies of scale and density for the fixed effects
model might be imprecise.

Interestingly, model I predicts a decreasing scale and density economies with
output, which appears to be consistent with the common perception that sources of
scale economies are exhaustible. However, models II to IV suggest that if the
unobserved heterogeneity is taken into account, this result may be reversed. As table 5
shows, according to these models, the unexploited scale and density economies are
greater in relatively large companies in the sample. This result can be explained by
some of the special features in relatively small network industries: Noting that the
smallest companies in our sample are bus companies with a single line and a few
employees, the potential gains of increasing the size are limited to savings in
distributing the same fixed costs over a higher output. On the other hand, large
companies have complex multiple-line networks. By increasing their size, such
companies can benefit not only from savings in the fixed costs but also from a better
possibility of reallocation of input factors over the network, thus reducing their
variable costs.
6. Conclusions

The application of alternative cost frontier models to a panel of rural bus companies in Switzerland indicates that the inefficiency estimates are sensitive to the adopted model. From a methodological point of view, the results largely depend upon how the unobserved firm-specific heterogeneity among firms is modeled. Our comparative analysis suggests that models that do not distinguish between unobserved network effects and inefficiency can overestimate the inefficiency scores. In particular, if the inefficiency estimates are derived from the firm-specific effects (cf. Schmidt and Sickles, 1984 and Pitt and Lee, 1981), they include an important part of the unobserved exogenous factors related to the network. Our sample shows that such factors can account for a considerable part of cost differences, thus bias the inefficiency estimates to implausibly high values.

This paper also highlights possible differences in cost function coefficients across models. A (pooled) cross-sectional model does not account for network heterogeneity. Since such heterogeneity is likely to be correlated with some of the explanatory variable, this model can give biased coefficients. A fixed effect model can solve the heterogeneity bias in the coefficients. However, because of the large number of parameters (incidental parameters problem), this model might lead to relatively large estimation errors especially for the second-order terms of the cost function. These latter coefficients might be important for the estimation of scale and density economies.

This study suggests that an econometric specification that includes separate stochastic terms for firm-specific effects and inefficiency can improve the estimations
regarding both inefficiencies and slopes. We considered a random-constant cost frontier model ("true" random effects model) proposed by Greene (2004, 2005). The results indicate that the main coefficients of the cost function are fairly close to the unbiased estimators obtained from the fixed effects model. Given that we do not have the true values of efficiency, we cannot conclude the validity of any model regarding inefficiency estimates. However, our analysis suggests that while conventional models could give implausible estimates, the true random effects model's estimates are within a reasonable range. These results underscore the importance of modeling unobserved firm-specific heterogeneity in efficiency measurement of network industries.

The results also indicate that the unexploited scale economies might be greater for relatively large companies, which can benefit from better possibilities of reallocation over larger networks. Such effects could be masked by unobserved network factors, which if neglected could lead to inaccurate results. It should be pointed out that the results of this paper are valid for the specific sample used here and cannot be directly extended to other cases.

From a policy point of view, this study suggests that the "true" random effects model could be a valuable alternative for setting a benchmark in regulating network industries. However, it has to be emphasized that a mechanical use of any of these models in regulation could be misleading. Since each industry has its specific cost characteristics that are not equally well reproduced by these models, establishing a reliable benchmark requires a careful analysis of the cost structure of the industry under consideration. Consequently, these models should be used as one among different instruments in the assessment of subsidy requests.
References


An Analysis of Efficiency and Productivity in Swiss Hospitals

MEHDI FARSI and MASSIMO FILIPPINI *

JEL classification: I180; I120; L330; L250
Keywords: Stochastic frontier; Cost efficiency; Scale economies; General hospitals

1. Introduction

The health care expenditure is growing rapidly in Switzerland. During the five-year period between 1997 and 2002, the national level of health care costs has grown with an average annual rate of about 4.5% attaining about 48 billion Francs in 2002. General hospitals1 incur a considerable part of health costs. In 2002, general hospitals (about 12.4 billion Francs) and specialized clinics (4.0 billion Francs) respectively accounted for about 25.8 and 8.3 percent of the total health care expenditures in Switzerland. In particular, the general hospitals sector shows an increasing growth rate rising form about 3.9 percent per year between 1997 and 1999 to an average of about 6.5 percent per year between 2000 and 2002. This increasing growth has raised the public interest in improving the performance of hospitals and determining the extent and identifying the sources of possible inefficiencies in this sector.

This paper studies the productive efficiency of the Swiss general hospitals. The financial data of 214 general hospitals over the four-year period between 1998 and 2001 are used. Specialized clinics are excluded from this study. After excluding the hospitals with less than twenty beds and the observations with missing and suspicious values, the final regression sample includes 459 observations of 156 general hospitals. The cost efficiency of hospitals is studied using stochastic cost frontier analysis. Several specifications are considered and the results are compared. The efficiency estimates of individual hospitals are also analyzed to test whether hospitals with different ownership and subsidization types are significantly different regarding efficiency. The results

* Department of Management, Technology and Economics, Swiss Federal Institute of Technology, Zurich, Switzerland and Department of Economics, University of Lugano, Switzerland. The authors wish to thank the editor and two anonymous referees for their helpful suggestions and André Meister, Luca Crivelli and Luca Stäger for their general support. This paper is based on extracts from the final report with the same title (June 2004) prepared for the Swiss Federal Statistical Office. The financial support of the SFSO and the Swiss Federal Office for Social Security is gratefully acknowledged. The original data are provided by the SFSO. The views expressed in this paper are those of the authors and do not necessarily reflect the positions of the sponsoring agencies.

1 In Switzerland hospitals are divided into two categories: general hospitals and specialized clinics. While general hospitals provide short-term medical care in any field, specialized clinics are restricted to one of the following care categories: psychiatrics, rehabilitation, surgeries, gynecology/neonatology, pediatrics, geriatrics, and other specialties. See SFSO (2001) for more details.
suggest considerable savings could be achieved through improvement of hospitals’
efficiency. On average, university hospitals and large regional facilities are the most
costly providers. However, part of these cost differences could be due to higher
expenses resulting from teaching and research activities. In small hospitals, one of the
main sources of excessive costs is related to lengthy hospital stays. The inefficiency
estimates do not provide any evidence of significant differences among hospitals with
different ownership/subsidy types. The results also point to unexploited economies of
scale.

The rest of the paper is organized as follows. Section 2 provides a general
description of the cost frontier approach followed by a discussion of the adopted
functional form and econometric specification. A descriptive analysis of the data is
given in Section 3. Section 4 describes the model specification. The estimation results
along with a discussion of cost and scale efficiency and the effects of ownership/subsidy
types are presented in Section 5. Section 6 concludes the paper with a summary of the
main results.

2. Methodology

There are several methods to estimate the cost efficiency of individual firms.
Two main categories are non-parametric methods which originated from operations
research, and econometric approaches namely stochastic cost frontier models. In non-
parametric approaches like Data Envelopment Analysis, the cost frontier is considered
as a deterministic function of the observed variables but no specific functional form is
imposed. Moreover, non-parametric approaches are generally easier to estimate.
Parametric methods on the other hand, allow for a random unobserved heterogeneity
among different firms but need to specify a functional form for the cost function. The
main advantage of such methods over non-parametric approaches is the separation of
the inefficiency effect from the statistical noise due to data errors, unobserved variables
etc. Another advantage of parametric methods is that these methods allow statistical
inference on the significance of the variables included in the model, using standard
statistical tests. In non-parametric methods on the other hand, statistical inference
requires elaborate and sensitive re-sampling methods like bootstrap techniques. Given
the above discussion we decided to focus on the stochastic cost frontier models.

Many authors have used cost frontier models to evaluate hospitals’ efficiency.
ZUCKERMAN ET AL. (1994), LINNA (1998) are two examples. The former paper used a
translog functional form while the latter used a Cobb-Douglas cost function. ROSKO
(2001) has also used the frontier approach with a translog cost function and with
instrumental variables to account for the potential endogeneity of capital and labor
prices. The use of cost frontier models to evaluate efficiency in the health-care sector
has been criticized by NEWHOUSE (1994) and SKINNER (1994). The main arguments
against these models are related to the unobserved heterogeneity due to differences in

2 See KUMBHAKAR and LOVELL (2000) for an extensive survey of parametric methods and COELLI ET AL.
(1998, chapter 6), and SIMAR (1992) for an overview of non-parametric approaches.
3 It should be noted that most non-parametric methods require convexity restrictions. See COELLI ET AL.
(2003) for more details on DEA. See also STEINMANN and ZWEIFEL (2003) for an application of DEA to
estimate the efficiency of Swiss hospitals.
4 These methods are available for rather special cases and have not yet been established as standard tests.
See SIMAR and WILSON (2000) for an overview of statistical inference methods in non-parametric
models.
case-mix and quality and the errors committed by aggregation of outputs as well as non-testable assumptions on the distribution of efficiency.

**Folland and Hofler (2001)** provide a discussion on the reliability of hospital efficiency estimates obtained from stochastic cost frontier models. These authors show that the individual efficiency estimates are rather sensitive to the adopted model specification and functional form. However, the results are robust when the comparisons are performed between hospital group mean inefficiencies. This finding is consistent with the results reported by **Hadley and Zucker (1994)** suggesting that the stochastic frontier analysis of hospitals efficiency is of practical use when applied for comparing group means. **Farsi, Filippini and Kuenzle (2005)** reached a similar conclusion in their study of the Swiss nursing homes.

A frontier cost function defines minimum costs given output level, input factor prices and the existing production technology. Theoretically, the perfectly efficient production units are located at the frontier. In stochastic frontier approach it is assumed that the cost frontier can differ across production units. The difference between a firm’s observed costs and its corresponding frontier costs is decomposed into two parts: The first part is a symmetric random error due to the unobserved differences between firms and the second component is related to the inefficiency of the firm. With certain assumptions on the distribution of these stochastic terms, individual inefficiencies can be estimated.

Cost frontier models also allow an estimation of scale efficiency. Scale efficiency indicates the degree to which a company is producing at optimal scale. The optimal size of a firm is defined as the amount of output that minimizes the average cost of producing one unit of output. Frisch (1965) defines the optimal scale as the level of operation where the scale elasticity is equal to one. The degree of returns to scale ($RS$) is defined as the proportional increase in output ($Y$) resulting from a proportional increase in all input factors, holding all input prices and output characteristic variables fixed (Caves et al., 1981). The $RS$ degree may also be defined in terms of the effects on total costs resulting from a proportional increase in output (Slik and Berndt, 2003). This is equivalent to the inverse of the elasticity of total cost with respect to the output.5

### 2.1 Functional form

The cost frontier is a function of output and input factor prices. Other hospital and output characteristics like quality indicators can also be included as independent variables. **Griffin et al. (1987)** provide a comprehensive list of alternative functional forms. These authors have also proposed a series of criteria for selecting the functional form in cost and production analyses. These criteria can be grouped in four categories corresponding to hypotheses, estimation methods, data and application. The first category concerns the restrictions imposed by the maintained hypotheses. In the absence of such hypotheses the unrestrictive forms are more appropriate. Second, the availability of data may restrict the choice of statistical estimation procedures. As the number of variables increase, most functional forms require a geometrically increasing number of parameters to be estimated, thus necessitate much larger samples. The third criterion

---

5 The inverse of cost elasticity of output is referred to by Chambers (1988), as the “economies of size” rather than economies of scale, which are defined in regards to production function. Scale and size economies are equivalent if and only if the production function is homothetic (see Chambers, 1988, p. 72). However, as for the purpose of this study we are more interested in the cost effects of output, we define the returns to scale in terms of cost elasticity.
concerns the conformity of the functional form to the data. Finally, in some applications, some properties are desired in the functional form, because for instance they might be used in simulations.

In this study the most important restrictions are related to the sample size and the estimation method. The best choice is therefore a functional form that can be estimated with available estimation procedures and limits the number of parameters while using as many relevant variables as possible. One of the most commonly used functional forms is the Cobb-Douglas (log-linear) model. A Cobb-Douglas cost function with \( M \) outputs, \( N \) input factors and \( K \) output characteristics can be written as:

\[
\ln TC = \beta_0 + \sum_{m=1}^{M} \beta_m \ln Y_m + \sum_{n=1}^{N} \gamma_n \ln P_n + \sum_{k=1}^{K} \omega_k Z_k
\]

where \( TC \) is the total costs; \( Y_m (m=1, \ldots, M) \) are the outputs; \( P_n (n=1, N) \) are the input factor prices; and \( Z_k (k=1, \ldots, K) \) are output characteristics and other exogenous factors that may affect costs.

The main advantage of this model is its simplicity. Thanks to its limited number of variables the Cobb-Douglas form has a practical advantage in statistical estimations over more complicated forms. The interpretation of the results is also easier because it does not include any interaction term. Another interesting characteristic of this function is self-duality. Namely, the corresponding production function of a Cobb-Douglas cost function is also log-linear. The main shortcoming of this model is the assumption of constant scale elasticity, which implies a constant rate of scale economies. This might be considered as restrictive because by using the same proportional increase in output, small companies usually gain more than large firms. However, in some industries, it might be the case that the scale elasticity does not vary much in the range of observed data.

The potential changes in scale elasticity with output can be analyzed using flexible functional forms. One of the main flexible forms is transcendental logarithmic (translog) model. This model is a second-order Taylor approximation of any arbitrary function. However, a translog model requires the estimation of a large number of parameters. Furthermore, the included interaction terms could cause multicollinearity. These problems can substantially affect the model’s statistical performance. As we will see later there are at least 15 important variables that are essential for our cost models. Compared to the sample size that is limited to about 500 observations, the number of parameters in more general functional forms can be excessively high. For instance the adopted specification with a general (non-homothetic) translog model could easily have more than 30 parameters. Moreover, the primary purpose of this study is hospitals’ cost efficiency and the scale economies come only as secondary results. A numerically feasible estimation of a translog cost frontier was only possible with simplified specifications that excluded some of the important output characteristics.\(^6\) We therefore decided to focus on the Cobb-Douglas functional form. Because of its simplicity, this functional form is commonly used in recent papers on cost-efficiency measurements such as Greene (2003, 2005) and Linna (1998). Nevertheless our main results

\(^6\) We estimated several specifications with translog form. The results (not reported here) indicate that when applied to our data, these models tend to converge to solutions in which one of the stochastic components degenerates to zero.
especially those related to scale economies are also confirmed by an additional analysis (not reported in this paper) with a parsimonious translog model with homothetic cost function.\footnote{The adopted specification is based on a simplification of model III (explained later) with 21 variables.}

It is generally assumed that the cost function is the result of cost minimization given input prices and output. Cost functions should therefore satisfy certain properties.\footnote{For more details on the functional form of the cost function see CORNES (1992, p.106).} Mainly, the cost function must be non-decreasing, concave, linearly homogeneous in input prices and non-decreasing in output. The linear homogeneity constraint is usually imposed by dividing total costs and input prices by one of the factor prices. However, as we see later, we do not impose this restriction because our models do not include all input factors. The other theoretical restrictions are usually verified after the estimation. In particular, the concavity of the estimated cost function reflects the fact that the cost function is a result of cost minimization. This latter condition is automatically satisfied in Cobb-Douglas functional form.

### 2.2 Econometric specification

There are a number of econometric approaches to estimate stochastic cost frontier models. KUMBHAKAR and LOVELL (2000) provide an extensive survey of these methods. A general form of a stochastic cost frontier can be written as:

\[
TC_{it} = f (Y_{1it}, \ldots, Y_{mit}; P_{1it}, \ldots, P_{nit}; Z_{1it}, \ldots, Z_{kit}) + u_{it} + v_{it} \tag{2}
\]

where subscripts \(i\) and \(t\) represent the firm and year respectively; \(u_{it}\) is a positive stochastic term representing inefficiency of firm \(i\) in year \(t\); \(v_{it}\) is the random noise or unobserved heterogeneity; and other variables are similar to those in Equation 1. Typically, it is assumed that the heterogeneity term \(v_{it}\) is normally distributed and that the inefficiency term \(u_{it}\) has a half-normal distribution that is, a normal distribution truncated at zero:

\[
u_{it} \sim N(0, \sigma_u^2), \quad u_{it} \sim \left| N(0, \sigma_u^2) \right|.
\tag{3}
\]

This model is based on the original cost frontier model proposed by AIGNER ET AL. (1977). The firm’s inefficiency is estimated using the conditional mean of the inefficiency term \(E[u_{it} | u_{it} + v_{it}]\), proposed by JONDROW ET AL. (1982).

An important variation of this model is PITT and LEE (1981)’s model in which the inefficiency term \(u_{it}\) is assumed to be constant over time, that is: \(u_{it} \sim N(0, \sigma_u^2)\). There is also another version of this model (proposed by SCHMIDT and SICKLES (1984)), that relaxes the distribution assumptions on both \(u_{it}\) and \(v_{it}\), and estimates the model using Generalized Least Squares (GLS) method. The advantage of these models is that they use the panel aspect of the data to estimate the parameters. In cases where the individual firm effects (\(u_{it}\)) are correlated with the explanatory variables, the estimated parameters may be biased. SCHMIDT and SICKLES (1984) proposed a fixed-effects approach to avoid such biases. In this model the inefficiency term is not random and is estimated as an intercept for each company.
There is however, an important practical problem with the fixed-effect model in that it requires the estimation of a large number of parameters, which limits its application to reasonably long panels with sufficient within-firm variation. Generally, in short panels the fixed effects are subject to considerable estimation biases, which directly reflect in the inefficiency scores. Given that our data is a rather short panel of four years, the fixed effects model is not a quite feasible approach. Moreover, our preliminary analysis shows that in virtually all the main variables, the between variations are dominant and the within variations are comparatively insignificant.

Another important issue is that in both fixed and random effects models discussed above, the inefficiencies are assumed to be constant over time. This is an unrealistic assumption in most practical cases, where the driving forces of cost-inefficiency are not generally persistent. In fact firms constantly face new problems and revise their strategies. Moreover, there exist incentive mechanisms (either through regulation and monitoring or through profit and career incentives) that induce managers to correct their past suboptimal decisions.

GREENE (2005, 2004) proposes a new approach that integrates the random and fixed effects approaches into the original AIGNER ET AL. (1977)’s model. Some of these models have been successfully used in other sectors like nursing homes (FARSI, FILIPPINI and KUENZLE, 2005) and public transport (FARSI, FILIPPINI and GREENE, 2005). These models can be written by adding a firm-specific stochastic term ($\alpha_i$) in the right-hand-side of Equation 2. This term is an i.i.d. random component in random-effects framework, or a constant parameter in fixed-effects approach. Such models have an important advantage in that they allow for time-variant inefficiency while controlling for firm-level unobserved heterogeneity through fixed or random effects. The main difficulty of these models is that they are numerically cumbersome. In particular, our experience suggests that in cases where the within variation in the data is low, these methods are numerically unstable. Our preliminary analyses show that with the available data, these models were not numerically feasible. This can be explained by the small number of periods in our sample and its relatively low within variations. As we see later in Section 3, our sample is an unbalanced data with maximum 4 periods but on average it has about three periods.

The data constraints and also the numerical restrictions bring us back to the original pooled frontier model in line with AIGNER ET AL. (1977). However, we also estimated PITTMAN and LEE (1981)’s model and checked if the results are consistent. Our analysis (not reported here) indicates that in terms of scale economies the two models provide comparable results. In terms of efficiency estimates the results show a quite high correlation. However, the results estimated from Pitt and Lee’s model were systematically higher than those of the pooled model. This difference can be explained by the fact that the inefficiency estimates from Pitt and Lee’s model capture other sources of heterogeneity across hospitals that are not necessarily related to inefficiency. In fact our analysis suggests that the firm-specific effects capture a significant part of between variations in costs and reflect them as inefficiency. Given that there may be a great amount of unobserved heterogeneity among hospitals, we contend that these estimates are likely to be exaggerated. Therefore, we restricted our analysis to the pooled model as shown in Equation 2.

---

9 See GREENE (2005, 2002) for more details. This author considers a panel of 5 years as a short panel.
3. Data

The data used in this study are extracted from the annual data reported by Swiss general hospitals to the Federal Statistical Office from 1998 to 2001. The sample consists of an unbalanced panel with 747 observations from 1998 through 2001. According to these data, overall 214 general hospitals have operated in Switzerland during this period. The Swiss Federal Statistical Office classifies general hospitals classified into five typologies based on their size and level of specialization. The details of this classification are given in SFSO (2001). Typology 1 includes only the five university hospitals, which provide a wide variety of services in a large number of specializations. At the other extreme, Typology 5 includes small general hospitals (mostly less than 100 beds), which provide basic medical care with few specializations. Accounting for more than 40 percent of Switzerland’s hospitals, this category has the highest number of hospitals in the sample. Table 1 lists the number of general hospitals available in the data by year and hospital typology. In line with the SFSO classification, we assume that hospitals with different typologies provide different levels of medical care.

| Type Code | Description | Number of hospitals |
|----------|-------------|---------------------|---------------------|
|          |             | 1998 | 1999 | 2000 | 2001 |
| 1 K111   | Centralized care level 1 (university hospital) | 5   | 5   | 5   | 5   |
| 2 K112   | Centralized care level 2 (regional hospital) | 20  | 20  | 21  | 21  |
| 3 K121   | Basic-care hospital level 3 (relatively large/specialized) | 27  | 28  | 29  | 30  |
| 4 K122   | Basic-care hospital level 4 (moderate size/specialization) | 53  | 53  | 56  | 53  |
| 5 K123   | Basic-care hospital level 5 (small size/low specialization) | 86  | 89  | 73  | 68  |
| Total    |             | 191 | 195 | 184 | 177 |

These data (administrative data) include variables such as total costs, total salaries and labor charges, hospital operating costs, capital expenditure, number of employees and paid hours, and number of hospitalizations. Capital costs are considered as the sum of the maintenance and repair costs for buildings and equipment, interest charges and all other investment charges and amortizations. Costs related to nursing staff’s salaries and physicians salaries and fees are also given separately. These variables allow that salaries for physicians, nursing staff and other employees be calculated separately. Among other variables are the total hospital revenue from medical services and its outpatient-related part. The reporting errors have been explored using an outlier analysis focusing on main variables used in the analysis such as capital and labor costs, numbers of beds, hospitalizations and paid hours. The observations with suspicious values have been excluded from the sample.\(^{10}\)

\(^{10}\) See FILIPPINI and FARSI (2004) for a detailed description.
Another data set used in this study is an aggregate extraction of the medical data of the Swiss hospitals from 1998 to 2001 with records for each individual admission. The extracted data used in this study consists of the number of cases by AP-DRG in each hospital. These data were merged with the cost weights from Swiss APDRG version 4.0 developed by Institut de Santé et d’Economie (2003). These cost weights are used as an official reference for cost reimbursement in several Swiss cantons that have adopted a DRG-based reimbursement system. These data have been used to calculate an average cost weight (AP-DRG adjustment ratio) for each hospital-year. The adjusted number of admissions is then calculated by multiplying these adjustment ratios by the number of admissions recorded in the administrative data.

The main trends in the number of hospitalizations are given in Table 2. This table shows that during the study period, while the number of hospitalizations has slightly increased (about 4%), the total number of semi-hospitalizations has significantly increased. Particularly, the semi-hospitalization cases have increased by about 35% from 1998 to 1999. Given that the distinction between full and semi-hospitalizations is not fully clear, more representative trend patterns can be seen through the numbers of admissions and patient-days. These numbers show that while the aggregate output of Swiss general hospitals have increased by about 10% in terms of admissions from 1998 to 2001, the number of patient-days has rather fluctuated around 9 million, suggesting shorter hospital stays over time. This pattern is confirmed by a continuous decrease in the average length of hospitalization from more than 12 days in 1998 to 10.7 days in 2001. The main observation here is the presence of a growing demand of hospital care shown by an overall increase in number of admissions. These numbers also point to a general trend in Switzerland’s hospitals to limit the hospital stays and to favor the short-term treatments like one-day surgeries and other semi-hospitalizations, over long-term hospitalizations.

Table 2 also lists the total hospital costs in the general hospitals. These numbers point to a significantly increasing trend of 3 to 6 percent per year. The ambulatory revenues account for a considerable portion (about 13%) of total costs. The aggregate numbers do not show any significant change in the share of ambulatory revenue over the study period. The average AP-DRG adjustment ratios are also given. These numbers do not change considerably over time. Finally, Table 2 indicates that the average size of general hospitals has slightly increased over the study period. This change can be explained by the decrease in the number of small hospitals (Typology 5) as shown in Table 1.

---

11 See SFSO (1997b) for more details on these data.
12 APR-DRG (All-Patients-Refined Diagnostic Related Groups) is a system of classification of diseases patented by 3M Health Information Systems www.3Mhis.com.
13 These cost weights were estimated based on a sample of about 200,000 acute short-term hospitalizations in 12 Swiss hospitals (including 3 university hospitals) during 1999-2001.
14 We observed some differences between the number of DRG records from the medical data and the number of hospitalizations from the administrative data, suggesting that some of the cases were not coded. Our method is based on the assumption that non-coded patients are not systematically different from the coded cases.
15 Usually, the planned hospitalizations of less than 24 hours such as one-day surgeries are referred to as semi-hospitalizations. See S.F.S.O. (1997) for more details. However, reporting an admission as semi-or full hospitalization is rather discretionary.
Table 2: Main trends in hospitalizations in Swiss general hospitals

<table>
<thead>
<tr>
<th></th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of full hospitalizations</td>
<td>918'972</td>
<td>938'525</td>
<td>972'244</td>
<td>955'729</td>
</tr>
<tr>
<td>Total number of semi-hospitalizations</td>
<td>114'309</td>
<td>158'604</td>
<td>179'870</td>
<td>186'064</td>
</tr>
<tr>
<td>Total number of hospitalizations</td>
<td>1'033'281</td>
<td>1'097'129</td>
<td>1'152'114</td>
<td>1'141'793</td>
</tr>
<tr>
<td>Total number of patient-days</td>
<td>8'977'192</td>
<td>9'180'478</td>
<td>9'000'636</td>
<td>8'733'425</td>
</tr>
<tr>
<td>Total hospital costs*</td>
<td>10'334</td>
<td>10'719</td>
<td>11'353</td>
<td>11'851</td>
</tr>
<tr>
<td>Total ambulatory revenues*</td>
<td>1'351</td>
<td>1'340</td>
<td>1'590</td>
<td>1'530</td>
</tr>
<tr>
<td>Average length of hospitalization</td>
<td>12.40</td>
<td>12.92</td>
<td>11.45</td>
<td>10.74</td>
</tr>
<tr>
<td>Average AP-DRG adjustment ratio</td>
<td>0.786</td>
<td>0.797</td>
<td>0.798</td>
<td>0.804</td>
</tr>
<tr>
<td>Average hospital capacity (beds)</td>
<td>163</td>
<td>162</td>
<td>166</td>
<td>172</td>
</tr>
</tbody>
</table>

* In million Swiss Francs deflated to May 2000 prices.

The number of general hospitals and their average capacity for six groups (by regulation /ownership) are listed in Table 3. According to these data, out of 177 general hospitals that operated in Switzerland in 2001, 88 hospitals (63.4 percent of hospital beds) were public (owned by government), 53 (23.5 percent of beds) were private non-profit, and 36 (13% of beds) were for-profit hospitals. All public hospitals and most private non-profit hospitals (about 80% of these hospital beds) are subsidized, whereas in the private for-profit sector, only 36% of the hospital beds are operated in subsidized hospitals. In fact only 8 for-profit hospitals benefited from government subsidies in 2001. We also studied the distribution of hospital regulation/ownership types across different typologies. It turns out that all university hospitals (Typology 1) and almost all regional hospitals (Typology 2) benefit from government subsidies. The distribution of different regulation/ownership types in the basic-care hospitals (Typologies 3 to 5) is not much different from the overall distribution shown in Table 3.

Table 3 also lists the average hospital size measured by the number of beds for each ownership/subsidy type. Public hospitals with 221 beds on average are by far the largest providers of health care, followed by private non-profit facilities with 135 beds and for-profit hospitals with 111 beds on average. This table also shows that the subsidized hospitals are considerably larger (an average capacity of 200 beds) than non-subsidized ones (average of 90 beds). Finally, the for-profit hospitals that benefit from subsidies (178 beds on average) are likely to be larger than the subsidized non-profit hospitals (156 beds on average).

---

16 According to our data for 2001, only one out of 21 regional hospitals was not subsidized.
### Table 3: Distribution of hospitals and average hospital size by ownership/subsidy type (2001)

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC</th>
<th>PRIVATE NON-PROFIT</th>
<th>PRIVATE FOR PROFIT</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBSIDIZED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td>88</td>
<td>37</td>
<td>8</td>
<td>133</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td>221</td>
<td>156</td>
<td>178</td>
<td>200</td>
</tr>
<tr>
<td><strong>NON SUBSIDIZED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td></td>
<td>16</td>
<td>28</td>
<td>44</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td></td>
<td>86</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td>88</td>
<td>53</td>
<td>36</td>
<td>177</td>
</tr>
<tr>
<td>Hospital size (beds)</td>
<td>221</td>
<td>135</td>
<td>111</td>
<td>172</td>
</tr>
</tbody>
</table>

The above descriptive analyses were based on the entire sample of general hospitals with 747 observations. Given that a number of variables used in the cost frontier analysis have missing values, this sample could not be entirely used for the regressions. The final sample used for the regression analysis consists of all the observations that have non-missing values for all the variables used in the specification of cost frontier models. We also excluded eight hospitals (27 observations) with less than 20 beds. This sample includes 459 observations related to 156 general hospitals. In this sample, there are only 69 hospitals that have non-missing values for all the four years. There are 24 hospitals that have information only for one year. In addition, there are respectively 30 and 33 other hospitals with non-missing values for 2 and 3 years. The regression sample is therefore an unbalanced panel with an average of 3 periods. This sample on average includes about 61 percent of all the general hospitals that operated in Switzerland from 1998 to 2001. A simple analysis (not reported here) using t-test, shows that the excluded 288 observations and the regression sample (with 459 observations) are not significantly different regarding hospital’s outputs, average LOS, total costs, labor costs and number of beds. Therefore, with a relatively high representation rate in all groups, the regression sample can be considered as a representative sample of all Swiss general hospitals in the study period. A descriptive summary of this sample is given in the next section (see Table 4).

### 4. Model specification

The efficiency of hospitals is studied using a total cost function with Cobb-Douglas functional form. Four different model specifications are considered. The cost functions used in this study are based on two outputs: hospitalizations and outpatient (ambulatory) care. Many authors such as LINNA (1998), ROSKO (2001) and HESHMATI (2002) used the DRG-weighted number of admissions as the hospital’s main output.ROSKO (2001) also controls for several case-mix adjusters such as the percentage of ER visits and outpatient surgeries out of all outpatient visits. Other authors like VITA (1990), EAKIN (1991) and STEINMANN and ZWEIFEL (2003) have considered unadjusted number of cases in several departments as multiple outputs. BROWN (2003) considers cases with DRG weight of lower than 1, between 1 and 2, and higher than 2 as three output categories.

---

17 We also observed unreasonable and suspicious values in a small number of observations, which were changed to missing. See FILIPPINI and Farsi (2004, Section 3.1), for more details.

18 The differences are never significant at 5% level. The following hospital outputs are considered: number of admissions, DRG-adjusted number of admissions, number of patient-days and ambulatory revenues.

19 ROSKO (2001) also controls for several case-mix adjusters such as the percentage of ER visits and outpatient surgeries out of all outpatient visits. Other authors like VITA (1990), EAKIN (1991) and STEINMANN and ZWEIFEL (2003) have considered unadjusted number of cases in several departments as multiple outputs. BROWN (2003) considers cases with DRG weight of lower than 1, between 1 and 2, and higher than 2 as three output categories.
Here, the main measure of hospitalization output is taken as the AP-DRG adjusted number of hospitalizations including both full and semi-hospitalizations. However, the unadjusted number of hospitalizations and the number of patient-days are also considered as alternatives. Other authors like VITA (1990), EAKIN (1991) and STEINMANN and ZWEIFEL (2003) have considered unadjusted number of cases in several departments as multiple outputs.

Since the number of outpatient cases is not available in the data, the ambulatory output is approximated by the corresponding revenues in real monetary terms (with May 2000 prices). This approximation is based on the assumption that the average unit price of ambulatory care is similar across hospitals. The ambulatory revenues are reported zero for about 5 percent of the observations. Since our econometric models are based on a logarithmic form, the zero values are replaced by a negligible value. This method has been used by KIM (1987) and GILLIGAN and SMIRLOCK (1984).20 As the minimum non-zero value in the regression sample is about CHF 120'000, we replace the zero values by one (less than .001 of the mean value) making the log values equal to zero.

Three input factors are considered: capital, physician labor services and all other employees’ labor services. Capital price is approximated by the hospital’s total capital expenditure divided by the number of available beds in the hospital. Therefore, similar to WAGSTAFF (1989) and ROSKO (2001) among others, the capital stock is proxied by the hospital capacity in terms of beds. Many authors have considered labor inputs in multiple categories.21 In this paper, similar to EAKIN (1991), physicians and non-physicians are considered as two labor categories. Physicians’ services constitute of interventions for medical treatments while other employees’ services are more continuous and aimed at nursing care, administration and maintenance. Furthermore, physicians’ wage rates are considerably higher and more variable than other employees.

Labor prices are calculated by dividing total salaries by the number of remunerated days for employed physicians and other employees. The physicians’ fees are not included.22 In fact, since these fees may also include payments to physicians who are not employed by the hospital, the regular salaries represent a more accurate measure of labor price. Labor prices are proportionally adjusted for social charges, which on average, account for about 8 percent of total costs. Namely, these charges are proportionally distributed to each one of the two groups (physicians, non-physicians), the proportions being the shares of each group’s salaries. This adjustment captures the potential variation in social charges across hospitals due to differences in pension funds as well as the age and seniority of the employees.

The three input factors considered in the models do not include all the hospital’s costs. In fact, capital and labor costs on average account for about 76 percent of a hospital’s total cost. Other expenses such as medical materials, food, cleaning, water and power etc. are on average, about 24 percent of total costs. Furthermore, the labor prices do not account for physicians’ fees and other personnel charges, which together, account for about 6.7 % of the total costs. This means that about 31 percent of the total costs. This means that about 31 percent of the total costs. This means that about 31 percent of the total

---

20 It should be noted that there exist other solutions (such as Box-Cox or hybrid functional forms) for the problem of zero values for one or several outputs in a cost function. See WENINGER (2003) for a recent review. Given that in our sample the zero values are quite limited, we adopted the simplest method.
21 For instance EAKIN (1991) considers physicians and other staff and FOLLAND and HOFLER (2001) consider nursing staff and other employees in separate categories. Others like STEINMANN and ZWEIFEL (2003), SCUFFHAM ET AL. (1996) and VITA (1990) used 3 or 4 labor categories.
22 Physicians’ fees (honoraires) account on average, for about 5.8% of total costs.
costs are related to input factors whose prices are not considered in the model. In fact, the available data do not allow an appropriate calculation of these prices. Given that the model specification does not include all input prices, the linear homogeneity cannot be imposed. The excluded prices are obviously not constant and neglecting their variation may affect the estimation results. However, some of these variations are probably captured by the three included factor prices. For instance, physicians’ fees are likely to be correlated with physicians’ salaries. Another concern is the accuracy of the price data. The measurement error in price variables may create bias in the price coefficients. However, to the extent that these measurement errors and the unobserved factor prices are randomly distributed across hospitals and over time, the other coefficients are not affected by any bias.

In addition to outputs and input prices, a series of hospital characteristics are included in the model. We included the year dummies to capture the technological progress and the variation in unobserved variables such as potential differences in reporting procedures and data collection from one year to another. For instance, some of the observed patterns in the data suggest that AP-DRG coding has improved over the years. The typology indicators are also included. The provided medical services vary across hospital types. In particular, university and regional hospitals provide a wide variety of services while other types provide basic medical care and do not have many specializations. Another difference is in teaching and research activities that are generally much less significant in basic-care hospitals. We considered three indicators for hospital typologies. Since there are too few university hospitals in the sample for having a meaningful separate indicator for these hospitals, a single indicator is considered for Typologies 1 and 2.

After a careful study of all other available characteristics, we concluded that as long as hospital typology is included in the model, additional variables can be limited to a few indicators representing important aspects of hospital output. The most important output characteristic is the average length of hospitalization. Many authors such as VITA (1990), SCUFFHAM ET AL. (1996) and CAREY (1997) have included this variable as an output characteristic. As we see later, variation in the length of stay is one of the main sources of cost differences between hospitals. One may argue that the DRG adjustment already controls for any justifiable variation in the length of stay. In this case, including the average length-of-stay in the model results in an underestimation of inefficiency in hospitals with lengthy stays. However, DRG adjustment is only an approximate way to control for severity variations. In fact, there are considerable cost variations among patients with the same DRG. For instance, the acceptable range of variation of hospital stays provided by the Swiss APDRG version 4.0 (I.S.E, 2003) is quite wide within a given DRG. Thus, considering a fixed LOS for all patients with the same DRG is at best a great approximation. Moreover, given that the length of hospital stays also represents hospital’s ‘hotel’ services like nursing care and accommodation rather than medical treatment, the LOS can be regarded as a separate output.

---

23 Though the expenditures on these input factors are available in the data the quantities are not. Moreover, these expenditures correspond to diverse items that could not be measured with similar units.
24 Following the suggestion of one of the referees we performed several t-tests to explore the input price variations across private and public hospitals. The average labor prices are not significantly different. As expected, capital prices are on average higher for private hospitals that have a lower access to subsidies and other tax benefits. We did not find any evidence of overstatement of prices in one or another category.
25 Others like FOLLAND and HOLFER (2001) have considered the LOS through the number of patient days.
26 See BREYER (1987) for a discussion.
Hospitals’ costs can also be affected by quality of care. The evidence on the effect of quality measures on hospital costs is rather mixed. Referring to his previous empirical literature, ROSKO (2001) concludes that the omission of quality indicators may not be as serious as commonly thought. For instance, ZUCKERMAN ET AL. (1994) controlled for several outcome measures of quality such as 30-day post-admission mortality rates. Their analysis suggests that none of those measures have significant effects. Similarly VITALIANO and TOREN (1996) report that most of their 12 quality measures showed insignificant effects on hospital costs. On the other hand, FOLLAND and HOFLER (2001) have considered two measures of structural quality (percentage of board-certified physicians and a measure of bed availability), both of which showed significant effects on total costs. In general the available evidence often points to a significant effect by structural quality measures, while outcome and process measures are more likely to appear unimportant. This may be explained by the fact that the structural quality is usually easier to measure whereas other quality indicators especially outcome measures are prone to measurement errors and outside factors. Given the data availability and measurement problems, we focused on one structural measure of quality, defined as the hospital’s nurse per bed ratio. We also included two binary indicators for emergency room (ER) and geriatrics department. Emergency services are usually costly and involve relatively severe cases, while geriatrics cases are less intensive in medical care.

It is assumed that all hospitals have similar cost functions and the hospital typology can only shift the costs without affecting the function’s shape and parameters. To study the validity of this assumption we used several tests of structural break. First, we considered the hypothesis that hospitals with different typologies have different cost function parameters. Four hospital groups have been considered, with the university hospitals and regional hospitals considered in a single group. Secondly, we grouped the hospitals in two groups: centralized general hospitals (Types 1 and 2) and basic-care hospitals (Types 3, 4 and 5). The third test is based on a break between small basic-care hospitals (Type 5) and other hospitals. The model specification includes the number of DRG-adjusted admissions and ambulatory revenues as output; capital price and a single labor price as input prices; LOS as output characteristics; and year dummies. None of the three tests can reject the hypothesis of no-structural break, suggesting that the cost function parameters are overall similar across different typologies. Finally, given that university hospitals might be completely different from other hospital types, we estimated the models on an alternative sample excluding the university hospitals. The results indicate that the presence of these hospitals does not affect the estimation results significantly.

Four specifications labeled as models I to IV, have been considered. The general model can be written as:

$$\ln TC_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 \ln AMB_{it} + \gamma_1 \ln PK_{it} + \gamma_2 \ln PL_{4it} + \gamma_3 \ln PL_{2it} + \omega_1 \ln LOS_{it} + \omega_2 \ln NB_{it} + \omega_3 \ln ER_{it} + \omega_4 \ln GER_{it} + \delta_1 D_{1it} + \delta_2 D_{2it} + \delta_3 D_{3it} + \delta_4 D_{4it}$$

+ $\delta_y Y_{99i} + \delta_y Y_{00i} + \delta_y Y_{01i} + u_{it} + \nu_{it}$

(4)

27 These tests are based on Chow test. The null hypothesis is that the regression coefficients are identical across different groups.
Subscripts \( i \) and \( t \) represent the hospital and year respectively. The stochastic components \( u_{it} \) and \( v_{it} \) respectively represent inefficiency and random noise as described in Equation 2. \( Y \) is the hospitalization output, which is taken as unadjusted number of hospitalizations in Model I, number of patient-days in Model III, and DRG-adjusted number of hospitalizations in Models II and IV. \( AMB \) is the ambulatory revenue; \( PK \), \( PL_1 \) and \( PL_2 \) are respective factor prices for capital, physicians and other employees; \( LOS \) is the average length of hospitalization (not included in Model III); \( NB \) the nurse per bed ratio; \( ER \) and \( GER \) are dummy variables for emergency room and geriatrics department respectively. \( D_{12} \) is a dummy for Typologies 1 and 2; and \( D_1 \) and \( D_4 \) are dummies for hospitals in Typologies 3 and 4. The small basic-care hospitals (Type 5) are the omitted typology. Finally, \( Y_{99} \), \( Y_{00} \) and \( Y_{01} \) are the year dummies for 1999, 2000 and 2001 respectively, 1998 being the omitted year.

The specification given in (4) summarizes Models I to III. Descriptive statistics of the main variables used in these models are given in Table 4. Model IV is similar to Model II with the difference that 13 additional binary indicators are also included for 14 cantonal groups.\(^28\) The idea here is to control for part of the unobserved heterogeneity that is specific to location. Populations in different cantons may differ in health and socio-economic status. Moreover, the hospitals are subject to different cantonal regulations that may affect their efficiency. Comparing this model with other models without canton dummies can indicate to what extent the inefficiency variations can be explained by differences in cantonal regulations. Finally, given that our measure of outpatient services is based on revenues rather than visits, the cantonal dummies could help capture some of the differences in outpatient unit prices across cantons.

The effects of ownership/regulation types on efficiency are studied using a two-stage method. This method is based on a non-parametric rank test on the efficiency estimates. The inefficiency scores for each hospital are considered as the average inefficiency values over the sample period. The hospitals that have apparently changed ownership status from one year to another are excluded from the analysis.\(^29\)

The two-stage approach has a disadvantage in that the first-stage estimation errors may affect the results of the test in the second-stage. These errors may lead to an under-rejection of the null hypothesis that cost-efficiencies are similar across different types.\(^30\) An alternative approach is to include ownership/subsidy indicators in the regressions and test the significance of the corresponding coefficients. We performed a GLS estimation of this alternative specification. The results (not reported here) generally confirm those obtained by the two-stage method. However, we decided to use the two-stage approach because it allows the use of non-parametric statistical tests like Kruskal-Wallis rank test.\(^31\) Such tests do not impose any distribution assumption on the efficiency scores. The KW test has an additional advantage in that it relies on efficiency

\(^{28}\) There are 23 cantons in the regression sample. Most of the cantons with less than 5% share in the sample are grouped with the neighboring cantons. Only two groups have less than 5% share in the sample (see Appendix).

\(^{29}\) According to our data out of 159 hospitals in the regression sample, there are 13 hospitals whose ownership has changed from one year to another. From these hospitals, 5 have changed status between public and FP, 5 between public and private NP and 3 between FP and private NP status. Some of these changes might be because of reporting errors. Moreover, previous studies like Farsi (2004) suggest that hospitals that undergo a change in ownership might be subject to gradual changes long before the ownership change occurs.

\(^{30}\) For a more detailed discussion of this point see Farsi and Filippini (2004).

\(^{31}\) This test is due to Kruskal and Wallis (1952). See Singh and Coelli (2001) and Farsi and Filippini (2004) for examples of application of this test to compare the efficiency across groups of firms.
ranks rather than efficiency magnitudes that are subject to relatively large estimation errors and sensitive to outlier values.

5. Results

Table 5 lists the regression results of the cost frontier analysis, with four different specifications. Some descriptive statistics of inefficiency estimates are also given at the bottom of the table. Most of the coefficients are statistically significant. Overall, the coefficients are generally reasonable and have the expected signs. The first two models (I and II) are based on the number of hospitalizations. In Model I the hospitalizations are not adjusted, whereas in Model II the hospitalization numbers are adjusted with AP-DRG cost weights. The first observation is that ignoring DRG adjustment slightly biases the coefficients. For instance, the output coefficient increases by about .03 and the first typology dummy by .04 when the number of hospitalizations is not adjusted. However, these biases appear to be insignificant for practical purposes. This can be explained by the fact that adjusted and unadjusted numbers of hospitalizations are highly correlated, with a correlation coefficient of about 0.99.
Table 4: Descriptive statistics of the regression sample

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hospital's total costs</strong> (SFr '000)</td>
<td>57'267</td>
<td>101'240</td>
<td>14'170</td>
<td>27'253</td>
<td>57'845</td>
</tr>
<tr>
<td><strong>Number of admissions</strong></td>
<td>5'692</td>
<td>6'399</td>
<td>1'571</td>
<td>3'761</td>
<td>7'779</td>
</tr>
<tr>
<td><strong>Number of admissions (AP-DRG adjusted)</strong></td>
<td>4'792</td>
<td>6'148</td>
<td>1'238</td>
<td>2'946</td>
<td>6'118</td>
</tr>
<tr>
<td><strong>Number of patient-days</strong></td>
<td>48'801</td>
<td>53'241</td>
<td>18'917</td>
<td>32'186</td>
<td>56'685</td>
</tr>
<tr>
<td><strong>Hospital's outpatient revenues (SFr '000)</strong></td>
<td>7'958</td>
<td>13'920</td>
<td>1'296</td>
<td>3'635</td>
<td>8'007</td>
</tr>
<tr>
<td><strong>P_k (capital price)</strong> SFr '000 per bed</td>
<td>23.60</td>
<td>20.15</td>
<td>10.41</td>
<td>16.21</td>
<td>27.73</td>
</tr>
<tr>
<td><strong>P_L - physicians (SFr per day)</strong></td>
<td>368.10</td>
<td>499.67</td>
<td>244.14</td>
<td>307.75</td>
<td>387.51</td>
</tr>
<tr>
<td><strong>P_L - others (SFr per day)</strong></td>
<td>181.74</td>
<td>137.28</td>
<td>153.14</td>
<td>169.49</td>
<td>194.03</td>
</tr>
<tr>
<td><strong>Nurse per bed</strong></td>
<td>0.882</td>
<td>0.354</td>
<td>0.633</td>
<td>0.839</td>
<td>1.067</td>
</tr>
<tr>
<td><strong>Average length of hospitalization (days)</strong></td>
<td>12.06</td>
<td>6.78</td>
<td>8.22</td>
<td>9.31</td>
<td>13.80</td>
</tr>
<tr>
<td><strong>Emergency Room</strong></td>
<td>0.847</td>
<td>0.360</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Geriatrics</strong></td>
<td>0.429</td>
<td>0.496</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Typology 1/2</strong></td>
<td>0.155</td>
<td>0.362</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Typology 3</strong></td>
<td>0.148</td>
<td>0.356</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Typology 4</strong></td>
<td>0.264</td>
<td>0.441</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Year 1999</strong></td>
<td>0.266</td>
<td>0.442</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Year 2000</strong></td>
<td>0.255</td>
<td>0.436</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Year 2001</strong></td>
<td>0.264</td>
<td>0.441</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The sample includes 459 observations from 156 general hospitals (1998-2001). All monetary values are deflated to May 2000 prices. Labor prices include charges. * calculated for physicians employed by the hospital. ** includes all hospital employees except physicians. *** calculated for hospitalizations of longer than 24 hours.

We consider Model II as our benchmark model because it is a complete model with DRG-adjusted output and all the relevant factors. According to this model the main output’s coefficient is 0.82, that is, a 1% increase in the adjusted number of
hospitalization will result in about 0.82% increase in total costs. As expected, the coefficient of ambulatory output is much smaller (.036), suggesting a .036% rise in total costs as a result of 1% increase in outpatient revenues, all other factors being constant.

The coefficient of LOS is about 0.45, suggesting that for instance, a 1% increase in the average length of hospitalization results in a .45% increase in total costs. Given that hospital stays are on average about 12 days, this result implies that a difference of one day in the hospital’s average LOS is approximately equivalent to 4% of total costs. The length of hospitalization is therefore an important predictor of hospital costs. Comparing the LOS coefficient between models I and II shows that if hospital output is not adjusted for severity, the effect of LOS is considerably higher (coefficient of 0.53). This result suggests that the average LOS captures part of the variations in severity.

In Model III the number of patient-days is considered as the hospital’s main output. The output coefficient in this model (0.81) is very close to the corresponding coefficient in Model II confirming the existence of unexploited scale economies. As expected the ambulatory output’s coefficient is higher in this model, because a patient-day is on average less costly than one case. According to this model the marginal cost of a relative increase in patient-day is on average about 11 times higher than that of a similar increase in outpatient visits.

As expected, the price coefficients are positive and significant. Since the three factor prices do not include all hospital inputs, the price coefficients do not add to one. The nurse per bed ratio has a positive and significant effect, indicating that quality of care is costly. As expected, the ER dummy has a positive coefficient, but its effect is statistically insignificant in most models. Similarly, the geriatrics dummy has expectedly a negative but insignificant coefficient in most models. As explained earlier, the effects of these indicators are partly captured by typology dummies.

All three typology dummies are positive, indicating that all other factors held constant, small basic-care hospitals are less costly than other hospitals. However, their difference with other basic-care hospitals (types 3 and 4) is significant only if the average LOS is not controlled for. This implies that the systematic cost differences between these hospital types are mainly due to their different hospitalization lengths. University and regional hospitals (typologies 1 and 2) are significantly more costly than basic-care hospitals. According to Model II compared to small basic-care hospitals the difference is strikingly high amounting to 35% in total costs. This difference can be partly explained by the additional expenses on medical equipment and also research and teaching activities. This result is in general consistent with the results documented in I.S.E. (2003) suggesting that the AP-DRG cost weights are on average 24% higher in university hospitals.
Table 5: Estimation results

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of admissions</strong></td>
<td>0.8532 *</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of admissions</strong></td>
<td>-</td>
<td>0.8180 *</td>
<td>-</td>
<td>0.7916 *</td>
</tr>
<tr>
<td>(AP-DRG adjusted)</td>
<td></td>
<td>(0.028)</td>
<td></td>
<td>(0.026)</td>
</tr>
<tr>
<td><strong>Number of patient-days</strong></td>
<td>-</td>
<td>0.8140 *</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Outpatient revenues</strong></td>
<td>0.0321 *</td>
<td>0.0357 *</td>
<td>0.0724 *</td>
<td>0.0363 *</td>
</tr>
<tr>
<td></td>
<td>(0.0071)</td>
<td>(0.0059)</td>
<td>(0.0063)</td>
<td>(0.0070)</td>
</tr>
<tr>
<td><strong>$P_K$ (capital price)</strong></td>
<td>0.1434 *</td>
<td>0.1552 *</td>
<td>0.1676 *</td>
<td>0.1866 *</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.019)</td>
<td>(0.018)</td>
</tr>
<tr>
<td><strong>$P_L$ - physicians</strong></td>
<td>0.0746 *</td>
<td>0.0764 *</td>
<td>0.0387</td>
<td>0.0507 *</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.017)</td>
<td>(0.023)</td>
<td>(0.020)</td>
</tr>
<tr>
<td><strong>$P_L$ - others</strong></td>
<td>0.2142 *</td>
<td>0.1981 *</td>
<td>0.2599 *</td>
<td>0.1445 *</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.041)</td>
<td>(0.061)</td>
<td>(0.044)</td>
</tr>
<tr>
<td><strong>Nurse per bed</strong></td>
<td>0.1875 *</td>
<td>0.1617 *</td>
<td>0.2236 *</td>
<td>0.1093 *</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.030)</td>
<td>(0.032)</td>
<td>(0.028)</td>
</tr>
<tr>
<td><strong>Average length of</strong></td>
<td>0.5346 *</td>
<td>0.4451 *</td>
<td>-</td>
<td>0.4759 *</td>
</tr>
<tr>
<td>hospitalization</td>
<td>(0.036)</td>
<td>(0.036)</td>
<td></td>
<td>(0.042)</td>
</tr>
<tr>
<td><strong>Emergency Room</strong></td>
<td>0.0369</td>
<td>0.0400</td>
<td>-0.0209</td>
<td>0.0953 *</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.038)</td>
<td>(0.036)</td>
<td>(0.039)</td>
</tr>
<tr>
<td><strong>Geriatrics</strong></td>
<td>-0.0591 *</td>
<td>-0.0423</td>
<td>-0.0395</td>
<td>-0.0767 *</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.028)</td>
<td>(0.032)</td>
<td>(0.034)</td>
</tr>
<tr>
<td><strong>Typology 1/2</strong></td>
<td>0.3915 *</td>
<td>0.3499 *</td>
<td>0.3766 *</td>
<td>0.3888 *</td>
</tr>
<tr>
<td></td>
<td>(0.075)</td>
<td>(0.079)</td>
<td>(0.096)</td>
<td>(0.077)</td>
</tr>
<tr>
<td><strong>Typology 3</strong></td>
<td>0.0974</td>
<td>0.0701</td>
<td>0.2801 *</td>
<td>0.1176 *</td>
</tr>
<tr>
<td></td>
<td>(0.056)</td>
<td>(0.058)</td>
<td>(0.076)</td>
<td>(0.052)</td>
</tr>
<tr>
<td><strong>Typology 4</strong></td>
<td>0.0135</td>
<td>0.0312</td>
<td>0.1316 *</td>
<td>0.0625</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.043)</td>
<td>(0.050)</td>
<td>(0.042)</td>
</tr>
<tr>
<td><strong>1999</strong></td>
<td>0.0349</td>
<td>0.0377</td>
<td>0.0447</td>
<td>0.0225</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.030)</td>
<td>(0.035)</td>
<td>(0.029)</td>
</tr>
<tr>
<td><strong>2000</strong></td>
<td>0.0884 *</td>
<td>0.0722 *</td>
<td>0.1222 *</td>
<td>0.0617 *</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.037)</td>
<td>(0.029)</td>
</tr>
<tr>
<td><strong>2001</strong></td>
<td>0.1426 *</td>
<td>0.1314 *</td>
<td>0.1830 *</td>
<td>0.1140 *</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.032)</td>
<td>(0.035)</td>
<td>(0.030)</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>-0.4354</td>
<td>0.2632</td>
<td>-1.1268 *</td>
<td>0.6754 *</td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.27)</td>
<td>(0.35)</td>
<td>(0.27)</td>
</tr>
</tbody>
</table>

**Inefficiency scores**

<table>
<thead>
<tr>
<th></th>
<th>$U_{mean}$</th>
<th>$U_{median}$</th>
<th>$U_{95perc}$</th>
<th>$U_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2261</td>
<td>0.2194</td>
<td>0.2164</td>
<td>0.1774</td>
</tr>
<tr>
<td></td>
<td>0.1879</td>
<td>0.1912</td>
<td>0.1930</td>
<td>0.1557</td>
</tr>
<tr>
<td></td>
<td>0.4355</td>
<td>0.4474</td>
<td>0.4190</td>
<td>0.3369</td>
</tr>
<tr>
<td></td>
<td>1.2845</td>
<td>1.1520</td>
<td>1.0849</td>
<td>0.8434</td>
</tr>
</tbody>
</table>

Standard errors are given in parentheses. * significant at .05
Model IV includes canton dummies, which are not shown in the table.
The year dummies indicate that over the study period the total hospital costs have grown about 4 percent by year. However, it should be noted that these dummies might capture other year-specific effects such as changes in quality of reporting DRG cases. Most probably, such changes in quality of data are not significant between 2000 and 2001. Therefore the difference between the coefficients of these two dummies is more representative of the annual growth in total costs. Interestingly, the growth in total costs from 2000 to 2001 is about 6% in all four models, while the differences with previous years vary considerably across different specifications. The robustness of this result to model specification confirms that the estimated growth after 2000 is not affected by changes in data quality. It should be noted that the year dummies should generally represent the technical change. Technical progress should in principle result in lower costs in usual production units that produce a similar output. However, an increasing growth in hospital costs and in the health-care sector in general is a common observation that is not contradictory to technical progress. In fact, with progress in medical technology, hospitals use increasingly more advanced methods and the quality of medical care constantly increases. All these changes result in higher costs. Many of these cost-increasing factors are not directly taken into account, thus are captured by the year dummies. Therefore, the estimated growth in costs should not be interpreted as a decline in technology.

Table 5 also provides the estimation results obtained from Model IV. This model is similar to Model II with the sole difference that 13 canton dummies are also included as explanatory variables. Comparing the results of Model IV with those of Model II indicates that the two models give quite similar results for the output coefficients and also the effect of LOS. In general the coefficients corresponding to time-variant factors (including year dummies) are not sensitive to whether or not the cantonal effects are controlled for. However, the coefficients of time-invariant factors, namely typology dummies and ER and geriatrics indicators, have considerably changed. In particular both ER and geriatrics dummies are significant when the canton dummies are included. As for inefficiency estimates, controlling for canton dummies decreases the inefficiency scores by about .04 (compare .22 in Model II with .18 in Model IV).

### 5.1 Scale economies and cost efficiency

The results listed in Table 5 indicate that the main coefficients are more or less similar across different specifications. In particular the coefficient of the hospital’s main output is about 0.8 in all specifications. This result implies that the returns to scale are on average significantly higher than 1 \( \left( \frac{\partial \ln T C}{\partial \ln Y} \right) = 1.2 \), suggesting that the majority of general hospitals in Switzerland do not fully exploit the potential scale economies. This implies that most of the hospitals in the sample do not reach the optimal size. This result is consistent with the empirical evidence in previous literature. In particular, CRIVELLI ET AL. (2001) who used a translog cost frontier model for Swiss hospitals between 1989 and 1991, suggest an optimal size of 300 beds, but conclude

---

32 See DRANO (2000) for an extensive study of such evolutions in the US health sector. In particular, this author studies how non-price competition between health care providers has resulted in higher quality and costs.  
33 The coefficients of canton dummies are listed in the appendix. Among these 13 indicators, 7 have significant effects (at .05 level). Canton Geneva has the most costly general hospitals while the least expensive general hospitals are located in canton Ticino. Compared to Bern (the omitted canton), Geneva’s hospitals are on average 29% more costly, and Ticino’s hospitals are on average 18% more economical.
that the unexploited scale economies are relatively low for hospitals with more than 135 beds. Other empirical results in the literature suggest an optimal size of about 200 beds.\footnote{See for instance ALETAS (1999), DRANO\'E (1998) and SCUFFHAM ET AL. (1996).} This implies that the unexploited scale economies could be considerable in typology 5, where a large majority of hospitals are smaller than 100 beds. On the other hand, in university and regional hospitals (types 1 and 2), where the capacity is generally higher than 200 beds and in the large basic care hospitals (type 3) with only about 10 percent of the hospitals smaller than 150 beds, such economies are likely to be fully exploited.

These results might appear in contradiction with the significantly higher costs in university and regional hospitals, suggested by their typology dummy’s coefficient. Whereas in typologies 4 and 5, with virtually all hospitals smaller than 200 beds the total costs are relatively low. However, it should be noted that the typology dummies should capture the specialization effect that while being correlated with size has a different effect on costs. The estimated effects of typology dummies suggest that hospitals with a higher number of service centers (departments such as surgical, pediatrics etc.) especially university and regional hospitals, are significantly more costly than other hospitals. At the same time the estimated output elasticity suggests that \textit{ceteris paribus} the larger hospitals can better exploit the scale economies and thus be less costly. A possible implication is that merging two small hospitals is economical if they provide similar departments after the merger, but might have additional costs if each one has some different departments.

Some statistics of the inefficiency estimates are given in the lower panel of Table 5. These results suggest that the average inefficiency score is not very sensitive to DRG-adjustment (22.6\% in Model I and 21.9\% in Model II). However, the maximum inefficiency score is significantly lower with DRG adjustment. This result suggests that the individual efficiency estimates especially the outliers can be biased if the output is not adjusted for case mix severity. In Model III in which the hospital output is measured as the number of patient-days, the average inefficiency score is quite similar to Model II, where the output is the number of hospitalizations. However, the maximum inefficiency estimate is about 7 percentage points lower in Model III, suggesting that part of the cost inefficiency in certain hospitals is due to the outlier cases that have longer than usual hospitalizations. The inefficiency estimates of Model IV are on average about 4 percentage points lower than those of Model II. The difference between the two models is more considerable at the tails with about 30 percentage points at the maximum. This result suggests that controlling for certain unobserved differences through canton dummies can considerably attenuate the estimates of individual hospitals’ inefficiencies.

The close similarity among average inefficiency estimates and the strong correlation between the individual scores obtained from different models suggest that the results are in general robust to specification.\footnote{All the pair-wise correlation coefficients between efficiency estimates from Models I, II and IV are higher than 85\%. The estimates of Model III show a correlation of 68 to 80 percent with those of the other three models.} Given that the inefficiency results are more or less similar across different models, we choose a single model to highlight some of the patterns in cost-efficiency. We consider Model II to present the results regarding the cost-efficiency, because this model controls for DRG variation. The variations in inefficiency scores by year and hospital typology are depicted in Figure 1. The first observation is that the inefficiency scores are different across hospital
typologies. In particular, the university hospitals (Type 1) have the highest scores with average values of 36% to 44%. The t-tests performed on the sample suggest that the efficiency difference with all other hospital types is statistically significant at 5%. However, this result should be considered with caution. University hospitals have the highest levels of research and teaching activities and provide a relatively wide range of medical services including most complex interventions. Given that there are only three university hospitals in our regression sample a separate dummy could not be included for these hospitals. Therefore, the inefficiency estimates inevitably capture some of these unobserved differences. A better estimation of cost inefficiency in university hospitals requires more information about the incurred costs of research and teaching activities and other medical interventions that are exclusively carried out in these hospitals.

Figure 1: Inefficiency by hospital typology and year (based on Model II)

Small basic-care hospitals (Type 5) with an average inefficiency of 24 to 25 percent have the second highest inefficiency scores. Similarly t-tests suggest these differences with type 1 and also with other types are statistically significant. Other hospital types (Typologies 2, 3 and 4) show a rather similar average inefficiency score of 18 to 20 percent. It should be noted that these inefficiency estimates are obtained after accounting for a potential shift of cost frontiers across hospital types. The second result from this figure is that the inefficiency has decreased over the sample period in all hospital types except large basic-care hospitals (Typology 3). The decrease of inefficiency is considerable in university hospitals (about 6 to 7 percentage points) but rather insignificant in small basic-care hospitals (Typology 5). These results also suggest that the inefficiency has slightly decreased in hospitals of Typologies 2 and 4.

36 The differences between hospital types 3 and 4 are not significant at 5%, but type 2 shows a significantly lower inefficiency compared to both types 3 and 4. The differences are however limited to 2 percentage points.
but slightly increased in type-3 hospitals. To explore the statistical significance of these changes we performed several t-tests between years 1998 and 2001. The results suggest that none of the above changes are significant at 5% significance level.

5.2 Effects of ownership/subsidy types

The inefficiency estimates obtained from different models do not show much difference insofar as the differences between ownership/regulation types are concerned. In order to avoid repetition the results are only reported for Model II, which we considered as the most realistic specification. The inefficiency estimates from Model II are given in Table 6. The numbers in this table point to slight efficiency differences among hospitals with different ownership or subsidies. For instance it appears that private NP hospitals are on average slightly more costly than FP and public hospitals, or subsidized hospitals are on average more cost-efficient than non-subsidized facilities. Particularly, this table suggests that among non-subsidized providers, the private NP hospitals are on average more costly than the FP ones. We used several tests to study whether these differences are statistically significant.

We used the Kruskal-Wallis test for several alternative sets of subgroups to test if the differences shown in Table 6 imply that different subgroups belong to different populations of hospitals in terms of their cost-efficiency. The first grouping is based on five ownership/subsidy subgroups as shown in Table 6, that is public, subsidized private NP, non-subsidized private NP, subsidized private FP, and non-subsidized private FP. The second grouping is related to ownership (public, private NP and FP) and the third is related to subsidies (subsidized versus non-subsidized). Finally the last set consists of three subgroups: public, private subsidized and private non-subsidized. In all cases, we also performed the test for all the possible pair-wise comparisons such as public vs. private, private NP vs. private FP etc. In order to see if the results are sensitive to the presence of university and regional hospitals (Typologies 1 and 2), similar tests were also performed on a sample excluding these hospitals.

The results indicate that in all the groupings and all pair-wise comparisons the Chi-squared test statistic is statistically insignificant even at 10% significance level.\(^{37}\) In the case of pair-wise comparisons, the results of the KW tests are confirmed with a simple t-test with equal variances. These results suggest that there is no statistically significant difference in efficiency among hospitals with different ownership or regulation types. These results are in general consistent with those reported by STEINMANN and ZWEIFEL (2003) who did not find any significant difference between private and public hospitals.

\(^{37}\) Only in one case the differences were significant at 10% but not significant at 5%. This was related to the comparison of public versus private hospitals using Model III. Note that this model does not adjust the output for AP-DRG (see Table 6).
Table 6: Average inefficiency estimates by ownership/subsidy type

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC</th>
<th>PRIVATE NON-PROFIT</th>
<th>PRIVATE FOR-PROFIT</th>
<th>OVERALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBSIDIZED</td>
<td>0.214</td>
<td>0.213</td>
<td>0.203</td>
<td>0.213</td>
</tr>
<tr>
<td>NON SUBSIDIZED</td>
<td>-</td>
<td>0.239</td>
<td>0.215</td>
<td>0.227</td>
</tr>
<tr>
<td>OVERALL</td>
<td>0.214</td>
<td>0.220</td>
<td>0.212</td>
<td>0.216</td>
</tr>
</tbody>
</table>

Notes: The inefficiency estimates are based on the results obtained from Model II (see Table 5). The inefficiency scores for each hospital are calculated as the average inefficiency values over the sample period (1998 to 2001). The results are based on 146 hospitals that have a constant ownership/regulation status over the sample period.

6. Conclusions

A panel data of all Swiss general hospitals over the four-year period between 1998 and 2001, including 747 observations of a total of 214 facilities, has been analyzed. These data show a significant increase in the total number of hospitalizations amounting to about 10 percent growth over the study period. In the same period, the total expenditures of Swiss general hospitals have increased by about 15 percent. Our descriptive analysis of the data shows that most hospitals while decreasing their average length of stay, have considerably increased the share of their outpatient revenues. The observed patterns in the data indicate that the small basic-care hospitals have the longest hospitalizations (on average about seven days longer than other hospitals) and that university hospitals treat the most severe cases shown by the highest average AP-DRG cost weight (20% higher than the overall average).

A sample of 459 observations corresponding to a total of 156 hospitals has been used for the econometric analysis of efficiency. A stochastic total cost frontier has been estimated using Cobb-Douglas functional form and several specifications. The main results of this analysis can be listed as follows:

- There are unexploited scale economies in the majority of Swiss general hospitals. Although we cannot clearly identify the optimal hospital size, our results along with the empirical evidence reported in the previous literature suggest that unexploited scale economies could be significant in hospitals with less than 200 beds.

- There are systematic cost differences among different typologies with hospital types with higher specialization levels being generally more costly. These differences remain considerable after controlling for severity through AP-DRG cost weights. In particular, the university and regional hospitals are the most costly hospitals (about 35% more costly than the small basic-care hospitals). This difference can be explained by the relatively wide range of medical specializations as well as research and teaching activities in those hospitals.

- Ignoring the severity adjustment by AP-DRG cost weights slightly biases the main coefficients. However, these differences are not significant for practical purposes, suggesting that most of the variation in DRGs among hospitals is random.
• A one-day decrease in the average length of hospitalization could lower the hospital’s total costs by up to about 4 percent. Given that the small basic-care hospitals have extremely long hospitalizations, considerable savings might be achieved by curtailing lengthy hospital stays.

• The marginal cost of ambulatory visits is much lower than that of inpatient care. To the extent that the insurers have more accommodating reimbursement plans for outpatient services, this result might partly explain the motivation behind the growing share of ambulatory care in most hospitals.

• Although our quality measures are limited the results suggest that the quality of medical services is an important factor in cost determination. Thus, some of the estimated cost differences could be due to unobserved variations in quality.

• There exists a considerable cost variation among hospitals operating in different cantons. Part of these differences may be related to different regulatory systems implemented in different regions.

• On average, the total costs of a typical general hospital have grown by about 4 percent per year. This can be explained by technological progress in medical care, which enables the hospitals to provide more advanced services to more severe cases resulting in higher costs.

The cost-efficiency analysis using several models indicates that the inefficiency scores are not sensitive to the adopted model specification. The resulting mean inefficiency score of about 20 percent suggests that there is a potential for cost saving in Switzerland’s general hospitals. However, it should be noted that part of these inefficiency estimates might be driven by unobserved factors. A better account of such factors would require a longer panel that is, more observations over time. The estimations also suggest that the cost-inefficiency has slightly but consistently decreased over the study period. Certain typologies show significantly different inefficiency estimates. In particular, the university hospitals show the highest inefficiency estimates. However, these estimates are partly because of the special activities like advanced medical research and complex medical interventions in these hospitals. The inefficiency estimates are also relatively high in small basic-care hospitals. This is probably related to extremely long hospitalizations in these hospitals. Given the methodological and data limitations of this study, the individual hospitals’ efficiency scores should be considered with caution. In particular, these estimates should not be directly used as a basis for rewarding or punishing specific hospitals. Rather, the present analysis provides an overall picture of inefficiency situation in Switzerland’s general hospitals.

Finally, the effect of different regulatory systems and ownership types on the hospital efficiency has been studied. The general hospitals are divided into five groups based on their ownership (public, private non-profit and for-profit) and subsidy status (subsidized, not subsidized). A large majority of Switzerland’s hospitals are owned by the State or benefit from government subsidies. Our data show that in 2001, 63 percent of general hospital beds were owned by the State, which together with the subsidized hospitals owned by the private sector, account for about 87 percent of the total general hospital beds in Switzerland. Our analysis of inefficiency estimates obtained from the stochastic frontier analysis suggests that the efficiency differences across different
ownership/subsidy types are not statistically significant. This result indicates that our data do not provide any evidence of a significant efficiency advantage of one type over another. However, it should be noted that this result is restricted to our specific data and cannot be generalized. Moreover, because of the potential correlation between ownership/subsidy types and other hospital characteristics such as typology and size, disentangling the actual effects of ownership/regulation may be difficult. Therefore, the presented results cannot be considered as conclusive evidence that different subsidy rules and ownerships induce similar cost efficiency.

In general the quality of the available data is acceptable for an econometric analysis of cost-efficiency. However, because of the limited number of available years with non-missing data (three in most hospitals), some of the advanced panel data econometric models could not be used. We contend that the data can be generally improved by minimizing the missing values and reporting errors and including more years. Moreover, potential data improvements can be considered in accounting capital investments and amortization, reporting average wage rates for hospital employees as well as coding DRGs and admission types. Furthermore, additional information on the resources allocated to research and teaching activities and hospital quality can be useful for an accurate analysis of costs. At the end, it should be noted that this paper is one of the first attempts in the analysis of efficiency in Swiss hospital using parametric methods. This issue requires further research. Especially as more data and detailed information become available, future studies should consider flexible functional forms and more elaborate panel data models and possibly include variables related to hospitals’ teaching and training activities.
Appendix

Regression coefficients for canton dummies (Model IV)

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Err.</th>
<th>Sample Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG</td>
<td>0.1599 *</td>
<td>0.0718</td>
<td>0.0501</td>
</tr>
<tr>
<td>BL / BS / SO</td>
<td>0.1368 *</td>
<td>0.0573</td>
<td>0.1111</td>
</tr>
<tr>
<td>FR</td>
<td>0.0682</td>
<td>0.0832</td>
<td>0.0610</td>
</tr>
<tr>
<td>GE</td>
<td>0.2902 *</td>
<td>0.0735</td>
<td>0.0196</td>
</tr>
<tr>
<td>GR</td>
<td>0.0599</td>
<td>0.0591</td>
<td>0.0479</td>
</tr>
<tr>
<td>LU / NW / OW / UR</td>
<td>0.1590 *</td>
<td>0.0793</td>
<td>0.0283</td>
</tr>
<tr>
<td>NE</td>
<td>-0.0106</td>
<td>0.1024</td>
<td>0.0523</td>
</tr>
<tr>
<td>SG / AI / SH / TG</td>
<td>-0.0090</td>
<td>0.0496</td>
<td>0.0588</td>
</tr>
<tr>
<td>TI</td>
<td>-0.1800 *</td>
<td>0.0542</td>
<td>0.1198</td>
</tr>
<tr>
<td>VD</td>
<td>0.0076</td>
<td>0.0461</td>
<td>0.1285</td>
</tr>
<tr>
<td>VS</td>
<td>-0.0089</td>
<td>0.0849</td>
<td>0.0523</td>
</tr>
<tr>
<td>ZG / SZ</td>
<td>0.2322 *</td>
<td>0.0603</td>
<td>0.0523</td>
</tr>
<tr>
<td>ZH</td>
<td>0.1285 *</td>
<td>0.0557</td>
<td>0.0850</td>
</tr>
<tr>
<td>BE</td>
<td>0</td>
<td>-</td>
<td>0.1329</td>
</tr>
</tbody>
</table>

- The omitted canton is Bern (BE).
References


SUMMARY

An Analysis of Efficiency and Productivity in Swiss Hospitals

This paper examines the productive efficiency of the hospital sector in Switzerland. A stochastic total cost frontier is estimated for a sample of 459 observations from 156 general hospitals between 1998 and 2001. Given the limited number of periods, a pooled cross-sectional model has been adopted. The severity of hospital patient mix is considered using the DRG cost weights. The analysis suggests a significant potential for improving efficiency. The results also point to unexploited scale economies in the majority of the studied hospitals. An analysis of efficiency estimates indicates that the differences among various ownership/subsidization types are not statistically significant.

JEL classification: I180; I120; L330; L250
Keywords: Stochastic frontier; Cost efficiency; Scale economies; General hospitals

ZUSAMMENFASSUNG

Effizienz- und Produktivitätsanalyse für Schweizer Krankenhäuser


JEL Klassifikation: I180; I120; L330; L250
Stichwörter: Stochastic frontier; Kosteneffizienz; Grössenvorteile; Krankenhäuser der Grundversorgung

RÉSUMÉ

Analyse de l’efficience et de la productivité des hôpitaux suisses

Cet article étudie l'efficience productive du secteur hospitalier en Suisse. L’on estime une frontière stochastique de coût total pour un échantillon de 459 observations de 156 hôpitaux de soins généraux entre 1998 et 2001. Etant donné le nombre limité de périodes, on a adopté un modèle cross-section. La sévérité du mélange de patient est considérée avec les poids relatifs de coûts de chaque DRG. L’analyse suggère un potentiel significatif pour améliorer l'efficience. Les résultats montrent également l’existence d’économies d’échelle inexploitées dans la majorité des hôpitaux étudiés. Une analyse des efficiences estimées indique que les différences entre divers types de propriété/subvention ne sont pas statistiquement significatives.

Classification JEL: I180; I120; L330; L250
Mots clefs: Frontière stochastique; Efficience de coût; Economies d’échelle; Hôpitaux généraux
ESTIMATING THE OUT-OF-HOSPITAL MORTALITY RATE USING PATIENT DISCHARGE DATA

Mehdi Farsi† ‡Geert Ridder† ‡

† Department of Management, Technology and Economics
ETH Zurich, Zurichbergstr. 18, Zurich 8032, Switzerland
‡ Department of Economics, University of Southern California
University Park, Los Angeles, CA 90089-0253, USA

ABSTRACT

This paper explores the hospital quality measures based on routine administrative data such as patient discharge records. Most of the measures used in the literature are based on in-hospital mortality risks rather than post-discharge events. The in-hospital outcomes are sensitive to the hospital’s discharge policy, thus could bias the quality estimates. This study aims at identifying out-of-hospital mortality risks and disentangling discharge and re-hospitalization rates from mortality rates using patient discharge data. It is shown that these objectives can be achieved without post-discharge death records. This is an example of the use of public use administrative data for estimating empirical relations when key dependent variables are not available. Using data on the lengths of hospitalizations and out-of-hospital spells, the mortality rates before and after discharge are estimated for a sample of heart-attack patients hospitalized in California between 1992 and 1998. The results suggest that the quality assessments that ignore the variation of discharge rates among hospitals could be misleading.

Keywords: Mortality; Hospital quality; Duration models; Survival analysis

Correspondence: Mehdi Farsi, ETH Zurich, Zurichbergstr. 18, 8032 Zurich, Switzerland.
Tel: +41 44 632 0656 Fax: +41 44 632 1050 Email: mfarsi@ethz.ch
1. Introduction

Assessing quality of health service providers is an important policy issue that has been subject of a great deal of research. Yet, measuring quality from routinely collected data is a challenging question for econometricians. In-hospital mortality has been widely used as a measure of quality of medical care. However, a major concern is that the in-hospital death outcomes do not necessarily reflect the long-term effects. Due to differences in discharge/transfer policies across hospitals the in-hospital mortality could be a biased measure, overstating the quality of hospitals with a relatively high discharge rate especially if low-quality hospitals discharge their patients prematurely or transfer their most severe cases to better hospitals.

In his study of ownership conversion, Sloan [1] reports that while the in-hospital mortality is not affected by conversions, the longer-term mortality probability has increased as hospitals converted to for-profit status. Those findings suggest that hospitals with shorter stays may have higher mortality rates after discharge. A complete measure of hospital-specific mortality risk should therefore include the post-discharge mortality risks. However longitudinal surveys that follow patients after discharge are expensive, hardly available or limited to certain groups of patients. Alternatively, external sources such as expert evaluations or other independent measures can be used to validate the quality measures based on in-hospital mortality risks [2]. Such validation studies are however very expensive. On the other hand, hospital administrative records such as Patient Discharge Data (PDD) provide a large number of observations at relatively low costs. Administrative data are often made available to researchers in public use files, which usually cannot be linked to external data such as death records. For instance in the PDD public use files, the patient’s identification is encrypted with a unique system that allows tracking any given patient
only within the discharge data. Such encrypted identifications allow for instance to identify later re-hospitalizations but not any out-of-hospital event.

Therefore, most of the mortality measures used in the literature are based on in-hospital events. In many cases, the long-term effects have been taken into account by complementary measures. Many studies [3-5] have used the probability of re-hospitalization in the future with or without complication, to account for long-term effects. However, other studies [6,7] have found that readmission risks are related to the patient's clinical conditions rather than hospital quality. Moreover, because of negative correlation between mortality and future readmission [4], the estimations based on readmission usually do not provide additional information on hospital quality. Another approach used in the literature is a censored duration model of in-hospital mortality [8-10]. These models to some extent control for the variations in hospital stays across hospitals, but do not give any information about probabilities of discharge, re-hospitalization and post-discharge mortality.

When the deaths out of hospital are not observed, the statistical inference about the out-of-hospital mortality is complicated. Nevertheless, the hospital discharge records can be used to determine the duration of out-of-hospital spells for all patients. For a fraction of these spells that do not end with a second hospitalization, the death outcome occurs but is not observed. This paper shows that the duration of the out-of-hospital spells can be used to derive information about the long-term survival rates after discharge. Given the importance and availability of PDD, an estimation procedure that can accommodate such an analysis can be very useful.

The data used in this paper are taken from the California Patient Discharge Data Version A. This data set contains the records of all individuals who were hospitalized in California from 1992 through 1998. Unlike the other version of PDD,
Version A provides a patient identification number. The PDD have been used to measure the quality of hospitals based on survival/death probabilities [3,8,11]. As far as we know, this paper is the first that estimates the mortality rate after discharge from the PDD. We show that out-of-hospital mortality rate is identified, even if deaths after discharge are not recorded. We apply the duration model framework to derive the distribution of hospital spells and out-of-hospital spells for this type of data. The model disentangles the discharge and re-hospitalization rates from mortality rates. In addition, using a simplified version of the derived distribution we evaluate the validity of the quality measures commonly used in the literature. It is shown that in most of the quality measures used in the literature the discharge rate is not disentangled from the mortality rate.

Another complication is that in the public use PDD the exact dates of admission have been omitted. Only the year and month in which the hospital stay started and the length of hospitalization in days are retained. Therefore, the derived distribution of out-of-hospital spells cannot be used directly. In this paper, we develop a statistical framework that deals with both problems, namely the censoring of death outcomes and the omission of exact dates. We show that, at the cost of loss of accuracy, the parameters of interest can be identified from the fragmentary data in the public use file. Therefore, this paper provides an example in which the data imperfections can be dealt with econometric modeling.

The plan of the paper is as follows. Section 2 introduces the statistical model. We discuss the identification of the out-of-hospital mortality rate if deaths after discharge are not recorded. We also show that most measures of in-hospital mortality that are commonly used in the literature do not fully separate the discharge outcome from survival. Section 3 provides more information on the Patient Discharge Data.
We also derive the distribution of the spells observed in the PDD. Section 4 contains the estimation results and section 5 concludes the paper with a brief discussion of the main results.

2. Identifying the out-of-hospital mortality rate

In this section we abstract from the problems created by the omission of the exact admission dates in the public use file. This shortcoming of the data will be discussed in section 3. Here, we address the problem of unobserved out-of-hospital death outcomes. We assume that for a member of the population the complete hospitalization history during the observation period \([0,T]\) is observed. A complete hospitalization history consists of a sequence of hospital stays and spells outside hospital (Figure 1) or equivalently, of a sequence of transitions between two states: hospitalized \((H)\) and discharged \((D)\). A hospital spell ends if the patient is discharged or if she dies. An out-of-hospital spell ends if the patient is admitted or if she dies. Death is thus considered as a transition to a third absorbing state.

Figure 1. Hospitalization record
Basically, if the out-of-hospital deaths were known, the problem would reduce to a three-state duration model similar to those models used in modeling unemployment and labor participation [12-14]. In line with this literature we use a proportional hazard framework. In our case however, the time of death is observed only if the patient dies in a hospital. The problem is to estimate the transition rates and in particular, the out-of-hospital mortality rate from the observed hospitalization records. The methodology used here is very similar to the approach used in ‘capture-recapture’ models for estimating demographic parameters of wildlife populations. Pollock [15] provides a survey of these models.

The hospitalization record has multiple time scales: the observation times (0 is the start of the observation period), the duration of hospital or out-of-hospital spells (0 is the start of the spell), the time since the onset of the disease, calendar time, and age. In the sequel both observation and duration time are used. It is clear from the context which time scale is used.

2.1. The in-hospital mortality and discharge rates

A hospital spell is denoted by $t_H$. As shown in Figure 1, a hospital spell ends with the death of the patient with intensity $\mu_H(t)$ or with the discharge of the patient with intensity $\lambda_D(t)$. A hospital spell could also end with the transfer of the patient to another hospital. This could be considered by introducing a transfer intensity. Here, only one hospital spell is considered and the patients who have been transferred to another hospital are excluded from the sample. In fact, as the estimated mortality rates are usually used as a hospital quality measure, it is difficult to distinguish the contribution of each one of the hospitals in survival rates. In some administrative records, transfers are not distinguished from other discharges. In such cases, $\lambda_D(t)$ can
be considered as a weighted average (with weights depending on the hospital spell) of the discharge and transfer densities.

Let $D_H$ be 1 if the spell ends with discharge and 0 if it ends with the death of the patient. The joint distribution of $t_H, D_H$ has the following pdf:

$$f_{H}(t, D_H) = e^{-M_H(t) - A_H(t)} \lambda_D(t)^{D_H} \mu_H(t)^{1-D_H}$$  \hspace{1cm} (1),

with $M_H(t) = \int_0^t \mu_H(s)ds$ and $A_H(t) = \int_0^t \lambda_D(s)ds$. $\mu_H$ and $\lambda_D$ are assumed to be piecewise constant over $k$ intervals $0 = t_0 < t_1 < ... < t_{k-1} < t_k = t_{\text{max}}$ where $t_{\text{max}}$ is the longest hospital stay. $\mu_H$ and $\lambda_D$ are also functions of covariates like patient and hospital characteristics. The covariates are assumed to be constant over time. If $X$ is the vector of covariates, the hazard functions can be written as:

$$\mu_H(t) = \exp(X \beta) \sum_{i=1}^{k} \mu_H^i I_{t_{i-1} < t \leq t_i}$$ \hspace{1cm} (2)

$$\lambda_D(t) = \exp(X \gamma) \sum_{i=1}^{k} \lambda_D^i I_{t_{i-1} < t \leq t_i}$$ \hspace{1cm} (3),

where $I(A)$ is the indicator function taking 1 if condition $A$ is satisfied and zero otherwise; $\mu_H^i$ and $\lambda_D^i$ are constants corresponding to interval $(t_{i-1}, t_i]$ with $\mu_H^0 = \lambda_D^0 = 1$; and $\gamma$ and $\beta$ are the vectors of coefficients corresponding to the independent variables. The pdf given in (1) is the basis for the likelihood function for the in-hospital spells.

2.2. **The out-of-hospital mortality and hospitalization rates**

For the identification of the out-of-hospital mortality rate the spell spent outside hospital denoted by $t_D$ is considered (Figure 1). This spell starts at the time of
discharge from the hospital. It ends if the patient returns to the hospital (not necessarily the same hospital) or if she dies. However, the death is not observed. Let \( \lambda_H \) denote the hospitalization rate and \( \mu_D \) the mortality rate outside hospital. These rates may depend on the time since the last hospitalization \( t \). For ease of exposition it is assumed that this spell starts at time 0 and that it is censored at time \( T \). The distribution of \( t_D \) is mixed discrete-continuous with a positive probability that \( t_D \leq T \).

To show this consider for \( t \leq T \):

\[
\Pr(t_D > t) = e^{-\Lambda_H(t) - \mu_D(t)} + \int_0^t \mu_D(s)e^{-\Lambda_H(s) - \mu_D(s)} \, ds \tag{4},
\]

with \( \Lambda_H(t) = \int_0^t \lambda_H(s) \, ds \) and \( M_D(t) = \int_0^t \mu_D(s) \, ds \). The first term on the right-hand-side is the probability that by \( t \) neither a death nor a hospitalization has occurred. The second term is the probability that during \([0, t]\) the individual has died. In this case since deaths outside hospitals are not observed, the observed spell is still in progress. In fact for all patients who die before re-hospitalization the observed spell \( t_D \) is of infinite length. This means that the distribution of \( t_D \) is defective and the probability of observing an infinite spell is the average of the probability of death before hospitalization, where the average is computed over the duration of the latent out-of-hospital spell, that is:

\[
\int_0^T \mu_D(s)e^{-\Lambda_H(s) - \mu_D(s)} \, ds.
\]

If the observation period is finite, \( t_D \) is observed if \( t_D \leq T \). Otherwise the event \( t_D > T \) is observed. Define \( D_D \) as the indicator of the event \( t_D \leq T \). The probability density of \( t_D \) given \( D_D = 1 \) is written as:

\[
f(t \mid D_D = 1) = \frac{\lambda_H(t)e^{-\Lambda_H(t) - \mu_D(t)}}{\int_0^T \lambda_H(s)e^{-\Lambda_H(s) - \mu_D(s)} \, ds}, \quad t \leq T \tag{5}
\]
Moreover:

\[
\Pr(D_D = 0) = \Pr(t_D > T) = e^{-\lambda_H(T) - \mu_D(T)} + \int_0^T \mu_D(s)e^{-\lambda_H(s) - \mu_D(s)} ds
\quad (6)
\]

The pdf given in (5) and the probability in (6) are the basis of the likelihood estimation of the out-of-hospital mortality and re-hospitalization rates. \(\mu_D\) and \(\lambda_H\) are assumed to be piecewise constant over \(k'\) intervals \(0 = T_0 < T_1 < ... < T_{k'-1} < T'\).

Similarly, the constant effects of covariates \((X)\) are included in a proportional hazard framework, resulting in the following hazard functions:

\[
\mu_D(t) = \exp(X\eta)\sum_{i=1}^{k'} \mu_D^i 1_{(T_{i-1}, T_i]}(t)
\quad (7),
\]

\[
\lambda_H(t) = \exp(X\zeta)\sum_{i=1}^{k'} \lambda_H^i 1_{(T_{i-1}, T_i]}(t)
\quad (8),
\]

where \(\mu_D^i\) and \(\lambda_H^i\) are constant rates corresponding to interval \((T_{i-1}, T_i]\) with \(\mu_D^0 = \lambda_H^0 = 1\), and \(\eta\) and \(\zeta\) are the vectors of coefficients.

To show that both hospitalization and mortality rates are identified consider first the special case where both rates are constant over time. In this case the conditional pdf (5) reduces to:

\[
f(t \mid D_D = 1) = \frac{(\lambda_H + \mu_D) e^{-(\lambda_H + \mu_D)T}}{1 - e^{-(\lambda_H + \mu_D)T}} , t \leq T
\quad (9),
\]

which is the pdf of a truncated (at \(T\)) exponential distribution with parameter \(\kappa=\mu_D + \lambda_H\). Hence, from the distribution of spells that end in hospitalization the sum of mortality and hospitalization rates is identified. Moreover, the probability of re-hospitalization before \(T\) is:

\[
\Pr(D_D = 1) = \frac{\lambda_H}{\lambda_H + \mu_D} (1 - e^{-(\lambda_H + \mu_D)T})
\quad (10)
\]
Since $\kappa = \mu_D + \lambda_H$ is identified from (9), $\lambda_H$ is identified using the probability given in (10). The joint distribution of $t_D, D_D$ has the following pdf:

$$f_D(t, D_D) = \left( \lambda_H e^{-\left(\lambda_H + \mu_D\right) t} \right)^{D_D} \left( \frac{\mu_D}{\lambda_H + \mu_D} + \frac{\lambda_H e^{-\left(\lambda_H + \mu_D\right) T}}{\lambda_H + \mu_D} \right)^{1-D_D}$$

(11)

The above argument can be extended to piecewise constant rates $\mu_D(t)$ and $\lambda_H(t)$. It suffices to first censor the out-of-hospital spells at $T_1$ (the first interval). The rates are constant over the interval thus identified using the censored spells. The spells that end with a hospitalization in the interval identify the sum of mortality and hospitalization rates and the fraction of spells that are censored identify the rates separately. Next, consider the out-of-hospital spells that end with a hospitalization in the second interval $(T_1, T_2]$. It can be shown that the distribution of these spells is such that $t_D - T_1$ has a truncated (at $T_2 - T_1$) exponential distribution with a parameter that is the sum of mortality and hospitalization rates on the second interval. Hence this distribution identifies the sum. The hospitalization and mortality rates are identified from the fraction of spells that are censored at $T_2$. This argument can be repeated for the remaining intervals.

### 2.3. Measures used in the literature

In this section the quality measures used in the literature are discussed using the proposed model. The measures can be divided into three categories: in-hospital mortality outcome, mortality outcome within a given period after admission, and readmission within a given period after discharge. For ease of exposition it is assumed that all transition rates are constant.
A number of papers [2,3,16] used the mortality outcome at discharge. This measure can be written as a function of in-hospital mortality and discharge rates:

$$\Pr(D_H = 0) = \frac{\mu_H}{\lambda_D + \mu_H}$$  \hspace{1cm} (12)

It can be shown that the in-hospital death probability is increasing in $\mu_H$ and decreasing in $\lambda_D$. To the extent that discharge practices differ across hospitals, this measure cannot be used as a hospital-specific mortality. An alternative used by Geweke et al. [11] is the in-hospital death probability within 10 days after admission. This in-hospital mortality probability within period $t$ after admission can be written as:

$$\Pr(D_H = 0, t_H \leq t) = \frac{\mu_H}{\lambda_D + \mu_H} (1 - e^{-(\lambda_D + \mu_H)t})$$  \hspace{1cm} (13)

In this case depending on the chosen value of $t$, the death probability can be decreasing or increasing in $\mu_H$. Therefore, even assuming a constant discharge rate across hospitals, this cannot be used as a measure of hospital-specific mortality.

Another commonly used measure is the death probability within a given period after admission. These deaths may occur inside hospitals or after discharge. Some studies [17-21] have used mortality within 30 days while others [4,22,23] used longer periods up to one year. This measure may seem appealing because it can represent a relatively long-term outcome that is seemingly independent of discharge rates.

The probability of death within $t$ days after admission can be written as the sum of the probabilities of the in-hospital and post-discharge death before $t$, that is:

$$\int_0^t \mu_H e^{-(\lambda_D + \mu_H)s} ds + \int_0^t \lambda_D e^{-(\lambda_D + \mu_H)s} \mu_H e^{-(\lambda_H + \mu_D)(t-s)} ds$$
It is easy to show that such measures are affected by discharge and hospitalization rates.

Another measure of quality is the re-hospitalization probability within a given period after discharge. Various authors [3,13,18,23] have considered different periods usually varying between 14 days to a few months. The readmission probability within \( t \) days after discharge can be written as:

\[
\Pr(t_D \leq t) = \frac{\lambda_H}{\lambda_H + \mu_D} (1 - e^{-(\lambda_H + \mu_D)t})
\]  

(14)

The problem with this measure is that for short readmission periods (small \( t \)), it is not increasing in \( \lambda_H \), and for relatively large periods the correlation between readmission risk and hospital quality is low [7]. Moreover, as it can be seen this measure depends on the out-of-hospital mortality rate. In fact, for short periods (small \( t \)) this measure is a decreasing function of \( \mu_D \). In cases where the out-of-hospital mortality is not observed, small rates of re-hospitalization may be associated with high mortality rates, hence not necessarily a higher hospital quality. The above problems provide an explanation to why the readmission measures of quality as used in the literature are inconsistent with other measures of hospital quality [6,7].

Ettner and Hermann [24] used the readmission within 30 days after discharge for psychiatric patients. Given that mortality rates are quite low for these patients, the re-admission measure may be appropriate. Assuming that \( \mu_D \) is close to zero, the probability given in (14) can be simplified as: \( (1 - e^{-\lambda_H t}) \), which is a non-decreasing function of \( \lambda_H \), and therefore can be used as a proxy for re-hospitalization rate.

3. The patient discharge data

3.1. Description of the data
The data used in this paper are extracted from California Hospital Discharge Data. The population considered in this paper are all individuals of 65 years of age or older who were hospitalized during 1992-1998 with Acute Myocardial Infarction (AMI) as their principal diagnosis and who were in an initial episode of treatment. This data set has been merged with data from California Hospital Disclosure Data on hospital characteristics such as size and ownership status. A detailed description of these data has been given elsewhere [25,26].

From the original sample including about 173,000 hospitalizations of 163,000 patients, we excluded the patients older than 95 years old and those who have been transferred from (or to) other hospitals leaving about 132,000 patients. The transferred cases have been excluded mainly because their survival probabilities cannot be related to a single hospital and distinguishing each hospital’s contribution is difficult, if at all possible. To further simplify the analysis we also excluded all the patients (less than 3 percent of the sample) who had multiple hospitalizations in their first admission month or whose first hospital spell was longer than a month. Moreover, since one of the parameters of interest is the effect of ownership status on hospital quality, in order to avoid the reporting errors of ownership changes and their complicated effects in quality [27], we focused our analysis to hospitals that had a stable ownership status over the sample period. The final sample consists of 115,805 AMI patients hospitalized in 387 California hospitals.

AMI is an acute condition and these patients are less subject to selection problems. Systematic selection of patients to specific hospitals may bias the estimates of hospital characteristics on mortality rates. Heart attack patients are likely to go the closest hospital. Moreover, a considerable part of deaths caused by AMI occur inside hospitals. The elderly age group is chosen because all these patients benefit from
Medicare and are less likely to be rejected by hospitals. The identification of post-discharge mortality relies on the assumption that an out-of-hospital spell ends in re-hospitalization or death. A third possibility is that the patient leaves the state of California. The migration is less likely for the elderly patients with an acute condition.

Using the patient identification numbers that are encrypted unique numbers, the patients in the sample have been linked to another data set including all the hospitalizations in California (for any reason) over the sample period. The latter data set including about 10 million patients has been extracted from the PDD files. For each patient in the sample the total lengths of hospitalizations in the first month and in the re-admission month were calculated. For each patient, the first month is the month in which her initial hospitalization for AMI has occurred. The second and later hospitalizations need not be for AMI and can be for any condition.

3.2. Implementation of the model

The estimation of hospital spells is straightforward and the joint distribution of $t_H, D_H$ with piecewise constant rates can be directly derived from equation (1) using (2) and (3). A complete derivation of the joint distribution and the likelihood function is provided in the appendix. For the out-of-hospital spells, because of the limitations of the data, the exact length of spell is not known. Instead, we derive bounds on the out-of-hospital spells that correspond to these data, and we use these interval data in our estimation. The data consist of a sequence of hospital spells together with the month in which each of these began. A typical realization for a given patient is illustrated in Figure 2. Suppose that the months in the sample period (1992-1998) are respectively numbered from 1 to $M$. Let $m_1$ denote the number of month in which the patient was first hospitalized for AMI, and $m_2$ the number of month in which she was
re-hospitalized (for any reason) after the initial discharge. Note that \( m_1 \) and \( m_2 \) have patient-specific values. Note also that a patient can have multiple hospitalizations in a given month, but the first AMI admission is uniquely identified for all patients in the sample.

Figure 2. In-hospital spells

For the spells that do not end in re-hospitalization the contribution to the likelihood function is given by (6) using expressions (7) and (8). The end of observation period (\( T \)) used in (6) is a patient-specific variable. \( T \) in days is given by:

\[
T = 30.5(M - m_1) - t_{H_0}
\]  

(15),

where \( t_{H_0} \) is the length of the initial hospitalization.

As for the cases that end in a re-hospitalization, the out-of-hospital spell \( t_D \) can be specified with the following lower and upper bounds:

\[
t_{D}^{\text{inf}} = 30.5(m_2 - m_1) - t_{H_0}
\]

(16),

\[
t_{D}^{\text{sup}} = 30.5(m_2 + 1 - m_1) - t_{H_0} - \left( \sum_{i=1}^{k} t_{H_i} - \max \{ t_{H_i}; i = 1, \ldots, k \} \right)
\]

(17),

where \( k \) is the number of hospital spells that started in the re-hospitalization month \( m_2 \), and \( t_{H_i} (i=1, 2, \ldots, k) \) is the length (in days) of these hospital spells. Note that the
above definitions can be readily extended to cases with multiple admissions in the first month, in which case $t_{HO}$ must be set equal to the first month’s longest hospitalization in (15) and (16), and the upper bound (17) must be reduced by the sum of the remaining hospital spells of that month.

The spells that end in re-hospitalization make the following contribution to the likelihood function:

$$\text{Pr}(t_D^\inf < t_D < t_D^\sup) = \int_{t_D^\inf}^{t_D^\sup} \hat{\lambda}_H(s)e^{-\lambda_D(s)-M_D(s)} ds$$

(18),

where the integrals $\Lambda_H(t)$ and $M_D(t)$ are obtained using the expressions in (7) and (8).

A complete derivation of the likelihood function is provided in the appendix.

4. Estimation results

The data on the first reported hospital spell are used to estimate the in-hospital mortality rate and the discharge rate. The in-hospital sample includes the entire sample of 115,805 elderly patients, hospitalized for an initial episode of AMI. The summary statistics are given in Table 1. The average hospital spell is about 6.4 days and about 17% of the spells end with the death of the patient. For 94,842 patients from this sample the out-of-hospital spells are calculated. Note that the patients who died in hospital are excluded from the out-of-hospital sample. About 65% of these patients were re-admitted after their first hospitalization and before the end of the observation period. For these patients the lower bound of out-of-hospital spells varies from 0 to about 2,465 days (with an average of 264 days) and the upper bound varies between 7 and 2,526 days (with an average of 321 days). Table 2 gives the summary statistics for out-of-hospital spells.
Table 1. Sample statistics for hospital spells (N=115,805)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital stay (days)</td>
<td>6.356</td>
<td>4.378</td>
</tr>
<tr>
<td>Discharged alive</td>
<td>0.828</td>
<td>0.377</td>
</tr>
<tr>
<td>For-Profit hospital</td>
<td>0.131</td>
<td>0.338</td>
</tr>
<tr>
<td>Public hospital</td>
<td>0.111</td>
<td>0.314</td>
</tr>
<tr>
<td>Number of beds /1000</td>
<td>0.288</td>
<td>0.167</td>
</tr>
<tr>
<td>Male</td>
<td>0.530</td>
<td>0.499</td>
</tr>
<tr>
<td>Black</td>
<td>0.048</td>
<td>0.214</td>
</tr>
<tr>
<td>Age 70-74</td>
<td>0.221</td>
<td>0.415</td>
</tr>
<tr>
<td>Age 75-79</td>
<td>0.214</td>
<td>0.410</td>
</tr>
<tr>
<td>Age 80-84</td>
<td>0.191</td>
<td>0.393</td>
</tr>
<tr>
<td>Age 85 +</td>
<td>0.184</td>
<td>0.387</td>
</tr>
<tr>
<td>Moderate severity</td>
<td>0.380</td>
<td>0.485</td>
</tr>
<tr>
<td>Major severity</td>
<td>0.300</td>
<td>0.458</td>
</tr>
<tr>
<td>Extreme severity</td>
<td>0.201</td>
<td>0.401</td>
</tr>
<tr>
<td>Year 1993</td>
<td>0.144</td>
<td>0.351</td>
</tr>
<tr>
<td>Year 1994</td>
<td>0.141</td>
<td>0.348</td>
</tr>
<tr>
<td>Year 1995</td>
<td>0.141</td>
<td>0.348</td>
</tr>
<tr>
<td>Year 1996</td>
<td>0.141</td>
<td>0.348</td>
</tr>
<tr>
<td>Year 1997</td>
<td>0.146</td>
<td>0.353</td>
</tr>
<tr>
<td>Year 1998</td>
<td>0.146</td>
<td>0.353</td>
</tr>
</tbody>
</table>
Table 2. Sample statistics for out-of-hospital spells (N=94,842)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Re-hospitalized before the end of observation period</td>
<td>0.647</td>
<td>0.478</td>
</tr>
<tr>
<td>Lower bound of out-of-hospital spell (days)</td>
<td>264.112</td>
<td>398.564</td>
</tr>
<tr>
<td>Upper bound of out-of-hospital spell (days)</td>
<td>321.340</td>
<td>400.398</td>
</tr>
<tr>
<td>Spell until the end of observation period (days)</td>
<td>1275.544</td>
<td>735.344</td>
</tr>
<tr>
<td>For-Profit hospital</td>
<td>0.126</td>
<td>0.331</td>
</tr>
<tr>
<td>Public hospital</td>
<td>0.109</td>
<td>0.312</td>
</tr>
<tr>
<td>Number of beds /1000</td>
<td>0.292</td>
<td>0.168</td>
</tr>
<tr>
<td>Male</td>
<td>0.538</td>
<td>0.499</td>
</tr>
<tr>
<td>Black</td>
<td>0.049</td>
<td>0.216</td>
</tr>
<tr>
<td>Age 70-74</td>
<td>0.230</td>
<td>0.421</td>
</tr>
<tr>
<td>Age 75-79</td>
<td>0.215</td>
<td>0.411</td>
</tr>
<tr>
<td>Age 80-84</td>
<td>0.182</td>
<td>0.386</td>
</tr>
<tr>
<td>Age 85 +</td>
<td>0.169</td>
<td>0.375</td>
</tr>
<tr>
<td>Moderate severity</td>
<td>0.435</td>
<td>0.496</td>
</tr>
<tr>
<td>Major severity</td>
<td>0.293</td>
<td>0.455</td>
</tr>
<tr>
<td>Extreme severity</td>
<td>0.131</td>
<td>0.338</td>
</tr>
<tr>
<td>Year 1993</td>
<td>0.144</td>
<td>0.351</td>
</tr>
<tr>
<td>Year 1994</td>
<td>0.142</td>
<td>0.349</td>
</tr>
<tr>
<td>Year 1995</td>
<td>0.142</td>
<td>0.349</td>
</tr>
<tr>
<td>Year 1996</td>
<td>0.144</td>
<td>0.351</td>
</tr>
<tr>
<td>Year 1997</td>
<td>0.150</td>
<td>0.357</td>
</tr>
<tr>
<td>Year 1998</td>
<td>0.138</td>
<td>0.345</td>
</tr>
</tbody>
</table>

Discharge and in-hospital mortality rates are assumed to be piecewise constant over 5 intervals: 0 to 2 days, 2 to 4, 4 to 6, 6 to 10, and more than 10 days. Table 3 gives a summary of the regression results for the hospital spells. For each listed variable the estimated coefficients represent the variable’s marginal effects on the
hazard rates of discharge and in-hospital mortality respectively. For instance, the results suggest that in-hospital mortality hazard rate in For-Profit (FP) hospitals is on average 8% higher than in Non-Profit (NP) hospitals (the omitted category). FP hospitals also show a 5% lower discharge hazard rate compared to NP hospitals. The results also indicate that compared to the base category (Non-Profit hospitals), public hospitals have higher rates in both mortality and discharge (by about 5%). Hospital size has a significant effect on both mortality and discharge rates with large hospitals having lower rates.

As expected, both severity and age have a positive effect on mortality. The discharge rate is negatively affected by severity and age, but the age effects on discharge are not uniform. This could be explained by the fact that very old patients might get discharged to nursing homes or long-term care centers. The calendar year effects indicate that there is no significant trend in the mortality rate, but there is a strong upward trend in the discharge rate suggesting a general tendency toward shorter hospitalizations. The significant changes in transition rates over the intervals show that the rates are time-variant. For instance the mortality rate in the first two days of the spell is significantly higher than in the rest of hospitalization. This result has an important health policy implication pointing to the crucial importance of the immediate stabilization of AMI patients.

The significant variation of discharge rates across hospitals with different ownership status supports the concern that lower in-hospital mortality rates may be associated to higher discharge rates. For instance, the results suggest that part of the difference in mortality between FP and NP hospitals could be associated with different discharge rates across the two hospital types. Therefore, the in-hospital mortality rate does not give a complete picture regarding hospital quality. On the
other hand relatively high discharge rates in the NP hospitals do not represent a lower quality in itself, as long as it does not lead to higher chances of post-discharge mortality.

### Table 3. Mortality and discharge rates for hospital spells

<table>
<thead>
<tr>
<th></th>
<th>Discharge rate</th>
<th>Mortality rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE Standard error</td>
<td>MLE Standard error</td>
</tr>
<tr>
<td>For-Profit hospital</td>
<td>-0.051* 0.010</td>
<td>0.080* 0.020</td>
</tr>
<tr>
<td>Public hospital</td>
<td>0.044* 0.010</td>
<td>0.050* 0.022</td>
</tr>
<tr>
<td>Number of beds /1000</td>
<td>-0.490* 0.020</td>
<td>-1.089* 0.049</td>
</tr>
<tr>
<td>Male</td>
<td>0.025* 0.007</td>
<td>-0.041* 0.014</td>
</tr>
<tr>
<td>Black</td>
<td>0.023 0.015</td>
<td>-0.123* 0.035</td>
</tr>
<tr>
<td>Age 70-74</td>
<td>-0.026* 0.010</td>
<td>0.148* 0.026</td>
</tr>
<tr>
<td>Age 75-79</td>
<td>-0.042* 0.010</td>
<td>0.257* 0.026</td>
</tr>
<tr>
<td>Age 80-84</td>
<td>-0.029* 0.011</td>
<td>0.484* 0.025</td>
</tr>
<tr>
<td>Age 85+</td>
<td>0.014 0.011</td>
<td>0.602* 0.025</td>
</tr>
<tr>
<td>Moderate severity</td>
<td>-0.355* 0.010</td>
<td>0.752* 0.064</td>
</tr>
<tr>
<td>Major severity</td>
<td>-0.908* 0.011</td>
<td>2.018* 0.062</td>
</tr>
<tr>
<td>Extreme severity</td>
<td>-1.733* 0.013</td>
<td>2.806* 0.061</td>
</tr>
<tr>
<td>Year 1993</td>
<td>0.111* 0.012</td>
<td>-0.045 0.025</td>
</tr>
<tr>
<td>Year 1994</td>
<td>0.241* 0.012</td>
<td>-0.044 0.026</td>
</tr>
<tr>
<td>Year 1995</td>
<td>0.327* 0.012</td>
<td>-0.017 0.026</td>
</tr>
<tr>
<td>Year 1996</td>
<td>0.421* 0.012</td>
<td>-0.092* 0.026</td>
</tr>
<tr>
<td>Year 1997</td>
<td>0.466* 0.012</td>
<td>-0.159* 0.026</td>
</tr>
<tr>
<td>Year 1998</td>
<td>0.462* 0.012</td>
<td>-0.060* 0.026</td>
</tr>
<tr>
<td>Interval 2 to 4 days</td>
<td>1.554* 0.015</td>
<td>-0.974* 0.020</td>
</tr>
<tr>
<td>Interval 4 to 6 days</td>
<td>2.220* 0.015</td>
<td>-1.273* 0.024</td>
</tr>
<tr>
<td>Interval 6 to 10 days</td>
<td>2.431* 0.015</td>
<td>-1.411* 0.025</td>
</tr>
<tr>
<td>More than 10 days</td>
<td>2.570* 0.016</td>
<td>-1.248* 0.025</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.220* 0.020</td>
<td>-4.833* 0.067</td>
</tr>
</tbody>
</table>

* Significant at 5%.

The estimation results for the out-of-hospital mortality and re-hospitalization rates are given in Table 4. Transition rates are assumed to be constant. This table does not show any significant effect of hospital ownership on the mortality and re-hospitalization rates. The hospital size shows a significant and negative effect on both re-hospitalization and mortality. Combining these results with those of Table 3, one
could conclude that mortality rates are on average lower in larger hospitals. As expected severity and age has a positive effect on both re-hospitalization and mortality rates. The results also indicate a significant growth in both hospitalization and mortality rates. These findings along with those of Table 3 suggest that over time, hospital stays have become on average shorter resulting in lower in-hospital mortality rates but higher re-hospitalization rates and higher out-of-hospital mortality probability.

Another interesting example that can be used to highlight the contribution of the proposed model is the analysis of quality differences between NP and FP hospitals, based on mortality outcomes. A few empirical studies on US hospitals suggest relatively high AMI mortality rates for FP hospitals [4,28]. This result is similar to our results suggesting relatively high re-hospitalization rate and in-hospital mortality for FP hospitals. However, our findings also show that a major part of the FP hospitals’ excess in-hospital death rate might be related to lower discharge rates. Moreover, although FP hospitals show slightly (but not significantly) higher re-admission rates, their out-of-hospital mortality risks are similar to NP hospitals. Finally, the relatively high discharge rate in NP hospitals (Table 3) is not associated with higher probability of out-of-hospital death or re-admission for these hospitals. Therefore, the results suggest that though being different in discharge and in-hospital mortality, NP and FP hospitals do not show any quality difference in this regard. However, public hospitals that indicate relatively high in-hospital mortality and discharge rates also have slightly (but not significantly) higher post-discharge death probability. This might be interpreted as a relatively low quality of care in these hospitals.
A shortcoming of the out-of-hospital analysis is that the transition rates are assumed to be constant. A version of the model with piece-wise constant rates with one cut-off point has been applied to a similar data set [25]. The high estimation errors of the possible changes in hazard rates in that analysis indicate that with the available data such an extension does not provide any significant improvement over the constant-rate model used in this study. This can be explained by the fact that with the available data we cannot calculate the exact duration of out-of-hospital spells.

Table 4. Mortality and re-hospitalization rates for out-of-hospital spells

<table>
<thead>
<tr>
<th></th>
<th>Re-hospitalization rate</th>
<th>Mortality rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>Standard error</td>
</tr>
<tr>
<td>For-Profit hospital</td>
<td>0.021</td>
<td>0.015</td>
</tr>
<tr>
<td>Public hospital</td>
<td>-0.001</td>
<td>0.016</td>
</tr>
<tr>
<td>Number of beds /1000</td>
<td>-0.078*</td>
<td>0.030</td>
</tr>
<tr>
<td>Male</td>
<td>-0.036*</td>
<td>0.010</td>
</tr>
<tr>
<td>Black</td>
<td>0.128*</td>
<td>0.022</td>
</tr>
<tr>
<td>Age 70-74</td>
<td>0.071*</td>
<td>0.015</td>
</tr>
<tr>
<td>Age 75-79</td>
<td>0.124*</td>
<td>0.015</td>
</tr>
<tr>
<td>Age 80-84</td>
<td>0.170*</td>
<td>0.016</td>
</tr>
<tr>
<td>Age 85+</td>
<td>0.204*</td>
<td>0.017</td>
</tr>
<tr>
<td>Moderate severity</td>
<td>0.337*</td>
<td>0.017</td>
</tr>
<tr>
<td>Major severity</td>
<td>0.588*</td>
<td>0.018</td>
</tr>
<tr>
<td>Extreme severity</td>
<td>0.829*</td>
<td>0.020</td>
</tr>
<tr>
<td>Year 1993</td>
<td>0.079*</td>
<td>0.016</td>
</tr>
<tr>
<td>Year 1994</td>
<td>0.165*</td>
<td>0.017</td>
</tr>
<tr>
<td>Year 1995</td>
<td>0.267*</td>
<td>0.017</td>
</tr>
<tr>
<td>Year 1996</td>
<td>0.416*</td>
<td>0.018</td>
</tr>
<tr>
<td>Year 1997</td>
<td>0.655*</td>
<td>0.019</td>
</tr>
<tr>
<td>Year 1998</td>
<td>1.037*</td>
<td>0.023</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.847*</td>
<td>0.023</td>
</tr>
</tbody>
</table>

* Significant at 5%.

5. Conclusions

This paper has explored the measures of hospital quality based on mortality risks estimated from hospital administrative data. These measures are commonly
based on in-hospital death outcomes, which might be affected by hospitals transfer/discharge policies. Using a transition model it has been shown that the out-of-hospital mortality rates can be identified using the patient discharge records without data on post-discharge deaths. This is an example of the use of public administrative data for the estimation of empirical relations when key independent variables are not available in the data. The paper shows that with certain assumptions, the data on the duration of hospitalizations and out-of-hospital spells can be used to estimate the mortality rates before and after discharge as well as discharge and re-hospitalization rates. The analysis is based on an important assumption that patients do not have access to hospitals outside the sample. The common measures of hospital quality based on mortality risks, used in the literature are studied. Most of these measures do not distinguish discharge from survival. Given the significant variation of discharge rates across hospitals, such measures of quality may be misleading.

The model has been applied to a sample of heart-attack patients hospitalized in California general hospitals from 1992 to 1998. The analysis has been performed for in-hospital and out-of-hospital spells separately. The in-hospital analysis indicates a considerable variation in the discharge rate of AMI patients among different hospital types. For instance, a low incidence of in-hospital mortality in a hospital type could be together with a high rate of discharge. Therefore, the use of such mortality outcomes as a measure of hospital performance, without considering the discharge rates could be misleading. In particular the results suggest that the relatively high in-hospital mortality rate in FP hospitals is partly due to their low discharge rate. However, public hospitals in the sample show relatively high rates in both in-hospital mortality and discharge.
As for the out-of-hospital analysis, an important complication of this data set is that the admission dates are identified only up to a month. The estimation procedure has been modified to accommodate this lack of data by writing the likelihood function based on upper and lower bounds rather than the exact length of the out-of-hospital spells. This comes at a loss in efficiency, which could potentially result in relatively high estimation errors in the parameters estimates. The results suggest that hospital ownership does not have a significant effect on out-of-hospital mortality or re-hospitalization rates. However, larger hospitals show on average lower incidences in both rates.

There are a few caveats in the present study, which are left for further research. First, in the epidemiological literature [29-31] additive covariate models are generally preferred over multiplicative forms such as proportional hazard framework used in this paper. The application the proposed model in the additive competing risks framework could be an interesting extension. Secondly, the restriction of piece-wise constant hazard rates could be relaxed by using semi-parametric models. Third, the unobserved heterogeneity can be taken into account by introducing stochastic variation in the model’s parameters. Finally and most importantly, a validation study using data with observed out-of-hospital deaths or a Monte Carlo simulation study should be used to validate the adopted methodology regarding the out-of-hospital mortality rates. Pending such validation studies, the results obtained in this paper cannot be directly used for any policy conclusions. Rather, the adopted methodology underscores the potential use of incomplete data for statistical inference about unobserved events.
Acknowledgements

The authors would like to thank the editor of this journal, two anonymous referees and the participants of the European Workshop on Econometrics and Health Economics, Dublin, 2005, for their helpful suggestions. They also appreciate the discussions on previous versions of the paper during seminars held at RAND, USC and UCLA. M. Farsi also wishes to thank Massimo Filippini, Janet Currie and Bentley MacLeod for their support. The data used in this study have been provided by the California Office of Health Planning and Development, which is gratefully acknowledged.

References


### Appendix: Derivation of the likelihood function

**In-hospital spells:**

Using piecewise integration and plugging relations (2) and (3) into Equation (1), the joint probability distribution corresponding to in-hospital spells can be written as:

\[
 f_H(t, D_H) = \exp\left\{\sum_{i=1}^{k} \left(I_{t \times t'}(t_i - t_{i-1}) \cdot M_i - \sum_{i=1}^{k} I_{t_i < t_{i+1}}(t_i - t_{i-1}) \cdot M_i' \right) \right\} \\
 \times \left[ \exp(X \gamma) \sum_{i=1}^{k} \lambda_D^i I_{t_i < t_{i+1}} \right]^{D_H} \\
 \times \left[ \exp(X \beta) \sum_{i=1}^{k} \mu_H^i I_{t_i < t_{i+1}} \right]^{1-D_H}
\]

where \( M' = \mu_H^i \exp(X \beta) + \lambda_D^i \exp(X \gamma) \).

The log-likelihood function is obtained by the following summation:

\[
 \log L(\beta, \gamma, \mu_H^1, ..., \mu_H^k, \lambda_D^1, ..., \lambda_D^k) = \sum_{n=1}^{N} \log f_H(t^n, D_H^n)
\]
where $N$ is the sample size and superscript $n$ denotes the observation number.

**Out-of-hospital spells:**

The probability that the spell does not end in a re-hospitalization can be obtained from Equation (6) by substituting mortality and re-hospitalization rates respectively from (7) and (8), and using piecewise integration:

$$
\Pr(D_D = 0) = \Pr(t_D > T) = \exp \left( - \sum_{i=1}^{k'} \lambda_{H} \exp(X\eta) \left[ \exp(-K_i t_{i-1}) + \exp(-K_i T_i) \right] \right)
+ \sum_{i=1}^{k'} \frac{\mu_D}{K_i} \exp(X\eta) \left[ \exp(-K_i t_{i-1}) + \exp(-K_i T_i) \right]
\tag{21},
$$

where $K_i = \lambda_{H} \exp(X\zeta) + \mu_D \exp(X\eta)$.

Similarly, the probability related to the spells that end in re-hospitalization can be obtained from Equation (18):

$$
\Pr(t_D^{inf} < t_D^{sup} = \Pr(t_D^{inf} < t_D^{sup}) = \sum_{i=1}^{k'} \lambda_{H} \exp(X\zeta) \left[ \exp(-K_i t_{i-1}) + \exp(-K_i T_i) \right]
+ \sum_{i=1}^{k'} \frac{\lambda_H}{K_i} \exp(X\zeta) \left[ \exp(-K_i t_{i-1}) + \exp(-K_i T_i) \right]
\tag{22},
$$

where $t_D^{inf}$ and $t_D^{sup}$ are respectively given in (16) and (17).

The joint likelihood corresponding to out-of-hospital spells can be written as:

$$
\ell_H(t_D^{inf}, t_D^{sup}, T, D_D) = \left[ \Pr(t_D^{inf} < t_D < t_D^{sup}) \right]^{D_D} \times \left[ \Pr(t_D > T) \right]^{1-D_D}
\tag{23},
$$

where the two probabilities are given in (21) and (22). The log-likelihood function of out-of-hospital spells can thus be written as the following summation:

$$
\log L(\zeta, \eta, \lambda_H, \ldots, \lambda_H^{k'}, \mu_D, \ldots, \mu_D^{k'}) = \sum_{n=1}^{N'} \log \ell_H(t_D^{inf^n}, t_D^{sup^n}, T^n, D_D^n)
\tag{24},
$$

where $N'$ is the sample size for out-of-hospital spells, and superscript $n$ denotes the observation number.
This paper applies a number of stochastic cost frontier models to a panel data set and compares their ability to distinguish unobserved heterogeneity from inefficiency variation among firms. The main focus is on Greene (2005)’s panel data model that incorporates firm-specific effects in a stochastic frontier framework. In cases where the unobserved heterogeneity is correlated with explanatory variables, while the random effects estimators can be biased the fixed effects model may overestimate inefficiency. In line with Mundlak (1978), a simple method is proposed to include such correlations in random effects specification. The sample includes 36 Swiss nursing homes operating from 1993 to 2001. The results suggest that the proposed specification can avoid the inconsistency problem while keeping the inefficiency estimates unaffected.
1. Introduction

Following the work of Aigner, Lovell and Schmidt (1977), stochastic frontier models have been subject of a great body of literature resulting in a large number of econometric models to estimate cost and production functions. Kumbhakar and Lovell (2000) provide an extensive survey of this literature. One of the most important issues in these models is adjusting for the unobserved heterogeneity among firms functioning in different production environments. Individual firms face different external factors that could influence their production costs but are not under their control. These factors may be environmental such as network effects in network industries or related to output characteristics such as the severity of illness in the health sector and the demand fluctuations in electricity utilities. Some of these factors are observed and can be controlled for in the analysis. However, in many cases the data are not available for all these variables. Moreover, the relevant factors are often too complex to be quantified by simple indicators. For instance, factors such as the patient case-mix of a hospital and the network’s shape of an electricity distribution company are hard to measure or require a great deal of information that is not usually available. Both these factors are generally beyond the firms’ control but affect their costs significantly.

A stochastic frontier model by definition includes a random error term that captures the idiosyncratic heterogeneity among different observations. In panel data where an individual firm is observed several times, the firm-specific unobserved variations can also be taken into account through fixed or random effects. This is an important practical advantage because in many cases the relevant environmental factors are location characteristics that vary among firms but are constant over time. For instance the natural obstacles in a railway network such as high slopes or forest
areas, or the average wealth of a community that may affect their health status, thus the operating costs of the neighboring hospital, are generally stable over a relatively long period of time.\footnote{Note that most of the panel data used in the literature cover periods from 5 to 10 years.}

The first use of panel data models in stochastic frontier models goes back to Pitt and Lee \citeyear{PittLee1981} who interpreted the panel data random effects as inefficiency rather than heterogeneity.\footnote{Pitt and Lee \citeyear{PittLee1981}'s model is different from the conventional RE model in that the individual-specific effects are assumed to follow a half-normal distribution. Important variations of this model} This tradition continued with Schmidt and Sickles \citeyear{SchmidtSickles1984} who used a similar interpretation applied to a panel data model with fixed effects. Both models have been extensively used in the literature. A main shortcoming of these models is that any unobserved, time-invariant, firm-specific heterogeneity is considered as inefficiency. In more recent papers the random effects model has been extended to include time-variant inefficiency. Cornwell, Schmidt and Sickles \citeyear{CornwellSchmidtSickles1990} and Battese and Coelli \citeyear{BatteseCoelli1992} are two important contributions in this regard. In particular in one of the models proposed in the former paper, inefficiency is assumed to be a flexible function of time with parameters varying among firms.

A common feature of all these models is that they do not fully separate the sources of heterogeneity and inefficiency at the firm level. Kumbhakar \citeyear{Kumbhakar1997} proposes a model that accommodates firm-specific variances in a heteroscedastic error term. An alternative approach is to consider two separate stochastic terms for efficiency and firm-specific heterogeneity. Theoretically, a stochastic frontier model in its original form \cite{AignerLovellSchmidt1977} can be extended to panel data models, by adding a fixed or random effect in the model. There are a few papers that have explored this possibility. The first development can be attributed to Kumbhakar \citeyear{Kumbhakar1991} who proposed a three-stage estimation procedure to solve the model with
time- and firm-specific effects. Similarly, Heshmati (1998) has used a two-step procedure in a random-effect framework to separate the firm-specific effects from efficiency differences. However, all these papers use a multi-step estimation procedure. Polachek and Yoon (1996) attempted to estimate a panel data frontier model with firm dummies using a one-step procedure. Greene (2002a) discussed the numerical obstacles that have apparently delayed such a development. He proposed numerical solutions for both models with random and fixed effects, which he respectively refers to as “true” fixed and random effects models. In this paper we use the Greene’s true RE model, which is basically the original cost frontier model with a random intercept.

Another problem arises when the firm-specific effects are correlated with the explanatory variables. In such cases, the random effects (RE) estimators are affected by heterogeneity bias, but the fixed effects (FE) model while being consistent regarding the cost frontier slopes, usually overestimates efficiency variations. Therefore in many cases these models do not provide a unified approach for estimating cost frontier and inefficiencies. An exception is one of the models proposed by Cornwell, Schmidt and Sickles (1990), which extends on Hausman and Taylor (1981)’s instrumental variable methodology. This model however requires that a sufficient number of explanatory variables be uncorrelated with random effects. This paper also proposes an alternative specification of RE models that controls for the correlation between firm-specific effects and explanatory variables. This model draws upon Mundlak (1978)’s formulation of a “within” estimator in the

---

were presented by Schmidt and Sickles (1984) who relaxed the distribution assumption and used the GLS estimator, and by Battese and Coelli (1988) who assumed a truncated normal distribution.

3 See also Heshmati and Kumbhakar (1994) and Kumbhakar and Hjalmarsson (1995) for two applications of this model. Note that in the latter paper, it is assumed that both time- and firm- specific effects are part of inefficiency.
random effects framework. When applied to the conventional RE model, the resulted GLS estimator is identical to the FE estimator, thus unbiased. The inefficiency estimates are however adjusted for the correlation with exogenous variables. A similar method can be applied to the true RE model to decrease the heterogeneity bias.6

The main purpose of this paper is to study if the alternative models can improve the estimates of cost frontier and inefficiency scores in the studied sample. The models are applied to a sample of 36 nursing homes operating in Ticino, the Italian-speaking region of Switzerland, over a nine-year period from 1993 to 2001. The models are compared regarding their performance on the cost function slopes and inefficiency estimates. The conventional FE estimators of the cost function are assumed to be unbiased, thus used as a benchmark to which other models are compared.

The results suggest that as far as the heterogeneity bias is concerned, while the random constant frontier model (true RE) slightly improves the results, the proposed Mundlak adjustment brings the estimates very close to the unbiased estimators. As for the inefficiency scores, the estimates obtained from alternative models show a generally weak correlation. As expected, the FE model gives extremely high inefficiency values. Our analysis suggests that these values capture at least partially, the firms’ heterogeneity that is correlated with exogenous variables. In fact, when these correlations are included in the model specification, the inefficiency estimates are systematically lower than comparable models. The results are in

---

4 The term “heterogeneity bias” is used by Chamberlain (1982) to refer to the bias induced by the correlation between individual effects and explanatory variables in a general RE model.
5 Notice that this model is different from those authors’ other model discussed earlier.
6 This argument is based on an analogy with a GLS model that can be transformed to a “within” estimator by using Mundlak’s specification. However, it should be noted that given that the residual
general promising in that the estimated cost frontier is similar to that of a conventional FE model thus unbiased, and the inefficiency estimates remain in a reasonable range. Our results also suggest that the average inefficiency scores and the their time trends are quite similar among comparable models with time-variant inefficiency.

The paper is organized as follows. Sections 2 and 3 present the model specification and the methodology respectively. The data are explained in section 4. Section 5 presents the estimation results and section 6 concludes the paper.

2. Model specification

A nursing home can be approximately represented as a production unit transforming labor and capital services into patient-days of residential health and social care for elderly people. Assuming that output level and input prices are exogenous, and that (for a given technology) firms choose input levels to minimize costs, the firm's total cost of operating a nursing home can be defined as a function of input prices and output. Moreover, in the cost model specification we take into account a number of output characteristics, which should capture, at least partially, the heterogeneity and quality dimensions of the nursing home’s output. Costs can also vary with a time trend. The total cost frontier can therefore be represented by the following cost function:

$$TC = f(Y, Q, R, P_K, P_L, \tau)$$

(1)
where $TC$ is the total annual cost and $Y$ is the output represented by the total number of resident-days of the nursing home. $PK$ and $PL$ are respectively the prices of capital and labor. $Q$ represents the average dependency index calculated annually by the Regional Department of Public Health. This index measures the average required assistance of a given nursing home’s patients with normal daily activities such as eating, personal care or performing physiological functions. $Q$ varies from 1 to 3, with 3 representing the most severe (dependent) case. $R$ is the nursing staff ratio, that is the ratio of the number of employed nurses in a nursing home to the number of nurses that should be employed according to the guidelines of the Regional Department of Public Health.\(^9\) Since the nursing care is a labor-intensive service and the quality of care depends on the time spent by nurses for each patient, this variable represents the quality of output and the production process.\(^10\) Finally, $\tau$ is a linear time trend that captures the changes in technical efficiency associated with technical progress.

It is generally assumed that the cost function given in (1) is the result of cost minimization given input prices and output and should therefore satisfy certain properties.\(^11\) Mainly, this function must be non-decreasing, concave, linearly homogeneous in input prices and non-decreasing in output. To estimate the cost function (1), a translog functional form is employed. This flexible functional form is a local, second-order approximation to any arbitrary cost function. It places no \textit{a priori} restrictions on the elasticity of substitution and allows the economies of scale to vary with the output level. The translog approximation to (1) can be written as:

\(^9\) These guidelines are only recommendations and the nursing homes are not required to exactly follow them.

\(^10\) See Cohen and Spector (1996) and McKay (1988) for a similar approach in cost model specification for nursing homes. Cohen and Spector measured quality of care by staff to resident ratios. McKay used “nursing hours per patient” to measure the nursing home’s quality.

\(^11\) For more details on the functional form of the cost function see Cornes (1992), p.106.
\[
\ln\left(\frac{TC_{it}}{P_{K_s}}\right) = \alpha_0 + \alpha_Y \ln Y_t + \alpha_Q \ln Q_{it} + \alpha_R \ln R_{it} + \alpha_L \ln \frac{P_{L_i}}{P_{K_s}} + \frac{1}{2} \alpha_{YY} (\ln Y_t)^2 + \frac{1}{2} \alpha_{QQ} (\ln Q_{it})^2 + \frac{1}{2} \alpha_{RR} (\ln R_{it})^2 \\
+ \alpha_{YL} \ln Y_t \ln \frac{P_{L_i}}{P_{K_s}} + \alpha_{YQ} \ln Y_t \ln Q_{it} + \alpha_{QR} \ln R_{it} \ln Q_{it} + \alpha_{LQ} \ln \frac{P_{L_i}}{P_{K_s}} \ln Q_{it} + \alpha_L \tau + \alpha_i + \epsilon_{it}
\]  

(2)

where subscripts \(i\) and \(t\) denote the nursing home and year respectively, \(\alpha_i\) is a firm-specific effect and \(\epsilon_{it}\) is an \(iid\) error term which can be symmetric or asymmetric depending upon the adopted econometric model. The models used in this paper are based on two general frameworks: Schmidt and Sickles (1984)’s model that assumes a zero-mean \(iid\) error term \(\epsilon_{it}\) without any further distribution assumption, and Aigner, Lovell and Schmidt (1977)’s original framework in which \(\epsilon_{it}\) is assumed to have a composite asymmetric distribution. Namely, the error term \(\epsilon_{it}\) is decomposed into a symmetric component \((v_{it})\) for the statistical noise and an asymmetric term \((u_{it})\) representing the inefficiency. In this paper, these components are assumed to have normal and half-normal distributions respectively.

All variables are normalized by the corresponding sample medians. Therefore, the translog form is considered as a second order approximation around the sample median.\(^{12}\) As it can be seen in equation (2), linear homogeneity in input prices is imposed by dividing total costs and input prices by capital price. All the monetary values (costs and prices) are deflated to year 2000 prices.\(^{13}\) The other theoretical restrictions are verified after the estimation. In particular, the concavity of the estimated cost function reflects the fact that the cost function is a result of cost

---

\(^{12}\) Translog functional form requires that the underlying cost function be approximated around a specific point. In our case this point is taken as the sample median. We choose the median rather than the mean, because it is less affected by outliers and thus the translog approximation will have a better precision.
minimization. However, this assumption may be unrealistic in non-profit firms. In such cases, the functions based on cost optimization may still be used as “behavioral” cost functions and can be helpful in studying the behavior of such firms.\textsuperscript{14} Especially, since all the nursing homes in our sample are non-profit, it can be reasonably assumed that they follow (or should follow) a similar objective function, implicitly set by the regulators. Given this assumption comparing costs among different firms can indicate which firms achieve these objectives with less costs.

Input prices and output are assumed to be exogenous, thus beyond the firm’s control. In a regulated industry these conditions are generally satisfied. Ticino’s non-profit nursing homes are fully regulated by the canton’s government. The residents are assigned to nursing homes by their community’s authorities, mainly based on their location, and the nursing homes’ costs are refunded on a cost-plus basis.

3. Methodology

The heterogeneity bias problem and its effect on inefficiency estimates are studied by a comparative analysis of pooled cross sections, random and fixed effects models. All models are based on the specification given in equation (2). The differences among the various specifications are related to the assumptions imposed on stochastic components $\alpha_i$ and $\varepsilon_{it}$. Table 1 summarizes the six models used in the paper. The first model is a fixed effects model. In this model the firm-specific effects are estimated as constant numbers, thus can be correlated with the explanatory variables. As is well known in the literature, the FE or “within” estimators are not

\textsuperscript{13} Switzerland’s global consumer price index has been used. However, since total costs and prices are normalized by the capital price this adjustment is not necessary for the regressions.

\textsuperscript{14} See Bös (1986), page 343.
influenced by heterogeneity bias.\textsuperscript{15} In the cost frontier literature the inefficiency scores are estimated as the distance from the firm with the minimum estimated fixed effect, that is $\hat{\alpha}_i - \min \{\hat{\alpha}_i\}$, as proposed by Schmidt and Sickles (1984). Model \textit{I} has two general shortcomings that can be dealt with in a random effects framework: First, the time-invariant factors are captured by the fixed effects, leading to an overestimation of inefficiencies. Secondly, because of the incidental parameters problem, the fixed effects model requires a sufficient level of within-firm variation to provide sensible estimation results.

\begin{table}[h]
\centering
\caption{Table 1}
\end{table}

Model \textit{II} is a random effects model, which is estimated using the GLS method. The inefficiency term is estimated following the approach proposed by Schmidt and Sickles (1984). The important limitation of this model is the assumption that the firm-specific stochastic term $\alpha_i$, here assumed to be the firm’s inefficiency, is uncorrelated with the explanatory variables. Although it is reasonable to assume that the firm’s cost-inefficiency\textsuperscript{16} is not correlated with exogenous variables, the firm-specific stochastic term may contain other unobserved environmental factors, which may be correlated with explanatory variables.

In both models (\textit{I} and \textit{II}), inefficiency indicators may include unobserved environmental factors, thus may overstate the firms’ inefficiency. There are however two factors that may lead to higher inefficiency estimates in the FE model. First, unlike the RE model, the firm-specific effects do not follow a single distribution, thus can have a relatively wide range of variation.\textsuperscript{17} Secondly, these effects can be

\textsuperscript{15} See Baltagi (2001) for an extensive discussion.
\textsuperscript{16} Note that here the cost-efficiency does not include scale efficiency.
\textsuperscript{17} This is a practical issue rather than a modeling problem. In fact the FE model is more general in that it does not assume a single underlying population for all the firms.
correlated with the explanatory variables, thus can also capture the heterogeneity factors that are correlated with the regressors. Whereas in the RE model in which the firm-specific effects are by construction uncorrelated with the regressors, these factors are partially suppressed through the “between” variations, into the regression coefficients.

Model III is the GLS model specified in line with Mundlak (1978)’s formulation. The correlation of firm-specific effects with explanatory variables are considered in an auxiliary regression given by:

\[ \alpha_i = \gamma X_i + \delta_i \]

\[ \bar{X}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} X_{it}, \delta_i \sim N(0, \sigma^2_\delta) \]  \hspace{1cm} (3)

where \( X_{it} \) is the vector of all explanatory variables and \( \gamma \) is the corresponding vector of coefficients.

Equation (3) actually divides the firm-specific stochastic term into two components: The first part can be explained by exogenous variables, whereas the remaining component (\( \delta_i \)) is orthogonal to explanatory variables. If the inefficiency is assumed to be constant over time, this latter part can be interpreted as the firm’s inefficiency. In this case, inefficiencies can be estimated by comparing each firm to the firm with the minimum \( \delta_i \), that is: \( \hat{\delta}_i - \min\{\hat{\delta}_i\} \). Equation (3) can be readily incorporated in the main regression equation (2). The GLS estimators of the resulting equation are identical to the FE estimators of the original equation (within estimators), thus unbiased. Therefore, one can expect that the proposed specification can avoid the heterogeneity bias and at the same time gives reasonable estimates of inefficiency. Moreover, time-invariant factors can also be included in the model.
Model \( IV \) is a pooled frontier model in which the firm-specific effect is assumed to be zero. Thus the sample is considered as a series of cross sectional sub-samples pooled together. The random error term is divided into two components: a normal error term \( v_{it} \), capturing heterogeneity and a half-normal random term \( u_{it} \), representing the inefficiency as a one-sided non-negative disturbance. This model is based on the original cost frontier model proposed by Aigner, Lovell and Schmidt (1977). The firm’s inefficiency is estimated using the conditional mean of the inefficiency term \( E[u_{it} | u_{it} + v_{it}] \), proposed by Jondrow et al. (1982).

Models \( V \) and \( VI \) are extensions to model \( IV \) in that they include an additional firm-specific effect \( (\alpha_i) \) to represent the unobserved heterogeneity among firms. In both models this effect is considered as a random effect. Model \( V \) is based on true RE model proposed by Greene (2002a,b).\(^{19}\) Finally model \( VI \) is the true RE model modified by Mundlak’s specification. This model not only includes a firm-level source of heterogeneity, potentially correlated with explanatory variables, it also allows for a time-variant inefficiency term. Mundlak’s adjustment is applied similar to model \( III \) as given in equation (3). Given that here the error term \( (\varepsilon_{it}) \) is a composite asymmetric term, the resulting coefficients are not within estimators as in model \( III \). However, since the correlation between individual effects and explanatory variables is at least partly captured in the model, the heterogeneity bias is expected to be minimal.

In our comparative analysis we consider two aspects of the models’ performance. The first dimension is the estimation of the cost function’s coefficients. In cases such as nursing homes (or in general health services), where the costs are

---

\(^{18}\) See also Hsiao (2003), pp. 44-46, for a proof of the identity of Mundlak’s GLS estimators and FE estimators.
affected by the case-mix severity, a number of location-related factors can affect both costs and explanatory variables. For instance, larger nursing homes are usually located in more populated urban areas where patients might be sicker (thus more costly) and the price of labor is potentially higher. Such relationships imply a positive correlation between the output level and labor price with the case-mix severity, which is not fully captured by the included factors in the model. The Hausman test is used to confirm that the firm-specific effects are correlated with the explanatory variables. In this case the FE estimators are unbiased, thus provide a benchmark to which other models can be compared. On the other hand, the GLS estimators are biased and therefore provide an indication for the direction of heterogeneity bias. Noting that the Hausman test statistics is based on an overall distance between the two estimators, for each model we compare the estimated cost function’s coefficients with the corresponding estimates from the FE and GLS models (models I and II here).

The heterogeneity bias is expected to be relatively low in models III and VI that directly control for correlation between individual effects and explanatory variables. In other models there is no way to predict the bias. One can argue that models with more general error structures have lower biases because the residuals can capture a larger part of the correlations between unobserved heterogeneity and explanatory variables, thus leaving the coefficients less affected. However, the residuals are by definition uncorrelated with explanatory variables and the extent to which they may confound such correlations with errors may significantly vary from one sample to another. Especially since the frontier estimators are non-linear, the

19 This model is a special case of a stochastic frontier model with random parameters (in this case random intercept).
prediction of the biases is not straightforward. This theoretical discussion is beyond
the scope of this paper. Here we rather focus on the evaluation of the models with
respect to our sample.

The second aspect of the models’ performance concerns the estimation of
inefficiency scores. Since they are based on certain interpretation of the stochastic
terms included in the model, the inefficiency estimates are considered as a separate
dimension of the model’s performance. In fact, an unbiased estimation of the cost
function is a necessary but not sufficient condition for consistent estimation of
inefficiency. In the first three models (I, II and III), the firm’s inefficiency is assumed
to be constant over time, thus captured by the firm-specific effects. In models IV, V
and VI on the other hand, the firm’s inefficiency can vary from one year to another.
In these models, the skewed stochastic error term is interpreted as inefficiency.
Except for the FE model (model I), in all these models it is assumed that the firm’s
cost efficiency is not correlated with the explanatory variables.\footnote{The average dependency index, which is included in the model, only measures the time required for nursing care, thus captures only one aspect of case-mix severity. Other factors like the need for medical treatment and drugs are not observed.} This assumption is
consistent with the requirement that the explanatory variables are exogenous.

The FE formulation in model I has two important limitations. First, the time
invariant variables are captured by the fixed effects and cannot be included in the
model. This implies that the inefficiency estimators include the variations in time-
invariant firm characteristics.\footnote{It is worth noting that here cost inefficiency is defined as the excess costs due to the firm’s technical problems or to suboptimal allocation of resources. Other inefficiency sources like scale inefficiencies, which are beyond the firm’s control are excluded.} Moreover, the estimated fixed effects include
unobservable firm-specific factors that are correlated with explanatory variables.
However, as is common in most frontier models, the firm’s inefficiency \emph{per se} is not
correlated with exogenous variables like output and input prices. Therefore, in cases
where unobserved environmental factors are likely to affect costs, model \( I \) appears to be inadequate regarding the estimation of inefficiencies. Model \( II \) is expected to have a better performance because the individual effects are by construction uncorrelated with explanatory variables, thus less affected by exogenous variables. However, the inefficiency estimates may still contain firm-specific heterogeneity that is not related to inefficiency. The Mundlak’s adjustment used in model \( III \) should take care of such heterogeneity to the extent that it is correlated with the explanatory variables.

In models \( IV, V \) and \( VI \), where the inefficiency can vary with time, one could expect to have higher inefficiency estimates compared to models \( II \) and \( III \). Model \( IV \) ignores the firm-specific heterogeneity, thus may overestimate inefficiency compared to models \( V \) and \( VI \). In model \( VI \), Mundlak’s adjustment may help to completely separate the correlation effects, thus leads to lower estimates of inefficiency.

Except for the FE model (Model \( I \)) that, for reasons mentioned above, is expected to have a poor performance, there is no general, clear-cut distinction among the studied models regarding their performance as to efficiency estimation. Rather, each model implies a different interpretation of cost-inefficiency. If the inefficiency is believed to be persistent, the models with time-invariant inefficiency may be more relevant. Similar to the FE model, the GLS models (\( II \) and \( III \)) assume that full efficiency corresponds with the minimum firm-specific effect. Thus, high levels of efficiency are considered as rare outcomes located at the tail of the probability distribution, limited to a few exceptional companies. This problem can be avoided at the expense of imposing certain distribution assumptions\(^{23} \) on the inefficiency term as in Pitt and Lee (1981).

\(^{22}\) As our specification does not include any time-invariant factor, this statement does not apply here. \(^{23}\) In distributions such as half-normal or exponential, perfect efficiency is at the mode of the probability distribution, thus the most likely outcome.
Most frontier models assume that inefficiency is uncorrelated with explanatory variables included in the cost function. While being practical for estimation purposes, this assumption can be justified based on the fact that the apparent excess costs that are correlated with explanatory variables may be due to factors beyond the firm’s control. To the extent that the firm’s inefficiency is not correlated with the explanatory variables Mundlak’s adjustment is likely to improve the estimations. The purpose of this paper is not to identify the most appropriate method, which could differ from one case to another. Rather, our comparative analysis should highlight in each one of the models, the relation of what is called inefficiency with other sources of heterogeneity as well as with the explanatory variables. In any case, a high correlation between the inefficiency estimates can be inferred as an indication of robustness and validity of individual approaches. Therefore, the correlation between the inefficiency scores estimated by different models is studied. Different models are also compared with respect to the average inefficiency of the whole sample. In cases where the inefficiency varies over time, the annual averages are also compared.

4. Data

The data set used in this paper is prepared based on the annual accounting reports of 36 non-profit nursing homes in Ticino, the Italian-speaking region of Switzerland, over the 9-year period from 1993 to 2001. The sample includes more than two thirds of Ticino’s nursing homes. All the nursing homes in the sample provide inpatient services. There are four missing observations in 1993, leaving a

24 The only exception is the FE models that interpret the effects as inefficiency.
25 There are some nursing homes that offer the possibility of nursing care in external residential apartments. The nursing care of this type is less intensive (thus less costly) than the care given to the home’s residents. For this reason we excluded four nursing homes whose share of external beds is
total of 320 observations. The variables include total costs, total number of employees (in terms of full-time equivalent units), average wage per employee per year, total number of beds and total number of resident-days. Other characteristics are the average dependency grade of the residents and the number of care personnel working for the nursing home.

Total cost is taken as the total annual expenditures of the nursing home. Output is measured as the total number of patient-days of the nursing home. Average yearly wage rates are estimated as the weighted mean of the average wage rates of different professional categories working in a nursing home, including nurses, administrative and technical staff and physicians. Following Friedlaender and Wang Chiang (1983) and Filippini (2001), the capital price is calculated from the residual costs divided by the capital stock. Residual cost is total cost minus labor cost. Similar to Wagstaff (1989), the capital stock is approximated by the number of beds operated by the nursing home. The quality indicators, $Q$ and $R$, (as described earlier) are calculated annually by the regional Department of Public Health.

The summary statistics of the main variables used in the analysis are given in table 2. As it can be seen in the table, there is a high variation in the costs of a patient-day care. The input prices show a great amount of variation as well. Part of these variations is associated with time variation. For instance the average cost of a patient-day care has increased from about 154 Francs in 1993 to 214 Francs in 2001. In the same period, the price of labor has increased about 15 percent in real terms and our measure of real capital price has increased about 20 percent. The last column of table 2 lists the fraction of the variance of each variable due to the sample’s variation

more than 10 percent of their total beds. In our final sample there remain two nursing homes that offered external care (less than 10 percent) for some years during the study period.

26 A more precise estimation of capital stock would require capital inventory data, which are not available to us.
between different nursing homes. These numbers suggest that all the variables show significant variations both within and between nursing homes. This result justifies the use of panel data models, especially the FE estimator that relies upon “within” variations.

5. Estimation results

The estimated parameters of the basic cost frontier models are listed in table 3. The regression results show that all the first-order terms are significant and in a reasonable direction. As expected, output and prices have a positive effect on costs, and the nursing homes with a more severe case-mix and/or with a higher quality of service are relatively more costly. Since total costs and the regressors are in logarithms and normalized by their medians, the first order coefficients are interpretable as cost elasticities evaluated at the sample median. The output elasticity is positive and implies that an increase in the supply will increase total cost. The results indicate unexploited scale economies in the production. Different models lead however to different results. A one percent increase in the number of patient-days of nursing home care will increase the total cost by about 0.75% to 0.92%. Other coefficients are also significantly different across different models.

Cost elasticities with respect to the output characteristics variables, \( Q \) and \( R \), are positive and imply that an increase in the average required assistance of a home's patients or an increase in the ratio of the number of nurses employed by a nursing
home and the number of nurses that should theoretically be employed will increase total cost. The coefficient of the linear trend suggests that the total costs have increased over time with a rate of about 0.9 to 1.8 percent per year. The growth of costs is a commonly observed phenomenon in labor-intensive industries such as health care, which usually face a persistent growth of labor price. The estimated cost functions do not however satisfy the concavity condition in input prices.\textsuperscript{28} This may suggest that the estimated cost functions are not resulted from a completely unconstrained cost-minimization strategy. Namely, the firms’ strategies are not responsive to changes in input factor prices. This can be explained by the fact that the input choices in Switzerland’s nursing homes are rather constrained by regulation.\textsuperscript{29}

The main observation on the results listed in table 3, is that the FE estimators (model I) can be singled out as extreme values for almost all the coefficients. The Hausman test rejects the hypothesis of no correlation between random effects and the explanatory variables quite significantly (Chi-square of 57.3 with 15 degrees of freedom). The FE model results are therefore unbiased and can serve as a benchmark for our estimations. This implies the inconsistency of all other models.

Table 4 lists the estimation results obtained from Mundlak’s formulation. As expected, when applied to the RE (GLS) model, the main equation coefficients are quite close if not identical, to the within estimators in model I. Some of the auxiliary equation coefficients are significant indicating that the random effects are actually

\textsuperscript{27} These findings are in line with the results obtained by Filippini (2001) using a shorter panel and a slightly different number of nursing homes.
\textsuperscript{28} Our results indicate that the Hessian matrix of the estimated cost function with respect to input prices (labor and capital) is not negative semi-definite, thus the concavity condition is not satisfied in any of the specifications.
\textsuperscript{29} See Farsi and Filippini (2003) for a more detailed discussion.
correlated with some of the explanatory variables. In particular the coefficients of output ($Y$) and the dependency ratio ($Q$) are highly significant and positive. This may suggest that the unobserved heterogeneity among nursing homes is partly due to the patients’ unobserved severity characteristics. As argued earlier such characteristics may well be positively correlated with the nursing home’s size. The last two columns of the table present the results of Mundlak’s specification applied to the true RE model. Most auxiliary coefficients are significant, confirming high correlation of random effects with explanatory variables. Interestingly, Mundlak adjustment has a similar effect on the true RE model, bringing the estimated coefficients closer to the unbiased results of model I.

\textless\textbf{TABLE 4 ABOUT HERE}\textgreater

The first-order coefficients obtained from different models indicate that the pooled model’s estimates (model \textit{IV}) are in general located relatively far from the FE estimates, suggesting that this model has the poorest performance with regard to heterogeneity bias. This is consistent with the fact that this model does not distinguish individual firms and may be strongly affected by the omitted variables bias. The coefficients estimated by the true RE model on the other hand, lie almost without exception, between those of the GLS model and the FE estimators. This result suggests that in our sample, compared to GLS, this model is less affected by heterogeneity bias. Finally, the estimates obtained from model \textit{VI} are quite close to the unbiased estimators, suggesting that controlling for correlations in the true RE model can decrease heterogeneity biases.

Table 5 provides a summary of the estimated inefficiency measures using different models. The inefficiency scores are taken equal to the inefficiency scores ($u_i$), obtained from the regression model. These measures represent the relative
excess cost of a nursing home compared to a minimum level that would have been achieved had the firm operated as cost-efficient as the “best practice” observed in the sample. Note that in the first three models (I, II and III) inefficiency is assumed to be constant over time with a single fully efficient firm, while in models IV, V and VI, the firm’s inefficiency is time-variant and most of the firms are expected to be fully efficient or close. Therefore, comparing the values across two groups should be done with caution.

As expected, the FE model predicts excessive inefficiency estimates averaging about 19% and up to a maximum of 38%. Model II’s results are less than half of these values, suggesting that the estimates in model I are confounding heterogeneity with inefficiency. This result suggests that both models are affected by the heterogeneity bias; while in the RE model, the coefficients capture most of the biases, in the FE model the bias appears only in the individual effects. The results also show that Mundlak’s adjustment in model III improves the results in that while keeping the coefficients unbiased, it decreases the bias in inefficiency estimates by separating the correlation effects. As seen in table 5, compared to the GLS model, the inefficiency estimates are on average about 40% lower when these correlations are taken into account.

Comparing models with time-variant inefficiency (last three columns of the table) shows that the inefficiency estimates are on average more or less similar. This implies that in these models, inefficiency estimates are not much sensitive to the specification of firm-specific heterogeneity. The differences however, point to a similar pattern in that a better control for firm-specific heterogeneity decreases the inefficiency estimates. Namely, the average inefficiency score decreases from 5.9%
in the pooled model (model IV) to 5.1% after controlling for firm-specific heterogeneity (model V), and to 4.5% with an additional Mundlak correction (model VI).

The pair-wise correlation coefficients between the inefficiency scores obtained from different models are presented in table 6. In order for the correlation coefficients to be comparable, they are calculated at the firm level using 36 observations (one observation for each firm). Namely, in models with time-variant efficiency, the inefficiency score is calculated as the firm’s average inefficiency score over the sample period. Although there is no clear threshold to evaluate these correlations, we consider that a coefficient less than 0.9 is indicative of quite significant differences in both individual scores and ranks across the models. According to this criterion, the correlations between the models in each group (time-variant and time-invariant) are rather weak.

This result is in contrast with the results reported by Greene (2002a) who applied a series of alternative models to a short panel of US banks sample (T=5). In that analysis the inefficiency estimates obtained from Pitt and Lee (1981)’s model and a standard FE model (as in Schmidt and Sickles (1984)), both with time-invariant inefficiency, are very close. Similarly, there is a quite high correlation between the estimates from the true FE and true RE models, with time-variant inefficiency. Greene’s results can however be explained by the fact that as suggested by the Hausman test, the heterogeneity bias is rather insignificant in that sample (see footnote 10).
Interestingly, the highest correlation coefficients are observed between models III and VI, and models II and V. Both these cases link a time-variant inefficiency model to a model that assumes constant inefficiency. The relatively high correlation between GLS estimates and the true RE model suggests that both models, although affected by heterogeneity bias in the coefficients, have a reasonable “mutual consistency” with regard to inefficiency estimation. On the other hand, the high correlation between two models with Mundlak’s specification suggests that the heterogeneity bias can be resolved without affecting the validity of inefficiency estimates.

Another observation on table 6 is that while the correlation between models I and II is fairly high (.849), both models show a weak correlation with model III, suggesting that Mundlak adjustment in has a significant effect on individual inefficiencies. This pattern is less evident in model VI compared to models IV and V. In fact, the Mundlak adjustment does not appear to cause a considerable change in the correlation with the pooled model (IV), which is fairly high (about 0.9). However, the correlation between models V and VI appears to considerably lower when the inefficiencies are averaged over the sample period. This result may suggest that the Mundlak adjustment is not just a shift at the firm level; rather, it causes a differential change in inefficiency estimates of a given firm over time. Similar correlation coefficients have been calculated for efficiency ranks. These coefficients (not shown in the paper) are generally close to the coefficients reported in table 6.

In figure 1, the average inefficiency score is plotted against time, as estimated by models IV, V and VI. All three models suggest that the cost efficiency of Ticino’s nursing homes has continuously improved since 1996. As expected, the

---

30 These models have also the highest correlation coefficients in efficiency ranks (0.98 between II and V, and 0.95 between III and VI).
pooled frontier model slightly overestimates the inefficiencies because it does not consider any firm-specific heterogeneity. This figure shows that the trends estimated by all three models are quite similar. These similarities are the more striking as these models result in significantly different estimates of the cost frontier coefficients (as shown in tables 3 and 4). These results, along with similar results in overall average inefficiencies (see table 5), suggest that although these models are different in individual inefficiency scores, the inefficiency estimates have robust average values as long as these values are taken over reasonably large subgroups. This implies that the considerable differences observed in individual scores are induced by sampling variation, rather than by differences in model specification. Therefore, these results points to a general conclusion that the inefficiency estimates in models with time-variant inefficiency are not much sensitive to the correlation between firm-specific heterogeneity and explanatory variables. Such correlations are captured by the cost function coefficients and therefore do not affect the residuals.

<FIGURE 1 ABOUT HERE>

6. Concluding remarks

The application of alternative cost frontier models to a panel of nursing homes in Switzerland suggests that the estimated cost frontier is sensitive to the adopted model. In particular, the results largely depend upon how the unobserved heterogeneity among firms is accounted for. Given that in our sample the within-firm variations are significant and that the Hausman test indicates a high risk of heterogeneity bias, the fixed effects (FE) model can be considered as a consistent

31 The expression “mutual consistency” is used by Bauer et al. (1998) in this context.
32 Note that each year subgroup has about 36 observations.
estimator while the random effects (RE) estimator is likely to be biased. Our analysis indicates that a frontier model with random constant (true RE model) slightly decreases these biases.

The results also point to the weak performance of the FE formulation in estimating inefficiencies in usually small-$T$ panel data samples and in presence of unobserved heterogeneity. Given that in many cases this model is the only unbiased estimator of the cost frontier, a modification that can improve the inefficiency estimates without affecting the model’s consistency can prove helpful. In this paper we propose a specification based on Mundlak (1978)’s formulation. This approach allows for a direct control for the potential correlation of firm-specific, latent heterogeneity with explanatory variables. The adjustment has been introduced to the conventional GLS model and the true RE model. The cost function’s coefficient estimates have been very close to those of the fixed-effects model, thus unbiased. The advantage over the FE model is that the time-invariant factors as well as other hidden correlations with exogenous variables are disentangled from the inefficiency estimates. Our empirical results suggest that this improvement can be quite significant, especially in models with time-invariant inefficiencies. Overall, the model resulted from combining Mundlak’s specification with the true RE model, provides a considerable advantage in that while avoiding heterogeneity bias and improving inefficiency estimates, it allows time-variant inefficiency.

Finally, our individual inefficiency estimates appear rather sensitive to the econometric specification. These differences are partly due to different specifications of inefficiency and heterogeneity across the models and partly to the large sampling errors incurred at the individual level. For instance when inefficiency is time-variant, we have only one observation for each inefficiency estimate; thus large errors can be
expected. This problem is documented by Horrace and Schmidt (1996), Street (2003) and Jensen (2000) in both cross-sectional data and small-\(T\) panels.\(^{33}\) Obviously, to the extent that inefficiencies remain constant over time a longer panel can help. Nevertheless, the assumption of constant inefficiency can be less realistic in longer panels. Our results indicate that when the inefficiency estimates are averaged over a fairly large number of observations, comparable models give rather similar results, or in case of different outcomes, the differences can be reasonably explained through econometric specification. In particular, the average inefficiency of the sample and the average annual inefficiency rates are consistently similar among three models with time-variant inefficiency.

**Acknowledgement**

The authors are grateful to the Ticino’s Department of Health and Social Services for providing the data and their general support. Ilaria Mosca, Chiara Gulfi and Giorgio Borradori provided an excellent assistance in preparing the final data set. An earlier version of this paper was presented at the 8\(^{th}\) European Workshop on Efficiency and Productivity Analysis, whose participants are thanked for their insightful discussion. The authors also wish to thank William Greene, Subal Kumbhakar, Robin Sickles, Luca Crivelli, the editor and an anonymous referee for their helpful suggestions. Any remaining errors are solely the responsibility of the authors.

---

\(^{33}\) Horrace and Schmidt (1996) show that a panel with 6 periods cannot provide a consistent estimation of individual efficiency scores.
References


Table 1. Econometric specifications

<table>
<thead>
<tr>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
<th>Model V</th>
<th>Model VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FE</td>
<td>RE (GLS)</td>
<td>RE (GLS) with Mundlak formulation</td>
<td>Pooled</td>
<td>True RE</td>
<td>True RE with Mundlak formulation</td>
</tr>
<tr>
<td>Firm-specific component $\alpha_i$</td>
<td>fixed</td>
<td>$\alpha_i = \gamma \bar{X}_i + \delta_i$</td>
<td>none</td>
<td>$\alpha_i = \gamma \bar{X}_i + \delta_i$</td>
<td>$\delta_i \sim N(0, \sigma_\delta^2)$</td>
</tr>
<tr>
<td>Random error $\varepsilon_i$</td>
<td>iid $(0, \sigma_\varepsilon^2)$</td>
<td>iid $(0, \sigma_\varepsilon^2)$</td>
<td>iid $(0, \sigma_\varepsilon^2)$</td>
<td>iid $(0, \sigma_\varepsilon^2)$</td>
<td>iid $(0, \sigma_\varepsilon^2)$</td>
</tr>
<tr>
<td>Inefficiency</td>
<td>$\hat{\alpha}_i - \min {\alpha_i}$</td>
<td>$\hat{\delta}_i - \min {\delta_i}$</td>
<td>$E[u_i</td>
<td>u_i + v_i]$</td>
<td>$E[u_i</td>
</tr>
</tbody>
</table>

Table 2. Descriptive statistics (320 observations)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
<th>Fraction of between variation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs per resident-day</td>
<td>184.05</td>
<td>28.92</td>
<td>183.10</td>
<td>111.85</td>
<td>279.81</td>
<td>.307</td>
</tr>
<tr>
<td>Total annual resident-days ($Y$)</td>
<td>23,176</td>
<td>9684</td>
<td>21,482</td>
<td>6,525</td>
<td>58,324</td>
<td>.848</td>
</tr>
<tr>
<td>Number of beds</td>
<td>66.23</td>
<td>26.81</td>
<td>61</td>
<td>28</td>
<td>162</td>
<td>.850</td>
</tr>
<tr>
<td>Average labor price ($P_L$) per employee per year</td>
<td>70,157</td>
<td>6,586</td>
<td>70,280</td>
<td>29,744</td>
<td>122,950</td>
<td>.099</td>
</tr>
<tr>
<td>Average capital price ($P_K$) per bed</td>
<td>11,008</td>
<td>2,579</td>
<td>10,714</td>
<td>3,466</td>
<td>22,426</td>
<td>.606</td>
</tr>
<tr>
<td>Average dependency index ($Q$)</td>
<td>2.575</td>
<td>.219</td>
<td>2.6</td>
<td>1.87</td>
<td>3</td>
<td>.387</td>
</tr>
<tr>
<td>Nursing staff ratio ($R$)</td>
<td>.963</td>
<td>.124</td>
<td>.97</td>
<td>.49</td>
<td>1.55</td>
<td>.235</td>
</tr>
</tbody>
</table>

* Fraction of variance due to between variation is defined as $\frac{Var(u_i)}{Var(u_i) + Var(\varepsilon_i)}$, where $u_i$ and $\varepsilon_i$ are the residuals of a GLS regression of the corresponding variable on a constant. $i = 1, 2, ..., N$; $t = 1, 2, ..., T$.

- All monetary values are in Swiss Francs (CHF) deflated to year 2000 prices, based on Switzerland’s global consumer price index.
<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model IV</th>
<th>Model V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>RE (GLS)</td>
<td>Pooled</td>
<td>True RE</td>
</tr>
<tr>
<td>( \alpha_Y )</td>
<td>.750*</td>
<td>.890*</td>
<td>.925*</td>
<td>.869*</td>
</tr>
<tr>
<td></td>
<td>(.028)</td>
<td>(.017)</td>
<td>(.014)</td>
<td>(.007)</td>
</tr>
<tr>
<td>( \alpha_Q )</td>
<td>.308*</td>
<td>.555*</td>
<td>.713*</td>
<td>.481*</td>
</tr>
<tr>
<td></td>
<td>(.097)</td>
<td>(.083)</td>
<td>(.082)</td>
<td>(.036)</td>
</tr>
<tr>
<td>( \alpha_R )</td>
<td>.317*</td>
<td>.382*</td>
<td>.435*</td>
<td>.350*</td>
</tr>
<tr>
<td></td>
<td>(.046)</td>
<td>(.046)</td>
<td>(.045)</td>
<td>(.021)</td>
</tr>
<tr>
<td>( \alpha_L )</td>
<td>.804*</td>
<td>.832*</td>
<td>.877*</td>
<td>.819*</td>
</tr>
<tr>
<td></td>
<td>(.027)</td>
<td>(.025)</td>
<td>(.023)</td>
<td>(.012)</td>
</tr>
<tr>
<td>( \alpha_{YY} )</td>
<td>-.149*</td>
<td>-.024</td>
<td>.050</td>
<td>-.085*</td>
</tr>
<tr>
<td></td>
<td>(.061)</td>
<td>(.053)</td>
<td>(.043)</td>
<td>(.022)</td>
</tr>
<tr>
<td>( \alpha_{QQ} )</td>
<td>-1.036</td>
<td>-1.558</td>
<td>-1.034</td>
<td>-1.440</td>
</tr>
<tr>
<td></td>
<td>(.91)</td>
<td>(.90)</td>
<td>(.96)</td>
<td>(.52)</td>
</tr>
<tr>
<td>( \alpha_{LL} )</td>
<td>.512*</td>
<td>.612*</td>
<td>.573*</td>
<td>.579*</td>
</tr>
<tr>
<td></td>
<td>(.076)</td>
<td>(.075)</td>
<td>(.061)</td>
<td>(.034)</td>
</tr>
<tr>
<td>( \alpha_{YQ} )</td>
<td>.078</td>
<td>-.011</td>
<td>.051</td>
<td>-.001</td>
</tr>
<tr>
<td></td>
<td>(.12)</td>
<td>(.12)</td>
<td>(.14)</td>
<td>(.056)</td>
</tr>
<tr>
<td>( \alpha_{YL} )</td>
<td>.004</td>
<td>-.00006</td>
<td>.050</td>
<td>-.020</td>
</tr>
<tr>
<td></td>
<td>(.045)</td>
<td>(.042)</td>
<td>(.036)</td>
<td>(.019)</td>
</tr>
<tr>
<td>( \alpha_{QY} )</td>
<td>.187</td>
<td>.034</td>
<td>.022</td>
<td>.094</td>
</tr>
<tr>
<td></td>
<td>(.17)</td>
<td>(.17)</td>
<td>(.19)</td>
<td>(.10)</td>
</tr>
<tr>
<td>( \alpha_{RR} )</td>
<td>-.200</td>
<td>-.113</td>
<td>-.304</td>
<td>-.176</td>
</tr>
<tr>
<td></td>
<td>(.20)</td>
<td>(.21)</td>
<td>(.193)</td>
<td>(.09)</td>
</tr>
<tr>
<td>( \alpha_{YR} )</td>
<td>.273*</td>
<td>.223*</td>
<td>.167</td>
<td>.193*</td>
</tr>
<tr>
<td></td>
<td>(.097)</td>
<td>(.098)</td>
<td>(.116)</td>
<td>(.047)</td>
</tr>
<tr>
<td>( \alpha_{LR} )</td>
<td>.395*</td>
<td>.348*</td>
<td>.408*</td>
<td>.412*</td>
</tr>
<tr>
<td></td>
<td>(.12)</td>
<td>(.12)</td>
<td>(.12)</td>
<td>(.045)</td>
</tr>
<tr>
<td>( \alpha_{QR} )</td>
<td>-.187</td>
<td>-.587</td>
<td>-.740*</td>
<td>-.447*</td>
</tr>
<tr>
<td></td>
<td>(.34)</td>
<td>(.34)</td>
<td>(.356)</td>
<td>(.16)</td>
</tr>
<tr>
<td>( \alpha_t )</td>
<td>.018*</td>
<td>.012*</td>
<td>.009*</td>
<td>.014*</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.001)</td>
</tr>
<tr>
<td>Constant</td>
<td>–</td>
<td>15.15*</td>
<td>15.10*</td>
<td>15.10*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.013)</td>
<td>(.014)</td>
<td>(.005)</td>
</tr>
<tr>
<td>R-square</td>
<td>.987</td>
<td>.975</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

* Significant at less than 5%.

- Standard errors are given in brackets.
- The sample includes 320 observations (36 nursing homes).
<table>
<thead>
<tr>
<th></th>
<th>Model III</th>
<th></th>
<th>Model VI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RE (GLS) with Mundlak formulation</td>
<td>True RE with Mundlak formulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**Main Equation</td>
<td><strong>Auxiliary Equation</strong></td>
<td><strong>Main Equation</strong></td>
<td><strong>Auxiliary Equation</strong></td>
<td></td>
</tr>
<tr>
<td>$\alpha_Y$</td>
<td>.750*</td>
<td>.184*</td>
<td>.762*</td>
<td>.175*</td>
</tr>
<tr>
<td></td>
<td>(.028)</td>
<td>(.041)</td>
<td>(.016)</td>
<td>(.019)</td>
</tr>
<tr>
<td>$\alpha_Q$</td>
<td>.303*</td>
<td>.583*</td>
<td>.341*</td>
<td>.564*</td>
</tr>
<tr>
<td></td>
<td>(.098)</td>
<td>(.184)</td>
<td>(.051)</td>
<td>(.071)</td>
</tr>
<tr>
<td>$\alpha_R$</td>
<td>.316*</td>
<td>.237</td>
<td>.331*</td>
<td>.203*</td>
</tr>
<tr>
<td></td>
<td>(.046)</td>
<td>(.193)</td>
<td>(.027)</td>
<td>(.070)</td>
</tr>
<tr>
<td>$\alpha_L$</td>
<td>.804*</td>
<td>.082</td>
<td>.802*</td>
<td>.096*</td>
</tr>
<tr>
<td></td>
<td>(.028)</td>
<td>(.062)</td>
<td>(.016)</td>
<td>(.024)</td>
</tr>
<tr>
<td>$\alpha_{YY}$</td>
<td>-.149*</td>
<td>.304*</td>
<td>-1.42*</td>
<td>.286*</td>
</tr>
<tr>
<td></td>
<td>(.061)</td>
<td>(.126)</td>
<td>(.033)</td>
<td>(.050)</td>
</tr>
<tr>
<td>$\alpha_{QQ}$</td>
<td>-1.048</td>
<td>3.95</td>
<td>-.876</td>
<td>2.72*</td>
</tr>
<tr>
<td></td>
<td>(.91)</td>
<td>(3.76)</td>
<td>(.63)</td>
<td>(1.37)</td>
</tr>
<tr>
<td>$\alpha_{LL}$</td>
<td>.513*</td>
<td>-.188</td>
<td>.515*</td>
<td>-.173</td>
</tr>
<tr>
<td></td>
<td>(.077)</td>
<td>(.339)</td>
<td>(.044)</td>
<td>(.13)</td>
</tr>
<tr>
<td>$\alpha_{YQ}$</td>
<td>.077</td>
<td>.436</td>
<td>.051</td>
<td>.471*</td>
</tr>
<tr>
<td></td>
<td>(.12)</td>
<td>(.592)</td>
<td>(.066)</td>
<td>(.212)</td>
</tr>
<tr>
<td>$\alpha_{YL}$</td>
<td>.040</td>
<td>.175</td>
<td>-.012</td>
<td>.191*</td>
</tr>
<tr>
<td></td>
<td>(.045)</td>
<td>(.141)</td>
<td>(.021)</td>
<td>(.052)</td>
</tr>
<tr>
<td>$\alpha_{LQ}$</td>
<td>.187</td>
<td>-.753</td>
<td>.203</td>
<td>-.780*</td>
</tr>
<tr>
<td></td>
<td>(.17)</td>
<td>(.819)</td>
<td>(.12)</td>
<td>(.303)</td>
</tr>
<tr>
<td>$\alpha_{RR}$</td>
<td>-.201</td>
<td>-.806</td>
<td>-.211</td>
<td>-.568</td>
</tr>
<tr>
<td></td>
<td>(.20)</td>
<td>(1.25)</td>
<td>(.11)</td>
<td>(.44)</td>
</tr>
<tr>
<td>$\alpha_{YR}$</td>
<td>.273*</td>
<td>.0011</td>
<td>.240*</td>
<td>-.020</td>
</tr>
<tr>
<td></td>
<td>(.097)</td>
<td>(.35)</td>
<td>(.054)</td>
<td>(.13)</td>
</tr>
<tr>
<td>$\alpha_{LR}$</td>
<td>.395*</td>
<td>.289</td>
<td>.365*</td>
<td>.402*</td>
</tr>
<tr>
<td></td>
<td>(.12)</td>
<td>(.531)</td>
<td>(.060)</td>
<td>(.19)</td>
</tr>
<tr>
<td>$\alpha_{QR}$</td>
<td>-.185</td>
<td>-1.53</td>
<td>-1.242</td>
<td>-.138*</td>
</tr>
<tr>
<td></td>
<td>(.34)</td>
<td>(1.45)</td>
<td>(.22)</td>
<td>(0.55)</td>
</tr>
<tr>
<td>$\alpha_t$</td>
<td>.018*</td>
<td>–</td>
<td>.018*</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>–</td>
<td>(.001)</td>
<td>–</td>
</tr>
<tr>
<td>Constant</td>
<td>15.12*</td>
<td>–</td>
<td>15.08*</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>(.014)</td>
<td>–</td>
<td>(.007)</td>
<td>–</td>
</tr>
<tr>
<td>R-square</td>
<td>.982</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

* Significant at less than 5%.
- Standard errors are given in brackets.
- The sample includes 320 observations (36 nursing homes).
Table 5. Inefficiency measures

<table>
<thead>
<tr>
<th></th>
<th>Model I FE</th>
<th>Model II RE (GLS)</th>
<th>Model III GLS with Mundlak formulation</th>
<th>Model IV Pooled</th>
<th>Model V True RE</th>
<th>Model VI True RE with Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>.191</td>
<td>.082</td>
<td>.050</td>
<td>.059</td>
<td>.051</td>
<td>.045</td>
</tr>
<tr>
<td>Median</td>
<td>.203</td>
<td>.089</td>
<td>.052</td>
<td>.054</td>
<td>.043</td>
<td>.040</td>
</tr>
<tr>
<td>Maximum</td>
<td>.379</td>
<td>.152</td>
<td>.104</td>
<td>.279</td>
<td>.251</td>
<td>.210</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.009</td>
<td>.006</td>
<td>.008</td>
</tr>
<tr>
<td>N</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>320</td>
<td>320</td>
<td>320</td>
</tr>
</tbody>
</table>

- Inefficiency measures represent the relative difference of a nursing home’s actual costs to minimum costs from the best practice in the sample.

Table 6. Correlation between inefficiency estimates from different models

<table>
<thead>
<tr>
<th></th>
<th>Model I FE</th>
<th>Model II RE (GLS)</th>
<th>Model III GLS with Mundlak formulation</th>
<th>Model IV Pooled</th>
<th>Model V True RE</th>
<th>Model VI True RE with Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model II</td>
<td>.849</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model III</td>
<td>.343</td>
<td>.670</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model IV</td>
<td>.534</td>
<td>.854</td>
<td>.834</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model V</td>
<td>.888</td>
<td>.939</td>
<td>.555</td>
<td>.806 [.902]</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Model VI</td>
<td>.260</td>
<td>.601</td>
<td>.941</td>
<td>.878 [.901]</td>
<td>.546 [.805]</td>
<td>1</td>
</tr>
</tbody>
</table>

- In models IV, V and VI the inefficiency estimates are the average values over the sample period.
- Correlation coefficients based on 320 observations are given in brackets.
Figure 1
Average inefficiency score

Year

- IV: Pooled
- V: True RE
- VI: True RE with Mundlak
CUT TO THE BONE? HOSPITAL TAKEOVERS
AND NURSE EMPLOYMENT CONTRACTS

Janet Currie
Mehdi Farsi
W. Bentley MacLeod

Working Paper 9428
http://www.nber.org/papers/w9428

This paper was prepared for a Festschrift in honor of Orley Ashenfelter. The authors thank Phillip Levine, and participants at the Festschrift conference in October 2002 for helpful comments. Ilya Berger provided research assistance, and we thank the Center for Law, Economics and Organization for research support. We are solely responsible for any errors. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

© 2002 by Janet Currie, Mehdi Farsi, and W. Bentley MacLeod. All rights reserved. Short sections of text not to exceed two paragraphs, may be quoted without explicit permission provided that full credit including, © notice, is given to the source.
ABSTRACT

This paper uses data from the 1990s to examine changes in the wages, employment, and effort of nurses in California hospitals following takeovers by large chains. The market for nurses has been described as a classic monopsony, so that one might expect increases in firm market power to be associated with declines in wages. However, we show that if one extends the monopsony model to consider effort, or if we apply a basic contracting model to the data, then we would expect to see effects on effort rather than on wages. This prediction is borne out by the data — nurses see few declines in wages following takeovers, but see increases in the number of patients per nurse, our measure of effort. We also find that these changes are similar in the largest for-profit and non-profit chains, suggesting that market forces are more important than institutional form.

Janet Currie
UCLA
Department of Economics
University of California, Los Angeles
90095-1477
and NBER
currie@simba.ss.cnet.ucla.edu

Mehdi Farsi
Department of Economics
University of Lugano
and Center for Energy Policy and Economics
Federal Institute of Technology
ETH Zentrum, WEC C20 8092
Zürich, Switzerland

W. Bentley MacLeod
Department of Economics and The Law School
University of Southern California
Los Angeles, CA 90089-0253
bentley@law.usc.edu
EFFICIENCY MEASUREMENT IN NETWORK INDUSTRIES:
APPLICATION TO THE SWISS RAILWAY COMPANIES *

Mehdi Farsi† Massimo Filippini† William Greene‡

† Center for Energy Policy and Economics
Swiss Federal Institute of Technology
ETH Zentrum, WEC, 8092 Zurich, Switzerland
and
Department of Economics, University of Lugano
Via Maderno 24, 6900 Lugano, Switzerland

‡ Department of Economics, Stern School of Business
New York University
44 West 4th St., New York, NY 10012, USA

November 2004
Original draft: May 2004

* The authors wish to thank Michael Crew and two anonymous referees for their helpful suggestions. We would also like to thank Aurelio Fetz for his assistance.
In 1996, following an alarming growth in government subsidies for railway transport in Switzerland, the federal government introduced a series of regulatory reforms. In particular, the subsidization of regional railway companies that was previously based on a full deficit coverage, has been replaced by an ex-ante fixed payment system. However, given that these subsidies are determined through a long series of negotiations and bargaining between railway companies and the corresponding cantonal governments, companies might use their local monopoly power to maintain high subsidies. Therefore, it is generally believed that the fixed payment system should be supported by a benchmarking practice to induce companies to minimize their costs. Benchmarking methods are mainly based on models of efficiency measurement. However, the reliability of such models in network industries has been often questioned. Network industries are characterized by a high degree of heterogeneity, much of which is network-specific and unobserved. The unaccounted-for heterogeneity can create bias in the inefficiency estimates. This paper examines the performance of several panel data models to measure cost and scale efficiency in network industries. The stochastic frontier models that include additional firm-specific effects, such as the random-constant frontier model proposed by Greene (2005), can control for unobserved network effects that are random but time-invariant. In cases like railway networks the unobserved heterogeneity is potentially correlated with other exogenous, but observed, factors such as network size and density. In such cases the correlation with explanatory variables may bias the coefficients of the cost function in a random-effects specification. However, these correlations can be integrated into the model using Mundlak’s (1978) formulation. The unobserved network effects and the resulting biases are studied through a comparative study of a series of stochastic frontier models. These models are applied to a panel of 50 railway companies operating over a 13-year period in Switzerland. Different specifications are compared regarding the estimation of both cost frontier coefficients and inefficiency scores.
1. INTRODUCTION

The railroad system in Switzerland consists of two sectors. The first sector includes the international and inter-regional transports. This sector is monopolized by the Swiss Federal Railways, which operates more than half of the railway networks in Switzerland. The second sector provides regional and local transport services that account for about a third of Switzerland’s railway passengers. Today, this sector consists of 49 Regional Railways (RR) with an average network length of approximately 40 kilometers. The main function of the RR is generally to link a rural region to an urban transport network or to the intercity railway network.

The RRs operate with a regional monopoly license given by the Confederation. This license defines the RR’s responsibilities, which imply an obligation to provide regular services according to a fixed time table and to apply the official tariff scheme. Moreover, the RR receives subsidies for their deficit in operating costs and the infrastructure investments from the Confederation and the cantons.

Given the strong increase of the subsidies for operations and infrastructures in 1996 the Swiss Government has introduced an important reform in the railways sector. The most relevant measures for the regional railway companies included in this reform are the change from the practice of ex-post deficit coverage to an ex-ante fixed payment system for transport services. The federal and cantonal governments commission transport companies for services

---

1 The regional railways are owned by different Swiss governments (Confederation, cantons and municipalities) and by some private investors. However, the share of private ownership is low.
2 Switzerland is a confederation composed of 26 cantons and approximately 3000 municipalities. Each canton has a high degree of autonomy in the organization, planning and regulation of the local public transport.
3 In 1995 the Swiss parliament approved the revision of the Railway Act (1995) which came into force on 1 January 1996. Moreover, in 1999, following the changes in the transport policy of the EC (Directive 91/440), further measures built on a European directive have been introduced: free access in freight transport, reinforcement of the fixed payment system and separation of infrastructure and transport services in terms of organization and accounting. The free access to freight transport does not apply to regional railways, which mainly provide short-distance passenger services.
on the basis of an estimated compensation defined in advance to cover the planned costs, which are not otherwise covered. Moreover, the reform introduced the possibility of organizing a competitive tendering procedure in the assignment of the licences, whereby the most performing railway companies would be incited to offer a public transport service satisfying the conditions imposed by the cantonal regulator. However, tendering is an optional measure and no cantons have used this possibility to date.

Although the ex-ante fixed payment rule represents an improvement with respect to the old subsidization practice, without a benchmarking analysis it does not contain incentives to minimize costs. Given that these subsidies are determined through a long series of negotiations and bargaining between railway companies and the corresponding cantonal governments, companies might use their local monopoly power to maintain high subsidies. Therefore, in the last years some cantonal authorities have begun to use simple benchmarking analysis of costs to determine the level of subsidies. The federal and cantonal authorities have been discussing the possibility to adopt more high-powered contracts based on yardstick competition model proposed by Shleifer (1985).4 In this context the use of cost frontier models could be useful as a complementary control instrument in the definition of the amount of subsidies granted to the regional railway companies.5

Since railway companies operate in different networks and environments, any method based on cost comparison has been subject to criticism. A high level of output heterogeneity is a general characteristic of network industries. Networks with different shapes have different organization and coordination problems, thus different costs. For instance, in the railway sector the production of 100 train-kilometers on a simple linear network is less costly than the same output in a Y-shaped network. Other factors such as the density of stops can also affect the costs. Furthermore, different environmental characteristics influence the production

---

process and therefore the costs. For instance, railway operation is more costly in a mountainous region than in a flat area. In general, the information is not available for all output and environmental characteristics. Many of these characteristics are therefore omitted from the cost function specifications.

Unobserved firm-specific heterogeneity can be taken into account with conventional fixed or random effects in a panel data model. In order to distinguish heterogeneities such as external network effects from cost efficiency, Greene (2005, 2004) proposed an approach that integrates an additional stochastic term representing inefficiency in both fixed and random effects models. In this paper we use a ‘true random-effects’ model, which is a random-constant frontier model, obtained by combining a conventional random-effects model with a skewed stochastic term representing inefficiency. The extended model includes separate stochastic terms for latent heterogeneity and inefficiency. Since many of the unobserved factors, especially those related to the network’s shape, are likely to be correlated with the output and perhaps other explanatory variables, the random-effect estimators of the cost function coefficients could be biased. To overcome this shortcoming, the ‘true random-effects’ model has been adjusted for correlation between unobserved heterogeneity and explanatory variables using Mundlak’s (1978) formulation.

The empirical results obtained from true random effects models in a variety of applications suggest that modeling unobserved heterogeneity could significantly decrease the inefficiency estimates. This could lend certain support to the application of benchmarking methods in the regulation of strongly heterogeneous network industries, in which the conventional inefficiency estimates appear to be overstated. Provided that they can

5 For an application of yardstick competition in the transport sector see Dalen and Gómez-Lobo (2003).
6 Kumbhakar (1991) proposed a similar approach using a three-stage estimation procedure. See Heshmati and Kumbhakar (1994) and Kumbhakar and Hjalmarsson (1995) for two applications of this model.
7 See Farsi, Filippini, Kuenzle (2003) for a discussion of Mundlak’s adjustment in frontier models.
sufficiently control for the unobserved heterogeneity across firms, these methods can be used to estimate an order of magnitude for the sector or individual companies’ cost-inefficiency. In addition, in the case of the Swiss regional railway sector, such analyses could be used to evaluate the subsidies for transport services.

The purpose of this paper is to study the potential advantages of these extended models in an application to Switzerland’s railway companies. In particular, our eventual interest is in models that can exploit the advantage of a fixed-effects model to have an unbiased estimate of the cost function without compromising the estimates of inefficiency scores. The models are estimated for a sample of 50 railway companies operating in Switzerland from 1985 to 1997. The alternative models are compared regarding the cost function slopes and inefficiency estimates. The conventional FE estimators of the cost function coefficients are assumed to be unbiased, thus used as a benchmark to which other models are compared. For the inefficiency estimates, the correlation between different models and the effect of econometric specification have been analyzed. The results suggest that the inefficiency estimates are substantially lower when the unobserved network effects are taken into account.

The rest of the paper is organized as follows: Sections 2 and 3 present the model specification and the methodology respectively. The data are explained in section 4. Section 5 presents the estimation results and discusses their implications, and section 5 provides the conclusions.

2. MODEL SPECIFICATION

A railway company can be considered as an aggregate production unit that operates in a given network and transforms labor and capital services and energy into units of transport services such as passenger-kilometers of public transport and ton-kilometers of freight. Given
the extremely high number and types of different transport services, the measure of output requires an aggregation of outputs in one way or another. A practical way of getting around this approximation is to include output characteristics such as network length or average haul in the model. Different strategies have been used in the literature. Caves et al. (1985) used passenger-miles and freight ton-miles as output, and controlled for the average lengths of trip for freight and passengers and the number of route miles as output characteristics. Filippini and Maggi (1993) have considered a single-output production function with the number of wagon-kilometers as a measure of output and included the network length in their model specification. In their international analysis, Cantos et al. (1999) considered the aggregate number of passenger-kilometers and ton-kilometers as two outputs. Todani (2001) considered three types of wagon-miles (high-valued, bulk and others) as three main outputs and accounted for average length of haul and the number of road miles as output characteristics.

In this paper a two-output production process is assumed. The outputs are transported passengers measured by the total number of passenger-kilometers in a given year, and the transported freight measured as the aggregate number of ton-kilometers. The length of network is included in the model as output characteristics. Three input factors are considered: labor, capital and energy. A total cost function has been considered.

---

9 In the case of railways each relation between any two points in the space could be defined as an output type. From a practical point of view it is not possible to estimate a multi-product cost function with so many outputs. Therefore, an aggregation process is inevitable.

10 Mizutani (2004) and Savage (1997) include in the cost model many kinds of output characteristics such as number of lines, load factor, station spacing. Unfortunately, information on these variables is not available to us.

11 In a preliminary analysis we also estimated a variable cost function. However, the results indicate a positive derivative of the variable cost function with respect to the capital stock, which violates the non-increasing regularity condition. Following Guyomard and Vermersch (1989) and Filippini (1996) we believe that this problem is due to the empirical difficulty in defining the capital stock variable. Due to lack of data we have used a physical measure of the capital stock, which is highly correlated with output and gives rise to a multi-collinearity problem. For this reason we preferred a total cost function, assuming that the companies can modify their capital expenses on a yearly basis. Insofar as this is equally applicable to all companies, the benchmarking analysis is not sensitive to such an assumption.
Based on the above specification the total cost frontier can be represented by the following cost function:

$$TC = f(Y, Q, N, P_K, P_L, P_E, d_i)$$  \(1\)

where \(TC\) is the total annual costs; \(Y\) and \(Q\) are the numbers of passenger-kilometers and ton-kilometers respectively; \(P_K\), \(P_L\) and \(P_E\) are respectively the prices of capital, labor and energy; \(N\) is the length of network and \(d_i\) is a vector including 12 year dummies from 1986 to 1997 (year 1985 is the omitted category). The year dummies capture the cost changes associated with technical progress as well as other unobserved year-specific factors.\(^{12}\)

It is generally assumed that the cost function given in (1) is the result of cost minimization given input prices and output and should therefore satisfy certain properties.\(^{13}\) Mainly, this function must be non-decreasing, concave, linearly homogeneous in input prices and non-decreasing in output. To estimate the cost function (1), a Cobb-Douglas (log-linear) functional form is employed. We also evaluated the possibility of applying a translog functional form that can account for variation of scale economies with output. However, we decided to exclude this model because it requires a relatively large number of parameters, which creates certain numerical problems in the simulated likelihood maximization for the random-constant model. Moreover, our preliminary estimations (not reported here) showed that this functional form resulted in counter-intuitive results for the sign of output variables. This is perhaps due to multicollinearity problems caused by strong correlation between the second order terms in translog form. Finally, given that most of the Swiss regional railway

\(^{12}\) In the cost function estimations it is common to use a linear trend for technical progress. However, our preliminary regressions indicated that the time-variation of costs is strongly non-linear. In fact there is a gradual increase in the beginning of the sample period followed by a decrease in costs. These variations can be explained by many unobserved factors (such as changes in collective labor contracts or seasonal composition of the demand) that change uniformly across companies.
companies are relatively small the assumption that the value of scale economies does not vary with output (implicit in Cobb-Douglas form) is not very restrictive.

The concavity assumption is automatically satisfied in Cobb-Douglas form. The linear homogeneity restriction can be imposed by normalizing the costs and prices by the price of one of the input factors. Here we considered the energy as the numeraire good. The other theoretical restrictions are verified after the estimation. The cost function can therefore be written as:

\[
\ln\left(\frac{TC_{it}}{P_{E_{it}}}\right) = \alpha_0 + \alpha_Y \ln Y_{it} + \alpha_Q \ln Q_{it} + \alpha_N \ln N_{it} + \alpha_S \ln S_{it} \\
+ \alpha_K \ln \frac{P_{K_{it}}}{P_{E_{it}}} + \alpha_L \ln \frac{P_{L_{it}}}{P_{E_{it}}} + \sum_{t=1986}^{1997} \alpha_d d_t + \alpha_i + \varepsilon_{it}
\]  

(2)

with \( i = 1, 2, \ldots, N \) and \( t = 1, 2, \ldots, T_i \)

Subscripts \( i \) and \( t \) denote the company and year respectively, \( \alpha_i \) is a firm-specific effect and \( \varepsilon_{it} \) is an \( iid \) error term. As we will explain in the next section, in the recent models proposed by Greene (2005), the stochastic term \( \varepsilon_{it} \) is composed of two parts: a skewed component representing inefficiency and a symmetric part for the random noise.

3. ECONOMETRIC MODELS

Stochastic frontier models have been subject of a great body of literature resulting in a large number of econometric models to estimate cost functions. Kumbhakar and Lovell (2000) provide an extensive survey of this literature. The main models used in this paper are based on Greene’s (2005) extension of the original frontier approach proposed by Aigner et

---

13 For more details on the functional form of the cost function see Cornes (1992), p.106.
In this framework, \( \varepsilon_{it} \) as given in specification (2), is assumed to be a composite stochastic term with a normal-half-normal distribution, including both idiosyncratic effects and inefficiencies. The additional firm-specific term, \( \alpha_i \) in equation (2), represents the unobserved heterogeneity and is assumed to have a normal distribution. This model is referred to as a “true” random-effects model.\(^{14}\) The estimation method is based on simulated maximum likelihood.

The results are compared with other alternative models such as the fixed-effects model proposed by Schmidt and Sickles (1984) and the random-effects model proposed by Pitt and Lee (1981). Both these models are covered by the general form given in (2) with the difference that in the former model \( \alpha_i \) is a fixed effect and \( \varepsilon_{it} \) is a zero-mean error term with no distribution restriction, and in the latter (Pitt and Lee) model \( \alpha_i \) is a random effect with half-normal (or truncated normal) distribution and \( \varepsilon_{it} \) is a normal random error term.

A summary of the five models used in the paper is given in table 1. The first model is a fixed effects (FE) model. In this model the firm-specific effects are considered as constant parameters that can be correlated with the explanatory variables. The coefficients are estimated through “within-firm” variations and therefore, are not affected by heterogeneity bias.\(^{15}\) In the cost frontier literature the inefficiency scores are estimated as the distance from the firm with the minimum estimated fixed effect, that is \( \hat{\alpha}_i - \min(\hat{\alpha}_i) \), as proposed by Schmidt and Sickles (1984).

\(^{14}\) The name “true” is chosen to show that the model keeps the original frontier framework and the extension is done only by including an additional heterogeneity term.

\(^{15}\) The term “heterogeneity bias” was used by Chamberlain (1982) to refer to the bias induced by the correlation between individual effects and explanatory variables in a random-effects model. See also Baltagi (2001) for an extensive discussion of fixed-effects (within) estimators.
Table 1. Econometric specifications of the stochastic cost frontier

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
<th>True RE with Mundlak adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>RE</td>
<td>Pooled</td>
<td>True RE</td>
<td></td>
</tr>
<tr>
<td>Firm-specific</td>
<td>Constant</td>
<td>Half-normal</td>
<td>None</td>
<td>N(0, σ²)</td>
<td></td>
</tr>
<tr>
<td>component αₗ</td>
<td></td>
<td>N(0, σ²)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random error</td>
<td>iid (0, σ²)</td>
<td>iid (0, σ²)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>εₗ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inefficiency</td>
<td>[\hat{\alpha}_l = \min{\hat{\alpha}_i} ]</td>
<td>[\alpha_l = \gamma \bar{X}_l + \delta_l]</td>
<td>[\delta_l \sim N(0, \sigma^2_\delta)]</td>
<td>[\bar{X}<em>l = \frac{1}{T} \sum</em>{t=1}^{T} X_{it}]</td>
<td></td>
</tr>
</tbody>
</table>

Model II is a random effects (RE) model proposed by Pitt and Lee (1981), which is estimated using the maximum likelihood method. The firm’s inefficiency is estimated using the conditional mean of the inefficiency term proposed by Jondrow et al. (1982), 16 that is:

\[ E[\alpha_l | \omega_{l1}, \omega_{l2}, \ldots] = E[\alpha_l | \omega_l] \] where \( \omega_{li} = \alpha_i + \varepsilon_{li} \) and \( \sigma_{\omega_l} = \frac{1}{T} \sum_{i=1}^{T} \omega_{li} \). A limitation of this model is the assumption that the firm-specific stochastic term \( \alpha_i \) is uncorrelated with the explanatory variables. Moreover, in both models (I and II), inefficiency indicators may include unobserved environmental factors, thus may overstate the firms’ inefficiency. There are however two factors that may exacerbate this problem in the FE model. First, unlike the RE model, the firm-specific effects do not follow a single distribution, thus can have a relatively wide range of variation. Secondly, these effects can be correlated with the explanatory variables, thus can also capture the heterogeneity factors that are correlated with the regressors. Whereas in the RE model in which the firm-specific effects are by construction

---

16 See also Greene (2002).
uncorrelated with the regressors, these factors are suppressed at least partially through the “between” variations, into the regression coefficients.

In the first two models (I and II), the firm’s inefficiency is assumed to be constant over time, thus captured by the firm-specific effects, while in other models inefficiency can vary across years. Model III is a pooled frontier model in that the sample is considered as a cross-section and its panel aspect is neglected. The random error term is divided into two components: a normal error term $v_{it}$ capturing the noise and a half-normal random term $u_{it}$ representing the inefficiency as a one-sided non-negative disturbance. This model is based on the original cost frontier model proposed by Aigner et al. (1977). The firm’s inefficiency is estimated using the conditional mean of the inefficiency term $E[u_{it}|v_{it}+v_{it}]$, proposed by Jondrow et al. (1982).

Models IV and V are extensions to model III that include an additional firm-specific random effect ($\alpha_i$) to represent the unobserved heterogeneity among firms. Model IV is Greene’s (2005) true RE model. In this model it is assumed that the unobserved cost differences across firms that remain constant over time, are driven by network-related unobserved characteristics rather than inefficiency. Given the relatively long period covered in the data (12 years on average), this is a realistic assumption. The inefficiency term is assumed to be an iid random variable with half-normal distribution. This implies that the inefficiency is not persistent and each period brings about new idiosyncratic elements thus new sources of inefficiency. This is a reasonable assumption particularly in industries that are constantly facing new technologies. Therefore there are two justifications for such a specification in network industries: The first one is a practical assumption that persistent cost differences are related to unobserved heterogeneity across networks and the second one is based on the conjecture that the sources of inefficiency in network industries are dominated by new technology shocks and the incomplete adaptation of managers facing them.
Model V is an extension of model IV that uses Mundlak’s (1978) specification to account for the potential correlation of unobserved firm-specific heterogeneity with the explanatory variables. Mundlak’s adjustment can be written as an auxiliary regression given by:

\[
\alpha_i = \gamma \bar{X}_i + \delta_i \quad \bar{X}_i = \frac{1}{T} \sum_{t=1}^{T} X_{it}, \quad \delta_i \sim N(0, \sigma^2_\delta)
\]

where \(X_{it}\) is the vector of all explanatory variables and \(\gamma\) is the corresponding vector of coefficients. Equation (3) actually divides the firm-specific stochastic term into two components: The first part can be explained by exogenous variables, whereas the remaining component \((\delta_i)\) is orthogonal to explanatory variables. The advantage of this model is that it allows for a time-variant inefficiency term while minimizing the heterogeneity bias. The heterogeneity bias can be avoided to the extent that the auxiliary equation can capture the correlations.

In our comparative analysis we consider two aspects of the models’ performance. The first dimension is the estimation of the cost function’s coefficients. In railway companies the operating costs are affected by network characteristics, which may be correlated with explanatory variables such as network’s size and input factor prices. For instance, larger networks are more likely to have more complex shapes. Denser networks are usually located in areas with higher population density, where wages are relatively high. Such relationships imply a positive correlation between the output level and labor price with the network complexity, which is not fully captured by the included factors in the model. The Hausman test is used to confirm that the firm-specific effects are correlated with the explanatory variables. In this case the FE estimators (model I) are unbiased, thus provide a benchmark to which other models can be compared.

See also Hsiao (2003), pp. 44-46, for an extensive discussion of Mundlak’s formulation.
The second aspect of the models’ performance concerns the inefficiency estimates. It is important to note that the consistency of slopes (coefficients) does not necessarily imply that inefficiency estimates are unbiased. Interestingly, the empirical results suggest that there is a trade-off in estimations. Namely, models (like the FE model) with a good performance on slopes have strongly biased inefficiency estimates. Roughly speaking, the heterogeneity bias may be suppressed into the slopes as it appears in the RE model, or into the efficiency estimates as observed in the FE model. Farsi et al. (2003) provide a discussion on this issue. The results of that study on a sample of nursing homes suggest that Mundlak’s formulation can be helpful to reduce the heterogeneity bias in both slopes and inefficiency estimates at the same time. In this paper we use a similar approach to study if such a conclusion can be applied to a network industry.

It should be noted that the inefficiency estimation requires a certain interpretation of the stochastic terms in the model. In the frontier literature, starting from Aigner et al. (1977), it is commonly accepted that the skewed stochastic term with a certain distribution represents inefficiency. Carree (2002) discusses some of the implications of such distribution assumptions. For instance a half-normal distribution through its zero mode, implies that any company is most likely to be completely efficient. Moreover, implicit in this model is the assumption that inefficiency is uncorrelated with all exogenous variables and also with the idiosyncratic variations reflected in the symmetric error term. In fact, through this assumption all the inefficiencies that are somehow related to exogenous variables such as factor prices and output are excluded from the firm’s productive inefficiency. Later studies like Cornwell et al. (1990) and Battese and Coelli (1992) extended the original framework to include exogenous variables in the distribution of the inefficiency term. However, in this

---

18 See Farsi et al. (2003) for a discussion of this point.
19 Here, cost inefficiency is defined as the excess costs due to the firm’s technical problems or to suboptimal allocation of resources. Thus, scale inefficiencies, which are related to suboptimal output, are excluded.
paper we maintain the original assumption such that the efficiency measures are restricted to the sources that are completely uncorrelated with all exogenous variables, which by definition are beyond the firm’s control. The only exception is the FE model (model I) that allows any correlation of inefficiency scores. Furthermore, we assume that the inefficiency can vary over time, thus for the inefficiency estimates we focus on models III, IV and V.

4. DATA

The data set used in this paper is extracted from the annual reports of the Swiss Federal Office of Statistics on public transport companies. The companies operating in main urban centers are excluded from the sample. Most of these companies operate inner-city tramways and buses, whose functioning is quite different from trains. We also excluded one other company whose extremely low total costs and energy expenses suggest the possibility of a reporting error. The final sample includes 50 railway companies over a 13-year period from 1985 to 1997. The sample is an unbalanced panel with number of periods ($T_i$) varying from 1 to 13 and with 45 companies with 12 or 13 years, resulting in 605 observations in total.\(^{20}\) The available information for any given year includes total costs, labor and energy expenses separately, total number of employees, the quantity of consumed electricity, network length, total number of seats, and total number of train-kilometers, passenger-kilometers and ton-kilometers.

Capital costs are calculated as the residual costs after deducting the labor and energy expenses from the total costs. These costs are mainly related to equipment and materials.

\(^{20}\) The average number of periods in the sample is 12 years. For 37 companies, the data are available for 13 years. Eight other companies have 12 years available. The number of years available for the remaining five companies is respectively 1, 3, 7, 7 and 10.
Total number of seats is used as a proxy for capital stock. Thus, the capital price is calculated as the residual expenses per seat. The passenger and freight outputs are respectively measured by the number of passenger-kilometers and ton-kilometers. In Switzerland, each railway company is required to run a certain minimum number of trips per day for any given connection, specified by the cantonal regulators. Therefore, the number of train-kilometers or wagon-kilometers could be also an appropriate measure of passenger output. However, in order to be consistent with the recent literature and also given that there is a high correlation between train-kilometers and passenger-kilometers (a correlation coefficient of 0.97 in our sample) we adopted the number of passenger-kilometers and ton-kilometers. All the costs and prices are adjusted for inflation using the Switzerland’s global price index and are measured in 1997 Swiss Francs.

Table 2 provides a descriptive summary of the main variables used in the analysis. As it can be seen in this table, the total costs show a high variation in the sample. The average cost of a passenger-kilometer varies from 0.3 to about 6 Swiss Francs. There is also a considerable variation in input prices and both outputs in the sample. Given the importance of within variations in most panel data models (especially the fixed-effect model), it is helpful to distinguish these variations from the variations across companies. Table 3 gives a summary of “within” and “between” variations for the main variables used in the regressions. As it can be seen in this table, the dependent variable and most explanatory variables show a fairly considerable amount of within variation, supporting the use of a fixed-effect model. As expected, the within variation of network length is relatively low (limited to 7 percent).

---

21 See Filippini and Prioni (2003) for a similar approach.
22 Some recent examples are Mancuso and Reverberi (2003), Estache et al. (2002), Cantos et al. (1999) and Banos-Pino et al. (2002).
Table 2. Descriptive statistics (605 observations)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs (TC) CHF million</td>
<td>26.73</td>
<td>49.88</td>
<td>8.83</td>
<td>2.12</td>
<td>307.43</td>
</tr>
<tr>
<td>Passenger output (Y) ×10⁶ passenger-kms</td>
<td>30.80</td>
<td>55.10</td>
<td>10.00</td>
<td>0.41</td>
<td>311.00</td>
</tr>
<tr>
<td>Average cost (CHF per passenger-km)</td>
<td>1.20</td>
<td>0.76</td>
<td>1.09</td>
<td>0.33</td>
<td>5.98</td>
</tr>
<tr>
<td>Freight output (Q) ×10⁶ ton-kilometers</td>
<td>10.20</td>
<td>52.70</td>
<td>0.27</td>
<td>0.00015</td>
<td>477.00</td>
</tr>
<tr>
<td>Network length (N) (km)</td>
<td>39.43</td>
<td>56.64</td>
<td>22.82</td>
<td>3.90</td>
<td>377.00</td>
</tr>
<tr>
<td>Capital price (P_K) per seat (CHF '000)</td>
<td>4.53</td>
<td>2.13</td>
<td>4.03</td>
<td>1.04</td>
<td>14.47</td>
</tr>
<tr>
<td>Average labor price (P_L) per employee per year (CHF '000)</td>
<td>86.05</td>
<td>6.48</td>
<td>86.09</td>
<td>60.93</td>
<td>104.93</td>
</tr>
<tr>
<td>Energy (electricity) price (P_E) CHF/ kWh</td>
<td>0.157</td>
<td>0.023</td>
<td>0.158</td>
<td>0.076</td>
<td>0.265</td>
</tr>
</tbody>
</table>

- All monetary values are in 1997 Swiss Francs (CHF), adjusted for inflation by Switzerland’s global consumer price index.
Table 3. Within and between variations (50 companies and 12 years on average)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Fraction of within variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>Between</td>
</tr>
<tr>
<td>( \ln \left( \frac{TC}{P_E} \right) )</td>
<td>11.31</td>
<td>1.10</td>
<td>1.12</td>
</tr>
<tr>
<td>( \ln (Y) )</td>
<td>16.32</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>( \ln (Q) )</td>
<td>12.49</td>
<td>2.72</td>
<td>2.78</td>
</tr>
<tr>
<td>( \ln (N) )</td>
<td>3.20</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>( \ln \left( \frac{P_E}{P_P} \right) )</td>
<td>10.18</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>( \ln \left( \frac{P_L}{P_E} \right) )</td>
<td>13.22</td>
<td>0.16</td>
<td>0.13</td>
</tr>
</tbody>
</table>

- For each variable \( (X) \) the between standard deviation is based on companies’ average values that is: \( \bar{X}_j = \frac{1}{T_j} \sum_{t=1}^{T_j} X_{jt} \); and the within standard deviation is based on deviations from companies’ averages \( (X_{jt} - \bar{X}_j). \) The overall and within statistics are calculated over 605 company-years and the between statistics are calculated over 50 companies. The fraction of within variation is calculated as the ratio of within to overall standard deviation.

5. ESTIMATION RESULTS

The estimation results for the five models are given in table 4. These results show that the output and input price coefficients are positive and highly significant across all models. The estimated coefficients show a considerable variation across different models. The estimates from the pooled model (III) are particularly different from those of other models. The year dummies are mostly significant and suggest that the cost variation over time is not linear. Again, the pooled model is an exception in which none of these dummies show any statistically significant effect. Noting that model III completely ignores the panel structure of the data, its estimates are likely to be strongly biased by omitted firm-specific variables. On
the other hand the fixed-effects estimators (model I) are derived from the within-firm variations and thus unbiased.

The year dummy coefficients (excluding model III) show that the total costs of railway companies rose almost linearly from 1985 to 1992 with an average annual growth rate of about 1.6%, but declined after 1992 with an average rate of about 1.5% per year. Since total costs and all the continuous explanatory variables are in logarithms, the estimated coefficients can be interpreted as average cost elasticities. For instance, the output coefficients suggest that on average a one percent increase in passenger-kilometers will increase the costs by 0.11 to 0.49 percent depending on the adopted specification. The marginal effect of ton-kilometers is about 10 times lower, suggesting substantially lower variable costs for freight transportation. The coefficient of network length indicates that the marginal cost of a one percent extension in the network keeping the output constant, is approximately equivalent to 0.4 percent increase in costs. These results are consistent with the previous empirical results regarding Switzerland’s railroad industry (cf. Filippini and Maggi, 1993) in that they suggest increasing returns to scale.

Table 4 also indicates that if the pooled model is set aside, the input price coefficients do not vary significantly across different models. The coefficient of labor price, varying between .55 and .57 (bar model III), is actually comparable to the average share of labor expenses, which is about 52% in the sample. The capital price coefficient varies between .31 and .32 (model III excluded), which is considerably below the average share of capital costs in the sample (44%). This result may suggest that the companies are not so responsive as a constantly cost minimizing behavior should be, to the changes in capital prices. This can be explained by the fact that in the short run railway companies cannot vary much of their capital stock such as equipment and machinery.
Comparing the results from different models in table 4 shows that excluding model III, all other models have reasonably comparable coefficients. In model III (pooled model) variations over time and within firms are treated exactly similar to those between different firms. Moreover, the unobserved firm-specific effects are completely neglected, which may bias the estimations. A Lagrange Multiplier test on an OLS model strongly rejects the hypothesis that the residuals of a given company are uncorrelated (test statistic of 2990 for a chi-square with 1 degree of freedom), suggesting that the pooled model is mis-specified. Moreover, the Hausman test rejects the hypothesis that the firm-specific effects are uncorrelated with the explanatory variables (test statistic of 61.5 for a chi-square with 17 degrees of freedom). This result suggests that models that do not account for these correlations can give biased results. Given the relatively high number of periods (on average 12 years) and the reasonable within-company variations (see table 3) in the sample, the fixed effects model’s results can be considered as unbiased estimates of the cost function parameters. Therefore, the coefficients estimated from model I are used as a benchmark for assessing the potential heterogeneity bias in other models.

Compared to model I, the parameter estimates in the pooled model (III) have the highest differences. The estimated coefficients in the remaining models are fairly close to those of the FE model, suggesting that heterogeneity biases in the coefficients are not substantial. This statement does not apply to the inefficiency estimates, which as we will see later, show considerable biases. As seen in table 4, there is no clear distinction between models II and IV concerning the heterogeneity biases. While in certain coefficients model IV is closer to the unbiased estimates (model I), in some others model II shows a ‘better’ performance.

The random effects specification in both models II and IV has however a shortcoming in that the firm-specific heterogeneity terms ($u_i$ in model II and $\alpha_i$ in model IV) are assumed to be uncorrelated with the explanatory variables. If we put any trust in the Hausman
specification test, this assumption is not realistic. Moreover, as discussed earlier, it is plausible that some of the unobserved network characteristics be correlated with the network length. Such correlations are taken into account in model $V$ through the auxiliary coefficients ($\gamma$). The results in table 4 indicate that model $V$ shows the smallest differences with the unbiased estimators of model $I$. This suggests that applying Mundlak’s (1978) adjustment to the TRE model (model $IV$) can decrease the heterogeneity biases. As shown in the table, the auxiliary coefficients ($\gamma$) are all significant. These coefficients can be interpreted as the correlation effect between the unobserved firm characteristics and the corresponding explanatory variable. For instance, the positive signs of $\gamma_Y$ and $\gamma_Q$ suggest that keeping all observed factors fixed, networks with higher outputs are more likely to belong to the ‘high-cost’ or ‘difficult’ networks; and the negative signs of $\gamma_N$, $\gamma_K$ and $\gamma_L$ suggest that larger networks and companies that have higher input prices are more likely to be in the ‘low-cost’ category.
<table>
<thead>
<tr>
<th></th>
<th>Model I FE</th>
<th>Model II RE</th>
<th>Model III Pooled</th>
<th>Model IV True RE</th>
<th>Model V True RE + Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_Y$</td>
<td>.114*</td>
<td>.200*</td>
<td>.492*</td>
<td>.133*</td>
<td>.106*</td>
</tr>
<tr>
<td></td>
<td>(.032)</td>
<td>(.030)</td>
<td>(.015)</td>
<td>(.023)</td>
<td>(.034)</td>
</tr>
<tr>
<td>$\alpha_Q$</td>
<td>.014*</td>
<td>.021*</td>
<td>.030*</td>
<td>.038*</td>
<td>.017*</td>
</tr>
<tr>
<td></td>
<td>(.006)</td>
<td>(.003)</td>
<td>(.006)</td>
<td>(.004)</td>
<td>(.003)</td>
</tr>
<tr>
<td>$\alpha_N$</td>
<td>.448*</td>
<td>.485*</td>
<td>.393*</td>
<td>.432*</td>
<td>.488*</td>
</tr>
<tr>
<td></td>
<td>(.051)</td>
<td>(.039)</td>
<td>(.026)</td>
<td>(.015)</td>
<td>(.035)</td>
</tr>
<tr>
<td>$\alpha_K$</td>
<td>.318*</td>
<td>.310*</td>
<td>.171*</td>
<td>.312*</td>
<td>.315*</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.010)</td>
<td>(.032)</td>
<td>(.008)</td>
<td>(.009)</td>
</tr>
<tr>
<td>$\alpha_L$</td>
<td>.546*</td>
<td>.548*</td>
<td>.592*</td>
<td>.568*</td>
<td>.562*</td>
</tr>
<tr>
<td></td>
<td>(.037)</td>
<td>(.029)</td>
<td>(.074)</td>
<td>(.036)</td>
<td>(.034)</td>
</tr>
<tr>
<td>$\gamma_Y$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.159*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.050)</td>
</tr>
<tr>
<td>$\gamma_Q$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.090*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.013)</td>
</tr>
<tr>
<td>$\gamma_N$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>-.150*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.056)</td>
</tr>
<tr>
<td>$\gamma_K$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>-.189*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.067)</td>
</tr>
<tr>
<td>$\gamma_L$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>-.193</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.180)</td>
</tr>
<tr>
<td>$\alpha_{1986}$</td>
<td>0.100</td>
<td>0.009</td>
<td>0.009</td>
<td>0.022</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>(.015)</td>
<td>(.041)</td>
<td>(.056)</td>
<td>(.027)</td>
<td>(.035)</td>
</tr>
<tr>
<td>$\alpha_{1987}$</td>
<td>.020</td>
<td>.012</td>
<td>.003</td>
<td>.032</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>(.015)</td>
<td>(.031)</td>
<td>(.056)</td>
<td>(.025)</td>
<td>(.031)</td>
</tr>
<tr>
<td>$\alpha_{1988}$</td>
<td>.039*</td>
<td>.028</td>
<td>.010</td>
<td>.051</td>
<td>.049</td>
</tr>
<tr>
<td></td>
<td>(.015)</td>
<td>(.044)</td>
<td>(.057)</td>
<td>(.037)</td>
<td>(.050)</td>
</tr>
<tr>
<td>$\alpha_{1989}$</td>
<td>.065*</td>
<td>.052</td>
<td>.036</td>
<td>.076*</td>
<td>.074</td>
</tr>
<tr>
<td></td>
<td>(.016)</td>
<td>(.046)</td>
<td>(.057)</td>
<td>(.033)</td>
<td>(.050)</td>
</tr>
<tr>
<td>$\alpha_{1990}$</td>
<td>.084*</td>
<td>.068</td>
<td>.024</td>
<td>.097*</td>
<td>.094*</td>
</tr>
<tr>
<td></td>
<td>(.016)</td>
<td>(.036)</td>
<td>(.058)</td>
<td>(.034)</td>
<td>(.044)</td>
</tr>
<tr>
<td>$\alpha_{1991}$</td>
<td>.098*</td>
<td>.078*</td>
<td>.030</td>
<td>.114*</td>
<td>.111*</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.029)</td>
<td>(.058)</td>
<td>(.028)</td>
<td>(.033)</td>
</tr>
<tr>
<td>$\alpha_{1992}$</td>
<td>.111*</td>
<td>.094*</td>
<td>.046</td>
<td>.130*</td>
<td>.122*</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.034)</td>
<td>(.058)</td>
<td>(.026)</td>
<td>(.034)</td>
</tr>
<tr>
<td>$\alpha_{1993}$</td>
<td>.100*</td>
<td>.081*</td>
<td>.015</td>
<td>.119*</td>
<td>.112*</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.034)</td>
<td>(.057)</td>
<td>(.026)</td>
<td>(.034)</td>
</tr>
<tr>
<td>$\alpha_{1994}$</td>
<td>.082*</td>
<td>.063</td>
<td>-.001</td>
<td>.103*</td>
<td>.091*</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.040)</td>
<td>(.056)</td>
<td>(.037)</td>
<td>(.039)</td>
</tr>
<tr>
<td>$\alpha_{1995}$</td>
<td>.059*</td>
<td>.048</td>
<td>.019</td>
<td>.081*</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td>(.016)</td>
<td>(.032)</td>
<td>(.057)</td>
<td>(.023)</td>
<td>(.034)</td>
</tr>
<tr>
<td>$\alpha_{1996}$</td>
<td>.037*</td>
<td>.028</td>
<td>.027</td>
<td>.066*</td>
<td>.043</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.024)</td>
<td>(.057)</td>
<td>(.022)</td>
<td>(.025)</td>
</tr>
<tr>
<td>$\alpha_{1997}$</td>
<td>.038*</td>
<td>.030</td>
<td>.019</td>
<td>.063</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td>(.018)</td>
<td>(.032)</td>
<td>(.060)</td>
<td>(.039)</td>
<td>(.032)</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>–</td>
<td>-4.90*</td>
<td>-8.31*</td>
<td>-3.89*</td>
<td>-1.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.57)</td>
<td>(.98)</td>
<td>(.51)</td>
<td>(2.66)</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.783*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.027)</td>
</tr>
<tr>
<td>$\sigma = \sqrt{\sigma_0^2 + \sigma_v^2}$</td>
<td>.807*</td>
<td>.464*</td>
<td>.109*</td>
<td>.095*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.14)</td>
<td>(.001)</td>
<td>(.005)</td>
<td>(.005)</td>
<td>(.005)</td>
</tr>
<tr>
<td>$\lambda = \sigma_v / \sigma_y$</td>
<td>11.37*</td>
<td>2.88*</td>
<td>2.58*</td>
<td>1.59*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.81)</td>
<td>(.30)</td>
<td>(.56)</td>
<td>(.031)</td>
<td>(.031)</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets. * means significant at less than 5%.
- The sample includes 605 observations (50 railway companies).
Table 5 provides a descriptive summary of the inefficiency estimates from different models (see table 1, last row). These estimates represent the relative excess cost of a given firm compared to a minimum level that would have been achieved if the firm had operated as efficiently as the ‘best practice’ observed in the sample. In comparing different models it should be noted that in the first two models (I and II), the inefficiency is assumed to be constant over time. Moreover, in these models all the unobserved firm-specific differences are interpreted as inefficiency. As expected, both models I and II, especially the FE model, predict rather unrealistic inefficiency scores averaging about .7 to .8 and up to a maximum of 2 to 2.5. According to these models, a typical company can save about a third of its costs by a more efficient allocation of resources. These high values indicate that the heterogeneity across companies is an important driver of cost differences and that neglecting it may create a substantial upward bias in inefficiency scores.

<table>
<thead>
<tr>
<th></th>
<th>Model I FE</th>
<th>Model II RE</th>
<th>Model III Pooled</th>
<th>Model IV True RE</th>
<th>Model V True RE with Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.813</td>
<td>0.696</td>
<td>0.343</td>
<td>0.078</td>
<td>0.063</td>
</tr>
<tr>
<td>Median</td>
<td>0.676</td>
<td>0.662</td>
<td>0.289</td>
<td>0.061</td>
<td>0.053</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.507</td>
<td>1.992</td>
<td>0.848</td>
<td>0.386</td>
<td>0.311</td>
</tr>
<tr>
<td>95 percentile</td>
<td>1.723</td>
<td>1.470</td>
<td>0.848</td>
<td>0.187</td>
<td>0.134</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.160</td>
<td>0.060</td>
<td>0.011</td>
<td>0.012</td>
</tr>
<tr>
<td>N</td>
<td>605</td>
<td>605</td>
<td>605</td>
<td>605</td>
<td>605</td>
</tr>
</tbody>
</table>

In model III the inefficiency estimates are in a more realistic range, with an average of .34 and a maximum value of .85. These values though still too high to be convincing, are
substantially lower than those predicted by models I and II; and this despite the fact that the pooled model (III) does not account for unobserved heterogeneity. This attenuation of inefficiency estimates can be explained by the structure of the inefficiency term in model III. Given that the inefficiency term \( u_{it} \) is assumed to be independently and identically distributed over time and across companies, it cannot fully capture the firm-specific differences that are time-invariant, thus such differences are partly suppressed into and bias the model’s coefficients.

Both models IV and V, which have separate stochastic terms for inefficiency and firm-specific heterogeneity, have quite reasonable inefficiency estimates about 6 to 8 percent on average and 31 to 38 percent on maximum. The substantial decrease in these values compared to other models, suggests that these models can separate to a considerable extent, the heterogeneity from the inefficiency. To understand the reasons behind these results, it is helpful to note that the sole difference between models III and IV is that model IV includes an additional firm-specific random term \( \alpha_i \). This term represents the variations across firms, which are about 7 times larger than the variation within firms (compare \( \sigma_\alpha \) to \( \sigma \) in the lower panel of table 4).

Given that the unobserved heterogeneity is potentially correlated with the explanatory variables, and that these correlations are not taken into account in model IV the resulting inefficiency scores may capture some of these differences. This issue can be explored by comparing models IV and V. In model V the time-invariant cost differences across companies are separated from inefficiency estimates (as in model IV). In addition, the possible correlations with explanatory variables are mitigated through auxiliary coefficients. The results in table 5 show that when such correlations are controlled for (model V), the inefficiency estimates slightly decline (by about .015 on average and by .075 on maximum). According to this model the average (median) company is only 6.3 (5.3) percent inefficient,
and the maximum inefficiency in 95 percent of the sample is limited to 13.4 percent. These results suggest that model \( V \) not only provides unbiased, or close to unbiased, estimates of the cost function’s coefficients, it can also better separate the heterogeneity from inefficiency.

The pair-wise correlation coefficients between the inefficiency estimates from different models are listed in table 6. In order for the correlation coefficients to be comparable, they are calculated at the firm level using 50 observations (one observation for each firm). Namely, in models with time-variant efficiency, the inefficiency score is calculated as the firm’s average inefficiency score over the sample period. For models with time-variant inefficiency the correlation coefficients are also given over the 605 observations.

As shown in table 6, models \( I \) and \( II \), and models \( IV \) and \( V \) show a relatively high correlation.\(^{23}\) However, except a few cases the correlation coefficients are quite low, suggesting substantial differences across models.\(^{24}\) Especially, models \( IV \) and \( V \) show a negative correlation with all other models. Given that the correlation coefficients are calculated on company-average inefficiency scores, the weak (and negative) correlations may suggest that the inefficiency estimates vary considerably from one year to another, in which case the correlation between models with constant and time-variant inefficiency should be weak. However, this can only partly explain the observed correlations. In fact the positive and fairly strong correlation between the pooled model \( III \) (with time-variant efficiency) and both models \( I \) and \( II \) (with time-invariant efficiency) indicates that averaging cannot explain the negative correlations.

The negative correlation coefficients (table 6) point to a striking distinction between the models \( IV \) and \( V \) and all other models, which do not distinguish unobserved heterogeneity from inefficiency. The negative correlations manifest especially in model \( V \) in which the

\(^{23}\) These results are consistent with Farsi et al. (2003) who used a similar methodology for a sample of nursing homes.

\(^{24}\) The rank correlations show similar patterns. These results are omitted to avoid repetition.
correlations with observed factors are taken into account. These values suggest that some of
the unobserved network characteristics may actually be negatively correlated with company’s
average inefficiency. One interpretation is that the relatively complex thus costly networks are
more likely to be operated by an efficient management. This is a plausible explanation
because the companies with complex networks are more likely to have a general awareness
and perhaps the required expertise for technical problems. Such expertise can directly or
indirectly contribute to the firm’s efficiency. The results in table 6 highlight the importance of
unobserved heterogeneity, as failure to account for such factors can result in a completely
misleading and even reverse picture of inefficiencies.

Table 6. Pair-wise correlation between inefficiency estimates

<table>
<thead>
<tr>
<th></th>
<th>Model I FE</th>
<th>Model II RE</th>
<th>Model III Pooled</th>
<th>Model IV True RE</th>
<th>Model V True RE with Mundlak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model II</td>
<td>.932*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model III</td>
<td>.497*</td>
<td>.614*</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model IV</td>
<td>-.247</td>
<td>-.256</td>
<td>-.158 [0.092*]</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Model V</td>
<td>-.334*</td>
<td>-.320*</td>
<td>-.197 [.105*]</td>
<td>0.948* [.971*]</td>
<td>1</td>
</tr>
</tbody>
</table>

- The correlation coefficients have been estimated over the firms (50 observations) that is, average
  values over the sample period are used in models with time-variant inefficiency (III, IV and V).
- Correlation coefficients based on 605 observations are given in brackets.
- * means significant at 5%.

The estimation of a cost function enables us to derive important characteristics of the
supply technology such as economies of density and scale. In line with Caves et al. (1985),
the economies of density are defined as the inverse of the elasticity of costs with respect to
outputs that is, the relative increase in total cost resulting from an increase in outputs, holding all input prices and the network size constant:

\[ ED = \frac{1}{\frac{\partial \ln TC}{\partial \ln Y} + \frac{\partial \ln TC}{\partial \ln Q}}. \]  

(4)

Economies of density exist if the above expression \((ED)\) has a value greater than one. For values of \(ED\) below one, we identify diseconomies of density. In the case of \(ED = 1\), the company’s output minimizes its costs given the network’s size.

Slightly different is the definition of economies of scale \((ES)\).\(^{26}\) Here, the increase in total costs is brought about by an increase in company’s scale that is in both outputs and the network size, holding the factor prices constant. The commonly used definition is the one proposed by Caves, Christensen and Tretheway (1984), which assumes that any increase in size raises the network size and the outputs with the same proportion. Based on this assumption, \(ES\) is defined as:

\[ ES = \frac{1}{\frac{\partial \ln TC}{\partial \ln Y} + \frac{\partial \ln TC}{\partial \ln Q} + \frac{\partial \ln TC}{\partial \ln N}}. \]  

(5)

\(^{25}\) See Caves et al. (1985).

\(^{26}\) It should be noted that the adopted definitions of scale and density economies do not necessarily correspond to the definitions based on the production function. In fact, only in homothetic production functions, where the optimal input bundles vary proportionately, the two definitions are equivalent. Here, we do not impose such an assumption. However, as in this paper we are interested in the cost effects of output, we define the scale and density economies as the inverse of the corresponding cost elasticities. See Chambers (1988) for more details about this issue.
Similarly, economies of scale exist if $ES$ is higher than 1. Table 7 shows the estimates of scale and density economies as given in equations (4) and (5), obtained from different models.

**Table 7. Economies of scale and density**

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
<th>Model V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$FE$</td>
<td>$RE$</td>
<td>$Pooled$</td>
<td>$True RE$</td>
<td>$True RE + Mundlak$</td>
</tr>
<tr>
<td>ED</td>
<td>7.79</td>
<td>4.51</td>
<td>1.91</td>
<td>5.82</td>
<td>8.18</td>
</tr>
<tr>
<td>ES</td>
<td>1.74</td>
<td>1.42</td>
<td>1.09</td>
<td>1.65</td>
<td>1.64</td>
</tr>
</tbody>
</table>

As can be seen in table 7, both economies of density and scale are greater than one in across all models, suggesting the presence of unexploited economies in most companies in the sample. As expected, the economies obtained from an increase in output density in a given network (density economies) are relatively higher than those gained by extending a company’s network (scale economies). The results listed in table 7 show a variation of the values of ED and ES between different models. This variation can be partially explained by the models’ differences with respect to the unobserved network effects. If these effects are correlated with explanatory variables (such as output and network length) the values obtained from the fixed effect model (Model I) and the Mundlak version of the True RE (Model V) are unbiased and those of the other three models are biased. Particularly, the values estimated by the pooled model (Model III) are likely to be biased downward. These results suggest that ignoring the unobserved firm-specific effects can bias the estimated coefficients. In fact such

---

27 It should be noted that the above definitions of scale and density economies are in terms of cost elasticity and do not necessarily correspond to the definitions derived from the production function. In fact, only in homothetic production functions, where the optimal input bundles vary proportionately, the two definitions are equivalent. Here, we do not impose such an assumption. However, as in this paper we are interested in the cost effects of output, we define the scale and density economies as the inverse of the corresponding cost elasticities. See Chambers (1988) for more details about this issue.
biases are driven by possible correlation of unobserved effects with output and network length.

6. CONCLUSION

The most relevant measure included in the railway reform of 1996 is the change from the practice of ex-post deficit coverage to an ex-ante fixed payment system for transport services. In this context cost frontier models could be useful as a complementary control instrument in determining the amount of subsidies granted to the regional railway companies. For this reason it is important to study the performance of different cost frontier models.

Alternative cost frontier models applied to a panel of Swiss railway companies indicate that the estimations particularly the inefficiency estimates, are sensitive to the adopted specification. The data show a considerable unobserved firm-specific heterogeneity that is likely to be correlated with explanatory variables. In such cases unbiased coefficients can be obtained from the fixed effects model. This model’s estimates of inefficiency are however unrealistic. In fact, comparing the results across different models suggest that the inefficiency estimates largely depend upon how the unobserved heterogeneity across firms is specified. Panel data models such as Pitt and Lee (1981) and Schmidt and Sickles (1984) that do not distinguish between unobserved firm-specific heterogeneity and inefficiency can overestimate the overall inefficiencies or even give misleading patterns of inefficiency. The cost frontier random effects model labeled as ‘true’ random-effects model (Greene, 2005) provides reasonable estimates of inefficiency confirming that the inefficiency estimates in other models are confounded with unobserved heterogeneity such as network effects. However, the problem of this model is that because of potential correlation between heterogeneity and
explanatory variables, the cost function coefficients may be biased (heterogeneity bias), especially as the Hausman specification test confirms the presence of such correlations.

Using an auxiliary equation in line with Mundlak (1978) can be helpful in this regard. This adjustment has been applied to the ‘true’ random effects. The resulted specification shows a very low level of heterogeneity bias, while slightly reducing the inefficiency estimates. The high correlation between the inefficiency scores across the two models suggests that in so far as the heterogeneity is accounted for, the correlation between heterogeneity and explanatory variables does not considerably affect the inefficiency estimates.

From a policy point of view, this study suggests that the Mundlak version of the “true” random effects model could be a valuable alternative for setting a benchmark in regulating the Swiss Regional Railways. However, it should be emphasized that a mechanical use of any of these models in regulation could be misleading. Since each industry has its specific cost characteristics that are not equally well reproduced by these models, establishing a reliable benchmark requires a careful analysis of the cost structure of the industry under consideration. Consequently, these models should be used as one among different instruments in the assessment of subsidy requests.

References


Federal Office of Transport (2003) Railway Reform 2, Swiss Federal Office of Transport Report, available in French (Réform des chemins de fer 2) and German (Bahnreform 2) at http://www.bav.admin.ch/.


AN EMPIRICAL ANALYSIS OF COST EFFICIENCY IN NON-PROFIT AND PUBLIC NURSING HOMES*  

Mehdi Farsi  Massimo Filippini

Department of Economics  
University of Lugano  
Via Ospedale 13, 6900 Lugano, Switzerland  
and  
Swiss Federal Institute of Technology  
ETH Zentrum, WEC, 8092 Zurich, Switzerland

September 2003

* We benefited from the suggestions of the editor and two anonymous referees, which are gratefully acknowledged. We would also like to thank Luca Crivelli for helpful comments and discussions, Ilaria Mosca and Chiara Gulfi for their excellent assistance, and Giorgio Borra tori for his help in understanding the data. This research has been partially financed by Ticino’s Dipartimento Sanità e Socialità. We thank this department for their support and also for providing the data. The views expressed in this paper are strictly personal and the responsibility for all remaining errors lies solely with the authors.
ABSTRACT
This paper examines the issue of cost-efficiency in Switzerland’s nursing homes, an issue of concern to policy makers because of the rapid growth of elderly care expenditure and the aging of the population. The fact that nursing homes in Switzerland exist in different institutional forms, private for-profit, government and non-profit status, raises the issue of their relative cost efficiency. A panel data of 17 public and 19 nonprofit nursing homes operating over the 9-year period from 1993 to 2001, in one of the 26 Swiss cantons, Ticino, is studied. Ticino’s nursing homes are heavily regulated and monitored by the canton’s authorities. However, given that in public firms there are more bureaucratic constraints and agency problems, one can expect a relatively low level of cost-efficiency. In this paper the effect of institutional form on efficiency is studied using a translog stochastic cost-frontier model. Several specifications are used to study the robustness of the results. The results suggest that the institutional form influences the efficiency of the studied nursing homes in that non-profit foundations are likely to be more cost-efficient than the nursing homes operated by government administration. The results also suggest that a great majority of the nursing homes in the sample do not fully benefit from scale economies. This implies that efficiency gains can be obtained with larger capacities or joint operations.

1. Introduction
Switzerland is a federal state composed of 26 cantons and approximately 3,000 municipalities. The federal state is characterized by a high degree of decentralization in the provision of public services. Article 3 of the Federal Constitution grants large autonomy to individual cantons in the sectors, such as health and social care, which are not directly regulated by the constitution. For instance, individual cantons are independent in the organization and regulation of the provision of long-term care for elderly people. While in most cases the provision of this service is organized at the cantonal level, some cantons such as Ticino considered in this analysis, decentralize the task to “communes” (local

---

1 Three cantons further divided into semi-cantons. The most populated canton is Zurich with a population of 1.2 million and the smallest one is Appenzell Interrhoden with slightly more than 15,000 inhabitants.
governments). The autonomy of the cantons creates a strong heterogeneity in the organization and regulation of the nursing home sector in Switzerland.

Generally the long-term care for the elderly is provided by non-profit nursing homes and private for-profit nursing homes. Non-profit nursing homes can be further distinguished in private and public nursing homes. From the institutional point of view, private non-profit nursing homes are generally foundations, whereas public nursing homes are so called “Municipalizzate”, e.g. firms without an own juridical status integrated directly in the local public administration. However, in few cases, local governments have chosen to create also a foundation to run their own nursing homes. In these cases, these types of public nursing homes are generally, more independent from the local political and public administration process than nursing homes directly included in the public administration. This mixed economy raises the interesting issue of the effects of different ownership and institutional forms on costs.

A number of recent studies have compared the cost efficiency of non-profit and for-profit hospitals and nursing homes. However, little empirical analysis has been done in the comparison of cost efficiency between public and private nonprofit nursing homes. In this paper we test the hypothesis that nursing homes operated by private nonprofit are more efficient than those managed by local governments. This hypothesis is based on the conjuncture that the managers in public organizations are faced with more bureaucratic constraints than those in private non-profit foundations. Moreover, compared to government employees, the managers in non-profit organizations have more intrinsic motivations to have a more efficient performance. Finally, as we see later in the paper the agency problems in a public organization are likely to occur in a higher number of levels and with more friction than in non-profit foundations.

For this purpose we estimate a cost function using data from a sample of 36 nonprofit nursing homes (public and private) operating in Ticino, a Swiss canton with a population of approximately 300,000 and an area size of 2,812 square km. Ticino is chosen because the planning of its new nursing homes is on the government’s agenda and therefore there is a need to determine the nursing home’s optimal capacity and the preferable ownership and organizational form. Moreover, Ticino’s nursing homes can be considered as a relatively

---

2 In Switzerland, nursing homes provide social care services as well as health care services to their patients. In fact, the residents’ average dependency and their need for medical care have been substantially rising during the last 15 years. Virtually all nursing homes have adapted the training of their staff and their infrastructure to provide health care services to their patients.

3 See for example Vitaliano et al. (1994), Bradford and Mobley (1997) and McKay (1988).
uniform sample regarding the quality of services as well as the financial incentives of their managers. Finally, Ticino’s cantonal administration regularly collects data from all the nursing homes operating in the canton and the data collected over the past years allow a panel data analysis. The hypothesis is tested using a two-stage methodology. In the first stage the nursing homes’ inefficiency scores are estimated using a cost model specification without controlling for organizational form. The estimated results are analyzed in a second stage to test for a significant difference between the two types. The cost frontier model used in the first stage is based on the random-effects model. Two specifications are used to study the effect of potential biases due to the endogeneity of the quality of care.

The paper is organized as follows. Section 2 presents the organization of the nursing home sector in Canton Ticino. A discussion on efficiency and institutional form is given in section 3. Section 4 discusses the cost model and its econometric specification. Section 5 presents the data, while section 6 illustrates the empirical results. Section 7 concludes the paper.

2. The organization of the nursing home sector in Ticino

The nursing home sector in canton Ticino, with 67 homes, represents an important element of the social care sector for elderly people. The most important characteristics of this sector are as follows:

- A variety of institutional forms operate in the market;
- The market is heavily regulated by the state;
- The state provides a large financial support in form of subsidies.4

In Ticino, as well as in other Swiss cantons, there are different institutional forms operating in the nursing home sector, i.e. private for-profit nursing homes, non-profit private institutions and public institutions. Crivelli, Filippini and Lunati (2002) provide a detailed description of different organization types in Switzerland’s nursing homes and their distribution across cantons. In all Swiss cantons there is a clear majority of non-profit nursing homes (i.e. public and private non-profit institutions). However, in some cantons there are more public institutions and in others more non-profit private institutions. Among 67 Ticino’s nursing homes, 24 are public, 32 are operated by private non-profit foundations, 6 are non-profit foundations created by municipalities, and 5 are private for-profit. The organizational structure of different nursing homes is discussed in the next section. The for-profit sector
accounts for about 5 percent of the nursing home beds in Ticino. For-profit nursing homes are not considered in this study.5

Ticino’s nursing homes are therefore dominated by non-profit and public firms that are strongly regulated and subsidized by the government. Economists generally regard competitive markets as the best way to promote an efficient allocation of resources. However, they recognize that there are circumstances where competitive markets fail to bring about an optimal allocation of resources. Three basic arguments for market failure in the Swiss nursing home industry are: (1) the local natural monopoly argument, which is typical of nursing homes in peripheral areas,6 and is supported by evidence of unexploited scale economies (cf. Filippini, 2001); (2) the lack of comprehensive information on the quality of care for the users of residential services, which creates a typical asymmetric information problem; and (3) the merit goods argument. In the latter case the state considers the services provided by nursing homes to be a merit good and therefore it recognizes the need to give them a financial contribution through the general fiscal system and to regulate their tariffs. For these reasons Ticino’s government has adopted the following forms of regulation: approval of the daily rates, definition of the minimum necessary infrastructure and staff requirements, control of the supply capacity in terms of beds, and concession of a financial contribution in the form of subsidies to public and non-profit nursing homes.

The regulation of the nursing homes in Ticino has shown at least three consequences:

- The subsidized price of residential care results in a significant increase in demand. The subsidized daily rates give families an incentive to transfer the burden of care of their elderly relatives to the state thus creating an excessive demand.
- To counteract this demand excess, the state controls the supply side by erecting barriers to entry to the market. For instance, in order to get subsidies for operating costs and health insurance reimbursement for medical care, a nursing home has to figure in the nursing home planning of the canton.
- The gap between supply and demand makes the use of a rationing tool necessary. In the case of Ticino this is represented by waiting lists.

---

4 However, in the majority of the Swiss cantons only public and non-profit nursing homes receive subsidies from the government.
5 The data on for-profit nursing homes are not available to the authors.
6 This argument should be considered in the context of Switzerland’s federal state, where each commune is in charge of supplying nursing care for its elderly residents. This implies that the elderly in need of nursing care are assigned to their local nursing homes, thus creating a local monopoly.
3. Institutional form and efficiency

The property rights theory has often been used to compare the efficiency of private non-profit organizations and for-profit firms. Following Alchian (1965) and Demsetz (1967), this theory postulates that for-profit firms provide a high-power incentive mechanism through clearly defined property rights. As residual claimants of the enterprise, owners have both the financial incentive and the means to induce an efficient performance in the firm.\(^7\) The distinction between non-profit and public organizations is however more subtle. In principle, there is no significant difference in the property rights defined in public and non-profit organizations. These two organization types are similar in several aspects: First, the directors do not have any direct pecuniary gain from profits, but may have non-pecuniary benefits or perquisites. These personal gains may however be more accessible in private non-profit organizations. Secondly, in both types there are intrinsic motivations to work efficiently. These incentives may be driven by the organization’s mission, organizational culture and trust, and long-term career objectives.\(^8\) One may argue however, that such motivations are more pronounced and clear in private non-profit foundations, driven by the very mission of those foundations.

There are other theoretical arguments that public non-profit foundations have a more efficient performance than organizations operated directly by government administration. The first argument is derived from public choice theory. The basic idea of these models proposed by Niskanen (1968, 1971) is that the bureaucrats responsible for delivering a fixed amount of output with a budget, have a tendency to over-budget and to produce more bureaucratic output than is socially optimal. To reach this goal they have incentives to respond to the demands of the political process with its different distribution of power. For politicians the incentive is to shift costs and benefits so that the net benefits to their constituency are positive. This leads to a “political optimisation” which may conflict with cost efficiency. On the other hand, managers of non-profit nursing homes may likewise have no strong incentive to be efficient, since efficiency cannot be lawfully rewarded. However, the incentive for these managers to be efficient may result from their personal satisfaction in providing a particular social service, which is likely to be stronger than their counterparts in the public sector.\(^9\)

Moreover, in the case of the Swiss nursing homes, the decision-making process of a private non-profit foundation has certain flexibility and is not influenced by the politicians. In

---

\(^7\) The owners can solve at least theoretically, the agency problems by sharing the risks with the managers.

\(^8\) See for instance DiIulio (1994) for the “principled” behavior of managers in public firms.

\(^9\) See Weisbrod (1997) for an interesting presentation of the efficiency incentives of nonprofit organizations.
the case of a state-owned firm, the taxpayers are the official owners of the firm and control
the way the firm conducts its activities through their representatives in the parliament and
government. These representatives are responsible for managing the company in order to
maximize the social welfare. Public representatives in turn, delegate the authority to a
commission that oversees the company’s management. Compared to private nonprofit firms,
public firms experience an attenuation of property rights resulting in a more significant
reduction in incentives for the management to minimize costs and to follow the owner’s
interests. Moreover, the agency problems within state-owned companies are more complex
than those in the private sector. In the case of a for-profit or nonprofit private firm, the
management itself answers only to the owners, and the employment relationship involves the
management and the employees. In the public sector, the chain of command from the
electorate to the management goes through the parliament, the government and the
government-appointed commission responsible for state-owned firms. Public firms involve
many principal-agent relationships, and agency problems can arise at each stage. In the face
of such agency problems monitoring managers is more difficult and the efficient performance
is harder to achieve. Therefore, compared to non-profit firms, the state-operated firms are
more likely to be away from efficiency. Moreover, the boards of directors in state-owned
firms are often political appointments and represent political parties whose objectives may not
be cost minimization. Private nonprofit firms are clearly less politicized than state-owned
firms. Finally, the nursing homes created by local governments in form of foundations are
also generally less influenced by the political process and experience less principal-agent
relationships than the nursing homes included in the local administration.

In this paper we study the effect of ownership and institutional form on the production
cost using a sample of Ticino’s nursing homes. As discussed above, one can expect that
nursing homes with different status are different regarding cost-efficiency. There are however,
several problems that may complicate the empirical analysis: First, public and non-profit
nursing homes may have different resident case-mixes thus different costs. Secondly, cost-
minimization is constrained by a minimum quality of care. It is not clear that public and non-
profit nursing homes have the same attitude towards quality. It is therefore necessary to
control for the differences in quality and case-mix.

4. Methodology

A stochastic cost frontier approach is adopted in this paper. A frontier cost function
defines minimum costs given output level, input prices and the existing production
technology. Due to technical and allocative inefficiencies it is unlikely that all firms operate at the frontier. The main advantage of the stochastic cost frontier approach compared to the deterministic approach is the separation of the inefficiency effect from the statistical noise.

The main approaches that can be used to estimate a frontier cost function with panel data are: random effects model without inefficiency distribution restriction, fixed effects model and random effects model with a restriction on the distribution of the inefficiency. In this paper a random effects model with time-invariant inefficiency in line with the model proposed by Pitt and Lee (1981) and Schmidt and Sickles (1984) is considered. We excluded the fixed effects model because the inefficiency indicators estimated by this approach may contain relatively high estimation errors due to the incidental parameters problem. Moreover, we did not use the third approach because it imposes a distribution on the inefficiency indicators.

In order to study the effect of institutional type on cost efficiency, a two-stage method is used. In the first stage, the cost frontier is estimated using a translog functional form and the nursing homes’ inefficiency scores are calculated. In the second stage the estimated inefficiency scores are analyzed with respect to the nursing home’s type. Basically the sample is divided into two sub-samples: public and non-profit nursing homes. The hypothesis that the inefficiency terms of the two sub-samples are from a single population is studied using Kruskal-Wallis test, a non-parametric rank test developed by Kruskal and Wallis (1952). A similar method is used to compare two other sub-samples: nursing homes (public and private) operated by a foundation and public nursing homes integrated in the local

10 We implicitly assume that public and private non-profit nursing homes have the same objective function. This is a reasonable assumption as the nursing care industry is strongly regulated in Switzerland. Since both service quality and output quantity are more or less narrowly determined by the regulator, different organizations are constrained to follow similar objectives.

11 See Battese (1992), Kumbhakar and Lovell (2000) and Simar (1992) for general overviews of these methods.

12 Notice that since the institutional form in our sample does not change over time, there is no need to consider models with time-variant inefficiency.

13 Note however that unlike Pitt and Lee we do not assume any form of distribution on the random effects. See also Kumbhakar and Lovell (2000) for more details.

14 For a discussion on the advantages and disadvantages of different approaches see Kumbhakar and Lovell (2000).

15 An alternative approach would be to estimate a cost function including one or two dummies for organization type. This method is free from the estimation errors incurred in the inefficiency estimates. These random errors may mask the transition between subsamples, thus may result in under-rejection (too few rejections) of the null hypothesis of similar cost-efficiencies across different types. However, since as we see later, this hypothesis is rejected anyways, our results do not appear to be sensitive to such errors. Moreover, our analysis (reported in the appendix) shows that both approaches lead to quite similar results.

16 See Singh and Coelli (2001) for an application of this test to compare the efficiency of Indian dairy plants in cooperative and private sectors.
administration. The analysis is performed with two specifications. This section provides a description of the cost frontier models and econometric specification used in the paper.

4.1. Specification of the Frontier Cost Function

A nursing home can be represented as a firm transforming labor and capital services into patient-days of residential health and social care for elderly people. Assuming that output level and input prices are exogenous, and that (for a given technology) firms choose input levels to minimize costs, the firm's total cost of operating a nursing home can be defined as a function of input prices and output. Moreover, in the cost model specification we take into account a number of output characteristics, which should capture, at least partially, the heterogeneity and quality dimensions of the nursing home’s output. Costs can also vary with a time trend. Since all the nursing homes in our sample are non-profit (whether public or private) and regulated, it can be reasonably assumed that they follow a similar objective function, implicitly set by the regulators. Given this assumption comparing costs among different firms based on the same function, can indicate which firms achieve these objectives with lower costs. The total cost frontier can therefore be represented by the following cost function:

\[ TC = f(Y, Q, R, P_K, P_L, \tau) \] (1)

where \( TC \) represents total annual cost and \( Y \) is the output represented by the total number of resident-days of the nursing home. \( P_K \) and \( P_L \) are respectively the prices of capital and labor. \( Q \) is the average dependency index calculated annually by the Regional Department of Public Health. This index measures the average required assistance of a given nursing home’s patients with normal daily activities such as eating, personal care or performing physiological functions. \( Q \) varies from 1 to 3, with 3 representing the most severe (dependent) case. \( R \) is the nursing staff ratio, that is the ratio of the number of employed nurses in a nursing home to the number of nurses that should be employed according to the guidelines of the Regional Department of Public Health. Since the nursing care is a labor-intensive service and the quality of care depends on the time spent by nurses for each patient, this variable represents

---

17 In Switzerland, in addition to the usual nursing care, nursing homes also provide basic medical services to their residents.
18 For a similar specification see Filippini (2001), who however used a smaller data set (38 nursing homes over the period 1993-1995) and did not estimate a cost frontier model.
19 These guidelines are only recommendations and the nursing homes are not required to exactly follow them.
the quality of output and the production process. 20 Finally, $\tau$ is a linear time trend that captures the changes in technical efficiency associated with technical progress.

Since the residents are assigned to the nursing homes by the canton’s authorities, variable $Q$ can be considered as an exogenous output characteristic. However, noting that this variable is obtained from a more or less subjective evaluation of the required caring time for each patient, one may as well argue that different nursing homes may have different evaluation or reporting criteria. Estimations would be biased should one type of nursing homes systematically over-report the dependency of their residents. To avoid any potential bias, we consider an alternative specification in which the dependency variable is excluded from the cost function.

It is generally assumed that the cost function given in (1) is the result of cost minimization given input prices and output and should therefore satisfy certain properties. 21 Mainly, this function must be non-decreasing, concave, linearly homogeneous in input prices and non-decreasing in output. To estimate the cost function (1), a translog functional form is employed. This flexible functional form is a local, second-order approximation to any arbitrary cost function. It places no a priori restrictions on the elasticity of substitution and allows the economies of scale to vary with the output level. The translog approximation to (1) can be written as:

$$\ln\left(\frac{TC_{it}}{P_{K_{it}}}\right) = \alpha_0 + \alpha_Y \ln Y_{it} + \alpha_Q \ln Q_{it} + \alpha_R \ln R_{it} + \alpha_L \ln \frac{P_{L_{it}}}{P_{K_{it}}} + \frac{1}{2} \alpha_{YY} (\ln Y_{it})^2 + \frac{1}{2} \alpha_{LL} (\ln P_{L_{it}})^2 + \frac{1}{2} \alpha_{QQ} (\ln Q_{it})^2 + \frac{1}{2} \alpha_{RR} (\ln R_{it})^2$$

$$+ \alpha_{YL} \ln Y_{it} \ln P_{L_{it}} + \alpha_{YQ} \ln Y_{it} \ln Q_{it} + \alpha_{YR} \ln Y_{it} \ln R_{it} + \alpha_{QY} \ln Q_{it} \ln Y_{it}$$

$$+ \alpha_{QR} \ln Q_{it} \ln R_{it} + \alpha_{RL} \ln R_{it} \ln P_{L_{it}} + \alpha_{LR} \ln L_{it} \ln R_{it} + \alpha_{LL} \ln L_{it} \ln L_{it} + \alpha_i + \nu_{it}$$  

with $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$

where subscripts $i$ and $t$ denote the nursing home and year respectively, $\alpha_i$ is a firm-specific effect and $\nu_{it}$ is the random error term. All variables are normalized by the corresponding sample medians. Therefore, the translog form is considered as a second order approximation

20 See Cohen and Spector (1996) and McKay (1988) for a similar approach when considering some measures of quality in a cost model for nursing homes. Cohen and Spector measured quality of care by case-mix adjusted staff to resident ratios. McKay used “nursing hours per patient” to measure the nursing home’s quality.

21 For more details on the functional form of the cost function see Cornes (1992), p.106.
around the sample median.\textsuperscript{22} This model can be estimated by the Generalized Least Squares method.\textsuperscript{23} In line with Schmidt and Sickles (1984) for any given firm \(i\), the cost inefficiency is estimated as: 

\[ u_i = \hat{\alpha}_i - \min\{\hat{\alpha}_i\} \]

where \(\hat{\alpha}_i\) is the estimated individual effect for that firm.

As it can be seen in equation (2), linear homogeneity in input prices is imposed by dividing total costs and input prices by capital price. The other theoretical restrictions are verified after the estimation. In particular, the concavity of the estimated cost function reflects the fact that the cost function is a result of cost minimization. However, this assumption may be unrealistic in public and non-profit firms, especially in sectors like health care, where other factors like quality may be as important as cost considerations. Moreover, since in Switzerland the labor markets (especially for nursing staff) are regulated, it is likely that the firms’ optimization strategies do not fully correspond to a perfectly minimal cost function. In such cases, the functions based on cost optimization may still be used as “behavioral” cost functions and can be helpful in studying the behavior of such firms.\textsuperscript{24}

Input prices and output are assumed to be exogenous, thus beyond the firm’s control. In a regulated industry these conditions are generally satisfied. Ticino’s non-profit nursing homes are fully regulated by the canton’s government. The residents are assigned to nursing homes by their community’s authorities, mainly based on their location, and the nursing homes’ costs are refunded on a cost-plus basis.

\section*{5. Data}

The data set used in this paper is based on the annual accounting reports of 36 nursing homes in canton Ticino over the 9-year period from 1993 to 2001. The sample includes more than two thirds of Ticino’s nursing homes. The sample includes 14 private non-profit homes, 5 public non-profit foundations and 17 public institutions. Ticino has a few for-profit nursing homes which are excluded from this study. All the nursing homes in the sample provide inpatient services.\textsuperscript{25} There are four missing observations in 1993, leaving a total of 320 observations. The variables include total costs, total number of employees (in terms of full-

\textsuperscript{22} Translog functional form requires that the underlying cost function be approximated around a specific point. In our case this point is taken as the sample median. We choose the median rather than the mean, because it is less affected by outliers and thus the translog approximation will have a better precision.

\textsuperscript{23} Alternatively, this model can be solved with a fixed effect estimator, assuming \(\alpha_i\)’s are fixed. Our analysis (not reported in the paper) indicates that our main results can also be obtained using this approach.

\textsuperscript{24} See Bös (1986), page 343.

\textsuperscript{25} There are some nursing homes that offer the possibility of nursing care in external residential apartments. The nursing care of this type is less intensive (thus less costly) than the care given to the home’s residents. For this reason we excluded four nursing homes whose share of external beds is more than 10 percent of their total beds.
time equivalent units), average wage per employee per year, total number of beds and total number of resident-days. Other characteristics are ownership form (public/non-profit) of the nursing home, the number of external apartment beds maintained by the home and the number of caring personnel working for the nursing home.

Total cost is taken as the total annual expenditures of the nursing home. Output is measured in total number of patient-days of the nursing home. Average yearly wage rates are estimated as the weighted mean of the average wage rates of different professional categories working in a nursing home, including nurses, administrative and technical staff and physicians. Following Friedlaender and Wang Chiang (1983), Filippini and Maggi (1993), and Filippini (2001), the capital price is calculated from the residual costs divided by the capital stock. Residual cost is total cost minus labor cost. Similar to Wagstaff (1989), the capital stock is approximated by the number of beds operated by the nursing home. The quality indicators, $Q$ and $R$, (as described earlier) are calculated annually by the regional Department of Public Health. The summary statistics of some of the main variables used in the analysis are given in table 1. All money values are converted to 2000 Swiss Francs using the global consumer price index.

As it can be seen in the table, there is a high variation in the costs of a patient-day care. The input prices show a great amount of variation as well. Part of these variations is associated with time variation. For instance the average cost of a patient-day care has increased from about 154 Francs in 1993 to 214 Francs in 2001. In the same period, the price of labor has increased about 15 percent in real terms and our measure of real capital price has increased about 20 percent. However, the organizational type of the nursing homes has not changed in our sample.

Table 2 lists means and standard deviations of the main variables by three organizational types. This table also reports the results of two-sided t-tests, testing the null hypothesis that the two groups are similar regarding each variable. The tests were performed separately between public and private type and between foundations and public administration nursing homes. The first observation on table 2 is that the public nursing homes are on average larger than the private ones and public-administration nursing homes have by far the largest size. Secondly, the average costs are higher in foundations (both private and public), but the difference is statistically significant at 10% level. This result however, does not bear any conclusion on cost-efficiency. For instance, as it can be seen in the table, public homes

26 A more precise estimation of capital stock would require capital inventory data, which are not available to us.
have access to a relatively lower capital price. Public nursing homes may have access to more governmental subsidies for equipments and buildings. Moreover, the patients’ case mix in non-profit nursing homes is on average more “dependent”, thus more costly than that of public-administration homes. Given that the patients are assigned to the nursing homes mainly based on proximity criteria, this finding may seem surprising.

Table 1. Descriptive statistics (320 observations)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total annual costs per resident-day (SFr)</td>
<td>184.05</td>
<td>28.92</td>
<td>183.10</td>
<td>111.85</td>
<td>279.81</td>
</tr>
<tr>
<td>Total annual resident-days ($Y$)</td>
<td>23175.7</td>
<td>9684.4</td>
<td>21482</td>
<td>6525</td>
<td>58324</td>
</tr>
<tr>
<td>Number of beds</td>
<td>66.23</td>
<td>26.81</td>
<td>61</td>
<td>28</td>
<td>162</td>
</tr>
<tr>
<td>Average labor price ($P_L$) in SFr per employee per year</td>
<td>70157.4</td>
<td>6586.4</td>
<td>70280.1</td>
<td>29744.3</td>
<td>122950.2</td>
</tr>
<tr>
<td>Average capital price ($P_K$) in SFr per bed</td>
<td>11008.3</td>
<td>2579.1</td>
<td>10714.1</td>
<td>3465.8</td>
<td>22426.3</td>
</tr>
<tr>
<td>Average dependency index ($Q$)</td>
<td>2.575</td>
<td>0.219</td>
<td>2.6</td>
<td>1.87</td>
<td>3</td>
</tr>
<tr>
<td>Nursing staff ratio ($R$)</td>
<td>.963</td>
<td>.124</td>
<td>.97</td>
<td>.49</td>
<td>1.55</td>
</tr>
</tbody>
</table>

- All monetary values are in 2000 Swiss Francs (CHF), adjusted for inflation by Switzerland’s global consumer price index.

A more careful look at the differences shows that on average, the non-profit foundations created by municipalities have the most dependent cases and public nursing homes have the least dependent case-mix. These differences may be related to demographic variations and health status of different locations in the canton. However, such a contrasting difference may also be at least partly, due to the variations in reporting dependency index by nursing homes to canton’s authorities. If a group of nursing homes systematically over-report the severity of their case-mix, this may create a bias in the estimation of cost-efficiency. In order to identify the extent of such potential biases, an additional specification without control for severity is used. Comparing the results between the model with and without severity can help determine the direction of this bias.
## Table 2. Means of the main variables by organizational type

<table>
<thead>
<tr>
<th>Mean (Standard Deviation)</th>
<th>t statistic (two-sided t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type I Private Foundations</td>
</tr>
<tr>
<td>Total annual resident-days ($Y$)</td>
<td>19,592 (7,802)</td>
</tr>
<tr>
<td>Number of beds</td>
<td>55.8 (22.2)</td>
</tr>
<tr>
<td>Average total costs per resident-day (SFr)</td>
<td>187.0 (33.6)</td>
</tr>
<tr>
<td>Average labor price ($P_L$) in SFr per employee per year</td>
<td>69,168 (8585)</td>
</tr>
<tr>
<td>Average capital price ($P_K$) in SFr per bed</td>
<td>12,718 (2,593)</td>
</tr>
<tr>
<td>Average dependency index ($Q$)</td>
<td>2.58 (.21)</td>
</tr>
<tr>
<td>Nursing staff ratio ($R$)</td>
<td>.963 (.144)</td>
</tr>
<tr>
<td>Share of total resident-days (%)</td>
<td>33.0</td>
</tr>
<tr>
<td>Share of beds (%)</td>
<td>32.9</td>
</tr>
<tr>
<td># of observations</td>
<td>125</td>
</tr>
<tr>
<td># of homes</td>
<td>14</td>
</tr>
</tbody>
</table>

- Standard deviations are given in brackets.
- * statistically significant at 5% significance level. ** statistically significant at 1% significance level.
- All monetary values are in 2000 Swiss Francs (CHF), adjusted for inflation by Switzerland’s global consumer price index.

### 6. Estimation results

The estimated parameters of the cost frontier are listed in table 3. Model $I$ is according to equation 2. In model $II$ the average dependency index ($Q$) is excluded from the explanatory
variables. This specification is used to study the extent to which the differences between two types of nursing homes are due to their different resident case-mixes.

Table 3. Estimated coefficients

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_Y )</td>
<td>.890** (.017)</td>
<td>.822** (.021)</td>
</tr>
<tr>
<td>( \alpha_Q )</td>
<td>.555** (.083)</td>
<td>-</td>
</tr>
<tr>
<td>( \alpha_R )</td>
<td>.382** (.046)</td>
<td>.256** (.045)</td>
</tr>
<tr>
<td>( \alpha_L )</td>
<td>.832** (.025)</td>
<td>.816** (.026)</td>
</tr>
<tr>
<td>( \alpha_{YY} )</td>
<td>-.024 (.053)</td>
<td>-.112 (.059)</td>
</tr>
<tr>
<td>( \alpha_{QQ} )</td>
<td>-.558 (.900)</td>
<td>-</td>
</tr>
<tr>
<td>( \alpha_{LL} )</td>
<td>.612** (.075)</td>
<td>.576** (.078)</td>
</tr>
<tr>
<td>( \alpha_{YQ} )</td>
<td>-.011 (.12)</td>
<td>-</td>
</tr>
<tr>
<td>( \alpha_{YL} )</td>
<td>-.00006 (.042)</td>
<td>-.016 (.042)</td>
</tr>
<tr>
<td>( \alpha_{LQ} )</td>
<td>.034 (.17)</td>
<td>-</td>
</tr>
<tr>
<td>( \alpha_{RR} )</td>
<td>-.113 (.214)</td>
<td>-.276 (.216)</td>
</tr>
<tr>
<td>( \alpha_{YR} )</td>
<td>.223* (.098)</td>
<td>.296** (.099)</td>
</tr>
<tr>
<td>( \alpha_{LR} )</td>
<td>.348** (.12)</td>
<td>.325** (.121)</td>
</tr>
<tr>
<td>( \alpha_{QR} )</td>
<td>-.587 (.34)</td>
<td>-</td>
</tr>
<tr>
<td>( \alpha_t )</td>
<td>.012 (.0022)</td>
<td>.024** (.0017)</td>
</tr>
<tr>
<td>Constant</td>
<td>15.15** (.013)</td>
<td>15.10** (.014)</td>
</tr>
<tr>
<td>R-square</td>
<td>.9754</td>
<td>.9607</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets.
- * statistically significant at 5% significance level.
- ** statistically significant at 1% significance level.
- The sample includes 320 observations (36 nursing homes)

Since total cost and the regressors are in logarithms and have been normalized, the first order coefficients are interpretable as cost elasticities evaluated at the sample median. The regression results show that all the first-order terms are significant and in a reasonable
direction. As expected, output and prices have a positive effect on costs, and the nursing homes with a more severe case-mix and/or with a higher quality of service are relatively more costly. The output elasticity is positive and implies that an increase in the supply will increase total cost. A 1% increase in the number of patient-days of nursing home care will increase the total cost by approximately 0.9% (Model I), 0.8% (Model II), respectively.

Cost elasticities with respect to the output characteristics variables, $Q$ and $R$, are positive and imply that an increase in the average required assistance of a home’s patients or an increase in the ratio of the number of nurses employed by a nursing home and the number of nurses that should theoretically be employed will increase total cost. The coefficient of the linear trend suggests that the total costs have increased over time with a rate of about 1 to 2 percent per year. The growth of costs is a commonly observed phenomenon in labor-intensive industries such as health care, which usually face a persistent growth of labor price.

The estimated cost functions do not however satisfy the concavity condition in input prices. This result suggests that the estimated cost functions are not resulted from a completely unconstrained cost-minimization strategy. Namely, the firms’ strategies are not responsive to changes in input factor prices. This can be explained by the fact that the input choices in Switzerland’s nursing homes are rather constrained by regulation. Labor contracts and quality regulations can be named among these restrictions. We contend therefore that the results of this paper should be interpreted in the behavioral cost framework proposed by Bös (1986), rather than an unconstrained minimum cost function.

**Cost-efficiency, ownership and institutional form**

Table 4 provides a summary of the estimated inefficiency measures for three groups of nursing homes. The inefficiency score is defined as $\exp(u_i)$, where $u_i$ is the inefficiency term obtained from the regression model. These measures represent the ratio of a nursing home’s actual costs to a minimum level that would have been achieved had the firm operated as cost-efficient as the “best practice” observed in the sample. Comparing the two specifications, with and without average dependency index, indicates that part of the difference between different nursing homes is related to the residents case-mix. When the dependency index is not taken into account, model II predicts that the private non-profit homes are on average 7 percent

---

27 These findings are in line with the results obtained by Filippini (2001) using a shorter panel and a slightly different number of nursing homes.
28 Our results indicate that the Hessian matrix of the estimated cost function with respect to input prices (labor and capital) is not negative semi-definite, thus the concavity condition is not satisfied in any of the specifications.
more cost-efficient (second column) than public-administration homes. However, when the differences in case-mix are controlled for, this difference reduces to about 3 percent (first column). This result is consistent with the results in table 2, suggesting that the public-administration nursing homes have on average a more severe case-mix.

As expected, the inefficiency estimates obtained from model II are higher than those of model I. The exclusion of variable $Q$ from the model specification increases the value of the composed error terms and also the inefficiency term. The inefficiency scores obtained from model II can therefore be overestimated. The results obtained from model I show relatively low values of inefficiency indicators. This means that the nursing homes included in our sample are comparatively efficient.

### Table 4. Inefficiency measures:

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Private Foundations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.07</td>
<td>1.21</td>
</tr>
<tr>
<td>St. Deviation</td>
<td>.038</td>
<td>.089</td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.05</td>
<td>1.19</td>
</tr>
<tr>
<td>Median</td>
<td>1.07</td>
<td>1.23</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>1.09</td>
<td>1.26</td>
</tr>
<tr>
<td><strong>Public Foundations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.08</td>
<td>1.23</td>
</tr>
<tr>
<td>St. Deviation</td>
<td>.054</td>
<td>.13</td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.10</td>
<td>1.23</td>
</tr>
<tr>
<td>Median</td>
<td>1.10</td>
<td>1.28</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>1.13</td>
<td>1.30</td>
</tr>
<tr>
<td><strong>Public Administration</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.10</td>
<td>1.28</td>
</tr>
<tr>
<td>St. Deviation</td>
<td>.028</td>
<td>.057</td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.03</td>
<td>1.13</td>
</tr>
<tr>
<td>Median</td>
<td>1.08</td>
<td>1.15</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>1.12</td>
<td>1.36</td>
</tr>
</tbody>
</table>

- Inefficiency measures represent the ratio of a nursing home’s actual costs to minimum costs from the best practice in the sample.

---

29 See also Breyer (1987) for a discussion of strengths and weaknesses of each approach.
The results of Kruskal-Wallis test on the differences in inefficiency scores between different groups are given in table 5. The results are in general robust to specification in the sense that the comparative cost-efficiency of different types of nursing homes does not depend on their reported case-mix severity. These results suggest that private nursing homes are on average, slightly more cost-efficient than the homes owned by governments. Moreover, the foundations, regardless of whether they are owned by public or private sector, are slightly more cost-efficient than the nursing homes integrated in the local public administration.\textsuperscript{30} Foundations, perhaps due to a relatively low influence of the political and bureaucratic process on management, seem to be an interesting institutional form even for public nursing homes. The Kruskal-Wallis tests results (not reported in the paper) do not indicate any significant difference between public foundations and private non-profit nursing homes. However, because of the relatively low number of public foundations in our sample, this result may be due to estimation errors. Our results are generally confirmed by an alternative approach including binary indicators for nursing home’s type in the regression model. The estimation results are reported in the appendix.

\begin{table}[h]
\centering
\caption{Kruskal-Wallis test statistics:}
\begin{tabular}{|l|l|l|l|}
\hline
 & Model I & Model II & More cost-efficient type \\
\hline
\text{Public} (22) \textit{vs} \text{Private} (14) & 5.308\text{*} & 3.665 & Private \\
\hline
\text{Foundation} (19) \textit{vs} \text{Public Administration} (17) & 6.501\text{*} & 4.575\text{*} & Foundation \\
\hline
\end{tabular}
\end{table}

- * statistically significant at 5% significance level.
- The statistics have a Chi-square distribution with one degree of freedom.
- The number of nursing homes in each group is given in parantheses.

\textit{Scale economies}

Scale efficiency indicates the degree to which a company is producing at optimal scale. Frisch (1965) defines the optimal scale as the level of operation where the scale elasticity is equal to one. The degree of returns to scale ($RS$) is defined as the proportional increase in output ($Y$) resulting from a proportional increase in all input factors, holding all

\textsuperscript{30} This result is consistent with the findings reported in Crivelli and Filippini (2002) who used a different sample of Ticino’s nursing homes and a different methodology.
input prices and output characteristic variables fixed (Caves et al., 1981). The RS degree may also be defined in terms of the effects on total costs resulting from a proportional increase in output (Silk and Berndt, 2003). This is equivalent to the inverse of the elasticity of total cost with respect to the output.31

The returns to scale can therefore be obtained from:

\[ RS = \frac{1}{\left( \frac{\partial \ln TC}{\partial \ln Y} \right)} \quad (3) \]

There are increasing returns to scale if RS is greater than 1, and conversely, there are decreasing returns to scale if RS is below 1. In the case of RS = 1 we have a constant returns to scale situation. Economies of scale exist if the average costs of a nursing home decrease as output increases. Table 5 shows the values of RS calculated for representative nursing homes of our sample.32 The results are given only for the main specification with control for dependency index (model I). The results obtained from the alternative specification II are in general very close to the reported results.

**Table 6. Returns to scale (RS):**

<table>
<thead>
<tr>
<th>Representative Nursing Home</th>
<th>Size (Number of Beds)</th>
<th>Resident-days per Year</th>
<th>Returns to scale (RS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Quartile</td>
<td>47</td>
<td>15,740</td>
<td>1.134 (0.030)</td>
</tr>
<tr>
<td>Median</td>
<td>61</td>
<td>21,480</td>
<td>1.124 (0.021)</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>80</td>
<td>28,790</td>
<td>1.116 (0.029)</td>
</tr>
<tr>
<td>Maximum</td>
<td>162</td>
<td>58,320</td>
<td>1.032 (0.060)</td>
</tr>
</tbody>
</table>

- Standard errors are given in brackets.

---

31 The inverse of cost elasticity of output is referred to by Chambers (1988), as the “economies of size” rather than economies of scale, which are defined in regards to production function. Scale and size economies are equivalent if and only if the production function is homothetic (see Chambers, 1988, page 72). Here, we do not impose this assumption. However, as for the purpose of this paper we are more interested in the cost effects of output, we define the returns to scale in terms of cost elasticity.

32 Equation (3) has been evaluated at the corresponding input prices and output characteristics $Q$ and $R$. For instance in order to estimate the returns to scale for the “1st quartile” nursing home, the equation is estimated at the 1st quartile of all the variables.
The estimated values of returns to scale range from a minimum of about 1 for the largest nursing home to a maximum of 1.3 for the smallest one. The results suggest that a great majority of the nursing homes in our sample do not fully exploit the scale economies. This is an important policy implication that is worth considering especially, in the organization of some of these nursing homes that operate in the same area, but are operated by different municipalities.

7. Conclusion

In this paper a translog cost function has been estimated for a panel of 36 Swiss nursing homes in canton Ticino over a nine-year period from 1993 to 2001. A random effects stochastic frontier model has been considered. The results suggest that government nursing homes have slightly higher costs than non-profit nursing homes when other factors are taken into consideration. In terms of efficiency, the non-profit foundations created by municipalities are rather similar to private non-profit nursing homes than those operated by public administration. These findings are robust to the model specification. The result that the foundations created and owned by local governments can almost be as efficient as private non-profit nursing homes suggests that the source of inefficiency is mainly in administrative restrictions rather than ownership differences. The relatively low efficiency of the nursing homes integrated in the public administration can therefore be explained by their strong bureaucratic constraints. Such constraints potentially limit the firms’ input choices as well as their flexibility in implementing incentive pay schemes. The results also indicate that in any case the cost differences are limited to a few percentage points.

Finally, this paper provides empirical evidence of unexploited scale economies at most output levels in Ticino’s nursing homes. The economies of scale should be taken into account for planning the size of new nursing homes. Theoretically, efficiency gains may also be obtained from merging smaller nursing homes operating in the same area or through partial mergers. For instance in many cases, joint purchase of food and medical supplies or sharing some clinical and administrative functions can result in significant savings. Capital costs can also be reduced thanks to the ability of large firms to negotiate lower interest rates. However, it should be noted that size may influence quality which is only partly considered in this paper. Especially because of relatively limited opportunities for human contact, the quality of service can be relatively low in large nursing homes. This is an important factor that should be considered in the assessment of optimal size and merger strategies.
References


Appendix

Alternative approach (GLS random-effects model including organization dummies)

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_Y$</td>
<td>.865** (.018)</td>
<td>.800** (.022)</td>
</tr>
<tr>
<td>$\alpha_Q$</td>
<td>.555** (.083)</td>
<td>–</td>
</tr>
<tr>
<td>$\alpha_R$</td>
<td>.380** (.045)</td>
<td>.255** (.044)</td>
</tr>
<tr>
<td>$\alpha_L$</td>
<td>.807** (.026)</td>
<td>.797** (.027)</td>
</tr>
<tr>
<td>$\alpha_{YY}$</td>
<td>-.022 (.053)</td>
<td>-.114* (.058)</td>
</tr>
<tr>
<td>$\alpha_{QQ}$</td>
<td>-.372 (.89)</td>
<td>–</td>
</tr>
<tr>
<td>$\alpha_{LL}$</td>
<td>.592** (.074)</td>
<td>.555** (.077)</td>
</tr>
<tr>
<td>$\alpha_{YQ}$</td>
<td>.001 (.12)</td>
<td>–</td>
</tr>
<tr>
<td>$\alpha_{QL}$</td>
<td>-.008 (.042)</td>
<td>-.016 (.042)</td>
</tr>
<tr>
<td>$\alpha_{LQ}$</td>
<td>.059 (.17)</td>
<td>–</td>
</tr>
<tr>
<td>$\alpha_{RR}$</td>
<td>-.106 (.209)</td>
<td>-.274 (.213)</td>
</tr>
<tr>
<td>$\alpha_{LR}$</td>
<td>.239* (.097)</td>
<td>.302** (.098)</td>
</tr>
<tr>
<td>$\alpha_{RL}$</td>
<td>.337** (.12)</td>
<td>.321** (.12)</td>
</tr>
<tr>
<td>$\alpha_{QR}$</td>
<td>-.465 (.34)</td>
<td>–</td>
</tr>
<tr>
<td>$\alpha_{R}$</td>
<td>.012 (.0022)</td>
<td>.024** (.0016)</td>
</tr>
<tr>
<td>Public Foundation</td>
<td>.013 (.024)</td>
<td>.020 (.036)</td>
</tr>
<tr>
<td>Public Administration</td>
<td>.061** (.018)</td>
<td>.079** (.027)</td>
</tr>
<tr>
<td>Constant</td>
<td>15.12** (.017)</td>
<td>15.06** (.021)</td>
</tr>
</tbody>
</table>

R-square | .9760 | .9602 |

- Private foundations are considered as the baseline.
- Standard errors are given in brackets.
- * statistically significant at 5% significance level.
- ** statistically significant at 1% significance level.
REGULATION AND MEASURING COST EFFICIENCY
WITH PANEL DATA MODELS: APPLICATION TO
ELECTRICITY DISTRIBUTION UTILITIES

Mehdi Farsi        Massimo Filippini

Center for Energy Policy and Economics
Federal Institute of Technology
ETH Zentrum, WEC, 8092 Zurich, Switzerland
and
Department of Economics
University of Lugano
Via Ospedale 13, 6900 Lugano, Switzerland

April 2004

Correspondence: Massimo Filippini, CEPE, ETH Zentrum, CH-8092, Zurich, Switzerland
Phone: +41-1-632-0649, Fax: +41-1-632-1050, E-mail: filippin@cepe.mavt.ethz.ch
ABSTRACT

This paper examines the performance of panel data models in measuring cost efficiency of electricity distribution utilities. Different cost frontier models are applied to a sample of 59 utilities operating in Switzerland from 1988 to 1996. The estimated coefficients and inefficiency scores are compared across different specifications. The results indicate that while the average inefficiency is not sensitive to the econometric specification, the efficiency ranking varies significantly across models. The reasonably low out-of-sample prediction errors suggest that panel data models can be used as a prediction instrument in order to narrow the information gap between the regulator and regulated companies.

Keywords: Cost efficiency; Electricity utilities; Incentive regulation; Yardstick competition.
1. INTRODUCTION

Transmission and distribution of electricity have been considered as natural monopolies, thus less affected by the recent waves of deregulation in power industry. However, as competition is introduced into generation sector, regulatory reform and incentive regulation of distribution utilities have become more common. In traditional cost-of-service regulation systems companies recover their costs with a risk-free fixed rate of return and therefore have little incentive to minimize costs. The incentive-based schemes on the other hand, are designed to provide incentive for cost-efficiency by compensating the company with its savings. A variety of methods are proposed in the literature. Main categories of incentive-based schemes used for electricity utilities are: price or revenue cap regulation schemes, sliding-scale rate of return, partial cost adjustment, menu of contracts, and yardstick regulation.\(^1\) Jamasb and Pollitt (2001) provide an extensive survey of different regulation practices in electricity markets around the world. Virtually all the models used in practice, are based on ‘benchmarking’ that is, measuring a company’s efficiency against a reference performance. Inefficiency can be resulted from technological deficiencies or non-optimal allocation of resources into production. Both technical and allocative inefficiencies are included in cost-inefficiency, which is by definition, the deviation from minimum costs to produce a given level of output with given input prices.

In benchmarking applications the regulator is generally interested in obtaining a measure of firms’ inefficiencies such as X-factors in price cap regulation, in order to reward (or punish) companies accordingly. The inefficiency estimates can have great financial sequences for the firms and therefore, their reliability is crucial for an effective regulation system. In particular, if the estimated inefficiency scores are sensitive to the benchmarking method, a more detailed analysis to justify the adopted model is required. However, in most

\(^1\) See Joskow and Schmalensee (1986) for a review of regulation models.
cases it is difficult to identify the ‘right’ model among the set of legitimate ones. Bauer et al. (1998) propose a series of criteria that can be used to evaluate if the results obtained from different methods are “mutually consistent”, that is, lead to comparable inefficiency scores and ranks. It is recommended that rather than using the inefficiency estimates in a mechanical way, the benchmarking analysis should be used as a complementary instrument in incentive regulation schemes.

Using a cross section of 63 power distribution utilities in Europe, Jamasb and Pollit (2003) show that there are substantial variations in estimated efficiency scores and rank orders across different methods.² There is a common perception that the estimation results can be improved using panel data. In contrast with cross-sectional data, panels provide information on same companies over several periods. Moreover, panel data models can better control for unobserved heterogeneity among companies. This perception is supported by suggestive evidence. For instance, after reviewing their previous empirical literature on measuring inefficiency with panel data, Kumbhakar and Lovell (2000)³ conclude that different approaches are likely to generate rather similar efficiency rankings, especially at the top and bottom of the distribution. However, none of the cited works is related to electricity distribution.⁴

These results raise an important question as to whether (or to what extent) the problems reported by Jamasb and Pollitt (2003) are due to the limitations associated with cross-section data models. This question is addressed in the present paper by using several alternative

---

² Other authors like Horrace and Schmidt (1996), Street (2003) and Jensen (2000) reported substantial errors and inconsistency problems in the estimation of individual efficiency scores in cross sectional data.

³ See page 107.

⁴ There is some evidence that the reliability of benchmarking analysis depends on the nature of production. For instance Gong and Sickles (1989), using Monte Carlo simulations, find that with complex production functions most models show a poor performance, while with simpler production forms the results are more reasonable.
models to a panel of power distribution utilities in Switzerland. Given that the validity of benchmarking methods in electricity industry has been put into question, studying the reliability of these methods using better data (panels in this case) can bring some light into a controversial debate. Moreover, as an increasing number of regulators throughout Europe and elsewhere have access to panel data sets, they will soon face the question if the longitudinal data can help evaluate the regulated companies’ inefficiency.

The main goal of this paper is to study the sensitivity of inefficiency estimates in panel data models. We focus on parametric cost frontier methods mainly because these methods are more easily adaptable to panel data. It should be noted that in practice, most of the regulators use deterministic frontier methods like Data Envelopment Analysis. This is mainly because such methods require a relatively low number of observations. However, the effects of unobserved differences among companies, which are particularly important in network industries, are completely ignored in these models.

In this paper, different stochastic frontier models are applied to a sample of distribution utilities in Switzerland. The sample is an unbalanced panel of 59 companies over a period of 9 years (a total of 380 observations). The inefficiency scores estimated from four different models are compared. In particular, the resulted efficiency rankings obtained from random effect and fixed effect models are analyzed. Although there is a reasonably good correlation between the estimates obtained from certain models, individual inefficiency scores and ranks change quite significantly from one model to another. The estimated measures of inefficiency are therefore sensitive to econometric specification and should not be used as a direct instrument in benchmarking. These results suggest that the sensitivity problems in

---

5 For instance see Shuttlewoth (2003) and Irastorza (2003) for criticisms of benchmarking approaches in electricity industry.

6 Other advantages of the parametric approaches are described in section 2.
benchmarking electricity utilities are not limited to cross-sectional data and cannot be completely resolved by using panel data models. However, our analysis of prediction errors indicates that panel data models can be used for predicting utilities costs with a rather reasonable error, suggesting that benchmarking analyses can be used as a complementary instrument for monitoring utilities’ performances.

The rest of the paper proceeds as follows: Section 2 provides a brief review of cost-frontier models and presents the specification used in this paper. The data are described in section 3. Section 4 presents the estimation results and discusses their implications. The main conclusions are summarized at the end.

2. METHODOLOGY

A frontier cost function defines minimum costs given output level, input prices and the existing production technology. It is unlikely that all firms will operate at the frontier. Failure to attain the cost frontier implies the existence of technical and allocative inefficiency. This section provides a description of the cost frontier models and the specification used in this paper. The adopted methodology is based on a comparison of different models with respect to the estimated cost function parameters, estimated inefficiency scores, and the out-of-sample prediction errors. The main goal is to study the limitations of different models in benchmarking and the sensitivity of inefficiency scores to model selection.

2.1. Cost frontier models

There are several cost frontier methods to estimate the cost efficiency of individual firms. Two main categories are non-parametric methods originated from operations research,
and econometric approaches namely stochastic cost frontier models.\(^7\) In non-parametric approaches like Data Envelopment Analysis, the cost frontier is considered as a deterministic function of the observed variables but no specific functional form is imposed. Moreover, non-parametric approaches are generally easier to estimate.\(^8\) Parametric methods on the other hand, allow for a random unobserved heterogeneity among different firms but need to specify a functional form for the cost function. The main advantage of such methods over non-parametric approaches is the separation of the inefficiency effect from the statistical noise due to data errors, omitted variables etc. The non-parametric methods’ assumption of a unique deterministic cost frontier for all companies is unrealistic. Another advantage of parametric methods is that these methods allow statistical inference on the significance of the variables included in the model, using standard statistical tests. In non-parametric methods on the other hand, statistical inference requires elaborate and sensitive re-sampling methods like bootstrap techniques.\(^9\) Given the above discussion we decided to focus on the parametric approaches.

In this paper we consider the estimation of a deterministic and three versions of a stochastic frontier cost function using panel data. It should be noted that the theoretical development of stochastic frontier models in panel data has been subject of a great body of literature.\(^10\) We contend that some of the recent developments like the models proposed by

\(^7\) See Kumbhakar and Lovell (2000) for an extensive survey of parametric methods and Coelli et al. (1998), chapter 6, and Simar (1992) for an overview of non-parametric approaches.

\(^8\) See Coelli et al. (2003) for more details.

\(^9\) These methods are available for rather special cases and have not yet been established as standard tests. See Simar and Wilson (2000) for an overview of statistical inference methods in non-parametric models.

\(^10\) See Kumbhakar and Lovell (2000) for a review and Greene (forthcoming) and Tsionas (2002) for some recent developments.
Greene (forthcoming) can be useful in benchmarking. However, a critical analysis of these models needs further research and is beyond the scope of this paper.

The deterministic approach adopted in this paper can be formulated as:

\[
\ln C_{it} = \ln C(y_{it}, w_{it}) + u_{it} \quad u_{it} \geq 0 \quad i = 1, 2, \ldots, N \quad \text{and} \quad t = 1, 2, \ldots, T
\]

(1)

where \( C_{it} \) is observed total cost in year \( t \), \( y_{it} \) is a vector of outputs, \( w_{it} \) is an input price vector and \( u_{it} \) is a positive one-sided disturbance capturing the effect of inefficiency. \( N \) represents the number of firms and \( T \) the number of years in the sample. Firms can therefore operate on or above the cost frontier but not below it. One interesting method proposed for estimating equation (1) is Greene's (1980) version of Richmond's (1974) Corrected Ordinary Least Squares model. A functional form for the cost function is assumed, and parameter estimates are obtained using ordinary least squares method. The intercept is corrected by shifting the value of the intercept such that all residuals are positive and at least one is zero.

The main shortcoming of this method is that it confounds inefficiency with statistical noise: the entire residual is classified as inefficiency. Nevertheless many studies have used this approach.\(^{11}\) It should be noted that the COLS method does not consider the panel aspect of the data because it considers the repeated observations of a given firm as independent observations. These problems can be partly overcome using the stochastic cost frontier approach suggested by Pitt and Lee (1981) who extended the original model of Aigner et al. (1977) to panel data setups. This model can be written as follows:

\[
\ln C_{it} = \ln C(y_{it}, w_{it}) + u_{it} + v_{it} \quad u_{it} \geq 0 \quad i = 1, 2, \ldots, N \quad \text{and} \quad t = 1, 2, \ldots, T
\]

(2)

In this specification the error term is composed of two uncorrelated parts: The first part \( u_{it} \), is a one-sided non negative disturbance reflecting the effect of inefficiency (including both

\(^{11}\) See for example Wagstaff (1989) and Filippini and Maggi (1993).
allocative and technical inefficiencies), and the second component \( v_{it} \), is a symmetric disturbance capturing the effect of noise. Usually the statistical noise is assumed to be normally distributed, while the inefficiency term \( u_i \) is assumed to follow a half-normal distribution.\(^{12}\) This model with a normal-half-normal composite error term can be estimated using Maximum Likelihood Estimation method. Consistent with Kumbhakar and Lovell (2000) we refer to this model as the MLE model.

Compared to a deterministic approach the main advantage of this stochastic cost frontier model is the separation of the inefficiency effect from the statistical noise. However, this method is subject to the potential criticism of having arbitrary assumptions about the distribution of the random terms. These assumptions can be relaxed by rewriting equation 2 as:

\[
\ln C_t = \ln C(y_{it}, w_{it}) + \alpha + u_i^* + v_{it} \quad \text{with} \quad u_i = u_i^* - \min\left\{u_i^*\right\},
\]

and using a feasible Generalized Least Squares method as proposed by Schmidt and Sickles (1984).\(^{13}\) The resulting model is referred to as the GLS model.

The remaining restrictive assumption is that the two random components be uncorrelated with each one of the explanatory variables. This implies that the firm’s inefficiency is uncorrelated with its observed characteristics included in the cost function. In the real world however, many of these factors may affect the firm’s inefficiency. Schmidt and Sickles (1984) propose a solution around this assumption.\(^{14}\) In their model the overall residual \( w_{it} \) is composed of two terms (\( w_{it} = u_i + v_{it} \)): a symmetric disturbance \( v_{it} \), like

\(^{12}\) Other extensions of this model have also considered exponential and truncated normal distributions for the inefficiency term. See for instance Battese (1992) and Battese and Coelli (1992).

\(^{13}\) See also Kumbhakar and Lovell (2000).

\(^{14}\) For a presentation of this method see also Simar (1992).
previous models, and a one-sided fixed component \( u_i \), that represents cost inefficiency. The latter component can be identified by a fixed effects specification with no assumption on the distribution of \( u_i \).\(^{15}\) Inefficiency scores are estimated as the distance to the firm with the minimum fixed effect, that is: \( u_i - \min \{ u_i \} \). The resulting model is a fixed-effects model and is labeled as the FE model in the rest of the paper.

The fixed effects approach controls for unobservable firm specific effects, such as inefficiency, that are not captured by control variables. There are however, two limits to this approach: First, the time invariant variables are captured by the fixed effects and cannot be included in the model. This implies that the inefficiency estimators include the variations in time-invariant firm characteristics. Moreover, inefficiency is assumed to be constant over time. Notice that this assumption can be relaxed in the random effects models discussed above.\(^{16}\)

The main advantage of the fixed-effects specification is that the estimations are unbiased even if explanatory variables are correlated with firm-specific dummies. However, the inefficiency measures may be confounded with time-invariant factors, which could not be included in the model. The choice between random effects and fixed effects models also depends on whether or not firms belong to the same population.\(^{17}\) The random effects model is a legitimate specification to the extent that the heterogeneity among companies is limited to a single population.

\(^{15}\) In this approach the term stochastic is referred to the fact that the model is stochastic (presence of a symmetric component of the disturbance \( v_{it} \)) but not the inefficiency term \( u_i \). In the approach suggested by Aigner et al. (1977) both the model and the inefficiency term are stochastic.

\(^{16}\) Battese and Coelli (1992) propose a method. See also Coelli, Rao and Battese (1998) for a summary.

\(^{17}\) See Baltagi (2001) and Hsiao and Sun (2000) for detailed discussions on fixed vs random effects.
2.2. Specification of the Cost Function

Electricity distribution utilities operate in networks with different shapes, which directly affect the costs. As discussed in Robert (1986), Salvanes and Tjøtta (1994) and Thompson (1997), the cost function should take into account differences in network characteristics, load factor and other factors that are unrelated to cost-efficiency but affect the costs. The specification used here draws basically from the model used by Filippini (1998). The output is measured by the total number of kWh delivered. Inputs to the electricity distribution process consist primarily of labor, capital and the power purchased from the generator. The costs of distribution utilities consist of two main parts: the costs of the purchased power and the network costs including labor and capital costs. There are therefore two alternatives for measuring cost efficiency in power distribution utilities: total costs approach and network costs approach. The network costs approach has a practical advantage in that the estimated average costs can be directly used in a price-cap formula. However, this approach neglects the potential inefficiencies in the choice of the generator. In this paper we use the first approach based on the total costs function. The firm's total cost of distributing electricity can be represented by:

\[ C = C(Y, P_K, P_L, P_P, LF, CU, AS, HGRID, DOT, DW, T) \]  

(4)

where \( C \) represents total cost; \( Y \) is the output in kWh; \( P_K \), \( P_L \) and \( P_P \) are respectively the prices of capital, labor and input power; and \( T \) is a time variable representing the linear trend in technological progress.

In addition to the above variables that are generally included in a cost function model, the following six output (and network) characteristics are included in the model: \( LF \) is the ‘load factor’ defined as the ratio of utility’s average load on its peak load; \( AS \) the size of the

---

18 Notice that the price cap is generally applied to the network access.
service area served by the distribution utility; and $CU$ is the number of customers. The load factor captures the impact of the intensity of use on costs.\textsuperscript{19} This variable is expected to have a negative effect on total costs, because a relatively high value of load factor represents a higher capacity utilization, thus lower fixed costs for producing a given output. Obviously, the service area and the number of customers are expected to have a positive effect on costs.

$HGRID$ is a binary indicator to distinguish the utilities that operate a high-voltage transmission network in addition to their distribution network. Some of the utilities in our sample are involved in auxiliary services such as installation of electric appliances. Both these types are expected to be more costly compared to other companies. The utilities whose share of auxiliary revenues is more than 25 percent of total revenues are distinguished by dummy variable $DOT$. The maintenance costs and damage risks of power lines are generally higher in forests. Binary indicator $DW$ represents the cases in which more than 40 percent of the service area is covered by forests.

It should be noted that there are other output characteristics that are not considered in the cost function (4). An important unobserved dimension of the output in our sample concerns the quality of service. Different utilities may deliver different levels of quality, which affect their costs. In power utilities, output quality is usually measured by the number of interruptions that are not related to rare natural accidents. In Switzerland however, there has been practically no such outage cases. The high level of quality is obviously related to the tight regulation and high quality standards applied in Switzerland. We therefore contend that the quality differences between the utilities in our sample are not significant. Moreover, the stochastic models used in our analysis control for unobserved variations that are not correlated with the included explanatory variables.

\textsuperscript{19} See Foreman-Peck and Waterson (1985) for a discussion of the role of load factor in cost models.
The regularity conditions require that the cost function in equation (4) be linearly homogeneous in input prices, non-decreasing in input prices and concave. The translog model and Cobb-Douglas form are two main functional forms commonly used in the literature. Translog form does not impose any technological restriction and allows the economies of scale, size and density vary with output. These values are assumed constant in the Cobb-Douglas functional form. In this paper, the Cobb-Douglas form is used for two main reasons. First, because of the large number of parameters in translog model there is a considerable risk of near-multicollinearity, especially given that different output variables $Y$, $AS$ and $CU$ are highly correlated. Moreover, the estimation of scale economies is of secondary importance in this paper. The assumption that scale-economies do not vary with output (implicit in Cobb-Douglas form) can therefore be justified to the extent that it does not affect the inefficiency estimators.

The Cobb-Douglas specification of the cost function in (4) can be written as:

$$
\ln\left(\frac{C}{P_p}\right) = \beta_0 + \beta_Y \ln Y + \beta_K \ln \left(\frac{P_k}{P_p}\right) + \beta_L \ln \left(\frac{P_L}{P_p}\right) + \gamma_1 \ln LF
$$

$$
+ \gamma_2 \ln AS + \gamma_3 \ln CU + \delta_1 HGRID + \delta_2 DOT + \delta_3 DW + \tau_T T
$$

(5)

Linear homogeneity in input prices is imposed by dividing money values by the price of the input power.

---


21 In our specification the number of parameters in translog model would be 40.

22 Our preliminary analysis suggests that this problem creates technical difficulties in our maximum likelihood estimations.
3. DATA

The data used in this paper consists of an unbalanced panel of 59 Switzerland’s distribution utilities over a 9-year period from 1988 to 1996. The sample includes 380 observations with a minimum of four observations per company. The original data set has been prepared and analyzed by Filippini (1998) and completed by Wild (2001) and Filippini and Wild (2001). These data are mainly based on the information from the annual reports of the Swiss Federal Statistical Office, the Swiss Federal Energy Office, and the Swiss Cities Association. A mail survey from the utilities has been used to complete the data.23 The sample does not include the utilities that generate more than 20 percent of their input power. There are about 900 electricity distribution companies in Switzerland. This sector is characterized by a large number of small companies along with a relatively small number of large firms. The 59 companies included in this study deliver about a third of Switzerland’s electricity consumption.24 The sample used in this study can therefore be considered as a representative sample of relatively large distribution utilities in the country. In spite of a considerable degree of variation in costs and other characteristics, the sample represents relatively similar companies compared to the entire sector.

Table 1 gives the summary statistics of the key variables used in the analysis. All money values are converted to 1996 Swiss Francs using the global consumer price index. The cost of purchased electricity is included in total costs. For those companies that produce part of their power the average price of input electricity is assumed to be equal to the price of purchased power. Labor price is defined as the average annual salary of the firm’s employees.

23 See Filippini and Wild (2001) for a more detailed description of data sources.

24 The total power delivered by the companies in our sample in 1993 is 13,250 GWh, which is about a third of the total 43,000 GWh electricity consumption in Switzerland in that year.
Capital expenditure is approximated by the residual costs that is, total costs minus labor and purchased power costs. Because of the lack of inventory data the capital stock is measured by the capacity of transformers.\textsuperscript{25}

| Table 1. Descriptive statistics (380 observations) |
|---------------------------------|------------|-------------|-------------|-------------|
|                                | **Mean**   | **Standard Deviation** | **Minimum** | **Maximum** |
| Total annual costs per kWh output (CHF) | .188       | .0303        | .128        | .323        |
| Annual output \((Y)\) in GigaWh      | 263.51     | 390.36       | 17          | 2301.5      |
| Number of customers \((CU)\)         | 26975.6    | 36935.8      | 2461        | 220060      |
| Load Factor \((LF)\)                | .5541      | .0727        | .3219       | .9817       |
| Service Area \((AS)\) in km\(^2\)    | 15,467     | 35,376       | 176         | 198,946     |
| Average annual labor price \((P_L)\) per employee (CHF 1000) | 101.27     | 32.55        | 43.36       | 253.89      |
| Average capital price \((P_K)\) in CHF per kVoltAmpere installed capacity | 95.06      | 39.35        | 32.08       | 257.98      |
| Average price of input power \((P_P)\) in CHF/kWh | .105       | .0210        | .0583       | .161        |
| High-voltage network dummy \((HGRID)\) | .35        | .4776        | 0           | 1           |
| Auxiliary revenues more than 25\% \((DOT)\) | .397       | .490         | 0           | 1           |
| Share of forest area more than 40\% \((DW)\) | .261       | .440         | 0           | 1           |

- All monetary values are in 1996 Swiss Francs (CHF), adjusted for inflation by Switzerland’s global consumer price index.

\textsuperscript{25} Transformer is the main device used to transfer electricity in the network. This is basically a device to convert current variations to voltage variations and vice versa.
4. ESTIMATION RESULTS

The estimated parameters of the cost frontier are listed in table 2. In the OLS model data are pooled across different years and the estimators are based on an implicit assumption that the unobserved random variations are not firm-specific. The other three models have the advantage of accounting for firms’ heterogeneity. In the random-effects model (GLS) it is assumed that firms’ unobserved heterogeneity is uncorrelated with their observed characteristics. The MLE model imposes an additional restriction that firms’ unobserved heterogeneity has a half-normal distribution. Both these assumptions are relaxed in the fixed-effects specification. This model however is mainly based on the variations within firms and cannot estimate the effect of time-invariant factors.

The first observation on the estimation results (table 2) is that virtually all coefficients are highly significant and have the expected signs. The within $R^2$ values are reported for the two panel models estimated by the least squares method. These values show that the adopted specification has a relatively high explanatory power. As it can be seen in the table, the fixed-effect estimators for output ($Y$) and costumers ($CU$) coefficients are quite different from other models. This contrasting difference suggests that the estimations could be sensitive to firm-specific characteristics. This result is not surprising in network industries. Any correlation between random effects and other explanatory variables may bias the estimation results. Therefore, in the absence of information regarding the unobserved heterogeneity among firms, the fixed-effect model can provide more reliable estimates for the factors that

26 It should be noted that the $R^2$ value in the OLS model does not consider the panel structure of the data and thus, cannot be used to evaluate the model’s explanatory power. In fact the extremely high value of $R^2$ in the OLS model can be explained by the relatively low within variations in costs.
vary over time. This advantage however, is hardly clear in our sample: A more careful examination of the results shows that other coefficients are rather similar among different models. Moreover, even though output coefficients $Y$ and $CU$ are different in the FE model, their sum is quite similar among different models. This result suggests that the value of economies of scale is robust to the econometric specification. Finally, the results of the Hausman specification test indicate that the FE and RE estimates are not significantly different at 5 percent significance level (p-value of .055). Overall, these results suggest that the cost function estimations are not sensitive to the unobserved heterogeneity among companies.

### Table 2. Cost frontier parameters

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Random-Effects (GLS)</th>
<th>Random-Effects MLE (Half-Normal)</th>
<th>Fixed-Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln Y$</td>
<td>0.851 0.017</td>
<td>0.780 0.032</td>
<td>0.767 0.040</td>
<td>0.677 0.055</td>
</tr>
<tr>
<td>$\ln CU$</td>
<td>0.084 0.017</td>
<td>0.153 0.033</td>
<td>0.163 0.048</td>
<td>0.251 0.096</td>
</tr>
<tr>
<td>$\ln AS$</td>
<td>0.044 0.004</td>
<td>0.051 0.009</td>
<td>0.047 0.013</td>
<td>-</td>
</tr>
<tr>
<td>$\ln LF$</td>
<td>-0.243 0.037</td>
<td>-0.239 0.039</td>
<td>-0.230 0.023</td>
<td>-0.213 0.044</td>
</tr>
<tr>
<td>$\ln PL$</td>
<td>0.067 0.011</td>
<td>0.041 0.014</td>
<td>0.039 0.014</td>
<td>0.038 0.016</td>
</tr>
<tr>
<td>$\ln PK$</td>
<td>0.200 0.009</td>
<td>0.174 0.010</td>
<td>0.171 0.005</td>
<td>0.169 0.010</td>
</tr>
<tr>
<td>HGRID</td>
<td>0.063 0.012</td>
<td>0.075 0.027</td>
<td>0.098 0.039</td>
<td>-</td>
</tr>
<tr>
<td>DOT</td>
<td>0.033 0.010</td>
<td>0.050 0.022</td>
<td>0.040 0.028</td>
<td>-</td>
</tr>
<tr>
<td>DW</td>
<td>0.014 0.010</td>
<td>0.012 0.023</td>
<td>0.062 0.028</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>0.001 0.002</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.002</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.236 0.233</td>
<td>-0.793 0.369</td>
<td>-0.653 0.430</td>
<td>-</td>
</tr>
<tr>
<td>R Square</td>
<td>0.995 0.711</td>
<td>0.711</td>
<td>0.714</td>
<td></td>
</tr>
</tbody>
</table>

- The reported $R^2$ values for GLS and FE models are based on the within variations.

17
It should be noted that both random and fixed effects models used in this analysis, assume that the firm-specific stochastic component represents the efficiency differences among firms.\(^{27}\) This assumption leads to an overestimation of inefficiency in the FE model for the following reasons. First, the fixed firm-specific effects capture both observed and unobserved time-invariant factors. Moreover, since the fixed effects do not follow any distribution and efficiency is estimated compared to the best observed practice (the firm with the minimum fixed effect), the estimators are sensitive to outliers. In fact, the problem of outlier firms is transferred from the cost function to efficiency estimators. To illustrate this fact, several specifications were compared. We started from a ‘naïve’ OLS model that ignores the time-invariant factors and refined this model in several steps until all time-invariant factors were included. The inefficiency scores obtained from these models were compared to those obtained from a fixed-effects model. The results (not given here) indicate that as the OLS model becomes more ‘refined’ the estimated inefficiency scores show less correlation with those obtained from the fixed-effects model.\(^{28}\) These results suggest that as far as inefficiency scores are concerned the performance of a fixed-effect model is quite poor (even compared to a naïve OLS model).

In order to see the limitations of these models we also study the inefficiency estimates obtained from different models. Table 3 gives the summary statistics of the inefficiency scores resulted from different models. The inefficiency score is defined as \(\exp(U_i)\), where \(U_i\) is the inefficiency term obtained from the regression model. In the COLS model where the inefficiency term is time-variant, the company’s inefficiency \(U_i\) is assumed to be the average

\(^{27}\) In fact with the exception of a few recent developments (cf. Greene, forthcoming; Tsionas, 2002), this is a general assumption in panel data frontier models.

\(^{28}\) This is also the case for the mean and median inefficiency scores (and other quartiles). That is, as the OLS model get more refined the summary statistics decrease and get farther from that of the fixed-effects model.
of $u_{it}$ over the entire sample period. The scores therefore represent the ratio of a company’s actual costs to a minimum level that would have been achieved had the company operated as cost-efficient as the ‘best practice’ observed in the sample. The excessively large values resulted from the fixed-effect model confirms the poor performance of this model in estimating inefficiencies.

### Table 3. Summary statistics of inefficiency scores

<table>
<thead>
<tr>
<th></th>
<th>COLS</th>
<th>GLS</th>
<th>FE</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1.07</td>
<td>1</td>
<td>1</td>
<td>1.07</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.46</td>
<td>1.38</td>
<td>2.14</td>
<td>1.36</td>
</tr>
<tr>
<td>Average</td>
<td>1.23</td>
<td>1.16</td>
<td>1.35</td>
<td>1.15</td>
</tr>
<tr>
<td>Median</td>
<td>1.22</td>
<td>1.16</td>
<td>1.31</td>
<td>1.13</td>
</tr>
<tr>
<td>95 percentile</td>
<td>1.41</td>
<td>1.32</td>
<td>1.66</td>
<td>1.30</td>
</tr>
<tr>
<td>Number of firms</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
</tbody>
</table>

In practice, benchmarking is usually based on efficiency ranking of companies. The correlation matrix between the ranks obtained from different models is given in table 4. The ranks are obtained by comparing the firms’ average inefficiency scores over the sample period. These results indicate a relatively high correlation between rankings from GLS and MLE models.

### Table 4. Correlation between inefficiency ranks from different models

<table>
<thead>
<tr>
<th></th>
<th>COLS</th>
<th>GLS</th>
<th>FE</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLS</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE (GLS)</td>
<td>0.936</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE</td>
<td>0.447</td>
<td>0.514</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>RE (MLE)</td>
<td>0.838</td>
<td>0.895</td>
<td>0.417</td>
<td>1</td>
</tr>
</tbody>
</table>

To see the individual differences the GLS and MLE models are compared regarding the inefficiency scores. The FE model is not considered here because as discussed above, its
efficiency scores are significantly different from other models. Table 5 summarizes the results. These results indicate that the maximum difference of cost-inefficiency between the two models is about 9 percent. This difference is quite significant noting that both models are based on the same Cobb-Douglas functional form and their only difference is in the distribution of the efficiency term. A closer look at the rankings highlights these differences. Our results show that changing from one model to another results in significant changes in rankings. For instance for more than half of the companies in the sample changing the model from GLS to MLE implies a change of 4 places or more in their ranks, and for about 25 percent of companies this means a change of 9 places or more. Even the ranking quartile changes considerably. Change of the model from MLE to GLS results in a change in ranking quartile for 20 companies out of 59. This change results in a change of quartile for about a third of the companies in the first quartile (the 25 percent most efficient firms). Moreover, different models do not identify the same companies as the best and worst practices. The best practice as identified by the MLE method is ranked 17 in the GLS model, whereas the GLS model’s best practice is ranked 7 by the MLE model. These differences are as more striking as the two models differ only in their assumption on the distribution of the inefficiency term.

Therefore, the mutual consistency conditions proposed by Bauer et al. (1998)\(^\text{29}\) are not satisfied. These results show the sensitivity of the benchmarking method in our sample. In contrast with the general contention in previous studies that different approaches give rather similar inefficiency rankings, this analysis suggests that rankings may be sensitive to the adopted model. Therefore, a direct use of inefficiency estimates in benchmarking regulation

\(^{29}\) According to Bauer et al., to be mutually consistent, different approaches should have the following conditions: inefficiency scores should have comparable means, standard deviations and other distributional properties; the ranking order of the firms should be approximately the same; and the ‘best-practice’ and ‘worst-practice’ firms should be mostly the same.
of network industries may be misleading. In usual cases where the choice of the appropriate model is not clear, a sensitivity analysis could be used to study the robustness of the results and the limitations of different models.

Table 5. Summary statistics of the absolute value of difference in inefficiency scores

<table>
<thead>
<tr>
<th>Model:</th>
<th>GLS and COLS</th>
<th>GLS and MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.14</td>
<td>0.09</td>
</tr>
<tr>
<td>Mean</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Median</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>95 percentile</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>N</td>
<td>59</td>
<td>59</td>
</tr>
</tbody>
</table>

Cost frontier models can alternatively be used by the regulator to predict the costs of individual companies. Three specifications, OLS, fixed effects (FE) and random effects (GLS), are compared with respect to their predictive power. Predictions are considered in two directions: out-of-sample prediction which consists of predicting the costs of a given firm using the estimations obtained from other firms, and forecasting that involves the prediction of costs in a year using the estimation based on the data previous to that year. One, two and three-year-ahead forecasts are considered. In all predictions the actual values of explanatory variables are used. Prediction errors are defined as the predicted total costs minus the actual costs divided by the actual costs. In the RE model the forecasts are based on the optimal predictor given in Bailli and Baltagi (2000).30 The prediction errors of different models are compared. The results are summarized in table 6. The two-year-ahead forecast errors are not listed.

30 See page 256 of Bailli and Baltagi (1999) for more details.
Table 6. Prediction errors

<table>
<thead>
<tr>
<th>Type of prediction</th>
<th>Out-of-sample</th>
<th>1-year-ahead</th>
<th>3-year-ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>GLS</td>
<td>FE</td>
</tr>
<tr>
<td>Average error</td>
<td>7.37</td>
<td>7.57</td>
<td>11.8</td>
</tr>
<tr>
<td>(absolute value)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum error</td>
<td>25.2</td>
<td>27.3</td>
<td>39.0</td>
</tr>
<tr>
<td>(absolute value)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95 percentile error</td>
<td>17.8</td>
<td>20.1</td>
<td>31.1</td>
</tr>
<tr>
<td>(absolute value)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average prediction</td>
<td>.34</td>
<td>.98</td>
<td>1.08</td>
</tr>
<tr>
<td>bias (average error)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of predictions</td>
<td>380</td>
<td>380</td>
<td>380</td>
</tr>
</tbody>
</table>

- Errors are given in percentage of the actual costs.

As expected, the out-of-sample estimation errors are significantly higher in the fixed-effects model. Interestingly, even the forecasting performance of the GLS model is comparable to that of the FE model. This implies that in our sample the FE model does not provide a significant predictive advantage over other models. The results show that from a practical standpoint, the prediction errors are generally within an acceptable range. The GLS model shows the best performance. As it can be seen in table 6 (see the middle column), the GLS model’s one-year-ahead prediction errors with an average absolute value of 3 percent and an average value of less than 0.1 percent are particularly low. While the maximum error in this case is 10.3%, for 95 percent of the companies the prediction error is limited to 6.8%.

These results suggest that the panel data models can predict individual companies’ total costs with a rather reasonable precision. Therefore, the regulator can use these models to predict a confidence interval for the costs of each one of the firms. Acceptable intervals for revenue and price caps can be calculated accordingly. Using such predictions along with other monitoring instruments, the regulator can hold the companies within a reasonably well-predicted range of cost-efficiency.
5. CONCLUSION

Four different parametric cost frontier models are applied to a panel data set of electricity distribution utilities in Switzerland. A comparison of the estimation results indicates significant differences among different models. This result can be explained by the strong unobserved heterogeneity among distribution companies, which is a common characteristic of network industries. These differences are particularly important for the estimates of inefficiency scores. While the summary statistics of inefficiency estimates are not sensitive to model specification, the efficiency ranks change quite significantly from one model to another. The alternative models are not found to be ‘mutually consistent’ with respect to inefficiency measures. These results point to an important shortcoming of the benchmarking methods in network industries. Furthermore, the results confirm that the sensitivity problems reported in the previous literature are not limited to cross sectional data. Given that the regulators actually use some of these methods in practice, the analysis in this paper has an important implication suggesting that benchmarking analysis should be applied with caution. In particular, it is recommended that several models be used and compared. A sensitivity analysis should be performed to identify the limitations of different models.

This paper also uses different cost frontier panel data models to predict the firms’ costs. In particular, the out-of-sample prediction performance of different models is studied. The prediction errors are within acceptable limits from a practical point of view. The results suggest that stochastic frontier models can be used to gain information about costs of individual firms. Moreover, certain models with apparent limitations in the estimation of cost-frontier parameters have a relatively good performance in predicting costs. Although this conclusion may be limited to the data used in this paper, the results suggest that cost frontier
models can be used as a complementary control instrument in order to narrow the information
gap between the regulator and regulated companies. An interesting example is provided by
Antonioli and Filippini (2001) in the regulation of water supply in Italy, where a yardstick
competition model in line with Schleifer (1985) has been applied. This regulation method is
based on an interactive approach: The company proposes its tariff in the first stage. The
regulator estimates a price cap for the firm using a benchmarking analysis and adjusting for
observed differences among companies. The proposed tariff is approved if it does not exceed
an acceptable range around the estimated price cap. Otherwise, the tariffs can be renegotiated
with the requirement that the company justify its excessive tariff before any revision. The
interaction between the regulator and companies may be helpful in the face of information
asymmetry. In order to provide a disincentive to renegotiate a penalty can be imposed on the
companies that do not accept the first-stage prices. In case of disagreement the regulator
performs a more detailed analysis with additional information from individual companies and
offers a new price. The probability of disagreement and the flexibility of the regulator depend
on the prediction power of the adopted econometric models.

Acknowledgements

The authors are grateful to Jörg Wild for useful discussions and his help in preparing the data.
We benefited from the suggestions of the editor and two anonymous referees, which are
gratefully acknowledged. We would also like to thank the participants of the Annual
Conference of the International Association for Energy Economics (Prague, June 2003)
particularly Tooraj Jamasb and Einar Hope for their helpful comments. The views expressed
in this paper are strictly personal and the responsibility for all remaining errors lies solely
with the authors.
References


An Analysis of Cost-Efficiency in Swiss Multi-utilities

Mehdi Farsi, Massimo Filippini

CEPE Working Paper No. 60
May 2008

CEPE
Zurichbergstrasse 18 (ZUE E)
CH-8032 Zurich
www.cepe.ethz.ch
AN ANALYSIS OF COST EFFICIENCY IN SWISS MULTI-UTILITIES *

Mehdi Farsi  Massimo Filippini
mfarsi@ethz.ch  mfilippini@ethz.ch

Department of Management, Technology and Economics, ETH Zurich
Zurichbergstr. 18, CH-8032 Zurich, Switzerland
Tel. +41-44-632 06 50, Fax. +41-44-632 10 50

and

Department of Economics, University of Lugano

Abstract

This study presents an empirical analysis of the cost efficiency of a sample of Swiss multi-utilities operating in the distribution of electricity, natural gas and water. The multi-utilities that operate in different sectors are characterized by a strong unobserved heterogeneity. Therefore the measurement of their performance poses an important challenge for the regulators. The purpose of this paper is to study the potential advantages of recently developed panel data stochastic frontier models in the measurement of the level of efficiency for multi-utility companies. These models are estimated for a sample of 34 multi-output utilities operating from 1997 to 2005. The alternative models are compared regarding the cost function slopes and inefficiency estimates. For the inefficiency estimates, the correlation between different models and the effect of econometric specification have been analyzed. The results suggest that the inefficiency estimates are substantially lower when the unobserved firm-specific effects are taken into account.

JEL Classification: C33, D24, L11, L25, L94, L95

Keywords: cost function; efficiency; panel data; multi-output utilities

* This paper is based upon the results of a research project (Farsi and Filippini, 2007) sponsored by Switzerland’s State Secretariat for Economic Affairs (SECO). We gratefully acknowledge their financial support but also the extraordinary cooperation of their staff especially Peter Balastèr, throughout the project. We also wish to thank Aurelio Fetz for compiling the data used in this study. The views expressed here do not necessarily reflect the position of any sponsoring agency.
1. Introduction

Along with the recent waves of liberalization and deregulation in network industries throughout Europe, the authorities are increasingly concerned about the productive efficiency of the utilities that, due to their natural monopoly characteristics, are not fully liberalized. In sectors such as power, gas and water distribution, because of the considerable economies of scale, a direct introduction of competition is not optimal. Instead, incentive regulation has been used to ensure (or maximize) the productive efficiency of the locally monopolistic companies. Everywhere in Europe, the traditional regulatory systems are being gradually replaced by incentive regulation schemes. Unlike the traditional contracting systems based on a reasonable rate of return, the incentive contracts are designed to induce incentives for reducing costs and increasing efficiency. Most incentive regulation schemes use benchmarking to evaluate the productive performance of the regulated companies in order to reward/punish them accordingly. Based on their efficiency performance, companies are allowed to retain part of their profits/savings through either differentiated price caps or adjustments in budget or network access fees.

Several OECD countries have already integrated a benchmarking practice in their regulation systems for electricity distribution networks (Farsi, Fetz and Filippini, 2007a; Crouch, 2007). A few countries have also introduced such incentive schemes based on performance in their water industry (Saal et al., 2007; Antonioli and Filippini, 2001).\(^1\) The application of benchmarking methods in the gas sector is probably not as advanced as that observed in the electricity industry. However, the use of incentive schemes based on performance has been proposed in several studies (cf. Casarin, 2007). In spite of a relatively common usage in each one of the distribution sectors, the direct application of benchmarking analysis in the regulation of multi-utilities has hardly been explored. This is especially interesting in Switzerland and some other

---

\(^1\) In Switzerland the distribution utilities are monitored and regulated by cantonal and federal governments. Although Switzerland has not yet implemented any incentive regulation system, the actual debates suggest that the regulators will probably follow similar reforms in the near future.
European countries, where multi-utilities dominate the distribution sectors in electricity, natural gas and water.

To our knowledge there is no reported empirical application of efficiency measurement in the multi-utility sector. This may perhaps be considered in line with arguments in favor of unbundling the multiple-utilities into separate legal entities. In fact, horizontal unbundling is a recurring subject of the public policy debates both in the EU and Switzerland. However, the dominance of multi-utilities in Switzerland is not expected to be affected by the ongoing reforms. According to the observed tendencies in the EU regulatory reforms, the multi-utilities especially those with moderate and small networks (less than 100,000 customers), will remain exempt from unbundling requirements.

Noting the importance of multi-utilities in many countries an important question is whether the benchmarking methods can be applied to multi-utilities as well as single-output distribution utilities. It is often argued that an accounting unbundling is sufficient for applying separate benchmarking analyses to each branch of a multi-utility. However, due to the fact that in certain situations only part of these sectors are regulated with incentive regulation schemes, a company could artificially shift part of the costs to the sector for which the regulation does not foresee a benchmarking process. Similarly, because of the different levels of incentive regulation across various sectors of a single firm, the management might focus their efforts in one sector, thus permit slackness in others. Moreover, extending single-sector benchmarking to multi-utilities requires pooling the data from single-output distributors with the corresponding branches of the multi-utilities. The latter units, benefiting from the economies of scope, are arguably not comparable to specialized firms, thus might bias the efficiency estimates. In these cases, a benchmarking across the entire operation of multi-utilities might be more relevant than separate benchmarking analyses for individual sectors. In many cases, with a mixture of mutli-output utilities and specialized distributors, the two types of analyses can also be combined to assess the potential differences among firms and also across sectors.

The effectiveness of the regulation systems relies upon the accuracy of estimated efficiency levels of individual companies. However, due to a great variety of available methods of efficiency measurement and the observed discrepancy of results
across different methods, benchmarking practice requires a methodology to adopt a single model among several legitimate approaches and specifications. This task is particularly complicated in network utilities in which unobserved firm-specific factors might be confounded with inefficiency. Obviously the problem of unobserved heterogeneity is more important in multi-output distributors that operate in several networks, each of which could have different types of cost drivers with specific characteristics.

Unobserved firm-specific heterogeneity can be taken into account with conventional fixed or random effects in a panel data model. In order to distinguish external heterogeneities from cost efficiency, Greene (2005a) proposed a model that integrates an additional stochastic term representing inefficiency in both fixed and random effects models. These models assume that the firm-specific heterogeneity does not change over time but sources of inefficiency vary both across firms and over time. In this paper we use a ‘true random-effects’ model, which is a random-constant frontier model, obtained by combining a conventional random-effects model with a skewed stochastic term representing inefficiency. The extended model includes separate stochastic terms for latent heterogeneity and inefficiency.

The empirical results reported in the literature obtained from true random effects models suggest that modeling unobserved heterogeneity could significantly decrease the inefficiency estimates. This could lend certain support to the application of benchmarking methods in the regulation of strongly heterogeneous network industries, in which the conventional inefficiency estimates appear to be overstated. Provided that they can sufficiently control for the unobserved heterogeneity across firms, these methods can be used to have a better estimate of cost-inefficiency in the sector or at individual companies.

---

2 Kumbhakar (1991) proposed a similar approach using a three-stage estimation procedure. See also Heshmati and Kumbhakar (1994) and Kumbhakar and Hjalmarsson (1995) for two applications

The purpose of this paper is to study the potential advantages of these extended models in an application to Switzerland’s multi-output utilities. The models are estimated for a sample of 34 companies operating in Switzerland from 1997 to 2005. The alternative models are compared regarding the cost function slopes and inefficiency estimates. For the inefficiency estimates, the correlation between different models and the effect of econometric specification have been analyzed. The results suggest that the inefficiency estimates are substantially lower when the unobserved firm-specific effects are taken into account.

The rest of the paper is organized as follows: Section 2 presents the model specification and the methodology. The data are explained in section 3. Section 4 presents the estimation results and discusses their implications, and section 5 provides the conclusions.

2. Stochastic frontier models for panel data

The methods used for measuring technical, allocative and cost efficiency are commonly referred to as frontier approaches, classified into two main categories of linear programing methods and econometric approaches. The latter group, also known as Stochastic Frontier Analysis (SFA) is easily adaptable to panel data structure and therefore used in this study. In SFA models, first developed by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), the regression residuals are decomposed into a symmetric component representing statistical noise and a skewed term one representing inefficiency.

As opposed to cross-sections, in panel data the repeated observation of the same company over time allows an estimation of unobserved firm-specific factors, which might affect costs but are not under the firm’s control. Individual companies operate in different regions with various environmental and network characteristics that are only partially observed. It is crucial for the regulator to disentangle such exogenous heterogeneities from inefficiency estimates. However the distinction between these

---

4 Murillo-Zamorano (2004) provides an account of advantages and shortcomings of each group. Other interesting surveys are Coelli et al. (2005), Simar (1992) and Kumbhakar and Lovell (2000).
two unobserved terms requires certain assumptions based on judgment. In early applications of SFA models to panel data (Pitt and Lee, 1981; Schmidt and Sickles, 1984; Battese and Coelli, 1988), the common assumption was that the productive efficiency is a time-invariant characteristic that can be captured by firm-specific effects in a random or fixed effects model.

A general form of a cost frontier based on these models can be written as:

$$\ln C_{it} = f(y_{it}, w_{it}) + u_{i} + v_{it}. \quad (1)$$

where subscripts $i$ and $t$ denote the firm and the operation year, $C$ is the cost variable usually in logarithms and $y$ and $w$ are respectively vectors of outputs and input factor prices. The time-varying error component $v_{it}$, typically a normal variable, represents the unobserved heterogeneity and random errors, whereas the time-invariant term $u_{i}$ is assumed to represent excess costs due to inefficiency. The latter term is considered with different distributions: While Pitt and Lee (1981) adopt a half-normal distribution that is, a normal distribution truncated at zero. Battese and Coelli (1988) extends the model to non-zero truncation points and Schmidt and Sickles (1984) propose two variations in which they relax the distribution assumptions respectively using Generalized Least Squares (GLS) and fixed-effect estimators. In particular, in the latter model, the individual effects $u_{i}$ can be correlated with the explanatory variables.

In more recent papers the random effects model has been extended to include time-variant inefficiency. Cornwell, Schmidt and Sickles (1990), Kumbhakar (1990), and Battese and Coelli (1992) are the important contributions that consider a time function to account for variation of efficiency. In particular the former paper proposes a flexible function of time with parameters varying among firms. In all these models, however, the unobserved external heterogeneity is suppressed in an iid error term across observations. This implies that the cost variations due to factors other than firm’s efficiency are randomly assigned to each observation. This could be a restrictive assumption in network industries in which certain external cost drivers specific to environment and/or network complexity remain practically unchanged over fairly long periods of time.

To the extent that environmental factors and network characteristics do not change considerably over time, associating the time-invariant excess costs to external
factors rather than inefficiency can be a sensible assumption. On the other hand, improvements in efficiency are usually linked to a dynamic learning process and adaptation to new technologies. Therefore, it can be assumed that inefficiencies are captured by the time-varying excess costs. These assumptions combined with the distribution assumption in line with the original frontier model allow a disentanglement of inefficiencies from firm-specific heterogeneity captured by panel’s individual effects.\(^5\)

In fact, the SFA model in its original form (Aigner, Lovell and Schmidt, 1977) can be readily extended to panel data models, by adding a fixed or random effect in the model. Although similar extensions have been proposed by several previous authors,\(^6\) Greene (2005a,b) provides effective numerical solutions for both models with random and fixed effects, which he respectively refers to as “true” fixed and random effects models. Several recent studies such as Greene (2004), Farsi, Filippini and Kuenzle (2005), Alvarez, Arias and Greene (2004) and Tsionas (2002) have followed this line. Some of these models have proved a certain success in a broad range of applications in network industries in that they give more plausible efficiency estimates.\(^7\) These results raise an important question as to what extent the panel-data-adapted models can be used to have a better understanding of the inefficiencies and whether they can provide a reliable basis for benchmarking and incentive regulation systems in industries characterized by strong heterogeneity. This question is especially important in the multi-utility sector, in which the companies operate in multiple networks, entailing several network-specific heterogeneity dimensions.

Greene’s (2005a) ‘true’ cost frontier model can be written as:

\[^5\] There are evidently other feasible econometric specifications that can incorporate these assumptions. A remarkable example is the flexible framework proposed by Sickles (2005).

\[^6\] In particular Kumbhakar (1991) proposed a three-stage estimation procedure to solve the model with time- and firm-specific effects, Polachek and Yoon (1996) estimated a panel data frontier model with firm dummies and Heshmati (1998) used a two-step procedure in a random-effect framework to separate the firm-specific effects from efficiency differences.

\[ \ln C_{it} = f(y_{it}, w_{it}) + \alpha_i + u_{it} + v_{it}. \] (2)

The term \((\alpha_i)\) is a normal \textit{i.i.d.} in random-effects framework, or a constant parameter in fixed-effects approach. \(u_{it}\) and \(v_{it}\) are respectively a half-normal variable representing inefficiency and a normal random variable that captures the statistical noise. In this study, we used the true random effect model, mainly because the numerical solution of the fixed effects model was cumbersome and did not converge to sensible results for the estimates of inefficiencies and individual intercepts. In order to provide a basis for comparing the results, three other models namely, Pitt and Lee (1981), Battese and Coelli (1992) and a GLS model in line with Schmidt and Sickles (1984) have also been considered. These models will be described in the next section.

3. Data and model specification

The data used in this study includes financial and technical information from a sample of electricity, natural gas and water distribution companies that have operated in Switzerland between 1997 and 2005. The data have been mainly collected from the annual reports. Information on the size of the firm’s distribution area is from the “Arealstatistik 2002” published by the Federal Statistical Office and the “Preisüberwacher”. The original data set covers about 90 companies covering about 42% of total electricity, 67% of total gas and 22% of total water distribution in Switzerland. That sample includes multi-utility firms as well as specialized companies in electricity, gas and water sectors and several double-output utilities, but excludes companies with more than 10% self-generation of total electricity distribution.

Since the focus of this study is on the horizontal integrated multi-utilities, we focused on a sub-sample of the data used by Farsi, Fetz and Filippini (2008),\(^8\) includ-

\(^8\) In that study we analyzed the economies of scope and scale in Swiss multi-utilities using a quadratic cost function without performing a frontier analysis. In contrast with the present study, the estimation of the economies of scope requires data from the integrated multi-utilities as well as specialized distributors. Pooling the data across different types of utilities is not appropriate for a benchmarking analysis that relies on comparing comparable companies.
ing observations from 34 companies. Moreover, as pointed out by Saal and Parker (2006) assuming a similar cost frontier among multi-output companies and specialized utilities is not a realistic assumption and might cause considerable distortion in efficiency estimates and ranking. Because the primary purpose of this analysis is the estimation of cost-efficiency, we did not pool the multi-utilities with specialized companies.

The final sample used in this analysis consists of an unbalanced panel data set including observations from 34 multi-utilities during the nine-year period spanning from 1997 to 2005. The sample represents about 60% of the integrated multi-utilities in Switzerland. According to our estimates based on the available information, the multi-utilities included in the sample cover about half of the national electricity and gas consumption provided by multi-utilities and about a fifth of the water distributed by multi-utilities. Overall, these companies cover approximately 13% of electricity, 38% of gas and 14% of water distribution in the entire country.

The model specification is based on a cost function with three outputs namely, the distributed electricity, gas and water and four input factors that is, labor and capital as well as the electricity and gas inputs. As in Sing (1987) customer density is introduced as a service area characteristic. This variable should capture, at least partially, the cost impact of the heterogeneity of the service area of the companies. In fact, differences in networks and environments influence the production process and therefore the costs. Obviously, the heterogeneity of the service area cannot be summarized into a single variable. However, the available data do not allow for any other environmental or network characteristic that is reasonably independent of the included explanatory variables. Given the risk of multi-collinerarity in the translog function, especially in the second-order terms, we preferred to retain a relatively simple specification. Thus, some of these characteristics are inevitably omitted from the cost function specification. As we see later these omitted factors are represented by firm-specific stochastic components in the adopted panel data econometric models.

Assuming that the technology is convex and the firm minimizes cost, the adopted total cost function can be written as:

$$ C = C(q^{(1)}, q^{(2)}, q^{(3)}, r, w^{(0)}, w^{(1)}, w^{(2)}, w^{(3)}, D_t), $$

(3)
where $C$ represents total costs; $q^{(1)}$, $q^{(2)}$ and $q^{(3)}$ are respectively the distributed electricity, gas and water during the year, $w^{(0)}$, $w^{(1)}$, $w^{(2)}$ and $w^{(3)}$ are respectively the input factor prices for capital and labor services and the purchased electricity and gas; $r$ is the customer density measured by the number of customers divided by the size of the service area measured in square kilometers; and $D_t$ is a vector of year dummies that represent technical change and other year-to-year variations with the first year of the sample (1997) as the omitted category.\(^9\) The technical change is assumed to be neutral with respect to cost minimizing input ratios, that is, it is represented by a cost shift that does not alter the optimal input bundles.

An important implication of the above specification is that the estimated economies of scale are based on the usual assumption (in line with Caves et al., 1981) that any change in the production scale entails a uniform proportional change in all outputs and network characteristics, thus retaining the same ratios in particular the same customer density. This assumption is consistent with many policy applications such as the economic assessment of mergers and acquisitions and the extension of local monopolists to new areas. However, the potential synergies could be understated in other cases such as the assessment of side-by-side competition, where considerable economies might also be achieved by increasing the density, namely the economies of density.\(^10\) Unfortunately, the sample’s independent variations in networks and outputs do not seem to be sufficient for a meaningful empirical distinction between the economies of scale and the economies of density. In fact, our preliminary analyses with several alternative specifications particularly, models including the size of the service area and/or the number of customers, indicated certain discrepancy in the signs and statistical significance of output coeffi-

\(^9\) As we will see later our regressions suggest that the time-variation of costs is not linear. These variations can be explained by many unobserved factors (such as changes in labor contracts or seasonal composition of the demand) that change uniformly across companies.

\(^{10}\) The economies of output (customer) density describe the effects of changes in output (number of customers) keeping all other network characteristics fixed (Caves et al., 1985; 1984). As illustrated in Farsi, Filippini and Kuenzle (2007, 2006), the economies of density are usually greater than the economies of scale.
cients, which can be explained by multicollinearity problems due to the strong correlation of output variables with those characteristics.

The variables for the cost function specification were constructed as follows. Total costs \((C)\) are calculated as the total firm’s expenditures in a given year. The outputs \(q^{(m)}\) are measured by the total quantity delivered to the customers. The measurement units are GWh for electricity and gas and million cubic meters for water. Input prices are defined as factor expenditures per factor unit. Following Friedlaender and Chiang (1983), we used the residual approach for estimating the capital prices. The residual costs are specified as the company’s total costs net of labor expenditures and purchases of electricity and natural gas. Capital price for each network is obtained by dividing the residual costs by the capital stock measured by the network length. The overall capital price \((w^{(0)})\) is then calculated as a weighted average of capital prices for each of the three sectors namely, electricity, natural gas and water. The weights, similar to Fraquelli, Piacenza et al. (2004), are proportional to the share of the residual costs in each sector out of the multi-utility’s total residual costs. Labor price \((w^{(1)})\) is defined as the ratio of annual labor costs to the total number of employees in terms of full time equivalent worker. In a few cases in which the full time equivalent was not available, in order to avoid the underestimation of labor price due to part-time employees, we considered a correction based on the mean labor price values within the same canton. The electricity and gas prices \((w^{(2)}, w^{(3)})\) are defined as the expenditures of purchasing the input factors divided by the amount purchased (in MWh).

Table 1 provides a descriptive summary of the variables included in the model. All the costs and prices are adjusted for inflation using consumer price index and are measured in year 2000 Swiss Francs (CHF). As can be seen in the table, the sample shows a considerable variation in costs and all three outputs.
Table 1: Descriptive statistics (237 observations from 34 companies)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>11.20</td>
<td>41.10</td>
<td>77.60</td>
<td>503.00</td>
</tr>
<tr>
<td>$q^{(1)}$</td>
<td>38.78</td>
<td>126.89</td>
<td>293.23</td>
<td>2'023.59</td>
</tr>
<tr>
<td>$q^{(2)}$</td>
<td>28.82</td>
<td>226.34</td>
<td>512.60</td>
<td>4'294.20</td>
</tr>
<tr>
<td>$q^{(3)}$</td>
<td>0.78</td>
<td>2.45</td>
<td>5.28</td>
<td>33.35</td>
</tr>
<tr>
<td>$r$</td>
<td>44.35</td>
<td>298.33</td>
<td>387.57</td>
<td>1'554.09</td>
</tr>
<tr>
<td>$w^{(0)}$</td>
<td>11'853</td>
<td>31'167</td>
<td>38'385</td>
<td>234'796</td>
</tr>
<tr>
<td>$w^{(1)}$</td>
<td>77'789</td>
<td>106'466</td>
<td>107'851</td>
<td>146'816</td>
</tr>
<tr>
<td>$w^{(2)}$</td>
<td>44.6</td>
<td>107.4</td>
<td>105.9</td>
<td>163.5</td>
</tr>
<tr>
<td>$w^{(3)}$</td>
<td>16.6</td>
<td>28.4</td>
<td>29.3</td>
<td>63.2</td>
</tr>
</tbody>
</table>

Following Christensen et al. (1973) we use a translog model which is probably the most widely used functional form in empirical studies of cost and production functions.\(^{11}\) This flexible functional form is a local, second-order approximation to any arbitrary cost function. The approximation point is usually set at the sample mean or median. Here the approximation point has been set at the sample median. Compared to the mean, the median values are less affected by outlier values. The translog form does not impose any restrictions on the elasticity of substitution and allows the economies of scale to vary with the output level. In order to avoid the excessive number of parameters we have considered a homothetic cost function in which the interaction terms between input price variables and output variables are

\(^{11}\) See Caves et al. (1980) on the advantages of translog form in multiproduct settings and Griffin et al. (1987) for a discussion of the criteria used for the choice of the functional form.
excluded. This brings about another assumption namely that marginal costs particularly cost complementarities and scale elasticities depend only upon the technological characteristics of the production, thus being independent of input prices. This is a valid assumption insofar as the input prices remain in a reasonable range, especially because the potential changes in the shape of the cost function can easily be dominated by other approximations entailed by the functional form.

It is generally assumed that the cost function is the result of cost minimization given input prices and output and should therefore satisfy certain properties. Mainly, this function must be non-decreasing in output and non-decreasing, concave and linearly homogeneous in input prices (Cornes, 1992). We imposed the latter condition by normalization of prices namely, by dividing the costs and all factor prices by one common factor price referred to as numeraire (cf. Farsi, Fetz et al., 2007b; Featherstone and Moss, 1994; Jara-Diaz, Martinez-Budria et al., 2003). Here we used the capital price as the numeraire. The remaining conditions can be tested based on the estimation results.

The general econometric specification of the cost function in (3) can be written as:

\[
\ln\left(\frac{C_{it}}{W_{it}(0)}\right) = \sum_{m} \alpha_{it}^{m} \ln q_{it}^{(m)} + \alpha_{iT}^{r} \ln r_{it} + \sum_{k} \beta_{ik}^{r} \ln \frac{w_{ik}^{(k)}}{w_{ik}^{(0)}} + \frac{1}{2} \sum_{m} \alpha_{it}^{mm} \left(\ln q_{it}^{(m)}\right)^{2}
+ \sum_{m(n \neq n)} \sum_{n} \alpha_{it}^{mn} \ln q_{it}^{(n)} \ln q_{it}^{(m)} + \frac{1}{2} \alpha_{it}^{rr} \left(\ln r_{it}\right)^{2} + \sum_{m} \alpha_{it}^{m} \ln q_{it}^{(m)} \ln r_{it}
\]

\[
+ \frac{1}{2} \sum_{k} \beta_{kk}^{r} \left(\ln \frac{w_{ik}^{(k)}}{w_{jk}^{(0)}}\right)^{2} + \sum_{k(k \neq m)} \sum_{l} \beta_{kl}^{r} \ln \frac{w_{ik}^{(k)}}{w_{jk}^{(0)}} \ln \frac{w_{il}^{(l)}}{w_{jk}^{(0)}}
+ \sum_{t} \delta_{it} D_{it} + \alpha_{0}^{0} + \alpha_{T0}^{0} + u_{it} + v_{it},
\]

12 We evaluated the possibility of applying a non-homothetic translog form. However, the relatively large number of parameters created certain numerical problems in some of the econometric models, especially the true random effects model that requires a simulated likelihood maximization method. This is perhaps related to problems due to the model’s over-identification and perhaps multicollinearity as suggested by the lack of significance and counter-intuitive signs for some of the main variables.
where subscripts $i$ and $t$ denote the company and year respectively; the parameters $\alpha^m, \beta^n, \alpha^{nm}, \beta^{kl}, \delta^t$ and $\alpha^0$ ($m,n,k,l = 1,2,3; t = 1998,\ldots,2005$) are the regression coefficients to be estimated; and all second-order parameters $\alpha^{nm}$ and $\beta^{kl}$, satisfy the symmetry conditions ($\beta^{kl} = \beta^{lk}; \alpha^{mn} = \alpha^{nm}$); $\alpha_i$ is a firm-specific effect; $u_{it}$ is an asymmetric stochastic component term that captures the time-variant inefficiency and $v_{it}$ is a symmetric term representing random noise and statistical errors.

We consider four variations of the above model. These models are summarized in Table 2. The first model (Model I) is a random effects model in line with Schmidt and Sickles (1984). The model is estimated using the Generalized Least Squares (GLS) method. The specification includes a firm-specific random effect $\alpha_i$, and a random noise term $v_{it}$, which are both assumed to be identically and independently distributed ($iid$) with any arbitrary distribution. In this model, the inefficiency is assumed to be constant over time, namely the term $u_{it}$ in Equation (4) is set equal to zero. A given company $i$’s inefficiency is considered as the difference between its estimated random effect $\alpha_i$ and that of the firm with the “best performance” namely, the minimum estimated random effect ($\min \{\alpha_i\}$).

The GLS model benefits from certain robustness in that no specific distribution assumption is imposed, except for the usual assumption that the random terms are uncorrelated with the explanatory variables. However, the very construction of this model implies that companies are compared to a single, fully efficient firm that has the lowest observed costs after adjusting for explanatory variables and allowing for random noise. This could be an unrealistic assumption that only one company is completely efficient. Moreover, there is always a probability of wrong identification of a single “best” company because of some firm-specific unobserved factor, in which case the efficiency estimates will be completely distorted. The advantage of imposing a distribution assumption on efficiency attenuates at least partly such seriously misleading outcomes. A commonly used distribution in the literature is the half-normal distribution which is obtained by a zero-mean normal distribution truncated at zero. This distribution assumption that dates back to the original frontier models (Aigner et al., 1977; Meeusen and van der Broek, 1997), implies that full efficiency is the most frequent outcome located at the mode of the distribution.
Table 2: Econometric specifications of the stochastic cost frontier

<table>
<thead>
<tr>
<th>Stochastic term</th>
<th>Model I</th>
<th>Model II</th>
<th>Model II</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLS (Schmidt-Sickles)</td>
<td>ML (Pitt-Lee)</td>
<td>ML (Battese-Coelli)</td>
<td>True RE (Greene)</td>
</tr>
<tr>
<td>Firm-specific effect</td>
<td>$\alpha_i \sim iid \ (0, \sigma_o^2)$</td>
<td>$\alpha_i \sim N^+ (0, \sigma_o^2)$</td>
<td>0</td>
<td>$\alpha_i \sim N(0, \sigma_o^2)$</td>
</tr>
<tr>
<td>Time-varying</td>
<td>0</td>
<td>0</td>
<td>$u_{it} = u_i \exp {-\eta(t-T)}$</td>
<td>$u_{it} \sim N^+ (0, \sigma_o^2)$</td>
</tr>
<tr>
<td>inefficiency $u_{it}$</td>
<td>$v_{it} \sim iid \ (0, \sigma_v^2)$</td>
<td>$v_{it} \sim N(0, \sigma_v^2)$</td>
<td>$v_{it} \sim N(0, \sigma_v^2)$</td>
<td>$v_{it} \sim N(0, \sigma_v^2)$</td>
</tr>
<tr>
<td>Random noise $v_{it}$</td>
<td>$\hat{\alpha}_i - \min {\hat{\alpha}_i}$</td>
<td>$E[u_{it}</td>
<td>\hat{\epsilon}_{it}]$</td>
<td>$E[u_{it}</td>
</tr>
<tr>
<td></td>
<td>with $\alpha_{it} = \alpha_i + v_{it}$</td>
<td>with $\epsilon_{it} = u_{it} + v_{it}$</td>
<td>with $\epsilon_{it} = u_{it} + v_{it}$</td>
<td>with $\epsilon_{it} = u_{it} + v_{it}$</td>
</tr>
</tbody>
</table>

The half-normal distribution not only provides a relatively solid benchmark performance observed in a relatively large number of cases, it is also more consistent with the economic theory. In fact the half-normal distribution implies that higher levels of inefficiency have lower incidence. This is aligned with the theory that predicts the prevalence of rational and cost-minimizing behavior and considers the non-optimal performance as sporadic and rare outcomes. Following this assumption in the other three models, we assume a half-normal distribution for inefficiency.

Model II is a random effects model proposed by Pitt and Lee (1981). Similar to the first model, the efficiency is assumed to be constant over time ($u_{it}=0$). As opposed to Model I that does not impose any distribution, here the stochastic terms are assumed to follow a composite normal-half-normal distribution: The firm-specific effect $\alpha_i$ that represents (time-invariant) inefficiency, follows a half-normal distribution, and the random noise $v_{it}$ is simply a normal variable with zero mean. This model is estimated using the maximum likelihood method. In line with Kumbhakar and Lovell (2000) we will refer to this model as the maximum likelihood (ML) model. The firm’s inefficiency is estimated using the conditional mean of the inefficiency term proposed.
by Jondrow et al. (1982),\(^\text{13}\) that is: 
\[ \hat{\alpha}_i = \hat{\alpha} = E[\alpha]\], 
where the hat symbol \(\hat{\cdot}\) is used to indicate the post-estimation predicted value; 
\[ \omega_i = \alpha_i + v_i; \]
and
\[ \hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\alpha}_i. \]

The assumption of the firm’s inefficiency being constant over time can be relaxed by assuming a parametric form. A commonly used functional form is the exponential decay function proposed by Battese and Coelli (1992). Model III is based on one of the specifications proposed by those authors. In this model the inefficiency is defined as 
\[ u_{it} = u_i \exp\{-\eta(t-T)\}, \]
where \(u_i\) is a firm-specific stochastic term, \(T\) is the end period and \(\eta\) is a positive constant to be estimated. The adopted functional form implies that a given company \(i\) starts with an initial level of inefficiency of 
\[ u_{i0} = u_i \exp(\eta T), \]
that declines over time with an exponential rate of \(\exp(-\eta)\) per period, reaching \(u_{iT} = u_i\) at the end of the sample period.\(^\text{14}\) This specification, while recognizing individual differences in efficiency, assumes a similar improvement rate for all companies. The firm-specific heterogeneity term \(\alpha_i\) in Equation (4), is set equal to zero.\(^\text{15}\)

This model is also estimated using the maximum likelihood method. The firm’s inefficiency is estimated using the conditional mean of the inefficiency term, namely:
\[ E[u_{it} | \hat{\epsilon}_{it}] = E[u_{i} | \hat{\epsilon}_{i1}, \hat{\epsilon}_{i2}, ..., \hat{\epsilon}_{iT}] \exp\{-\eta(t-T)\}, \]
where \(\hat{\epsilon}_{it} = u_{it} + v_{it}\).

In both models I and II, it is assumed that all the unobserved differences across firms that do not vary over time are related to inefficiency. Model III relaxes the time-invariance by imposing a deterministic form of evolution that is uniform among all companies. In all three models, all the unobserved differences that cannot be captured by the random noise \((v_{it})\) are assumed to be due to inefficiency. As we have seen in

\(^{13}\) See also Greene (2005a).

\(^{14}\) Note that a more general notation \(T_i\) is usually used for the end of sample period \((T)\) that can be specific to company. Here we dropped the subscript for simplicity.

\(^{15}\) Battese and Coelli (1992, 1995) have proposed variations of this model with different distributions for \(u_i\), including truncated normal distribution. In this study we assume a half-normal distribution.
the previous section this could be a restrictive assumption in network industries especially in multi-utilities, which might entail a considerable cost variation through unobserved factors that vary from one network to another but are more or less constant over time and cannot be changed by the management. This implies that in these cases some of the unobserved heterogeneity, for instance, the complexity of the distribution network that is mainly determined by the topology of the service area, can be identified as inefficiency.

Model IV allows for a separate stochastic term that captures the time-invariant unobserved heterogeneity. This model is the ‘true random effects’ frontier specification proposed by Greene (2005a,b), which extends the original frontier model (Aigner et al., 1977) in a panel data framework with random effects. The stochastic components \( \alpha_i, u_i, \) and \( v_i \) respectively represent the firm-specific random effect, inefficiency term and random noise: \( \alpha_i \sim N(0, \sigma^2_\alpha), u_i \sim N(0, \sigma^2_u) \) and \( v_i \sim N^+(0, \sigma^2_v) \). This model is estimated using Simulated Maximum Likelihood (SML) method. We use quasi-random Halton draws to minimize the potential sensitivity of the results to simulation process. Number of draws has been fixed to 1000. Our sensitivity analysis using several options suggested that the estimation results are not sensitive when the number of draws is higher than a few hundred. The inefficiency is estimated using the (simulated) conditional mean of the inefficiency term \( u_i \) given by \( \hat{E}[u_i | r_i] \), where \( r_i = \alpha_i + u_i + v_i \) is the regression residual. The above conditional expectation is also calculated by Monte Carlo simulations.\(^{16}\)

With two heterogeneity terms, Model IV is expected to provide a better distinction between inefficiency and other unexplained variations. This advantage is especially important in network industries, in which a significant part of unobserved differences is due to time-invariant factors. All the adopted models assume that the stochastic terms namely, cost-efficiency and unobserved heterogeneity are independent from each other and are both uncorrelated with the explanatory variables included in

\(^{16}\) See Greene (2005b) for more details. A general discussion of the SML estimation method is also provided by Greene (2007).
the model. There are several methods to relax these assumptions. For instance the correlation between firm-specific effects and explanatory variables can be allowed by Mundlak’s specification (Farsi, Filippini and Greene, 2005; Farsi, Filippini and Kuenzle; 2005) or the impact of explanatory variables on efficiency can be modeled by specifying the truncation point of the normal distribution as a function of observed factors (Kumbhakar et al., 1991; Battese and Coelli, 1995) or as a general functional form (Wang and Schmidt, 2002). However, such elaborations can only be achieved through more complicated and often arbitrary assumptions that might compromise the clarity of the original assumptions and make the interpretations more difficult. Moreover, including explanatory variables in several forms in the model specification could cause over-identification and multi-collinearity issues. Such problems could bias the estimated coefficients or lower their accuracy, and eventually cause misleading estimates of cost-efficiency as well as technological characteristics such as the economies of scale. Finally, most of these “refinements” cannot be combined with the true random effects model that provides an already rich structure of the stochastic terms.

4. Empirical results

Table 3 lists the regression results of the cost frontier analysis, using the four alternative models as presented in Equation (4) and Table 2. The estimated coefficients of the first-order terms generally have the expected signs and are statistically significant across all models. Given that all the variables except the dummy variables are in logarithmic form, these coefficients can be directly interpreted as elasticities. The coefficients of first-order output variables represent the cost elasticities with respect to the corresponding outputs at the sample median. These coefficients indicate that the marginal costs of electricity distribution are considerably higher than those of natural gas, which in turn are substantially greater than those of water distribution.
Table 3: Estimation results

<table>
<thead>
<tr>
<th></th>
<th>Model I GLS (Schmidt-Sickles)</th>
<th>Model II ML (Pitt-Lee)</th>
<th>Model III ML (Battese-Coelli)</th>
<th>Model IV True RE (Greene)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha^1 ) (Electricity output)</td>
<td>0.505** (.053)</td>
<td>0.460** (.069)</td>
<td>0.418** (.063)</td>
<td>0.527** (.020)</td>
</tr>
<tr>
<td>( \alpha^2 ) (Gas output)</td>
<td>0.317** (.032)</td>
<td>0.298** (.041)</td>
<td>0.245** (.045)</td>
<td>0.258** (.012)</td>
</tr>
<tr>
<td>( \alpha^3 ) (Water output)</td>
<td>0.092** (.039)</td>
<td>0.178** (.053)</td>
<td>0.212** (.047)</td>
<td>0.146** (.015)</td>
</tr>
<tr>
<td>( \alpha^r ) (Customer density)</td>
<td>0.064** (.027)</td>
<td>0.043 (.038)</td>
<td>0.026 (.037)</td>
<td>0.007 (.009)</td>
</tr>
<tr>
<td>( \beta^1 ) (Labor price)</td>
<td>0.242** (.057)</td>
<td>0.229** (.054)</td>
<td>0.236** (.058)</td>
<td>0.201** (.027)</td>
</tr>
<tr>
<td>( \beta^2 ) (Electricity price)</td>
<td>0.326** (.059)</td>
<td>0.317** (.051)</td>
<td>0.333** (.052)</td>
<td>0.370** (.033)</td>
</tr>
<tr>
<td>( \beta^3 ) (Gas price)</td>
<td>0.234** (.043)</td>
<td>0.243** (.039)</td>
<td>0.223** (.038)</td>
<td>0.215** (.024)</td>
</tr>
<tr>
<td>( \alpha^{11} )</td>
<td>0.646** (.197)</td>
<td>0.368* (.221)</td>
<td>0.218 (.193)</td>
<td>0.231* (.086)</td>
</tr>
<tr>
<td>( \alpha^{22} )</td>
<td>0.234** (.055)</td>
<td>0.154* (.080)</td>
<td>0.067 (.071)</td>
<td>0.093** (.023)</td>
</tr>
<tr>
<td>( \alpha^{33} )</td>
<td>0.287** (.141)</td>
<td>0.042 (.176)</td>
<td>0.186 (.167)</td>
<td>0.089* (.052)</td>
</tr>
<tr>
<td>( \alpha^r )</td>
<td>0.019 (.061)</td>
<td>-0.063 (.095)</td>
<td>-0.233** (.089)</td>
<td>-0.146** (.026)</td>
</tr>
<tr>
<td>( \alpha^1 )</td>
<td>-0.273** (.086)</td>
<td>-0.182* (.105)</td>
<td>-0.048 (.091)</td>
<td>-0.099** (.041)</td>
</tr>
<tr>
<td>( \alpha^1 )</td>
<td>-0.327** (.149)</td>
<td>-0.124 (.158)</td>
<td>-0.214 (.148)</td>
<td>-0.133** (.058)</td>
</tr>
<tr>
<td>( \alpha^r )</td>
<td>-0.215** (.070)</td>
<td>-0.220** (.097)</td>
<td>0.074 (.104)</td>
<td>-0.119** (.030)</td>
</tr>
<tr>
<td>( \alpha^{13} )</td>
<td>-0.002 (.059)</td>
<td>0.049 (.072)</td>
<td>0.051 (.068)</td>
<td>0.037 (.026)</td>
</tr>
<tr>
<td>( \alpha^r )</td>
<td>0.123** (.059)</td>
<td>-0.002 (.079)</td>
<td>-0.147* (.080)</td>
<td>-0.065** (.027)</td>
</tr>
<tr>
<td>( \delta_{1998} )</td>
<td>0.085* (.050)</td>
<td>0.120 (.081)</td>
<td>0.104 (.076)</td>
<td>0.122** (.020)</td>
</tr>
<tr>
<td>( \delta_{1999} )</td>
<td>0.419 (.279)</td>
<td>-0.031 (.270)</td>
<td>0.051 (.248)</td>
<td>0.384** (.121)</td>
</tr>
<tr>
<td>( \delta_{2000} )</td>
<td>0.695** (.205)</td>
<td>0.524** (.172)</td>
<td>0.565** (.167)</td>
<td>0.758** (.110)</td>
</tr>
<tr>
<td>( \delta_{2001} )</td>
<td>-0.243** (.120)</td>
<td>-0.291** (.106)</td>
<td>-0.278** (.110)</td>
<td>-0.217** (.108)</td>
</tr>
<tr>
<td>( \delta_{2002} )</td>
<td>-0.701** (.221)</td>
<td>-0.419** (.197)</td>
<td>-0.460** (.189)</td>
<td>-0.724** (.102)</td>
</tr>
<tr>
<td>( \delta_{2003} )</td>
<td>0.294** (.147)</td>
<td>0.422** (.137)</td>
<td>0.386** (.136)</td>
<td>0.351** (.096)</td>
</tr>
<tr>
<td>( \delta_{2004} )</td>
<td>-0.096 (.135)</td>
<td>-0.154 (.118)</td>
<td>-0.156 (.115)</td>
<td>-0.136 (.092)</td>
</tr>
<tr>
<td>( \delta_{2005} )</td>
<td>-0.004 (.019)</td>
<td>-0.005 (.015)</td>
<td>0.011 (.016)</td>
<td>-0.005 (.032)</td>
</tr>
<tr>
<td>( \delta_{2006} )</td>
<td>-0.003 (.020)</td>
<td>-0.002 (.016)</td>
<td>0.028 (.019)</td>
<td>-0.005 (.021)</td>
</tr>
<tr>
<td>( \delta_{2007} )</td>
<td>-0.015 (.021)</td>
<td>-0.013 (.018)</td>
<td>0.035 (.024)</td>
<td>-0.006 (.025)</td>
</tr>
<tr>
<td>( \delta_{2008} )</td>
<td>-0.014 (.023)</td>
<td>-0.015 (.020)</td>
<td>0.049* (.029)</td>
<td>-0.012 (.022)</td>
</tr>
<tr>
<td>( \delta_{2009} )</td>
<td>-0.037* (.021)</td>
<td>-0.036* (.018)</td>
<td>0.036 (.030)</td>
<td>-0.040* (.022)</td>
</tr>
<tr>
<td>( \delta_{2010} )</td>
<td>-0.041* (.021)</td>
<td>-0.044* (.018)</td>
<td>0.039 (.033)</td>
<td>-0.039* (.023)</td>
</tr>
<tr>
<td>( \delta_{2011} )</td>
<td>-0.064** (.023)</td>
<td>-0.069** (.020)</td>
<td>0.032 (.038)</td>
<td>-0.067** (.024)</td>
</tr>
<tr>
<td>( \delta_{2012} )</td>
<td>-0.059** (.026)</td>
<td>-0.065** (.023)</td>
<td>0.046 (.043)</td>
<td>-0.073** (.022)</td>
</tr>
<tr>
<td>( \sigma_{\delta} )</td>
<td>7.164** (.029)</td>
<td>6.989** (.032)</td>
<td>6.917** (.046)</td>
<td>7.120** (.019)</td>
</tr>
<tr>
<td>( \sigma_{\eta} )</td>
<td>0.53</td>
<td>0.217** (.034)</td>
<td>0.114** (.005)</td>
<td>0.081** (.030)</td>
</tr>
<tr>
<td>( \sigma_{\eta} )</td>
<td>0.054</td>
<td>0.054** (.003)</td>
<td>0.052** (.003)</td>
<td>0.024** (.006)</td>
</tr>
<tr>
<td>( \eta )</td>
<td>0.048** (.015)</td>
<td>0.081** (.030)</td>
<td>0.024** (.006)</td>
<td>\text{Not Applicable (R}^2=0.982)</td>
</tr>
</tbody>
</table>

logL 296.785 299.355 303.786

** and * refer to 5% and 10% significance levels respectively. Standard errors are given in parentheses.
Approximately, the results suggest that by adding electricity output by 10 percent, the total costs will increase by about 5 percent on average, but the same relative increase in other outputs will raise the company’s total costs by about 2.5 to 3 percent for gas and only about 0.9 to 2 percent for water output. These predictions vary slightly across different models. Many of the second-order terms are also statistically significant, implying that the assumption of constant elasticities is unrealistic. The coefficients of the squared output terms ($\alpha_{11}$, $\alpha_{22}$, $\alpha_{33}$) are positive and mostly significant across all models. This suggests that a marginal increase in a given output increases the cost elasticity of that output. Therefore, as expected, the (product-specific) economies of scale are decreasing in output.

As we see in Table 3 the output cross-interaction terms ($\alpha_{12}$, $\alpha_{13}$, $\alpha_{23}$) are mostly negative across the models. In particular, the cross effect between electricity and other two outputs (natural gas and water), is statistically significant. This suggests that the multi-utilities with higher electricity output have a relatively low marginal cost for distributing water and gas. This cost complementarity also applies to companies with high gas or water output, which according to the estimation results, have lower marginal cost for electricity output. The results show however that the cost complementarity between gas and water outputs (as shown by coefficient $\alpha_{23}$) is not statistically significant. If we interpret this as a zero effect, this result suggests that the marginal cost of distributing gas (water) is not related to the volume of water (gas) output. This is a weak form of cost complementarity, implying that the marginal costs of one output will not increase in the amount of the other output.

As for the effect of customer density, the results show that the first order term is positive but statistically insignificant in most models. This suggests that the effect at the median company is probably not important. However, the mostly negative coefficient of the square term ($\alpha''$) suggests that higher densities could have a decreasing effect on costs. At first impression, this can be considered as counter-intuitive because increasing the customer density may be economical in low-density areas, but could create extra costs in congested areas. However, the statistically significant interaction terms between customer density and outputs, suggest that the density has a strongly non-linear effect depending on the output combination across the three services.
For instance the interaction term with electricity output ($\alpha_{1r}$) is mostly negative and significant, suggesting that the marginal cost of electricity output is lower in networks with higher customer density. This cannot be said for gas and water outputs. Especially the corresponding interaction term for water distribution ($\alpha_{3r}$) is mostly on the positive side, suggesting that an increase in customer density will increase the marginal cost of water distribution. These results can be related to different costs of network connection for various outputs, and also different amount of extra cables and pipes required for the provision of greater volumes of electricity, gas and water, depending on the actual customer density. For instance, in a dense and crowded area providing more electricity might be handled easier than a considerable increase in gas and water output. Moreover, connection of new customers to electricity networks is probably less costly than that of water and gas distribution networks.

The coefficients of the first-order terms of input prices are an indicator of the share of each factor price at the sample median. Based on the regression results, the shares of labor, electricity and gas inputs respectively amount to about 22, 33 and 23 percent of the total costs. These numbers are comparable to the sample mean of the observed factor shares which is 12, 35 and 17 percent of the company’s total costs, respectively for labor, electricity and gas inputs. As we see the share of electricity and gas expenses are quite close the average observed values. The remaining costs have been considered as ‘capital’ costs that are 36 percent on average, but about 22 percent from the regression results. Therefore in the model, the share of labor costs is overstated compared to that of the residual capital costs.

We explored if the estimated cost functions satisfy the theoretical properties implied by cost-minimization. As shown by the positive coefficients of the first order terms (Table 3), all the estimated cost functions are non-decreasing in output and input prices at the approximation point (sample median). However, our calculations showed that the Hessian matrix defined by the second derivatives of the translog cost

17 Note that in translog form, any statement about sample points other than the approximation point (here, sample median), should consider the second-order terms in addition to the main effects.
function with respect to log of input prices, is not negative semi-definite. The violation of this necessary condition\textsuperscript{18} for concavity might be considered as an indication that the concavity in input prices is not satisfied. This result can be explained by the fact that the multi-utilities are probably not as sensitive to price changes as the textbook economic theory might predict. Theoretically the companies are expected to substitute labor with capital or capital with energy in response to changes in the relative prices. However, in practice these substitutions are not feasible in many cases. For instance if the relative price of electricity to gas increases, the companies cannot substitute electricity input with gas input, because these inputs are mainly determined by the demand side.

In any case, even if we consider the lack of concavity in input prices as an indication that the companies do not fully minimize their costs the estimated cost functions can be useful to study the marginal effects of different factors on costs and also to compare the companies’ performance. In such cases, as pointed out by Bös (1986) and Breyer (1987), functions based on cost optimization can still be used as ‘behavioral’ cost functions and can be helpful in studying the firms’ behavior. Moreover, we should keep in mind that we are estimating a cost frontier function, which allow the possibility that some companies do not minimize their costs.

\textit{Cost efficiency}

The estimates of inefficiency scores obtained from the four models are summarized in Table 4. As expected, compared to all other models, the True RE model’s estimates provide generally lower inefficiency. According to this model the multi-utilities have on average about 6 percent excess costs compared to the fully efficient production whereas the other models predict from 18 to 21 percent excess cost on average. The median inefficiency for the True RE model is about 5%, while being about

\textsuperscript{18} As pointed out by Diewert and Wales (1987), even with a negative semi-definite Hessian matrix for the translog cost function, the costs might be concave with respect to input prices. So applying such a condition on the coefficient matrix of a translog cost function is too strong for concavity in input prices.
20% for all other models. It should be noted that the True RE model’s estimates do not include the persistent inefficiencies that might remain more or less constant over time. To the extent that there are certain sources of inefficiency that result in time-invariant excess costs, the estimates of the True RE model should provide a reasonable lower bound for the companies’ inefficiency. On the other hand, in all the three other models, it is assumed that all the time-invariant cost differences due to exogenous heterogeneity are accounted for by the observed explanatory variables included in the model, and whatever remains can be interpreted as inefficiency. Therefore, the overall estimates of inefficiency obtained from these models can be considered as a kind of upper bound for the actual level of inefficiency in the sector.

Table 4: Descriptive summary of inefficiency estimates

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLS (Schmidt-Sickles)</td>
<td>ML (Pitt-Lee)</td>
<td>ML (Battese-Coelli)</td>
<td>True RE (Greene)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.184</td>
<td>0.183</td>
<td>0.216</td>
<td>0.063</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>0.079</td>
<td>0.119</td>
<td>0.143</td>
<td>0.043</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.013</td>
<td>0.014</td>
<td>0.010</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>0.144</td>
<td>0.060</td>
<td>0.075</td>
<td>0.031</td>
</tr>
<tr>
<td>Median</td>
<td>0.202</td>
<td>0.207</td>
<td>0.214</td>
<td>0.050</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>0.251</td>
<td>0.275</td>
<td>0.303</td>
<td>0.082</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.303</td>
<td>0.401</td>
<td>0.699</td>
<td>0.277</td>
</tr>
</tbody>
</table>

The distribution of the inefficiency estimates in the sample is depicted in Figure 1. The distribution densities have been smoothed using Kernel density method. As seen in the figure the extent of inefficiency in the True RE model is considerably narrower than in other models. Moreover, the distribution of the GLS estimates suggest a negative skewness, which contradicts the usual assumption of positive skewness in cost-inefficiencies. Moreover, both Models II and III indicate a tendency to-
ward a bimodal distribution, which goes against the underlying half-normal distribution assumption in these models. These peculiar patterns might be indicative that the econometric specification of the error term in the first three models could be insufficient to capture the inefficiencies in a coherent way. This can be explained by unobserved cost differences that are not due to inefficiency but to other external factors.

Figure 1: Distribution of inefficiency estimates

In order to explore if the efficiency estimates provide a consistent ranking pattern across different modes, we studied the correlation coefficients between these estimates. Table 5 provides the correlation matrix of inefficiency scores across the four models. The results suggest a high positive correlation among the first three models. There is however a relatively low correlation between each one of these models and the True RE model. The Spearman rank correlation matrix shows slightly lower correlation in general but confirms the above pattern namely low correlation between Model $IV$ and the other three models, and high correlation among the latter models. This result suggests that even if we are only interested in
efficiency ranking rather than the numerical level of inefficiency, using the inadequate model can give a misleading ordering of individual companies.

Table 5: Pearson correlation matrix between inefficiency estimates

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLS (Schmidt-Sickles)</td>
<td>ML (Pitt-Lee)</td>
<td>ML (Battese-Coelli)</td>
<td>True RE (Greene)</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>0.863**</td>
<td>0.715**</td>
<td>0.124*</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1</td>
<td>0.793**</td>
<td>0.140**</td>
</tr>
<tr>
<td>III</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.128**</td>
</tr>
</tbody>
</table>

** and * refer to 5% and 10% significance levels respectively.

5. Conclusions

This study presents an empirical analysis of cost inefficiency in a sample of Swiss multi-utilities operating in the distribution of electricity, natural gas and water. The issues addressed in the study involve an important question related to the application of benchmarking analysis in incentive regulation schemes for multi-utilities. In general, the benchmarking of multiple-output companies is more complicated than in utilities with a similar output. Multi-utilities that operate in several different sectors, are characterized by a strong unobserved heterogeneity making the measurement of their performance an important challenge for the regulators.

It is shown that the recent methodological developments in the estimation of cost frontier functions using panel data methods can be helpful to achieve more reliable estimates of inefficiency in presence of unobserved and omitted factors. The previous studies have used some of these methods in single-network distributors such as electricity and gas. However to our knowledge there is no reported empirical application in the multi-utility sector. The present analysis serves as a first illustration of the difficulties involved in the estimation of efficiency in multi-network utilities.
Using a translog cost function and several stochastic frontier models this analysis indicates the presence of unexploited global scale economies in the majority of the companies included in the sample. The efficiency estimates are sensitive to the econometric specification of unobserved factors through the model’s stochastic components. While highlighting the potential problems in benchmarking multi-utilities, this study shows that adequate panel data models can be used to identify the inefficient companies and determine to certain extent, which part of their excess costs has been persistent and which part has varied over time.

Combining several frontier models also allows two types of inefficiency estimates: a “lower bound” estimate that includes only the transient part of the firm’s excess costs assuming that all persistent cost differences are due to unobserved factors rather than poor efficiency performance, and an “upper bound” that associates all the firm-specific unaccounted cost differences to their productive efficiency and neglects the effect of external unobserved factors. Both estimates could be useful for the regulator, as they can use them to identify the companies that are persistently more costly than others and those that have high time-variant inefficiency. The regulator should perform further detailed and possibly case-by-case studies to assess to what extent the excessive costs of the former group can be associated with productive inefficiency and identify the potential external factors and peculiarities that might have caused such excessive costs.

References


Mehdi Farsi, Massimo Filippini, Diego Lunati

Economies of scale and efficiency measurement in Switzerland’s nursing homes

Quaderno N. 08-01

Decanato della Facoltà di Scienze economiche
Via G. Buffi, 13 CH-6900 Lugano
ECONOMIES OF SCALE AND EFFICIENCY MEASUREMENT IN SWITZERLAND’S NURSING HOMES

January 2008

Authors:

Mehdi Farsi§ Massimo Filippini§ Diego Lunati#

# Department of Economics
University of Lugano
Via Buffi 13, 6904 Lugano

§ Department Management, Technology and Economics
ETH Zürich
Zürichbergstrasse 18, 8092 Zurich
ECONOMIES OF SCALE AND EFFICIENCY MEASUREMENT IN THE SWISS NURSING HOMES INDUSTRY

ABSTRACT

This paper examines the cost efficiency in the nursing home industry, an issue of concern to Swiss policy makers because of the explosive growth of national expenditure on elderly care and the aging of the population. A stochastic cost frontier model with a translog function has been applied to a balanced panel data of 1780 observations from 356 nursing homes operating over five years (1998-2002) in Switzerland. We compare the estimation results from different panel data econometric techniques focusing on the various methods of specification of unobserved heterogeneity across firms. In particular, the potential effects of such unobserved factors on the estimation results and their interpretation have been discussed. The paper eventually addresses three empirical issues: (1) the measurement of economies of scale in the nursing home sector, (2) the assessment of the economic performance of the firms by estimating their cost efficiency scores, and (3) the role of unobserved heterogeneity in the estimation process. The findings suggest that the economies of scale are an important potential source of cost reduction in a majority of Swiss nursing homes. Taking the size as given the efficiency performance of most individual units is practically very close to the estimated best practice. Nevertheless, the efficiency estimates suggest that some of the nursing homes can significantly reduce their costs by improving their operations.

KEY WORDS: COST EFFICIENCY; ECONOMIES OF SCALE; NURSING HOMES; STOCHASTIC FRONTIER; PANEL DATA.

JEL CLASSIFICATION NUMBERS: C13, C21, D24, H70, I11, I18, L30
1. INTRODUCTION

Health care costs are steadily growing in all industrialized countries as a result of several supply/demand factors such as the cost-increasing technological progress and the rise in health care prices, the increase in health care demand induced by higher real incomes, and the aging of the population. A low rate of economic growth in conjunction with the increase in public social and health care spending during the last decade induced the Swiss government to restrain the fiscal deficit, contain costs and achieve efficiency gains. Given that a considerable share of the elderly care in the State’s health care budget, a special attention could be paid to the possibility of improving the efficiency of the nursing home sector. Moreover, the importance of this sector, with total expenditures approaching two percent of GDP, is expected to grow as the population grows older.

The nursing home industry, with over 1300 providers spread throughout the whole country, represents an important element of the Swiss national system of social welfare and health care. The sector includes various kinds of facilities with different origins and aims. A number of nursing homes have been established at the local level as non-medical centers, whereas others, which serve a higher number of potential users at the district level, are equipped for basic medical services. An important characteristic of the sector is the presence of several institutional forms and various regulation systems. The main ownership categories are private for-profit, non-profit private and public. Thanks to their autonomy in social welfare and health care legislations, Swiss cantons apply different subsidy models to their respective nursing homes.1 It is important to point out that in most cantons the state intervenes in both public and private providers by regulating the daily rates, setting quality standards, determining the minimum necessary infrastructure and staff requirements and, above all, by granting a financial contribution in form of subsidies.

With the growing financial pressure on the federal state and cantonal governments, many policy-makers have voiced concerns about the efficiency of the nursing homes. In particular, a dominant perception is that most providers operate at a suboptimal size. The sector has been also criticized for the lack of initiatives to adopt adequate management strategies such as operating multiple-home networks, in order to exploit scale economies. Moreover, it has been argued that as opposed to private firms, heavily subsidized and public facilities do not have strong incentives for saving costs. Much of the ongoing debates remain however qualitative and lack the support of sufficient empirical evidence. The few existing

---

1 Switzerland (7.4 million inhabitants) is a federal State composed of 26 cantons.
studies in the context of Switzerland are limited to a specific region or lack the data from recent years.

This paper attempts to throw some light on the above policy debates through an econometric analysis of the cost structure of the Swiss nursing homes sector. The presented analysis is based on several stochastic cost frontier models. The paper’s focus is on the assessment of the scale economies and the estimation of cost-efficiency. The econometric models have been applied to a national sample of 356 nursing homes operating over the five-year period from 1998 to 2002. Compared to previous research on Swiss nursing homes, this study benefits from a larger data set and an extended set of variables. Moreover, the adopted methodology is based on some of the recent developments in stochastic frontier panel data models, which provide a better account of unobserved heterogeneity across firms.

The remaining part of this paper is organized as follows: A brief review of the relevant literature is provided in section 2. Section 3 presents the cost model and the specification of the included variables. In section 4 the adopted econometric models are discussed. The data set and the estimation results are described in section 5. Section 6 takes a closer look at the estimates of scale economies and cost efficiency. The conclusions are drawn at the end.

2. REVIEW OF THE LITERATURE

The literature on nursing home industry, focusing on the empirical evaluation of the production and cost structure, is well established and dates back to the 80s with the appearance of first econometric studies based on data from the long term care sector. Since then the estimation of a cost function has been performed from many authors with a broad scope of research objectives, but sharing a common problem: the specification of a feasible cost model. In what follows we cite some papers that trace and summarize the evolution in model design and technique choice of the relevant literature for this work: McKay (1988), Vitaliano and Toren (1994), Filippini (2001), Crivelli et al. (2002) and Farsi et al. (2005). McKay (1988) estimated a total cost function for a sample of 82 NH operating in Texas in 1983. The cost model is a neoclassical cost function including output (patient days) and input prices (nursing, aide, building & equipment and services).

Since the main goal of the paper is the measurement of economies of scale in the industry for policy reasons, the problem of quality level is considered by adjusting the model with a quality proxy given by the nursing hours per day ratio. For both models a translog cost function is estimated using the system equation and the Zellner’s (1962) seemingly unrelated
regression model. The main finding of this study is the presence of scale economies (evaluated at sample mean corresponding to capacity of 70 beds) in the nursing home industry, a result in contrast with the previous studies of that time. Vitaliano and Toren (1994) estimated a stochastic cost frontier for a panel of 164 New York NH in two distinct years (1987 and 1990). The cost model is designed to capture differences in quality, patients turnover and ownership structure.

The estimation technique is based on the composed error term (Aigner et al. 1977) which is the core of a stochastic frontier approach. In this paper the focus is on the comparison between the inefficiency levels achieved by different kinds of nursing home and the empirical results show that there’s no difference in efficiency over time and among the ownership types. The paper of Filippini (2001) is the first empirical estimation of a translog cost function on Swiss data. The goal of the paper is the assessment of economies of scale and the data structure utilized is a panel of nursing homes operating in Ticino – a Swiss Canton (State) – over the years 1993-1995. The cost model includes two indicators to partially capture the heterogeneity in the output characteristics and in the quality level. The paper suggests the presence of scale economies at the sample median which correspond to a capacity of 60 beds.

In the paper Crivelli et al. (2002) the authors extend the empirical analysis to a Swiss sample of 886 nursing home for the year 1998 covering almost all the 26 Cantons (States) of the Swiss Confederation. In order to assess the efficiency level of nursing home industry and the impact of ownership and regulation environment on the cost performance a stochastic cost frontier is estimated. The main findings of the paper are: (1) around the 60% of the Swiss nursing homes operate close to the national standard for efficiency, achieving scores of 15% or lower in terms of cost difference to the empirical frontier; (2) there is no statistical evidence of efficiency difference between ownership types and regulatory environment; (3) economies of scale are present and relevant but tend to expire when the capacity reaches 50-60 beds. The work of Farsi et al. (2005) deals with the econometric problem due to unobserved heterogeneity in the stochastic frontier estimation, by applying several methods to the same data set and cost model. The translog cost function is fit to a sample of 36 Swiss nursing homes operating in Ticino over 9 years period (1993-2001). The model includes in the regressors some proxy for quality and patient case-mix (observed heterogeneity) point out that the choice of the econometric approach can influence the empirical results in a substantial way, by considering or not in the efficiency scores the underlying heterogeneity.
### 3. COST MODEL FOR SWISS NURSING HOMES

A nursing home can be represented as a production unit transforming two major inputs (capital and labor) into resident-days of long-term care. Such an aggregate measure of output should be complemented with additional variables representing the severity of the case mix of the residents. Moreover, the cost model specification should take into account a number of variables describing output characteristics as well as regional differences, which should capture the heterogeneous dimension of the output of a nursing home. For instance, the costs of operating a nursing home may depend on the type and quality of care provided per resident-day, the level of assistance required by the residents in normal daily activities such as eating, personal care or performing physiological functions and the level of medical assistance required by the residents.

Assuming that output level and input prices are exogenous[^2], and that (for a given technology) firms adjust input factors in order to minimize costs[^3], the firm's total operating costs of a nursing home can be represented by the following long run cost function:[^4]

\[
TC = f(Y, P_K, P_L, H, R, D_{fo}, D_{qc}, T)
\]

where \( TC \) represents the total cost and \( Y \) is the output, represented by the total number of resident-days of nursing care.

\( P_K \) and \( P_L \) are the prices of capital and labor, respectively. Unfortunately the data which would allow us to calculate the capital stock using the capital inventory method are not available. According to Wagstaffs (1989) and Filippini (2001) the capital stock is approximated by the number of beds owned and operated by a nursing home. The cost of capital is represented by all the expenses apart from labor cost, following the residual capital approach suggested by Friedlaender and Wang Chiang (1983). Hence the capital price is approximated by the capital expenses normalized by the number of available beds in the nursing home. The price of labor is computed as the total labor expenses divided by the total number of full time equivalent employees. Unfortunately the available data do not allow any distinction of labor prices by professional categories.

\( H \) is the average assistance time (expressed in hours per day) per resident including both normal daily activities (eating, personal care or performing physiological functions) and

---

[^2]: These assumptions are the subject of debate in the literature. For an extensive discussion see Breyer (1987).

[^3]: This assumption implies an input orientation for efficiency measurement.

[^4]: The adopted specification is an extension of that used by Crivelli et al. (2002) in a cross-section of Swiss nursing homes. The additional advantage of this study to that paper is the use of panel data.
medical care. Following McKay (1988), this variable is introduced in the model to control for differences in the output characteristics. This variable can be interpreted as a proxy for the quality of care as well as a measure of case mix severity through the average required assistance. $R$ is the average medical expenses per patient reimbursed by the health insurance system and is expressed in Swiss Francs. This variable can be considered as a proxy variable for case-mix since there is a negative correlation between the health condition of a resident and the medical expenses reimbursed by the health insurance system.

To better explain total cost differences among Swiss nursing homes, we included a set of dummy variables for different types of nursing homes. The nursing homes that provide part of their care through in-law apartments (less restricted living conditions to residents with a high degree of independence) are distinguished with a dummy variable ($D_{fo}$). These are expected to be less costly compared to nursing homes with only rooms located in the main building. The binary indicator $D_{qc}$ represents the high-quality nursing homes with respect to care person resident ratio. This dummy is equal to 1 whenever the ratio of the medical and nursing staff to total residents is bigger than 0.424 (median value) implying that for every two residents there is at least one member of the medical staff (either a nurse or a doctor). This variable should distinguish nursing homes operating with different levels of medical and nursing staff and can be viewed as a crude proxy for the quality of care. The model include a linear time trend ($T$).

The theoretical restrictions related to cost optimization require that the cost function expressed in (1) be concave and linearly homogeneous in input prices and non-decreasing in input prices and output.\(^5\) The parametric estimation of the cost function (1) requires the specification of a functional form. The translog function offers an appropriate flexible form for answering questions about economies of scale. Being a second-order approximation for any function, translog form does not impose a priori restrictions on the nature of technology. In particular, the values for economies of scale can vary with output. The linear homogeneity restriction is imposed by setting the price of capital as the numeraire. The complete cost frontier model function results as:

$$\ln( TC / P_K ) = \alpha_0 + \alpha_y \ln y + \alpha_L \ln P_L / P_K + \alpha_H \ln H + \alpha_R \ln R$$

$$+ \alpha_{yL} \frac{1}{2} \ln^2 y + \alpha_{yH} \frac{1}{2} \ln^2 P_L / P_K + \alpha_{yR} \frac{1}{2} \ln^2 H + \alpha_{yR} \frac{1}{2} \ln^2 R$$

$$+ \alpha_{yL} \ln y \ln P_L / P_K + \alpha_{yH} \ln y \ln H + \alpha_{yR} \ln y \ln R$$

$$+ \alpha_{LH} \ln P_L / P_K \ln H + \alpha_{LR} \ln P_L / P_K \ln R + \alpha_{HR} \ln H \ln R$$

$$+ \alpha_{fo} D_{fo} + \alpha_{qc} D_{qc} + \alpha_T T + \alpha_i + \varepsilon_i$$

Considering the sample median as the approximation point, all the variables have been normalized to their respective sample median values. Subscripts \( i \) and \( t \) respectively denote the nursing home and year, \( \alpha_i \) is a firm-specific effect and \( \varepsilon_{it} \) is an iid error term. As we will explain in the next section, in the recent models proposed by Greene (2005), the stochastic term \( \varepsilon_{it} \) is decomposed into two parts: a skewed component representing inefficiency and a symmetric part for the random noise.

4. METHODOLOGY

Cost frontier analysis can be used to estimate indicators of cost-efficiency and scale economies. The use of cost frontier models to evaluate efficiency in the health-care sector has been criticized by Newhouse (1994) and Skinner (1994). The main arguments against these models are related to the unobserved heterogeneity due to differences in case-mix and quality and the errors committed by aggregation of outputs as well as non-testable assumptions on the distribution of efficiency. Folland and Hofler (2001) provide a discussion on the reliability of hospital efficiency estimates obtained from stochastic cost frontier models. These authors show that the individual efficiency estimates are rather sensitive to the adopted model specification and functional form. However, the results are robust when the comparisons are performed between hospital group mean inefficiencies. This finding is consistent with the results reported by Hadley and Zuckerman (1994) suggesting that the stochastic frontier analysis is of practical use when applied for comparing group means of hospital efficiency. Crivelli et al. (2002) and Farsi et al. (2005, 2007) reached a similar conclusion in their studies regarding nursing homes and general hospitals. Although nursing homes have generally a more uniform case-mix than hospitals, the above arguments apply more or less to these providers as well.

There are several parametric cost frontier methods to estimate the cost efficiency of individual firms. Kumbhakar and Lovell (2000) provide an intensive survey of various models proposed in the literature. The main models used in this paper are based on Greene’s (2005) extension for panel data of the original frontier approach proposed by Aigner et al. (1977). In this framework, \( \varepsilon_{it} \) in equation (2) is assumed to be a composite stochastic term with a normal-half-normal distribution, including both idiosyncratic effects and inefficiencies. The additional firm-specific term, \( \alpha_i \) in equation (2), represents the unobserved heterogeneity and
is assumed to have a normal distribution. This model is referred to as the “true” random-effects model. The estimation method is based on simulated maximum likelihood.

The results are compared with two alternative models namely, the random-effects model proposed by Schmidt and Sickles (1984) and the original pooled model proposed Aigner et al. (1977). A summary of the three models used in the paper is given in Table 1.

The first model is a pooled frontier model in that the sample is considered as a cross-section and its panel aspect is neglected. In this model there is no firm-specific component $\alpha_i$ and the random error term is divided into two components: a normal error term $\nu_i$ capturing the noise and a half-normal random term $u_i$ representing the inefficiency as a one-sided non-negative disturbance. This model is based on the original cost frontier model proposed by Aigner et al. (1977). The firm’s inefficiency is estimated using the conditional mean of the inefficiency term $E[ u_i | u_i + v_i ]$, proposed by Jondrow et al. (1982).

**Table 1. Econometric specifications of the stochastic cost frontier**

<table>
<thead>
<tr>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pooled</strong></td>
<td><strong>RE-GLS</strong></td>
<td><strong>True RE</strong></td>
</tr>
<tr>
<td>Firm-specific component $\alpha_i$</td>
<td>None</td>
<td>$iid (0, \sigma_{\alpha}^2)$</td>
</tr>
<tr>
<td>Random error $\epsilon_{it}$</td>
<td>$\epsilon_{it} = u_{it} + \nu_{it}$, $u_{it} \sim N(0, \sigma_u^2)$, $\nu_{it} \sim N(0, \sigma_\nu^2)$</td>
<td>$\epsilon_{it} = u_{it} + \nu_{it}$, $u_{it} \sim N(0, \sigma_u^2)$, $\nu_{it} \sim N(0, \sigma_\nu^2)$</td>
</tr>
<tr>
<td>Inefficiency</td>
<td>$E[ u_{it}</td>
<td>u_{it} + \nu_{it} ]$</td>
</tr>
</tbody>
</table>

Model II is a random effects (RE) model that is estimated using GLS method. In this model the inefficiency scores are estimated as the distance from the firm with the minimum estimated random effect, that is $\hat{\alpha}_i - \min \{ \hat{\alpha}_j \}$, as proposed by Schmidt and Sickles (1984). In this model the firm’s inefficiency is assumed to be constant over time, thus captured by the firm-specific effects, while in other models inefficiency can vary across years. The RE model assumes that all the cost variation across firms that cannot be explained by the included explanatory variables, are due to efficiency differences. This assumption is rather restrictive.

---

6 The name “true” is chosen to show that the model keeps the original frontier framework and the extension is done only by including an additional heterogeneity term.
in the health sector, where the services are generally characterized by strong heterogeneity across the users. Especially in the case of long-term care, many of these characteristics could be time-invariant over the relatively short periods covered in the usual samples, thus could bias the efficiency estimates.

Model III is an extension to model I which includes an additional firm-specific random effect ($\alpha_i$) to represent the unobserved heterogeneity among firms. Model III is Greene’s (2005) true RE model. In this model it is assumed that the unobserved cost differences across firms that remain constant over time, are driven by firm-specific unobserved characteristics rather than inefficiency. In the context of nursing homes, the unobserved factors such as differences in quality and case-mix severity across providers could be very important. Therefore, it is crucial to disentangle such differences from inefficiency. However, the distinction between the two requires certain non-testable assumptions. The true RE model assumes that persistent cost differences are related to unobserved heterogeneity across firms rather than efficiency differences. On the other hand, the inefficiency term is assumed to be an iid random variable with half-normal distribution. This implies that inefficiencies are not persistent and each period brings about new idiosyncratic elements thus new sources of inefficiency. This is a reasonable assumption particularly in industries that are constantly facing new technologies or time-variant external factors such as changing labor prices and regulation constraints.

In the RE model II, the inefficiencies are assumed to be constant over time. This could be an unrealistic assumption in most practical cases, where the driving forces of cost-inefficiency are not generally persistent. In fact firms constantly face new problems emerging from the implementation of new techniques, or from dealing with new regulation systems, or other external constraints. Moreover, there exist incentive mechanisms (either through regulation and monitoring or through profit and career incentives) that induce managers to revise their strategies and correct their past suboptimal decisions. To what extent this is the case in the context of nursing home sector in Switzerland, is open to debate. However, the presence of time-invariant unobserved output characteristics that are specific to location and case-mix but beyond the management’s control is undeniable.

Another important issue is that in all the above models it is assumed that the unobserved firm-specific output characteristics are uncorrelated with the explanatory variables. This assumption might be restrictive and its violation could bias the coefficients of the cost function, hence distort the estimates of scale economies. The effect of such biases on the efficiency estimates depends on whether or not the “true” inefficiencies are correlated with unobserved heterogeneity or the included explanatory variables. If inefficiencies are not
allowed to have such a correlation, then the bias on coefficients does not affect the efficiency estimates. Since the true values of efficiency are not known, the assumption of no correlation between efficiency and other factors is not testable. We contend however that such an assumption is consistent with the original spirit of the frontier models in which the inefficiency is defined as cost-differences that cannot be explained by other factors.

As for the potential bias in the cost function’s slopes, fixed effects models based on within estimators can be helpful. However, given that our data is a rather short panel of 5 years, the statistical efficiency of fixed effects models depends on the level of variation of costs and explanatory variables especially output, within firms. Our preliminary analysis with several fixed effects specification indicates high estimation errors and lack of statistical significance for many of the important explanatory variables. This can be explained by the fact that in the main variables, the within variations are comparatively insignificant compared to the dominating between variations. Moreover, the inefficiency estimates from the fixed effects models are implausibly high (reaching easily to 100% of costs), suggesting that they capture some of the between variations that are not related to inefficiency. Therefore, we decided to exclude the fixed effects model.

5. THE DATA AND ESTIMATION RESULTS

The data used for this study have been provided by the Swiss Federal Statistical Office. The original sample consisted of 3474 observations from 1070 nursing homes operating from 1998 to 2002. By nursing homes, we mean the facilities that provide full-time long-term care as well as basic medical care to the elderly people. After excluding the observations with missing values and the invalid data, the final sample includes 1780 observations from 356 nursing homes resulting in a balanced panel of 5 years. The sample includes of 73 public, 218 private non profit and 65 private for profit nursing homes.

Excepting the quality measures the data set was sufficiently detailed to ensure an adequate specification of the cost model. All the financial data have been deflated according

---

7 See Greene (2005, 2002) for more details. This author considers a panel of 5 years as a short panel. In contrast with “between” variations that are related to the differences across companies, “within” variations correspond to the changes in a given company over time. Roughly speaking, even a long panel data with low within variation is equivalent, for econometric purposes, to a short panel data.

8 See Farsi et al. (2005) for more details on the problems of fixed effects models in estimating efficiency.

9 The Swiss nursing home industry is composed by several heterogeneous facilities. In this study we used the definition of nursing home proposed by the Federal Statistical Office. Moreover, in order to keep a relatively homogeneous sample we excluded facilities that do not provide any medical care, identified by zero reimbursement from health insurance systems.
to the Swiss consumer price index based on May 2000 prices. In table 1 we present the main descriptive statistics for the continuous variables included in the regression.

Table 1: Descriptive statistics for cost model variables

<table>
<thead>
<tr>
<th>Variable (Label)</th>
<th>unit</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>1st qu.le</th>
<th>Median</th>
<th>3rd qu.le</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cost (TC)</td>
<td>CHF</td>
<td>4563181</td>
<td>3646319</td>
<td>2087600</td>
<td>3606600</td>
<td>6070200</td>
</tr>
<tr>
<td>Output (Y)</td>
<td>Patient-days</td>
<td>22501</td>
<td>15806</td>
<td>11960</td>
<td>18802</td>
<td>27968</td>
</tr>
<tr>
<td>Labor price (PL)</td>
<td>CHF /worker</td>
<td>69765</td>
<td>13371</td>
<td>62228</td>
<td>69375</td>
<td>76867</td>
</tr>
<tr>
<td>Capital price (PK)</td>
<td>CHF / bed</td>
<td>21325</td>
<td>10401</td>
<td>13967</td>
<td>18491</td>
<td>25485</td>
</tr>
<tr>
<td>Care giving (H)</td>
<td>care hours / resident</td>
<td>2.793</td>
<td>1.118</td>
<td>1.950</td>
<td>2.646</td>
<td>3.617</td>
</tr>
<tr>
<td>Reimbursement (R)</td>
<td>CHF / resident</td>
<td>20540</td>
<td>12011</td>
<td>12219</td>
<td>18011</td>
<td>25822</td>
</tr>
<tr>
<td>Average Cost</td>
<td>CHF/resident-day</td>
<td>201.8</td>
<td>62.6</td>
<td>155.3</td>
<td>191.9</td>
<td>239.7</td>
</tr>
<tr>
<td>Number of nursing home beds</td>
<td></td>
<td>64.2</td>
<td>45.4</td>
<td>35</td>
<td>53</td>
<td>81</td>
</tr>
</tbody>
</table>

The results of the estimates for the total cost frontier function (3) under different econometric specifications are presented in Table 2. Most of the first and second order terms are statistically significant in all the specifications. Since total costs and all the regressors are in logarithms, the first order coefficients can be interpreted as cost elasticities evaluated at the approximation point that is the sample median. All these coefficients have the expected signs and are highly significant.

The output elasticity is positive and implies that an increase in supply will raise the total cost. The increase in total cost, as a response to a 10% increase in the number of patient-days, corresponds to 9.8% for the pooled model and is just slightly smaller (9.4%) for the other two models. These results, showing some variability in the parameter estimates, are fairly similar across different econometric specifications. This suggests that the heterogeneity biases due to potential correlation between explanatory variables and firm-specific unobserved factors are not of a considerable order.

The labor and capital cost shares are positive, implying that the cost function is monotonically increasing in input prices. The cost elasticities with respect to the output characteristics (H and R) are, as expected, positive. This result suggests that a ceteris paribus

---

10 We exclude all the observations with evidently unrealistic values such as extreme outliers. In addition, we excluded nursing homes with less than 10 beds and also those with a bed occupancy rate of less than 10%.

11 In the estimation process we also evaluate several alternative specifications for each of the three presented models. The estimation results were quite stable across the three econometric specifications.
increase in the average assistance time and in the reimbursement per patient will lead to an
increase in total cost.

Table 2  Estimation results

<table>
<thead>
<tr>
<th>Model</th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pooled</td>
<td>RE-GLS</td>
<td>True RE</td>
</tr>
<tr>
<td>( \alpha_y )</td>
<td>0.985</td>
<td>0.935</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>( \alpha_L )</td>
<td>0.640</td>
<td>0.591</td>
<td>0.576</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.009)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>( \alpha_H )</td>
<td>0.101</td>
<td>0.052</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>( \alpha_R )</td>
<td>0.069</td>
<td>0.038</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>( \alpha_{yy} )</td>
<td>0.036</td>
<td>0.238</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.017)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>( \alpha_{LL} )</td>
<td>0.068</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.020)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>( \alpha_{HH} )</td>
<td>0.113</td>
<td>0.097</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.026)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>( \alpha_{RR} )</td>
<td>0.044</td>
<td>0.063</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>( \alpha_{yL} )</td>
<td>0.033</td>
<td>0.035</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>( \alpha_{yH} )</td>
<td>-0.004</td>
<td>-0.035</td>
<td>-0.036</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>( \alpha_{yR} )</td>
<td>0.055</td>
<td>0.084</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>( \alpha_{LH} )</td>
<td>0.102</td>
<td>0.044</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.015)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>( \alpha_{LR} )</td>
<td>0.019</td>
<td>0.034</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.013)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>( \alpha_{HR} )</td>
<td>-0.126</td>
<td>-0.119</td>
<td>-0.102</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.014)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>( \alpha_{fo} )</td>
<td>-0.055</td>
<td>-0.072</td>
<td>-0.066</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.013)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>( \alpha_{qc} )</td>
<td>0.174</td>
<td>0.077</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.006)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>( \alpha_T )</td>
<td>0.007</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>( \alpha_0 )</td>
<td>14.894</td>
<td>14.961</td>
<td>14.887</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>( \sigma_u )</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.178</td>
<td>-</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>-</td>
<td>(0.002)</td>
</tr>
<tr>
<td>( \lambda = \sigma_u / \sigma )</td>
<td>1.251</td>
<td>-</td>
<td>2.137</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>-</td>
<td>(0.173)</td>
</tr>
</tbody>
</table>

All the coefficients except those in shaded fields are statistically significant at 1% significance level. Standard errors are given in brackets.

The dummy variable \( D_{fo} \) that distinguishes nursing homes with in-law apartments is negative and statistically significant. This result can be explained by the fact that generally in the residents of these apartments require a lower degree of assistance compared to those living in ordinary nursing home. The dummy variable \( D_{qc} \), introduced to explain the quality of

13
provided care, is positive and suggests that a higher ratio of medical staff per patient raises costs. Finally the time trend (T) is positive indicating an upward cost shift over time.

6. ECONOMIES OF SCALE AND COST EFFICIENCY

In this section we turn to a detailed analysis of estimation results regarding scale and cost efficiency. The analysis of scale economies is aimed at identifying if and to what extent the nursing homes exploit the potential cost reductions related to higher output levels. The concept of economies of scale indicates the degree to which a company operates at the optimal scale. Frisch (1965) defines the optimal scale as the level of operation where the average costs are minimum, which coincides with a unit value for the scale elasticity. Hence, the economies of scale (ES) are defined as the inverse of scale elasticity, which is the proportional increase in total costs resulting from a proportional increase in the output (Y), holding all input prices and other explanatory variables fixed:

\[
ES = \frac{1}{\partial \ln TC / \partial \ln Y}
\]

There are unexploited economies of scale if \( ES \) is greater than 1 and, conversely, there are diseconomies of scale if \( ES \) is lower than 1. In other words economies (diseconomies) of scale exist if the average cost of a nursing home decreases (increases) as output increases.

Table 3 presents in more details the results in term of economies of scale for the different specifications. The output levels (total resident-days over a year) have been transformed to an equivalent number of beds assuming a full occupancy rate. The value of \( ES \) for the small (33 beds) and medium (51 beds) sized nursing homes is greater than 1 for all specifications and this means that economies of scale still exist at the sample median. In other words, half of the units observed in the sample could reduce total costs by increasing their output.

The optimal size (\( ES=1 \)) for a nursing home corresponds to the minimum of the estimated average cost frontier which depends on the days of care offered (Y) but also on other regressors. The optimal size figures correspond to the minimum of the average cost frontier.
Table 3: Economies of scale

<table>
<thead>
<tr>
<th>Economies of Scale</th>
<th>( Y = 1^{st} \text{ Quartile} ) (Size = 33 beds)</th>
<th>( Y = \text{median} ) (Size = 51 beds)</th>
<th>( Y = 3^{rd} \text{ Quartile} ) (Size = 77 beds)</th>
<th>Optimal size ((ES = 1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Model I:}</td>
<td>1,038</td>
<td>1,015</td>
<td>1,007</td>
<td>93 beds</td>
</tr>
<tr>
<td>\textit{Pooled}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{Model II:}</td>
<td>1,267</td>
<td>1,069</td>
<td>1,008</td>
<td>79 beds</td>
</tr>
<tr>
<td>\textit{RE-GLS}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{Model III:}</td>
<td>1,221</td>
<td>1,065</td>
<td>1,014</td>
<td>82 beds</td>
</tr>
<tr>
<td>\textit{True RE}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Number of beds is computed as the output in patient-days \((Y)\) divided by 365 days.

The cost minimizing size is given by an extension of output \((Y)\) keeping all the other variables of the model constant to their respective median values. The number of beds corresponding to the optimal size varies according to the different econometric specifications (table 3) but we can safely consider the bed interval 75-95 as a reliable indication for the optimal size of a representative nursing home. This optimal level can be compared to the actual distribution of the nursing homes capacity presented in figure 1. This distribution is characterized by a strong concentration around the median value of 53 beds.¹²

Figure 1: Box-plot of the actual size (number of bed) based on the final sample

The estimation results (table 3) show that in the Swiss nursing home industry most of the firms are operating at a sub-optimal size. This widespread scale inefficiency could be justified with two arguments: quality and access. The quality argument relies on the fact that often a nursing home’s priority is to maintain a comfortable and friendly living environment, which is more easily achieved in a small institution. The access argument is based on the importance of providing a social service to elderly people also in rural and border area. In the Swiss case, another explanation for the small dimensions of many nursing homes could be fiscal federalism. In the past municipalities were responsible for providing care of their elderly populations and, of course, this situation produced a huge numbers of small nursing homes.

¹² The high concentration around the median value as shown in figure 1 suggests that the sample median can be considered as an appropriate representative nursing home for the sector.
throughout the country. It should be noted that because of capacity constraints, an existing small nursing home cannot simply increase its output to exploit the scale economies. We contend that such nursing homes could however exploit certain economies of scale through cooperation with neighboring nursing homes in different areas such as sharing their staff, joint purchase of drugs and materials, laundry and other services.

The results reported in table 3 are confirmed by the results obtained in previous studies using data on Swiss nursing homes. Table 4 illustrates the results obtained in these studies. The study by Filippini (2001) and Farsi and Filippini (2004) estimate a cost function using panel data for a sample of nursing homes operating in Canton Ticino. Crivelli et al. (2002) estimate a cost frontier model using a cross-section for a sample of Swiss nursing homes. The values of the optimal size reported in the studies by Farsi and Filippini (2004) and Filippini (2001) are higher than those reported in this study and in the study by Crivelli et al. (2002). This difference may be due to the different data set and to the different econometric approaches used. For instance, Farsi and Filippini (2004) were able to use a rich panel data set composed of 36 nursing homes operating in Ticino over the 9-year period from 1993 to 2001.

Table 4: Previous results regarding scale economies in Swiss nursing homes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frontier</td>
<td>Frontier</td>
<td>SUR</td>
</tr>
<tr>
<td></td>
<td>All Switzerland</td>
<td>Ticino</td>
<td>Ticino</td>
</tr>
<tr>
<td></td>
<td>MLE (Aigner)</td>
<td>REM (Pitt-Lee)</td>
<td></td>
</tr>
<tr>
<td>Optimal size at median</td>
<td>79 beds</td>
<td>120 beds</td>
<td>120 beds</td>
</tr>
</tbody>
</table>

Table 5 shows summary statistics of inefficiency scores, calculated for the nursing homes included in the sample. The value of inefficiency can be read as a percentage of cost beyond the frontier value obtained by a fully efficient unit. For instance, a score of 0.24 indicates that the actual cost of a nursing home is 24% higher than the (estimated best practice) fully efficient level. The inefficiency scores are relatively sensitive to different model specifications.

As discussed earlier, assuming that the unobserved firm-specific factors are not correlated with efficiency differences the estimates obtained from the true RE are pure inefficiency separated from unobserved heterogeneity. Nonetheless this model cannot detect the presence of time-invariant inefficiency, which is captured by the firm-specific term, leading to a (possible) downward bias in efficiency estimates. On the other hand, the RE-GLS model considers any time-invariant variation in the model’s residuals as inefficiency. This
model is therefore likely to create an upward bias in the inefficiency scores. Finally, the pooled model does not assume any distinction between time-variant and time-invariant residuals. The separation of inefficiency and heterogeneity is based on a simple distribution assumption considering the symmetric residual component as heterogeneity and the skewed part as inefficiency.

**Table 5  Inefficiency scores statistics**

<table>
<thead>
<tr>
<th>Inefficiency</th>
<th>Model I Pooled</th>
<th>Model II RE-GLS</th>
<th>Model III True RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.117</td>
<td>0.621</td>
<td>0.081</td>
</tr>
<tr>
<td>Median</td>
<td>0.105</td>
<td>0.616</td>
<td>0.067</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.670</td>
<td>1.400</td>
<td>1.184</td>
</tr>
<tr>
<td>90 percentile</td>
<td>0.184</td>
<td>0.942</td>
<td>0.135</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.021</td>
<td>0</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Looking at table 5 and noting the above discussion it is clear that identifying the inefficiency level in the sector is rather contentious. Due to the importance of unobserved factors related to quality of care we contend that the true RE model’s estimates are on the safe side. According to this model, the median inefficiency score is 0.067 suggesting that about 50% of the nursing homes in the sample are operating with a total cost that is 6.7% higher than the minimum possible cost, and in 90 percent of the cases, the inefficiency is limited to 13.5%. This result shows that the majority of Swiss nursing homes are operating relatively close to the fully efficient cost frontier.

Regarding the efficiency differences across nursing homes with different ownership types, we applied several statistical tests on the inefficiency scores estimated from the three models. We used a non-parametric rank test (Kruskal-Wallis) as well as t-tests across three groups of nursing homes: public, private non-profit (NP) and private for profit (FP). The estimation results suggest that on average the public nursing homes are the least efficient group followed by NP providers and that the FP nursing homes are on average the most cost-efficient category. This result is confirmed across all three econometric specifications. However, the practical and statistical significance of the differences varies across the models. The true RE model does not provide any evidence of significant differences. The null hypothesis of no difference cannot be rejected at any significance level. On the other hand the pooled and RE-GLS models favor the statistical significance of efficiency differences.
However, it should be noted that the practical significance of these differences is questionable. According to the pooled model the average efficiency differences across the ownership categories is always below 1%. In the case of RE model the predicted efficiency differences can go up to 7-8 percent with the maximum difference between public and FP categories.

7. CONCLUSIONS

The purpose of this study was to analyze the cost structure of a sample of Swiss nursing homes in order to assess economies of scale and cost inefficiency. Policy-makers are particularly interested in cost information in order to determine the optimal size of a nursing home. Moreover, this paper measures cost efficiency under different econometric specifications. Four different stochastic frontier cost models have been considered. A translog cost function was estimated using an unbalanced panel of 356 nursing homes operating in Switzerland from 1998 to 2002. The main conclusions of this study with respect to the adopted methodology and the empirical implications can be shortly summarized.

Empirical evidence indicates that economies of scale are exhausted up to a capacity of beds ranging from 75 to 95 beds. This result suggests that the effects of size on costs should be taken into account in building new nursing homes. This result also points to the economic advantages of joint activities between small nursing homes, and eventually consolidation of small facilities through mergers and acquisitions. The outcome of this analysis, according to the most favorable results, shows that over 75% of the nursing homes included in our sample operate close (not more than 10%) to the national (relative) best practice for efficiency. The average inefficiency is about 7% of cost excess above the (estimated) Swiss best-practice. The results also point to possible efficiency differences across different types of ownership. However, depending on the adopted econometric model and the assumptions related to unobserved factors, these differences are either statistically insignificant or mostly of no practical significance.

We would like to stress that with the available short panel, the unobserved heterogeneity at the firm level cannot be sufficiently considered and the empirical results especially concerning the individual efficiency estimates should be considered with caution. Secondly, we note that the two purposes of our estimations require somewhat different conditions in the model. As we have seen earlier the pooled model ignores the panel structure of the data, thus may result in biased estimates of the slopes, whereas, the panel data models control to some extent, for the unobserved heterogeneity, thus should have a lower bias in the slope estimates. Therefore, the estimated scale economies from the panel model are more
reliable, or at least they can indicate the direction of such biases. Finally, there are potentially important quality differences across different ownership categories, which have not been accounted in this analysis and requires further research and more elaborate data. These points should be kept in mind in the interpretation and application of the results presented in this paper for policy purposes.

REFERENCES


Appendix II

Articles related to Consumers Choices


Risk-Aversion and Willingness to Pay for Energy Efficient Systems in Rental Apartments

Mehdi Farsi

CEPE Working Paper No. 55
July 2008
First draft: February 2007

CEPE
Zurichbergstrasse 18 (ZUE)
CH-8032 Zürich
www.cepe.ethz.ch
Risk Aversion and Willingness to Pay for Energy Efficient Systems in Rental Apartments*

Mehdi Farsi
ETH Zurich, Switzerland

July 2008

Abstract

This paper uses a random utility model with a non-linear utility function to estimate the consumers’ valuation for energy efficient insulation and ventilation systems in apartment buildings. The proposed model is applied to data from a choice experiment conducted among 264 apartment tenants in Switzerland. The model relaxes the assumption of constant rate of substitution between income and energy-saving measures. These amenities are considered as non-market goods whose advantages are little known thus entailing certain risk-aversion in consumers’ preferences. The non-linear formulation can accommodate the common cases when the non-market attributes are measured by discrete variables. The analysis indicates that assuming constant rate of substitution could bring about misleading estimates of the willingness to pay, especially when the system is combined of several components. The findings provide evidence in favor of consumers’ risk-averse behavior facing choice situations regarding energy-efficient systems. The estimated risk premiums suggest that risk considerations remain a central issue in dealing with energy efficiency in residential buildings.

Keywords: choice experiment; willingness to pay; risk aversion; energy efficiency; housing
JEL classification: Q51, C25, D12, C91

Correspondence: Mehdi Farsi, Centre for Energy Policy and Economics (CEPE), Department of Management, Technology and Economics, ETH Zurich, Zurichbergstr. 18, Zurich, 8032 Switzerland; E-mail: mfarsi@ethz.ch.

*I am grateful to Martin Jakob for providing the data and for numerous insights and discussions. I also thank Anna Alberini, Silvia Banfi, Massimo Filippini and Lester Hunt for helpful suggestions and the participants of the 16th annual conference of the European Association of Environmental and Resource Economics, Gothenburg, June 2008, especially Christian Gollier for their inspiring comments on a previous version. Any remaining errors and omissions are mine.
1. Introduction

The enhanced insulation and energy-efficient ventilation of residential buildings are new technologies that can considerably reduce the energy consumption for indoor heating and cooling. Cost-benefit analyses point to the economic viability of these systems even if the comfort co-benefits such as improvements in indoor air quality and protection against noise are not taken into account (Jakob, 2006; Ott et al., 2006). Moreover, the empirical analyses of the few available choice experimental surveys (e.g. Poortinga et al., 2003; Jaccard and Denise; 2006) assessed a relatively high Willingness To Pay (WTP), which would easily justify the investment in energy-saving systems. However, the actual investment in these systems is relatively rare. For instance, in Switzerland, notwithstanding several subsidy programs, the reported usage of the Minergy label (used for highly efficient insulation and ventilation systems) is less than 5% in new apartment buildings, and virtually non-existent in renovated buildings (Banfi et al.; 2008). The reasons for this ‘energy-efficiency gap’ are often linked with the limitations of the financial markets, the frictions in the housing markets caused by legal restrictions as well as transactions costs due to lack of information (cf. Jakob, 2006; Ott et al., 2006).

One of the factors, hardly addressed in the previous studies, is the effect of risk-aversion in consumers’ and investors’ behaviors: Due to lack of information about the private and social benefits of new technologies, consumers might show a greater degree of risk-aversion as compared to ordinary systems with a widespread usage. Moreover, given the great fluctuations in energy prices the benefits of energy-saving technologies bear a relatively high risk compared to other assets.

The risk-averse behavior can be relatively easily identified in an empirical context, for instance through the concavity of the utility function. Stated Preference
(SP) methods especially choice experiments provide an interesting basis to assess the extent of risk-aversion in the individual preferences regarding new commodities such as energy efficient insulation and ventilation. Unlike the revealed preferences, in which the data on available alternatives is usually lacking, the stated choices among pre-defined hypothetical alternatives can help identify the utility function.

Previous analyses of WTP for energy-saving systems in buildings are based on linear Random Utility Models (RUM) assuming a risk-neutral behavior. In line with the few previous studies such as Herriges and Kling (1999) and Layton and Lee (2006), this paper deals with the non-linear effects in RUMs. A novelty of this paper is in applying the non-linear models within the expected-utility framework, thus allowing an inference on risk behavior. We argue that due to the inherent risks in new technologies and the uncertainty regarding their benefits, the individuals’ assessment relies on their expected utility. Given that the risks are not explicitly measured and no data is available about the consumers’ perception of those risks, the model relies on the concept of Certainty-Equivalent\(^1\) (CE) to elicit the consumer’s behavior facing the involved risks.

The adopted methodology also differs from most previous studies in two aspects: First, rather than the non-linear income effects, the non-linear forms are used for quality attributes of the new technologies. Namely, the utility function is linear in income and ordinary market goods. Secondly, unlike most previous studies, here the attributes cannot be represented by continuous variables that could allow an easy integration of non-linear functions. In order to overcome this problem, it is assumed that the CE value of certain ‘non-market’ attributes is a continuous function of the dummy variables representing different levels of those attributes. This assumption

\(^1\) For a given risky asset and a given utility function, the CE value is defined as a risk-free asset that provides the same expected utility.
allows the specification of a ‘link’ or index function between the continuous utility function and the discrete attribute variables.

Using the data from a choice-experimental survey conducted in Switzerland, this paper analyzes the consumers’ preferences regarding energy saving systems in apartment buildings. The paper’s objectives are twofold: First, using several functional forms, we test the hypothesis of risk-neutrality and assess the potential effects of mis-specification on the WTP estimates. Second, using an appropriate functional form, we illustrate the effects of potential risks on the consumer’s or investor’s valuation of energy-efficient systems. The results suggest that risk aversion is a determinant aspect of the consumers’ behavior regarding enhanced insulation and ventilation of their residence. The risk premium entailed by the concavity of the utility function can easily attain a few percentage points with conceivable levels of variability. The WTP analysis shows that misspecification by a linear models does not create significant estimation errors for each attribute when considered separately. However, the results indicate a diminishing marginal rate of substitution suggesting a dislike for combining several attributes together, which is ignored in the linear model.

The remainder of the paper is organized as follows. After providing the general background in Section 2, the methodology and utility functional forms are presented in Section 3. Section 4 describes the data and the specification of the variables. The estimation results are presented and discussed in Section 5, and Section 6 concludes the paper.

2. Background

SP methods based on choice experiments or vignettes are increasingly commonplace in the economic evaluation of environmental goods. In these methods
the experimental data fitted with a RUM are used to elicit individuals’ preferences and estimate their WTP\(^2\) for a non-market commodity or its attributes (Louviere, Hensher and Swait, 2000; Holmes and Adamowics, 2003). The SP approach can also be used for goods that have a limited or incomplete market (Bateman et al., 2002). In particular, a few studies such as Poortinga et al. (2003), Sadler (2003), Jaccard and Denise (2006) and Banfi et al. (2008), used this method to assess the consumers’ valuation of energy efficient systems of heating and insulation in residential buildings.

As in most applications of the stated choice models, these studies use a linear utility function\(^3\) usually including dummy variables for various attributes and at least one ‘continuous’ variable for monetary values. In linear utility models, any individual’s WTP is equal to her rate of substitution of non-market commodity with the numeraire market good or money (Heshner, Rose and Greene, 2005). This is often referred to as the marginal WTP, which can be directly obtained from a given utility function (Freeman, 2003). Random utility models (RUM) are used to estimate the individuals’ WTP by estimating their utility function (\textit{cf.} Birol, Smale and Gyovai, 2006; Heshner, Shore and Train, 2005; Carlsson and Martinsson, 2001; Hanley, Mourato and Wright, 2001). In these models the utility function is elicited by comparing the random utility of chosen offers versus the not-chosen alternatives (Train, 2003; Ben-Akiva and Lerman, 1985).

The linearity assumption facilitates the estimation of WTP and the resulting welfare estimates, mainly because the effect of initial utility is canceled out and the

---

\(^2\) WTP is defined as the maximum amount that an individual is willing to bid for a public good without losing any utility, or alternatively, as the compensating variation that equates the utility with and without the non-market good, thus ensuring the same indifference curve.

\(^3\) By linear functions I mean those functional forms whose second-order derivatives are either zero or undefined. This definition also includes the functions with dummy variables and interaction terms that are commonplace in these applications, yet are not helpful for inference about risk aversion. A better term would be first-order functions. However, in order to avoid any confusion with the notion of first-order risk aversion (more on this later), I prefer to use the term linear in this context.
WTP remains independent of the initial levels of consumption and income.\(^4\) However this assumption might be unrealistic as it implies a constant rate of substitution between the non-market commodity and market goods. Especially in many cases, the marginal utility of the non-market good might decline drastically after certain threshold. For instance, the consumer’s WTP for any additional measure against pollution would quickly approach zero below certain levels of pollution. Moreover, the consumer’s attitude might be different facing the risks involved in new commodities that are not widely available in the markets as opposed to those related to market goods. Moreover, some of the widely observed disparity between WTP and willingness to accept (WTA)\(^5\) can be related to the non-linearity of the utility function due to risk-aversion rather than loss-aversion or irrational behavior (Coursey et al., 1987). For instance a risk-averse individual, whose utility function is concave in income, will have a higher monetary equivalent for a given income gain than for a loss of the same magnitude, thus a greater WTA compared to WTP.

There is a great body of economics literature on risks and their effects on investment and consumption (cf. Gollier, 2001). In addition to risk-aversion, the economic literature on risks and uncertainty has brought about other notions such as prudence and aversion to ambiguity, which are likely to affect the individuals’ behavior regarding environmental commodities such as energy efficiency. While prudence implies a preference for relative risks in best outcomes compared to those of worst events, ambiguity-aversion is referred to an additional dislike of assets whose risks are not known. While certain types of behavior can be relatively easily identified

\(^4\) Especially the independence from initial income is helpful in estimating the aggregate welfare: Hanemann (1984) shows with linear utilities, even if the individuals have different valuations of the non-market good, the expected value of WTP can be directly obtained from the ratio between the coefficients in a logit regression of the individuals’ binary response for accepting/rejecting offers with given costs.

\(^5\) WTA is defined as the minimum monetary compensation in order for an individual to give up a non-market benefit.
in an empirical context, others are more difficult to identify. For instance, risk-aversion can be identified by the concavity of the utility function as well as the consumer’s systematic rejection of a mean-preserving transformation with higher variance compared to the initial asset, whereas prudence is usually associated with the signs of the utility function’s third and fourth derivatives (positive and negative respectively) or systematic preference of statistical translations to the right that preserve both mean and variance (Eeckhoudt et al., 1995).

Given that the data on perceived risks is difficult to obtain and clearly defined levels of risks are difficult to implement in hypothetical choice experiments, the form of the estimated utility function and its non-linearity can be used to assess the risk-averse behavior and its effects on the adoption of new technologies. A number of studies have explored the non-linear effects of income and attributes in RUMs. An important example is Herriges and Kling (1999) that provides a review of the few preceding studies. Applying several functional forms including translog and generalized Leontief to sportfishing survey data, Herriges and Kling (1999) find that the possible non-linear effects are generally small and mostly dominated by the changes due to various econometric specifications. Layton and Lee (2006) and Layton (2001) are two other studies that considered several non-linear functional forms and highlighted the differences among various specifications. In particular, the former paper recommends a strategy based on averaging across different models in order to improve the WTP estimates.

Other studies focused on the non-linear effects of income: Aiew, Nagya and Woodward (2004) estimate the WTP for irradiated ground beef from a choice experiment. Using several functional forms for the income variable those authors applied several functional forms including translog and generalized Leontief to sportfishing survey data, Herriges and Kling (1999) find that the possible non-linear effects are generally small and mostly dominated by the changes due to various econometric specifications. Layton and Lee (2006) and Layton (2001) are two other studies that considered several non-linear functional forms and highlighted the differences among various specifications. In particular, the former paper recommends a strategy based on averaging across different models in order to improve the WTP estimates.

6 A important contribution is McFadden (1999) who proposed a Monte Carlo simulation method for calculating aggregate welfare measures based on compensating variation in the non-linear cases. Alternative closed-form solutions have been recently proposed by Morey and Rossman (2008).
show that the differences across different specifications are not statistically significant. Using several data sets Cooper (1991) reports that the estimated mean WTP could be sensitive to the adopted functional form. Cooper (2002) used semi-parametric methods to estimate the WTP from choice data with dichotomous response, but remained inconclusive regarding the non-linear effects. Semi-parametric methods allow a fully flexible functional form, but as Cooper (2002) points out, this flexibility comes at a loss of statistical efficiency and the difficulty in the economic interpretation of the estimated coefficients. Morey et al. (2003) propose a method for incorporating income effects through a piece-wise linear function that allows a differentiation between income categories or several representative consumers.

In virtually all the above studies, the non-linearity is considered for continuous variables such as income and those attributes that could be approximated as continuous variables. Moreover, the main focus of these studies lies upon the robustness of the WTP measures, and the consumers’ risk attitudes toward the non-market attributes have received little attention. This shortcoming is probably related to the fact that the concept of risk-aversion is often associated with twice-differentiable and continuous utility functions. In such a framework, it might appear that the concept of risk aversion based on expected utility does not apply to discontinuous or dummy variables representing non-market attributes. Yet, it is important to note that even with dummy variables the expected utility can be defined (or approximated) as a continuous and twice-differentiable function of the CE value. Namely, even though the utility function is discontinuous or piece-wise linear, the expected utility incurred from having a commodity is a continuous variable.

While Cooper (2002) applied his model to contingent valuation method, the proposed semi-parametric framework can also be used in choice experiments.
The risk-aversion concept for piece-wise linear utility functions has a fairly long history in the economics literature. This type of risk-aversion, identified by Stiglitz (1969) in the behavior of firms facing progressive tax rates, is occasionally referred to as ‘first-order risk aversion,’ a term coined by Segal and Spivak (1990). The latter define the first-order risk aversion as the cases in which the risk premium has a non-zero derivative with respect to the risk measure (e.g. variance of a payoff), which implies that even at very small risks, the decision-maker cannot be considered as almost risk-neutral.\footnote{As pointed out by Segal and Spivak (1990), the twice-differentiability of the utility function implies that at sufficiently small risks, any risk-averse person accepts a lottery with positive expected value. This restriction (countered by observations such as demand for full insurance) does not apply to a decision-maker with first-order risk-aversion, who systematically rejects sufficiently small gambles, but obviously when risks are away from zero, might have a lower or higher risk premium compared to an ordinary (2\textsuperscript{nd} order) risk-averse individual.} An example of 1\textsuperscript{st} order risk aversion is the risk-averse behavior observed when the utility function is a piece-wise linear function, but the discontinuities and the resulting nonlinearities are such that the mean-preserving spreads are systematically rejected by a seemingly risk-neutral firm (Eeckhoudt \emph{et al.}, 1997).\footnote{Eeckhoudt \emph{et al.} (1997) explain how otherwise risk-neutral firms can show risk-aversion in their investment decisions should their expected net profits is a non-smooth function (due for instance, to differential tax treatment of losses and gains). The observed risk aversion could be decreasing or increasing depending upon the location of discontinuities with respect to the risk-free profit curve.}

The form of the risk aversion considered in this paper is in some ways similar to that analyzed by Eeckhoudt \emph{et al.} (1997) for piece-wise linear utility functions: In both cases, while in the underlying utility function the Arrow-Pratt index of absolute risk-aversion is zero everywhere except at discontinuity points, there is a positive risk premium. However, the adopted approach to specify the expected utility function differs from that paper. Here instead of calculating the expected utility by integrating probabilities with the piece-wise constant utility function, we use the concept of CE value, to ensure the continuity of the utility function. The CE value is then specified as a first-order link function of the attributes (more on this later). CE is considered as
a continuous variable but as we see later, given the discrete nature of the attribute variables it can only take a finite number of values.

Focusing on the risk aversion in quality attributes rather than the non-linear income effects we propose a method to counter the difficulties entailed by discrete variables included in the utility function, while retaining linear terms for income variable and other market goods. The non-linear effect of income variable in the utility function can be interpreted as risk-aversion with respect to income or generally speaking, market goods. Given that the WTP in practical examples of public goods is generally a rather small fraction of the person’s income, one could expect that the non-linear effect of income should not be significant. That is, the small changes in the utility function due to costs of such non-market goods can be reasonably approximated by a linear function of income. Therefore, for all practical purposes to the extent that the WTP remains a sufficiently small fraction of the income, linearity (or risk-neutrality) with respect to income could remain a reasonable assumption.

Such an argument however does not apply to the benefits of the non-market good, which are generally bounded. Often times, especially in environmental goods, the marginal value of such benefits diminish considerably with the usage level. Moreover, for the benefits that are scantily known to the consumers, one can expect a risk-averse behavior namely, concavity of the utility function with respect to those attributes. Such a behavior can be considered as a relatively low marginal utility for higher levels of attributes. For instance, a consumer who does not benefit from an adequate insulation might have a relatively high marginal utility for an insulated window system or an air renewal system separately, but she might be less willing to accumulate the two systems.
3. The Model

We assume that the utility underlying the preferences of a typical individual facing the choice of a rental apartment can be defined as a function of the person’s monthly income ($y$), the monthly rent ($R$), and the characteristics of the apartment. We assume that these characteristics can be classified into two groups: The first group consists of “ordinary” attributes that are reasonably well known to the consumer and are readily available in the marketplace with a competitive price. The second group includes “non-market” attributes, namely those attributes for which markets are not fully developed and/or consumers have little information. Let $X$ and $Z$ denote the vectors respectively corresponding to non-market and ordinary attributes of a rental apartment. These are typically binary-valued vectors consisting of zeros and ones, representing the presence or lack of the corresponding attributes. In the context of this paper, non-market attributes are enhanced insulation and ventilation systems that are relatively new in the market, while the market attributes can be considered as the conventional forms of insulation.

An offer $j$ proposed to person $i$ will therefore be specified as a set of three components $(R_y, X_y, Z_y)$ respectively representing the levels of monthly rent, non-market attributes and ordinary amenities. Assuming a multinomial logit specification the probability that individual $i$, selects a specific alternative $j$ among $J$ options denoted by subscript $j = 0, 1, 2, \ldots, J-1$ can be written as a function of utility values obtained from each option, namely:

$$P_y = \frac{\exp(U_y)}{\sum_{j=0}^{J-1} U_y} , \text{ with } U_y \equiv E[U(y_i, R_y, X_y, Z_y)].$$  \hspace{1cm} (1)

The linear utility model assumes a constant rate of substitution between income and the non-market good. In this case $U(.)$ can be replaced by a linear
function. Here the idea is that in the utility function only the market attributes are readily exchangeable with money at a constant rate of substitution. In the context of risks, this implies that the inherent risks involved in market goods are insignificant and comparable to those related to holding cash. It should be emphasized that the risky assets in this model namely the non-market attributes have a stochastic cash value with an expected payoff with certain risk, while the market goods are considered as risk-free assets. The valuation of non-market attributes enter with a non-linear (concave) form, implies that the consumer can show certain aversion to the inherent risks of new commodities knowing that their consequent value (comfort, cost saving, etc.) is uncertain at least relative to other conventional commodities.

Assuming additive separability between the market and non-market components of the utility function, and denoting the status quo by subscript $j=0$, we write the utility function of Equation (1) as:

$$U_j = E\left[\tilde{u}\left(X_j\right)\right] + w(y_i - R_{i0}) + \beta(R_j - R_{i0}) + Z_y'\gamma,$$

where $\tilde{u}\left(X_j\right)$ is the utility of non-market attributes $X_j$ with $\tilde{u}\left(0\right) = 0$, $w(y_i - R_{i0})$ is the utility from the disposable income at the status quo rental situation, $\beta$ is a the marginal value of a unit increase in rent, $\gamma$ is a vector of parameters representing the valuation of the apartment’s market attributes. Note that the expectation operator only applies to the non-market attributes that entail certain risks. We can assume that the utility of these attributes namely $\tilde{u}\left(X_j\right)$ at any given state $X_j = x$ is a continuous

10 By contrast with many WTP studies based on ‘conditional’ indirect utility as a function of income and unit prices (Small and Rosen, 1981), here we use the utility function directly. Therefore, income can be interpreted as a composite commodity or a numeraire good.

11 This assumption is consistent with the findings reported by Knutson et al. (2007), suggesting that the rewards associated with a product and the losses associated with prices are processed in two distinct parts of the brain (respectively nucleus accumbens and insular cortex) and are consequently synthesized in the prefrontal cortex before individuals make a purchase decision.
random variable with mean $\mathbb{E}[\tilde{u}(x)]$ and a variance that measures the uncertainty about the benefits of that state. The above expected utility can readily be transformed in certainty-equivalent terms as:

$$U_{ij} = u\left[\theta(X_{ij})\right] + w(y_i - R_{i0}) + \beta(R_{ij} - R_{i0}) + Z_{ij} \gamma,$$

where $\theta(X_{ij})$ is a scalar function (link or index function) that satisfies the following equality:

$$u\left[\theta(X_{ij})\right] = \mathbb{E}[\tilde{u}(X_{ij})].$$

(4)

Notice that if $u$ and $\tilde{u}$ are the same functions the above equality defines the certainty-equivalent value of non-market attributes $X_{ij}$. In order to define the risk-aversion measures, we need to impose regularity conditions on $u(.)$ namely, being continuous and twice-differentiable. However, because of a discontinuous support due to dummy variables included in $X_{ij}$, the utility function $\tilde{u}(.)$ cannot be twice differentiable or continuous for that matter. An important point here is that for function $\theta(X_{ij})$ to be the CE value, it is not necessary that $\tilde{u}(.)$ and $u(.)$ are exactly the same function. Rather, the sufficient condition is that the certainty-equivalent condition in Equation (4) holds for all possible combinations of non-market attributes.

In fact for all practical purposes if $\tilde{u}(.)$ can be approximated by a continuous function $u(.)$ such that the certainty-equivalent condition holds for all relevant values, the decision-maker represented by the utility function in Equation (2) can be equivalently modeled by that of Equation (3), with a continuous and twice-differentiable utility function. Therefore, in order to proceed with the analysis we require two assumptions: First, CE function is a linear function of attributes defined
by: \( \Theta(X_y) = X_y' \alpha \), where \( \gamma \) is a parameter vector representing the valuation of the apartment’s non-market attributes. Secondly, there exists a twice-differentiable function \( u(.) \) that can satisfy Equation (4) at least approximately for all possible values of \( X_y \).

By comparing the utility values on the basis of the status quo utility \( (U_{i0}) \) it is easy to see that the function \( w(.) \) cancels out in the econometric model implied by Equation (1). In fact, in multinomial discrete choice models it is the utility differences compared to one of the alternatives that matter, not the absolute levels. Here, the status quo can be comfortably taken as the base alternative. Equation (3) can therefore be re-written as:

\[
U_{ij} = U_{i0} + u[\Theta(X_y)] - u[\Theta(X_{i0})] + \beta(R_{ij} - R_{i0}) + \gamma(Z_y - Z_{i0}),
\]

with \( j = 1, 2, ..., J - 1 \).

Five different functional forms are considered for the utility function \( u(\Theta) \):

- linear, quadratic, logarithmic, power function and exponential:

\[
\begin{align*}
u(\Theta) &= \Theta; & u(\Theta) &= \Theta + \lambda \Theta^2; & u(\Theta) &= \ln(1 + \Theta); \\
u(\Theta) &= \Theta^r; & u(\Theta) &= -\exp(-\Theta); 
\end{align*}
\]

where \( r \) and \( \lambda \) are the model parameters to estimate. There are a few issues in the specification of the parameters in \( u(.) \). An obvious restriction is \( u(0) = 0 \), implying that the additional utility from not having the non-market attributes is zero. Another restriction is an identification issue. Given that \( \Theta \) is defined by a fully parameterized function, \( \Theta(X_y) = X_y' \alpha \), no additional multiplying factor can be included in the specification of \( u(.) \).

Each one of the functions in (6) have their specific feature in the risk-averse behavior. Based on the common measures namely, Arrow-Pratt coefficients of
absolute and relative risk aversion (RA) defined respectively as: 
\[ c_A \equiv -\frac{u''}{u'} \] and 
\[ c_R \equiv -\frac{u''\theta}{u'} \], the exponential function represents a constant \( c_A \), while the power and logarithmic functions have a constant relative RA, but a decreasing absolute RA. All three forms are commonly used in the microeconomics literature. The quadratic and power functions can cover both risk-averse and risk-loving behavior as special cases, depending upon the sign of parameters \( r \) and \( \lambda \). The quadratic function has an increasing absolute risk-aversion, which appears to be counter-intuitive in an ordinary consumption problem. However, we retain this form, exactly because the increasing risk-aversion property might be appealing in considering the decisions about energy-efficiency. In fact, because the efficiency has an upper bound (perfect efficiency), attaining which might entail unusual sources of uncertainty due to technological constraints. These factors might justify an increasing risk aversion with the level of efficiency.

The WTP for each attribute \( X_k \) at a given initial level \( X_0 \) is defined by the corresponding increase in the utility divided by the coefficient of the rent, that is:
\[
wtp_k = \frac{u(X_0'\alpha + \alpha^k) - u(X_0'\alpha)}{\beta},
\]
(7)
where \( \alpha^k \) represents the marginal value of attribute \( k \). The WTP expression can be readily generalized to the difference between any pair of states \((X_0, X_1)\), with the corresponding WTP expressed as:
\[
\frac{1}{\beta} \left[u(X_0'\alpha) - u(X_1'\alpha)\right].
\]
A relative measure of risk premium at any given point \( X_0 = x \) can be defined in utility terms as the difference between the expected utility and the utility of the expected attributes that is:
\( \pi \propto \frac{\tilde{u}(x) - E[\tilde{u}(x)]}{E[\tilde{u}(x)]} = \frac{\tilde{u}(x) - u[\theta(x)]}{u[\theta(x)]}. \) (8)

\( u^{-1}[\tilde{u}(x)] - \theta(x) \) However, in our case as we do not estimate the utility function \( \tilde{u}(.) \) a more convenient measure can be specified as:

\[ \pi = \frac{E[m(x)] - \theta(x)}{\theta(x)}, \] (9)

where \( m(x) \) is a monetary measure of \( x \), with the same units as \( \theta(x) \), that can be specified as: \( u^{-1}(\tilde{u}(x)) \), where \( u^{-1}(.) \) is the inverse function of \( u(.) \).

The above expression is especially useful because with some assumptions numerical values can be estimated for the risk premium. In fact, when functions \( u \) and \( \theta \) are identified, for a given level of attributes \( X_i = x \), we can simulate a random variable \( \tilde{u} = \tilde{u}(x) \) with mean \( \mu = E[\tilde{u}] = u[\theta(x)] = u[x', \alpha] \), and standard deviation \( \sigma = \delta, \mu \) for a given variability ratio \( \delta \). Assuming a normal distribution (for instance), the expected value of the attributes at \( x \), can be calculated as:

\[ E[m(x)] = \int_{-\infty}^{\infty} u^{-1}(t) \varphi\left(\frac{t-\mu}{\sigma}\right) dt, \] (10)

where \( \varphi(.) \) is the pdf of the standard normal variable. The numerical values of the above integral can be calculated using the Monte Carlo simulation technique and the value of relative risk premium can be estimated for different variability ratios.

4. Data and Specification

The data used in this paper are based on the choice experiment conducted in 2003 in Switzerland and reported by Banfi et al. (2008) and Ott et al. (2006). This paper focuses on the tenants of apartment buildings. The respondents were selected
among households that had recently moved in their dwellings. The sampling procedure used in the survey is based on stratified sampling to ensure the existence of a sufficient number of new buildings that are more relevant for energy-efficient technologies, and also of those buildings equipped with these systems. On average about 20 percent of the respondents in the sample have already certain experience in using the energy-efficient systems. It is therefore expected that compared to the Swiss population the sample has an over-presentation of individuals with a fairly good knowledge of the benefits of these technologies. This might result in an understatement of the risk-averse behavior, but also might cause an overstatement of the WTP for policy conclusions.

The respondents were repeatedly offered an alternative housing with various levels of energy-saving systems and were asked if they would prefer the offered alternative to their Status Quo (SQ). In each choice situation the respondent was provided with a choice card including the characteristics of the offered alternative along with those of their actual housing. These characteristics consist of monthly rent, window and facade insulation each defined in four levels (none, low, standard, enhanced) and ventilation (with or without air renewal). The alternatives are constructed by improving or deleting some of the actually available amenities. The alternative’s monthly rent is specified based on the modifications of the status quo considering a decrease or increase of 0 to 25 percent of the actual rent (ranging mostly from -300 to 300 Francs per month). A factorial random design has been used to assign the levels of attributes and rents in various alternatives.\footnote{See Banfi et al. (2008) for more details about the experiment design.}

The final sample consists of 3,861 observations from 264 respondents. Table 1 provides the descriptive statistics of the main variables included in the analysis. The
non-market attributes (vector $X$) consists of three dummy variables representing energy-efficient ventilation and enhanced window and facade insulation. The ordinary attributes (vector $Z$) include four dummy variables for the six remaining insulation categories for windows and facade. It is interesting to note that because of the linearity of the effects of rent and market attributes ($R$ and $Z$), for these variables it is the relative differences not the initial levels that matter.

Table 1: Descriptive statistics

<table>
<thead>
<tr>
<th>Status Quo</th>
<th>Hypothetical Offers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Monthly rent (Swiss Francs)</td>
<td>1'660</td>
</tr>
<tr>
<td>Enhanced window insulation (triple glazing)</td>
<td>0.144</td>
</tr>
<tr>
<td>Standard window insulation (rubber sealing)</td>
<td>0.682</td>
</tr>
<tr>
<td>Low window insulation (old)</td>
<td>0.144</td>
</tr>
<tr>
<td>Non-insulated windows (very old)</td>
<td>0.030</td>
</tr>
<tr>
<td>Enhanced facade insulation</td>
<td>0.212</td>
</tr>
<tr>
<td>Standard facade insulation</td>
<td>0.394</td>
</tr>
<tr>
<td>Low facade insulation (newly repainted)</td>
<td>0.121</td>
</tr>
<tr>
<td>Non-insulated facade (old)</td>
<td>0.273</td>
</tr>
<tr>
<td>Ventilation (air renewal system)</td>
<td>0.193</td>
</tr>
<tr>
<td>New building (constructed after 1995)</td>
<td>0.409</td>
</tr>
<tr>
<td>Alternative rent is lower than the S.Q. rent</td>
<td>–</td>
</tr>
<tr>
<td>Alternative offer strictly dominates S.Q.</td>
<td>–</td>
</tr>
<tr>
<td>Positive response (alternative offer chosen)</td>
<td>–</td>
</tr>
<tr>
<td>Number of choice cards per respondent $^a$</td>
<td>14.63</td>
</tr>
<tr>
<td>Number of cards with positive response $^b$</td>
<td>3.28</td>
</tr>
<tr>
<td>Number of observations</td>
<td>264</td>
</tr>
</tbody>
</table>

$^a$ Number of choice cards varies from 11 to 18.

$^b$ Number of cards with positive response varies from 0 to 14.

An asymmetry in the respondents’ preferences observed in the experimental data used in this paper has been reported in Banfi et al. (2008), in that the individuals who are currently using an attribute show a relatively high valuation of that attribute. These results are consistent with several previous studies (Horowitz and McConnell, 2002 and Sayman and Öcüer, 2005) that observed a disparity between WTP and WTA. Moreover, individuals have a tendency to choose their SQ over the
hypothetical offers even in cases that the offers seem to be favorable, suggesting a disutility from changing the SQ.

In this paper, assuming that the asymmetry effect is driven by an attachment to SQ as well as a difference in marginal effects of income (rent), the effect is modeled through differentiating the regression coefficient of the monthly rent depending on the location (in the space of price and attributes) of the SQ compared to the hypothetical offer. Namely, in addition to a dummy variable representing the SQ, two interaction terms with the rent variable are also included. By including these interactions, we expect that the main price coefficient be abstracted from the endowment effect and the SQ inertia. The first interaction term is applied to cases when the hypothetical offer has a lower price compared to the SQ rent. Whereas the second term is used to differentiate the marginal value of money when the (hypothetical) alternative strictly dominates the SQ situation, that is, an apartment with strictly better attributes is offered at a strictly lower price. The data indicate that probably because of the disutility of any change (moving apartments), these cases do not systematically receive positive response. However, these cases cannot be pooled with those cases that appeal to endowment effect. In fact, the regression results suggest that the respondents show a relatively high responsiveness to prices in these cases compared to the endowment-effect situation.

13 See Scarpa et al. (2005) for a discussion of various methods of modeling the status-quo effects.

14 The endowment effect is the commonly observed effect in which the decision maker shows a relatively high valuation of a commodity (or low value for money) when she is to lose that commodity.

15 Following Scarpa et al. (2005), we also considered several alternative specifications that include additional interaction terms for various attribute variables. The results (available upon request) show that these effects are generally insignificant, unless the price interactions were excluded. This might suggest that the available data does not allow a sensible differentiation of WTA from WTP for each attribute. Noting that the WTA discussions are beyond the scope of this paper, we decided to restrict the interaction terms to price variables.
5. Results

The regression results obtained from models 1 to 5 are provided in Table 2. These results are obtained by maximum likelihood method. Model 1 is the conventional linear model. Overall, the results in terms of sign and significance are plausible across all the models. As expected the negative effect of the SQ variable indicate a significant reluctance against change, the positive effect of the first interaction term indicates a considerably lower responsiveness to decreasing prices, suggesting a strong endowment effect. However, when the alternative offer dominates the SQ, the respondents show relatively more responsiveness to prices, as indicated by the negative value of the second interaction term.
Table 2: Regression results

<table>
<thead>
<tr>
<th></th>
<th>Model 1 Linear</th>
<th>Model 2 Quadratic</th>
<th>Model 3 Logarithmic</th>
<th>Model 4 Power</th>
<th>Model 5 Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly rent (Swiss Frs)</td>
<td>-0.0063 **</td>
<td>-0.0063 **</td>
<td>-0.0059 **</td>
<td>-0.0063 **</td>
<td>-0.0054 **</td>
</tr>
<tr>
<td></td>
<td>(.0006)</td>
<td>(.0006)</td>
<td>(.0006)</td>
<td>(.0006)</td>
<td>(.0005)</td>
</tr>
<tr>
<td>Monthly rent × (alternative rent is lower than the SQ rent)</td>
<td>.0052 **</td>
<td>.0051 **</td>
<td>.0052 **</td>
<td>.0052 **</td>
<td>.0054 **</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td>(.0008)</td>
<td>(.0008)</td>
<td>(.0008)</td>
<td>(.0008)</td>
</tr>
<tr>
<td>Monthly rent × (alternative offer strictly dominates the status-quo)</td>
<td>-0.0018 **</td>
<td>-0.0018 **</td>
<td>-0.0027 **</td>
<td>-0.0018 **</td>
<td>-0.0047 **</td>
</tr>
<tr>
<td></td>
<td>(.0006)</td>
<td>(.0006)</td>
<td>(.0005)</td>
<td>(.0006)</td>
<td>(.0005)</td>
</tr>
<tr>
<td>Ventilation (air renewal system)</td>
<td>.5302 **</td>
<td>.5718 **</td>
<td>.7610 **</td>
<td>.5391 **</td>
<td>.3652 **</td>
</tr>
<tr>
<td></td>
<td>(.081)</td>
<td>(.093)</td>
<td>(.164)</td>
<td>(.095)</td>
<td>(.114)</td>
</tr>
<tr>
<td>Enhanced window insulation (triple glazing) a</td>
<td>1.647 **</td>
<td>1.948 **</td>
<td>6.373 **</td>
<td>2.132 **</td>
<td>13.556</td>
</tr>
<tr>
<td></td>
<td>(.155)</td>
<td>(.283)</td>
<td>(.1419)</td>
<td>(.373)</td>
<td>(.111)</td>
</tr>
<tr>
<td>Standard window insulation (rubber sealing) a</td>
<td>1.288 **</td>
<td>1.2823 **</td>
<td>1.1329 **</td>
<td>1.2866 **</td>
<td>.6170 **</td>
</tr>
<tr>
<td></td>
<td>(.130)</td>
<td>(.130)</td>
<td>(.127)</td>
<td>(.130)</td>
<td>(.483)</td>
</tr>
<tr>
<td>Low window insulation (old) a</td>
<td>.5916 **</td>
<td>.5876 **</td>
<td>.4845 **</td>
<td>.5894 **</td>
<td>.1695 *</td>
</tr>
<tr>
<td></td>
<td>(.123)</td>
<td>(.123)</td>
<td>(.119)</td>
<td>(.122)</td>
<td>(.103)</td>
</tr>
<tr>
<td>Enhanced facade insulation b</td>
<td>.7973 **</td>
<td>.9583 **</td>
<td>1.5444 **</td>
<td>.9972 **</td>
<td>.4019 *</td>
</tr>
<tr>
<td></td>
<td>(.155)</td>
<td>(.206)</td>
<td>(.459)</td>
<td>(.224)</td>
<td>(.228)</td>
</tr>
<tr>
<td>Standard facade insulation b</td>
<td>.6098 **</td>
<td>.6112 **</td>
<td>.3613 **</td>
<td>.5954 **</td>
<td>.1088</td>
</tr>
<tr>
<td></td>
<td>(.117)</td>
<td>(.116)</td>
<td>(.091)</td>
<td>(.116)</td>
<td>(.075)</td>
</tr>
<tr>
<td>Low facade insulation (newly repainted) b</td>
<td>.3751 **</td>
<td>.3728 **</td>
<td>.2762 **</td>
<td>.3625 **</td>
<td>.1720 *</td>
</tr>
<tr>
<td></td>
<td>(.104)</td>
<td>(.104)</td>
<td>(.100)</td>
<td>(.104)</td>
<td>(.095)</td>
</tr>
<tr>
<td>Status quo indicator</td>
<td>-1.323 **</td>
<td>-1.337 **</td>
<td>-1.301 **</td>
<td>-1.344 **</td>
<td>-1.192 **</td>
</tr>
<tr>
<td></td>
<td>(.096)</td>
<td>(.097)</td>
<td>(.094)</td>
<td>(.097)</td>
<td>(.091)</td>
</tr>
<tr>
<td>( \lambda ) (coefficient of the quadratic term)</td>
<td>-0.0438 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.024)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r ) (exponent of the power function)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.831 **</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.082)</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-1691.9</td>
<td>-1691.0</td>
<td>-1698.9</td>
<td>-1689.9</td>
<td>-1725.4</td>
</tr>
<tr>
<td>AIC (Akaike Information Criterion)</td>
<td>3405.9</td>
<td>3406.1</td>
<td>3419.8</td>
<td>3403.9</td>
<td>3472.7</td>
</tr>
<tr>
<td>BIC (Bayesian Information Criterion)</td>
<td>3474.7</td>
<td>3481.2</td>
<td>3488.6</td>
<td>3479.0</td>
<td>3541.6</td>
</tr>
<tr>
<td>HQC (Hannan-Quinn Criterion)</td>
<td>1713.1</td>
<td>1712.1</td>
<td>1720.0</td>
<td>1711.1</td>
<td>1746.5</td>
</tr>
</tbody>
</table>

* The omitted window category is non-insulated windows (very old).
* The omitted facade category is non-insulated facade (old).
** significant at \( p<.05 \); * significant at \( p<.10 \); Standard errors are given in parentheses.

The estimation results indicate a reasonable explanatory power for all adopted models. The overall rate of correct prediction\(^{16}\) of the respondents’ choices in the sample is about 80% for the three models with linear, quadratic and power functional forms and 73% for the remaining two models. The exponential model showed some numerical problems and certain sensitivity to the initial values, which could be explained by the fact that the likelihood function can have undefined values at zero

\(^{16}\) A predicted probability of greater than \( \frac{1}{2} \), of accepting the offer is considered as positive response.
values for attributes. This model (Model 5) along with the logarithmic model (Model 3) show a relatively poor performance as regards to log-likelihood and the conventional diagnostic criteria as listed in Table 2. The remaining models are quite comparable in terms of results and prediction power. However, Model 4 (power function) can be singled out: The likelihood ratio test rejects at 5% significance level, the linear model in favor of this model ($\chi^2=4.025$). Moreover, according to two of the three conventional criteria namely, AIC and HQIC Model 4 outperforms all other models.

The results (Table 2) also provide evidence against the risk-neutrality hypothesis, for the null hypotheses $H_0: \lambda=0$ and $H_0: r=1$ are both rejected respectively in Models 2 and 4. Especially the evidence is stronger in Model 4 that rejects the null at 5% significance level. Considering these results, Models 4 along with the linear model are retained for the rest of our analysis. The estimated values of WTP based on these two models are listed in Table 3. The upper panel provides the WTP for amenities that are relevant to new buildings (approximately those constructed after 1995), which we labeled as non-market or new technologies, whereas the lower panel lists the numbers related to the ordinary insulation systems that exist in virtually all Switzerland’s residential buildings constructed after 1995.
Table 3: Estimates of willingness-to-pay

<table>
<thead>
<tr>
<th>Improvements in new buildings</th>
<th>Initial status</th>
<th>Min.</th>
<th>Mean</th>
<th>Max.</th>
<th>Min.</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ventilation Enhanced</td>
<td>2.8</td>
<td>4.2**</td>
<td>5.6</td>
<td>1.2</td>
<td>2.9**</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>Ventilation Standard</td>
<td>2.8</td>
<td>4.2**</td>
<td>5.6</td>
<td>3.1</td>
<td>4.7**</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Window enhanced insulation Enhanced</td>
<td>1.5</td>
<td>2.8**</td>
<td>4.2</td>
<td>0.2</td>
<td>1.8**</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Window enhanced insulation Standard</td>
<td>1.5</td>
<td>2.8**</td>
<td>4.2</td>
<td>2.3</td>
<td>4.7**</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>Facade enhanced insulation Enhanced</td>
<td>0.1</td>
<td>1.5**</td>
<td>2.8</td>
<td>-0.8</td>
<td>0.7</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>Facade enhanced insulation Standard</td>
<td>0.1</td>
<td>1.5**</td>
<td>2.8</td>
<td>0.9</td>
<td>3.2**</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>Window &amp; facade enhanced insulation Standard</td>
<td>2.4</td>
<td>4.3**</td>
<td>6.3</td>
<td>3.3</td>
<td>5.5**</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>Full enhanced insulation &amp; ventilation Standard</td>
<td>6.1</td>
<td>8.5**</td>
<td>11.0</td>
<td>6.0</td>
<td>8.4**</td>
<td>10.9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Improvements in old buildings</th>
<th>Initial status</th>
<th>Min.</th>
<th>Mean</th>
<th>Max.</th>
<th>Min.</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window standard insulation Low</td>
<td>5.7</td>
<td>8.5**</td>
<td>11</td>
<td>5.7</td>
<td>8.5**</td>
<td>11.3</td>
<td></td>
</tr>
<tr>
<td>Facade standard insulation Low</td>
<td>-0.1</td>
<td>2.9*</td>
<td>5.8</td>
<td>-0.1</td>
<td>2.8*</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Window low insulation Poor</td>
<td>4.2</td>
<td>7.2**</td>
<td>10.2</td>
<td>4.2</td>
<td>7.2**</td>
<td>10.2</td>
<td></td>
</tr>
<tr>
<td>Facade re-paint Poor</td>
<td>1.9</td>
<td>4.6**</td>
<td>7.2</td>
<td>1.8</td>
<td>4.4**</td>
<td>7.0</td>
<td></td>
</tr>
</tbody>
</table>

** significant at p<.05; * significant at p<.1.
WTP estimates are given as percentage of the average monthly rent (New: 2000 Frs; Old: 1300 Frs).
The minimum and maximum values correspond to the 95% confidence interval.

Jakob (2006) provides a summary of the actual costs of each one of these improvements, which can be used as a comparison basis for the plausibility of the WTP estimates. However, as pointed out by that author transforming the costs to monthly annuities relies on several assumptions about the discount rate and the building’s lifetime as well as the apartment’s characteristics. While such a discussion is beyond this paper’s scope, it is worth noting that the estimated WTP values are quite comparable to the estimated costs. It is also interesting to note that while for the ordinary improvements (from low to standard) the WTP estimates are generally (but slightly) higher than the corresponding costs, for energy-efficient systems the ordering could be easily reversed depending on the adopted assumptions.17

17 For instance a typical cost estimate of updates as a percentage of the monthly rent is about 5% for window (or facade) enhanced insulation and 8% for ventilation, and about 4% for updating low to standard insulation of windows (or facade). Repainting facade costs only about 1 or 2 percent, in which case the high WTP reflects the consumers’ valuation of aesthetics. Because of the large variety of old windows, such comparisons are not sensible for updates in poorly insulated windows. See Jakob (2006) and Ott et al. (2006) for more details.
As shown in Table 3, the linear model’s estimates of WTP are independent of the initial status. The non-linear model however, can distinguish between the WTP for each attribute separately. For instance, if we consider an improvement from standard insulation to a fully efficient system, the total WTP estimate from both models is about 8.5 percent. According to the non-linear model (Model 4) being initially at standard insulation, the consumer values both ventilation and window (enhanced) insulation at about 4.7%, but the facade improvement comes at only 3.2%. Now if the individual decides to improve the windows, further improvement of the facade does not provide any statistically significant value (WTP=0.7%) in the non-linear model, but has 1.5% WTP in the linear model. The non-linear model shows that the same individual will have 2.9% WTP for ventilation system (as opposed to 4.2% from the linear model). If these numbers can be used as a guide for policy programs for promoting energy-efficiency, the state should subsidize only certain forms of the improvements (for instance facade insulation) or sets the subsidies conditional on an entire improvement in all attributes.

Overall the results suggest that misspecification of a risk-averse behavior with a linear utility model, might result in misleading estimates for certain attributes, especially if individuals consider the attributes separately rather in single packages. In general, the non-linear model highlights the diminishing marginal utility of the attributes. The non-linear model shows that as the level of energy efficiency increases the valuation of further improvements decreases substantially. In principle modeling the varying rate of substitution would be alternatively possible by constructing appropriate interaction terms with the SQ levels of attributes. However, from a practical viewpoint, depending on the number of categories such a strategy might lead
to an excessively large number of regression parameters\textsuperscript{18} and in most cases, could cause difficulties in interpretation of the results.

Figure 1 depicts the estimated relative Risk Premium (RP) based on Model 4 (power function) for enhanced insulation and ventilation\textsuperscript{19}. The numerical values are obtained from Equations (9) and (10) assuming a normal distribution for the utility $\tilde{u} = \tilde{u}(x)$, of having all the attributes, with mean $\mu = (x'\alpha)'$, and standard deviation $\sigma = \delta \mu$, where the variability ratio $\delta$, can be considered as a measure of risk. This risk is the inherent risk in the decision maker’s perceived utility (cash payoff) of the energy-saving attributes. RP values are plotted as a function of $\delta$, for several values of $r$. The solid line shows the estimated values based on the data with $r=.831$.

**Figure 1: Risk premium for enhanced insulation and ventilation**

\textsuperscript{18}For instance for 3 binary variables the number of additional interaction terms will be 15, including $3 \times 3$ second-order and $3 \times 2$ third-order terms.

\textsuperscript{19}RP values for other categories of attributes (as in the upper panel of Table 3) show a very similar pattern to that shown in Figure 1.
As illustrated in the figure, the model predicts about 0.5% risk premium for variability of 20% in the perceived utility around its expected value. But the RP increases quickly as this variation increases, with a value of 1.2% for 30% risk to a staggering value of 25% risk premium for a rather extreme variability of 80%. What values can be considered as reasonable risk is an interesting question for further research. However, it should be noted that the risk involved in the perceived benefits of energy-efficient systems most probably depends on the decision maker’s knowledge about these systems as well as the risks involved in energy prices. A variability of 50% in a normal distribution implies that the utility can vary with 95% confidence, roughly between zero and two times the expected utility. To understand the importance of the issue, let us consider this risk as a reasonable value, in which case the predicted risk premium (6.2%) has important policy implications. If we extend this risk-aversion to investment decisions, this implies that energy efficient systems in residential buildings need to have about 6% extra return (compared to interest rate on safe assets), in order to be economically viable.

Figure 1 also shows that the RP values can vary considerably with the model parameter $r$, which represents the degree of relative risk aversion (Arrow-Pratt coefficient $c_R=1-r$). For instance, if we consider $r=.7$, even 30% variability leads to a risk premium of 2.9%, whereas 50% variability will result in 11% risk premium. Even if we consider a value close to risk-neutrality (say $r=.95$), 50% variability implies 3.6% risk premium. While these results indicate the sensitivity of RP estimates, suggesting excessive premium in worst cases, the overall result remains valid that even with optimistic assessments, the risk aversion is an important issue for energy-efficiency.
6. Conclusion

This paper proposes a methodological framework to consider the non-linearity of the utility function in terms of non-market goods that are defined by qualitative discrete variables. This is the case of many non-market, public and environmental goods that are not divisible and are consumed only once. The proposed models have important applications in choice experiments conducted for the evaluation of new goods because the consumers could show a risk-averse behavior due to lack of information on the potential benefits. Moreover, such models can solve the general shortcoming of the linear models in assuming constant rate of substitution between non-market goods and other commodities. Such behaviors should be modeled by non-linear functional forms. In particular, the linearity assumption appears to be too restrictive for exploring some of the peculiarities observed in choice experiments such as disparity between willingness-to-pay and willingness-to-accept.

The proposed model is applied to experimental data from a survey about the use of energy-efficient ventilation and insulation measures in apartment buildings. Most of these systems are new technologies that are not widespread in the markets. The purpose of this exercise is to estimate the consumers’ willing-to-pay for these systems, but also to explore if the data provides any evidence of risk-averse behavior. An important assumption is that the non-linear effects of income and all other market goods are considered as insignificant compared to the risks involved by the non-market goods, namely, the enhanced insulation and ventilation systems. This implies that the decision of adopting these technologies has only a marginal effect on the individual’s overall income, thus justifying a linear effect for income.

Using the results one can reject the null hypothesis of risk-neutrality. The results also suggest that the assumption of constant rate of substitution between
market goods and new technologies implied by linear models could result in misleading estimates of willingness-to-pay especially if the valuation of the attributes is considered separately. The estimated results are also used to assess the extent of potential risk premium that the consumers consider for the energy efficient systems. The results point to a significant relative risk premium even if the perceived risks are within reasonably low margins.

References


An Empirical Analysis of Child Care Demand in Switzerland

Silvia Banfi\textsuperscript{a,*}, Mehdi Farsi\textsuperscript{a} and Massimo Filippini\textsuperscript{a,b}

\textsuperscript{a} ETH Zurich, Department of Management, Technology and Economics
Zürichbergstrasse 18, CH - 8032 Zurich, Switzerland

\textsuperscript{b} University of Lugano, Department of Economics
Via G. Buffi 6, CH – 6904 Lugano, Switzerland

Abstract

This paper analyzes the demand of Swiss families for child care facilities. A choice experiment is used to study the effects of the facilities’ characteristics as well as socio-economic factors on the selected child care mode. The experimental data are analyzed using a discrete choice model with multinomial logit specification. The results suggest that the demand for extra-familial day care could be considerably higher than that observed from the actual choices constrained by insufficient provision of affordable day care. The price, access, and the quality of service as well as parents’ income and education have important impacts on the choice of the mode of care.

\textit{JEL classification code:} C25, D12, J13
\textit{Keywords:} Child care; Choice experiment; Discrete choice

Acknowledgements

This paper is based on the results of a joint research project (Stern et al., 2006) with two consulting groups: INFRAS and Tassinari Beratung. The authors gratefully acknowledge the financial support of the Swiss National Science Foundation. They also wish to thank Rolf Iten, Susanne Stern, Sergio Tassinari and Ria Schrottman for their invaluable support throughout the project and Daniela Pióro for her excellent assistance. The authors are responsible for all remaining errors and omissions.

\textsuperscript{*} Corresponding author
1. Introduction

Although child care services are gaining some importance in Switzerland, the provision of extra-familial day care has not been fully developed (OECD, 1994). Day care facilities, organized as child-care centers or in family homes, are usually run privately. Mostly subsidized by the local governments, these services are very limited especially in the countryside and small towns. The non subsidized private facilities are usually limited to large city centers. The pre-school care centers and the allocation offices for family day care homes have long waiting lists, particularly for subsidized providers and for children below two years of age. The non subsidized day care facilities are hardly affordable for most families. The high prices are often considered as a result of the strict regulatory framework and the attempt to guarantee a high quality of care (Stutzer and Dürsteler, 2005). According to statistics about 9% of children younger than 6 years have regular visits to a child care centre while 6% are taken care of in family day care homes. However, given the limited availability and the uncertainty of obtaining a placement for their children, many parents who would potentially demand child care services might be excluded from the market.

The lack of services for supporting parents in their children’s day care can have various negative social and economic consequences, for instance on the fertility rate (Schröder 2005), on women’s labour supply (Stebler 1999), and on the integration possibilities of disadvantaged children. In order to improve the provision of child care centers and family day care homes, the Swiss government initiated an incentive programme for start-up financing of such ser-

---

2 For instance, the prescribed number of children per care-giver and the maximum size of the groups are significantly lower than in the contiguous countries.
3 Stebler (1999) provides empirical evidence that the provision of child care facilities has a crucial impact on the working behavior of mothers in Switzerland. OECD (2004) reports that in Switzerland a relatively high share of working women work less than 30 hours a week (44.9% as opposed to the overall OECD average of 18.8%). Lanfranchi (2002) shows that schooling results of children of immigrants are highly related to their attendance of pre-school facilities.
services. This program is aimed at providing the greatest possible number of families with the access to day care, support and instruction of their children by qualified care givers. Besides the positive effect on the cognitive and social development of children, such a program may have a positive impact on mothers’ participation in the labour market and attenuate the problems related to aging of the population by increasing fertility rates.

From a policy standpoint a statistical estimation of potential child care choices can be very useful given the aim of the State to improve the provision of affordable child care facilities in Switzerland:

1. First, an effective promotion of child care facilities requires information about the parents’ potential demand for different types of child care services. Especially as in certain cases, the alternative modes of day care could be considered to avoid the relatively costly child care centers.

2. Secondly, in order to achieve an optimal provision of day care system, it is essential to assess the sensitivity of the demand in response to variations in the service attributes such as price, distance and quality characteristics.

3. Finally, it is believed that the existing heterogeneity in the provision and types of (subsidized) child care across different areas creates an equity problem regarding both access and variety of choices. An equitable provision would require an estimation of potential demand based on demographic and socio-economic characteristics pertaining to various locations.

In principle, the data on child care choices can be based on revealed or stated preferences. The former method, used by virtually all previous studies, focuses on the households’ actual deci-

---

This heterogeneity is documented in Stern at al. (2006); Parlamentarische Initiative Anstossfinanzierung (2002); and different evaluations of the Swiss Labour Force Survey, Swiss Federal Statistical Office.
sions. The stated preference method on the other hand, draws upon individuals’ choices in hypothetical situations defined by the researcher.\footnote{For an overview of the general advantages and drawbacks of the two methods see for example Verhoef and Franses (2002) or Louviere et al. (2000).}

Given the actual state of day care provision in Switzerland, the observed utilization of child care services does not provide a realistic picture of potential demand. In fact, most often parents do not have several alternatives to choose from. In many cases only a single type of care is available to them. Therefore, the adopted mode of child care is not representative of the families’ real preferences with sufficient provision. In fact, in this situation, the revealed approach could lead to biased predictions, as actual choice behavior reflects a combination of consumer preferences and often prevailing constraints, induced by current market conditions. Moreover, it is often difficult to identify the available choice set from which the actual care has been chosen. Therefore, the revealed preferences based on actual usage are not much helpful in eliciting consumers’ preferences, especially for assessing the potential responses to future provisions and extensions. In the stated preference approach, these problems are solved through preset choice conditions. However, this solution entails hypothetical choice sets, which might bring about a loss of precision, due to the potential risk of careless and inconsistent decisions in vignette surveys.

Following the stated preference approach, this paper simulates families’ decisions with hypothetical choice situations (so called vignettes or choice experiments) in which several modes of child care are offered and the respondents are asked to choose the alternative that suits them best. The data have been collected for a sample of about 600 families with at least one child of pre-school age. Each family has been provided with six different choice situations. A discrete choice econometric model has been applied to the recorded decisions in order to estimate the effect of various household characteristics and care attributes. In particular, the effects of price and quality of care as well as the impacts of other child care possibilities avail-
able to the family and the parents’ current work status have been analyzed. This latter variable is considered to be exogenous to the demand on (hypothetical) child care facilities.\footnote{We consider the current employment situation of the parents as an explanatory variable for the hypothetical (future) choice of child care mode.}

The results point to a considerable variation in demand depending on several demographic characteristics. They also indicate that price and distance from home are the two most important factors that determine the families’ choices. The data confirm that the potential demand at the present subsidized prices (for low and medium income families) is considerably higher than the actual provision of child care services. Especially, the demand for family day care homes is comparable to that of day care centers, which suggests that this option could be considered as an effective substitute.

The paper continues with a description of the methods and the model specification in Section 2. The experiment design and survey procedures are presented in Section 3. Section 4 describes the data and the regression sample and Section 5 provides the estimation results. The paper ends with a summary of the results, an interpretation of the main results and some comments on their policy relevance.

\section{Model}

There is a great quantity of papers applying the choice experiment approach for the estimation of demand for public (and private) goods.\footnote{Actually, the choice experiment approach, called also conjoint analysis, has its conceptual foundation in Lancaster’s theory of consumer demand. This technique was used at first by marketing researchers, who recognized the importance of commodity attributes when designing new products (Champ et al. 2003).} For a review of the history and an overview of the fields of application of choice experiment we refer to Champ et al. (2003). This approach has been frequently applied for the estimation of demand for public services, for example the demand of consumers for different health care plans or public health programs and services (Jan et al 2000; Gyldmark et al., 2001; Harris, 2002), for the analysis of university choices
(Oosterbeek et al., 1992; Soutar and Turner, 2002) or the evaluation of services offered by state-owned cultural sites (Mazzanti, 2003). To our knowledge such an experimental approach has never been applied in the analysis of child care demand.

Concerning the child care demand, there is a great body of international economic literature analyzing the demand for day care facilities, focusing mainly on the impact of service attributes especially prices, on demand (Hofferth and Wissoker, 1992; Chaplin et al., 1996; Van Horn et al. 2001). In particular a stream of this literature that focuses on the sensitivity of women’s labor participation to changes in the price of day care services, bears a considerable policy interest (Cleveland et al., 1996; Pungello and Kurtz-Costes, 1999; Powell, 2002; Del Boca et al., 2004). However, exploring the impact of child care provision on labor supply is beyond the scope of this paper. Here, the focus is on the evaluation of child-care demand and identifying its main determining factors through households’ preferences.

A large part of the analyses on child care demand is based on data from national surveys. The studies use cross sectional (Chaplin et al., 1996; Connelly and Kimmel, 2003) as well as longitudinal data (Leibowitz et al., 1992; Anderson and Levine, 2000). These surveys collect data on the actual child care choices of parents. In Switzerland, there are no national data available on this topic which could be used for forecasting demand. Further, since actual child care choices are often restricted by insufficient provision and access problems, they could be a poor indicator of the families’ preferences.

The general empirical results reported in the child care studies highlight the importance of several factors such as cost of care (Anderson and Levine, 2000; Chaplin et al., 1996), family income and child care tax credits (Hofferth and Wissoker 1992, Michalopoulos et al. 1992), children’s age (Leibowitz et al. 1992), mother’s working hours (Connelly and Kimmel, 2003)

---

8 The only data available are those of the Swiss Labour Force Survey. This survey collects amongst others data on the child care choices of employed parents. Information on the prices of child care services and their availability are not collected.
as well as other socio-economic characteristics on child care choice. Due to the great variation of the data used in these studies, comparison between the results of the studies is not always feasible.

From the econometric point of view, most of the papers employ a discrete choice analysis such as multinomial logit (Kreyenfeld and Hank 2000, Michalopoulos and Robins 2002, Del Boca et al. 2004) or probit (Anderson and Levine, 2000; Chevalier and Viitanen, 2002; Connelly and Kimmel, 2003; Del Boca et al., 2004). Further, a combination of discrete and continuous models has been used when the objective was to estimate the demand quantities (for example, the demand for hours of child care), for a specific child care mode (Powell, 1997).

With reference to the random utility theory\(^\text{9}\), this paper models the choice of child care services for families with children younger than 5 years (before kindergarten). The underlying assumption is that families evaluate the characteristics of different child care services and then choose the service, which maximizes their utility. It is assumed that households consider the tradeoffs between benefits gained from day care services based on care attributes and the incurred costs including service prices and other opportunity costs depending on the household characteristics. According to the random utility theory, the utility of a service or good is considered to depend on observable (deterministic) components, including the attributes of the services and individual characteristics, plus a stochastic element that captures the influence of unobserved factors (cf. Louviere et al. 2000).

We represent the utility function of a child care mode \(j\) for family \(i\) as:

\[
U_{ij} = X_{ij}\beta_j + Z_{ij}\gamma_j + \epsilon_{ij}
\]  

\(^{9}\) For a description of the random utility theory see Louviere et al. (2000) or Ben-Akiva and Lerman (1985).
where $X_{ij}$ is the vector of attributes of alternative $j$ for household $i$; $Z_i$ is the vector of household characteristics; $\beta_j$ and $\gamma_j$ are the parameter vectors to be estimated; and $\epsilon_{ij}$ is an independently and identically distributed stochastic error term that represents the unobserved heterogeneity across households and alternatives. The adopted model in this paper is based on a multinomial logit model in which the error term $\epsilon_{ij}$ is assumed to follow a type I extreme value (Gumbel) distribution. In this model, the probability of choosing alternative $j$ can be written as:

$$
\Pr(Y_i = j) = \frac{e^{X_i \beta_j + Z_i \gamma_j}}{\sum_{j=0}^{J+1} e^{X_i \beta_j + Z_i \gamma_j}} \quad \text{for } j = 0, 1, 2, \ldots, J
$$

(2),

where $J+1$ is the number of alternatives and $Y_i = 0, 1, \ldots, J$ is the individual $i$'s response. As the model in equation (2) is indeterminate, it requires a normalization assumption, which can be obtained by setting $\beta_0$ and $\gamma_0$ equal to zero. Thus, equation (2) can be written as:

$$
\Pr(Y_i = j) = \frac{e^{X_i \beta_j + Z_i \gamma_j}}{1 + \sum_{j=1}^{J+1} e^{X_i \beta_j + Z_i \gamma_j}} \quad \text{for } j = 0, 1, 2, \ldots, J, \beta_0 = \gamma_0 = 0
$$

(3),

where alternative $j=0$ is considered as the comparison outcome.

It is worth noting that in this study the households are offered repeated choice situations and a more accurate presentation of the model should consider an index for the choice situation (card). Moreover, the number of alternatives is set equal to four. The model in equation (3) can thus be written as:

$$
P_{icj} = \Pr(Y_{ic} = j) = \frac{e^{X_{ic} \beta_j + Z_{ic} \gamma_j}}{1 + \sum_{j=1}^{3} e^{X_{ic} \beta_j + Z_{ic} \gamma_j}} \quad \text{for } j = 0, 1, 2, 3, \beta_0 = \gamma_0 = 0
$$

(4),

10 For more details about the multinomial logit model see Greene (2003), chapter 21.
where \( c \) is the choice-situation (card) number. Notice that the choice attributes vary across different cards, but the parameters are alternative-specific.

The marginal effects of the continuous explanatory variables are calculated as the partial derivative of the probability of outcome \( j \), that is: \( P_j \), with respect to the explanatory variable \( x \), which is an element of the explanatory vector \( [X, Z] \). The marginal effect and elasticity of a continuous variable \( x \) can thus be obtained respectively from:

\[
\frac{\Delta P_j}{\Delta x} \approx \frac{\partial P_j}{\partial x} = P_j \left[ \beta_j^* - \sum_{k=1}^{3} P_k \beta_k^* \right] \quad \text{for } j = 0, 1, 2, 3 , \beta_0^* = 0
\]

\[
e_x = \frac{\partial P_j}{\partial x} \left( \frac{x}{P_j} \right)
\]

where \( \beta_k^* \) represents the coefficient related to outcome \( k \), of explanatory variable \( x \), that is the corresponding element of the parameter vector \( [\beta, \gamma] \). Similarly, the marginal effects for dummy variable \( x \) can be obtained from the following equation:

\[
\frac{\Delta P_j}{\Delta x} = P_j(x = 1) - P_j(x = 0)
\]

The model’s explanatory variables include several child care attributes such as price, distance from home and quality of the service. In line with previous empirical studies the household’s socio-economic characteristics such as parents’ education, income and work status are also included. Since any additional variable requires three more parameters in the model, we tried to limit the number of parameters to a reasonable number. The final model specification was selected using a series of Wald tests to identify and exclude the variables that have no statistically significant effect in any outcomes. Therefore, some of the variables in the data, which
would have otherwise required several dummy variables, have been reduced to a single dummy.\textsuperscript{11} A list of the variables and their definition are provided in Section 4.

It is reasonable to assume, as in any grouped data, that the errors can be correlated across the observations that belong to the same household. Here, the correlation within household observations is considered by robust standard errors with the cluster option in Stata program.\textsuperscript{12} In this method the errors are only required to be independent across groups and can be correlated within groups. Consequently, the variations within groups contribute little to the estimation precision. The standard errors are therefore more realistic than those obtained with the independence assumption, which may be under-estimated.

3. Experiment design and data description

The data used in this paper are collected using a choice experiment approach. This approach initially proposed by Louviere and Hensher (1983), consists of asking a number of respondents to choose one among several alternatives characterized by various attributes. Within the range of non-market valuation techniques, choice experiment is most appropriate for capturing the implicit values of a good or service as a whole or its given attributes (Birol et al., 2005).

Of course, the choice experiment approach has also some limitations.\textsuperscript{13} One disadvantage is linked to the cognitive skills which are required from respondents when choosing the utility maximizing alternative from a complex choice situation. This complexity can lead to decisions which do not reflect a utility maximization process but rely on short-cuts. This is the

\textsuperscript{11} For instance, the mother’s education is available in 13 categories, but after controlling for other variables only mothers with university degrees showed significant difference from others. Child’s age and gender categories had no significant effect on choice probabilities. Thus, only one dummy variable with relatively important effect has been included. Similarly, the scheduling flexibility, which has been defined in 5 categories, was shown to be significant only when one-month-ahead scheduling is required. Finally, the measures related to opening hours of child-care centers did not appear to be significant.

\textsuperscript{12} See Moulton (1990) for more details about heteroscedasticity in grouped data, and Rogers (1993) for the clustering approach.

\textsuperscript{13} For a more detailed overview see for example Bateman et al. (2002).
case for example when respondents decide by considering just one attribute instead of the entire set of attributes. In order to reduce the cognitive difficulties, the hypothetical choice situation should be defined by a limited number of attributes. As a consequence, the characteristics used in our choice experiment have been chosen with particular accuracy and were collected with a separate survey which allowed identifying the most important characteristics for the choice of child care services. Some evidence for the importance of the child care characteristics selected for the choice experiment was also given by the literature reviews.

The study design, in particular the choice of the attributes as well as their levels, can have an impact on the choice experiment’s results, which is another disadvantage of the method. In this regard it has to be considered that this is the case with all stated preference techniques. Another problem mentioned in the literature\textsuperscript{14} is linked with the assumption that the sum of the attributes’ values sum up to the value of the whole good, although not all attributes can be considered in the choice experiment\textsuperscript{15}. Although it is important to consider these limitations when choosing an evaluation technique, we believe that in our case the advantages of the use of a choice experiment method prevail over the drawbacks. Overall, we judge this method as interesting and appropriate for estimating the demand for hypothetical\textsuperscript{16} child care facilities.

In this paper the experiment simulates a choice situation in which the respondent is asked to choose one care mode among several options. Each option is characterized by a series of attributes. The range of parameters and attributes are chosen within a realistic range comparable to the actual state in Switzerland.

The extra-familial day care for children can be classified in three main categories:

1. **Child care center**: Day care provided by professional staff with several children in a facility, other than private residence, which is specifically equipped for this purpose.

\textsuperscript{14} Bateman at al. (2002), Hanley et al. (2001)
\textsuperscript{15} In this case, the value is captured in the constant term.
\textsuperscript{16} The child care facilities are hypothetical in terms of not available for a large share of parents.
2. **Family day care home**: Day care provided by a parent who has one or more children of their own. The children are looked after in the caregiver’s private residence.

3. **Baby-sitter**: Care provided by a private individual at home.

In the experiment the alternatives are organized in four modes including the above alternatives plus a fourth option labeled as *private care*, that represents parental care as well as all other options arranged within the circle of relatives and friends. In contrast to the other types of care, the private care is unpaid. In each choice situation the respondent has access to one option from each the three alternatives. This is more or less similar to the actual situations, in which the availability of multiple options of a single mode happens very rarely.

The external child care modes were characterized by the following attributes:

1. **Price for half a day care**: In order to simulate the customary pricing policy in Switzerland the prices are selected from an interval proportional to the household’s income. Thus, the hypothetical prices consider automatically income difference between the rural and urban areas. The price of the child care center has been set between 0.3 and 0.6% of the family’s income per half-day of care. The price of the family day care home was set slightly lower that is, between 0.2 and 0.5% of the family’s income for half a day. The average price level corresponds approximately to the price set currently by the child care facilities. Finally, the price for the nanny option was selected in a range similar to the actual market rates namely, between 60 and 100 Swiss francs (CHF)

2. **Distance from home**: The distance was set between 5 and 25 minutes without specifying the transport mode.

---

17 1 CHF ≈ 0.6 €

18 The families were asked to assume that they took their usually preferred transport mode. We wanted to avoid the possibility of refusal of an alternative only because of the suggested transport mode. The distance for the nanny alternative is naturally set to zero.
3. **Opening hours**: For each alternative five different levels of opening hours were defined. Child care centers are usually open from Monday through Friday. For the family day care home and the baby-sitter, some choice cards considered availability of care on Saturday and Sunday. The opening hours varied between 9 and 14 hours a day.

4. **Number of children per staff member**: This characteristic represents a quality aspect of the care. The number of children per care-giver varied between 3 and 7 children for the child care center and between 3 and 6 for the family day care home.

5. **Flexibility**: This attribute represents the scheduling flexibility. In the most restrictive form, the child care service is available only on certain days with the possibility of rescheduling on a monthly basis. In the most flexible form there is the possibility to use the service at short notice and without restriction on the number of hours.

Table 1 shows an example of a choice situation that has been presented to the families. Each family has been presented six different choice cards. Respondents were asked to imagine that the three offered alternatives to private care are available in their residence area and that can be obtained without the usually required registration in the waiting lists.

A full fractional design of all levels of the attributes would require a very high number of cards. Therefore, the different levels of the characteristics were combined using an orthogonal factorial design (Louviere et al. 2000, Champ et al. 2003). Using this approach redundant combinations of the levels of the characteristics are omitted. Thus it was possible to cover the whole space of attribute combinations with a limited number of alternatives. This allows maximizing the information obtained by the choice experiment, without presenting all combination possibilities to the respondents.
Table 1: Example of a choice card

<table>
<thead>
<tr>
<th>SITUATION 1</th>
<th>Child care alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CHARACTERISTICS</strong></td>
<td><strong>ALTERNATIVE 1</strong></td>
</tr>
<tr>
<td>Price for half a day</td>
<td>CHF 40 per child 10 minutes</td>
</tr>
<tr>
<td>Distance from home</td>
<td>Monday-Friday 7 a.m.– 6 p.m.</td>
</tr>
<tr>
<td>Opening hours</td>
<td>5</td>
</tr>
<tr>
<td>Number of children per staff-member</td>
<td>Fixed day, monthly scheduling</td>
</tr>
<tr>
<td>Flexibility of the service</td>
<td></td>
</tr>
<tr>
<td>My choice is:</td>
<td></td>
</tr>
<tr>
<td>Days per week (e.g. 1 day, 2.5 days...)</td>
<td></td>
</tr>
</tbody>
</table>

The respondents were initially a random sample from the population of families living in nine Swiss cantons that participated in the study, reflecting all three linguistic regions of Switzerland. The municipalities in the selected cantons currently offer a number of child care facilities. Thus, we could assume that some of the parents have already used this service. In order to attain a balanced sample across rural and urban areas (according on the distribution of the sample frame population), special attention was put on the regional distribution of the households in the sample. Within the cantons, the parents of children aged below four years were chosen randomly from a database of the market research company commissioned with the survey. The families were first contacted by phone and asked about their family composition and the age of the children. The families with any children of four years old or younger were asked further questions on their actual child care choices as well as some socio-economic characteristics, including income. In a second stage, the families were mailed six choice cards with the alternatives day care modes and the related instructions. In a third stage, they were...

---

19 The participating cantons are: Bern, Luzern, Zug, Baselstadt, Aargau, Ticino, Vaud, Wallis and Jura. Switzerland has 26 cantons.
contacted by phone and asked to reveal their choices. The average length of the interview was about 24 minutes. The survey was carried out between October 2003 and July 2004.

From the 694 households that participated at the first stage of the survey 88% have completed the choice cards and participated at the second part of the survey. The final sample including the valid observations used for this study consists of 2972 records from 597 families. Thus it is possible that the final sample is not representative of the initial population. However, a primary analysis of several household characteristics such as income, household size and parents’ age and work status suggests that the composition of the households included in the final sample is not significantly different from that of the initial sample in regards to these variables. Considering this and in view of the relatively high participation rate, we contend that the sample can be considered as a fairly representative sample for the participating cantons.

The families include households living in both rural and urban areas with about 61% in the latter group, and both German-speaking (58% of households) and the French and Italian-speaking (42%) parts of Switzerland.

A descriptive summary of the sample used in the econometric analysis is given in table 2.
Table 2: Descriptive statistics of socioeconomic characteristics (N=597)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child is one-year old or younger</td>
<td>0.268</td>
<td>0.443</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>One parent's 1st nationality is not Swiss</td>
<td>0.206</td>
<td>0.405</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Rural household</td>
<td>0.390</td>
<td>0.488</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>French/Italian speaking region</td>
<td>0.414</td>
<td>0.493</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Additional child(ren) younger than 5</td>
<td>0.382</td>
<td>0.486</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Additional child(ren) of age 5-12</td>
<td>0.405</td>
<td>0.491</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Additional child(ren) of age 13-18</td>
<td>0.049</td>
<td>0.215</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mother's age</td>
<td>33.662</td>
<td>4.188</td>
<td>22</td>
<td>49</td>
</tr>
<tr>
<td>Mother has a university degree a</td>
<td>0.152</td>
<td>0.360</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mother works 50% or more</td>
<td>0.256</td>
<td>0.437</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Household monthly income in CHF 1000 b</td>
<td>6.015</td>
<td>2.178</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>The respondent is the father c</td>
<td>0.095</td>
<td>0.294</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Father's job is not a normal daily job d</td>
<td>0.186</td>
<td>0.389</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Main child-care provided by parents</td>
<td>0.591</td>
<td>0.492</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Main child-care by relatives/friends</td>
<td>0.258</td>
<td>0.438</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number of choice situations (cards)</td>
<td>4.978</td>
<td>0.960</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

*a University degree means an education level of University or professional college.

*b Monthly income is available as a multiple of thousand Swiss Francs (e.g. 3 means between 3000 and 4000).

*c In these households the father is the main person in charge of child-care arrangements.

*d Also includes cases in which father does not have any employment.

In the initial part of the questionnaire respondents were asked about their actual day care choices. The respondents reported the two options they have used most often.\(^{20}\) Only 10.2% of the families use day care centers and 4.7% use the services of family day care homes as their primary option. As for the nanny option, only one household reported to have used this option as their main child care. The rest of the sample, namely 84.9 percent of the families report a private alternative as their main care option. Overall, considering the two reported main child-care options, the data show that 78.7 percent of the families do not actually use

---

\(^{20}\) Among the families who reported non-zero values, the median number of hours of care used for the main child-care option is 15 hours per week and that of the secondary choice is about 5 hours per week.
extra-familial day care; 14.6% reported to use a child care center; 6.2% family homes; and only 0.5% (3 families) use baby-sitters.\textsuperscript{21}

The distribution of hypothetical choices shows that the child care center and family home alternatives have been chosen respectively in 28.0% and 22.9% of the cases. While the private option has been selected in 44.8% of the cases, only in 4.3% of the hypothetical observations the baby-sitter alternative has been chosen. 181 households (about 30% of the respondents) have always chosen the private alternative. This suggests that these households have probably access to a private source of care and are not responsive to any changes in the attributes of other alternatives. On the other hand, about 69 percent of the households have chosen at least once, a child-care center or a family home option. Only a small fraction of families (about 7%) have always chosen the same non-private option, suggesting that there is no strong preference for any one of these choices.

Given that in the experiment design the values of the choice attributes are simulated based on the real world, the data suggest that the positive response to the child-care center and family home options could be increased to levels as much as twice or three times the actual utilization rates. However, given that many families use the external care as their complementary day care option, the primary and secondary actual choices may understate the actual usage of the extra-familial care.

A descriptive summary of the explanatory variables used in the econometric analysis is given in table 3.

\textsuperscript{21} The number of observations of parents using formal care modes is too low for carrying out estimations using the actual choices. In addition, there is no information on alternative care modes (presence of such alternatives, characteristics).

\textsuperscript{24} The results of the alternative regressions are available upon request from the authors.
Table 3: Descriptive statistics of the characteristics of the chosen alternatives (N=2972)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day Care Center price (CHF/half-day)</td>
<td>29.103</td>
<td>11.362</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>Family Home price (CHF/half-day)</td>
<td>23.093</td>
<td>10.307</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>Baby Sitter price (CHF/half-day)</td>
<td>75.760</td>
<td>13.877</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>Number of children per person (DCC)</td>
<td>5.041</td>
<td>1.399</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Number of children per person (FH)</td>
<td>4.489</td>
<td>1.124</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Distance from DCC (multiples of 5 minutes)</td>
<td>3.002</td>
<td>1.418</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Distance from FH (multiples of 5 minutes)</td>
<td>3.029</td>
<td>1.397</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>FH is open at least 1 week-end day</td>
<td>0.401</td>
<td>0.490</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DCC requires 1-month-ahead scheduling</td>
<td>0.397</td>
<td>0.489</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FH requires 1-month-ahead scheduling</td>
<td>0.392</td>
<td>0.488</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

5. Estimation results

The regression results are given in table 4. The model is based on four alternatives. Before discussing the results, we turn to two modeling issues that we considered. First, the multinomial logit model assumes that the ratio of odds between a pair of alternatives does not depend on the third alternative. This assumption referred to as independence of irrelevant alternatives (IIA) can be violated particularly when the decision is made in a nested manner. One might argue that families first decide whether they would like to keep their child at home (babysitter option) and then depending on this first-stage decision they examine the external options (child care center and family day care home). In this case, the IIA assumption is not satisfied because the exclusion of one external care is likely to result an increase in the other external care but no considerable change in the home option. We used a series of Hausman tests to test the IIA hypothesis with respect to each one of the alternatives. The results are in favor of the independence assumption with relatively high p-values.
Our second concern was the fact that the response to the baby-sitter outcome is quite low (about 4.3%), which might decrease the model's statistical efficiency through small sample errors. We therefore decided to consider an alternative analysis in which we ignore the baby-sitter option and exclude the households that have chosen this option from the sample. The results in terms of marginal effects on the external alternatives are very similar to those reported in the paper.\textsuperscript{24}

The coefficients listed in table 4 show the effects of the explanatory variables on the probability of choosing child care center, family day care home and the baby-sitter option compared to the base category (private option). Many of the coefficients are statistically significant and the model shows a reasonable explanatory power as indicated by a 54.8% rate of correct prediction of the choices in the sample. Most of the coefficients have the expected sign and the main choice attributes such as price and distance are highly significant in both choices.

Table 5 provides the marginal effects along with the elasticity estimates for the continuous variables at the sample mean.\textsuperscript{25} Only the significant effects at 5% significance level are included in the table.

The results indicate that many of the household characteristics have a significant effect on choice probabilities. For instance immigrant families are on average about 10 percentage points more likely to choose child care center. Compared to German-speaking households, families residing in French/Italian-speaking regions are about 5% more likely to use non-private external child-care.

The above results are more or less consistent with the actual choice of these families. Among the 597 households in our sample, about 10% are actually using a child care center, and about 5% a family day care house as their main child-care option. These numbers increase to about

\textsuperscript{25} The marginal effects and elasticities were also estimated for each observation and then averaged over the sample. The results (available upon request) are not much different from the estimated values at the sample mean.
15 and 10 percent (respectively for child care center and family day care home) among the 123 immigrant families (with at least one foreign parent) in the sample. As for the actual choice among the 233 ‘Latin’ families in the sample, the distribution changes to about 15% for the child care center and 8% for the family day care home alternative.

Table 4 shows that there is no statistically significant difference between the choices of the residents of rural and urban areas. However, among the 233 rural households in the sample, only about 9% reported to actually use an external child-care (child care center or family day care home) as their main option. This can be explained by the fact that access to child-care services is relatively limited in rural areas.

The presence of a sibling has a significant effect in choice probabilities. The estimation results suggest that the households with additional children older than 5 are less likely to use a non-private care option. However, the effect depends on the age category of the sibling(s): Households with teenage children are on average 14% less likely to choose a family day care home whereas the presence of children between 5 and 12 decreases the child care center incidence by 0.09 on average. Families with more than one preschool-age children are 2 percent more likely to hire a baby-sitter.

Older mothers are significantly more likely to choose an external day care especially a child-care center. Mothers with university degrees are on average 9% more likely to use child care centers. The results also suggest that the demand for non-private child care increases with family income. However, the effect for the family home option is not statistically significant. As for child care centers and baby-sitters the demand elasticity of income is respectively 0.56 and 0.92. As expected because of relatively high prices of baby-sitters the demand for this is relatively sensitive to income.
Table 4: Regression results

<table>
<thead>
<tr>
<th></th>
<th>Child care center</th>
<th></th>
<th>Family day care home</th>
<th></th>
<th>Baby sitter</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Child is one-year old or younger</td>
<td>-.046 .191</td>
<td>.303 .194</td>
<td>.167 .324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One parent’s nationality is not Swiss</td>
<td>.564** .216</td>
<td>.285 .233</td>
<td>-.197 .373</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household living in rural area</td>
<td>-.305 .173</td>
<td>-.024 .178</td>
<td>-.381 .313</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French/Italian speaking region</td>
<td>.341* .175</td>
<td>.399* .180</td>
<td>-.297 .317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) younger than 5</td>
<td>.073 .184</td>
<td>.199 .188</td>
<td>.585* .299</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) of age 5-12</td>
<td>-.554** .204</td>
<td>-.198 .202</td>
<td>-.057 .328</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) of age 13-18</td>
<td>-.573 .375</td>
<td>-1.385** .521</td>
<td>-.247 .567</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother’s age</td>
<td>.071** .023</td>
<td>.049* .023</td>
<td>.070 .042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother has a university degree</td>
<td>.520* .259</td>
<td>.295 .257</td>
<td>-.132 .370</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother works 50% or more</td>
<td>.092 .222</td>
<td>-.145 .223</td>
<td>.154 .342</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monthly income in 1000 CHF</td>
<td>.163** .060</td>
<td>.083 .060</td>
<td>.222* .094</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent is the father</td>
<td>-.729* .310</td>
<td>-.757* .321</td>
<td>.143 .450</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father has not a normal daily job</td>
<td>-.228 .219</td>
<td>.011 .234</td>
<td>-.404 .386</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main child-care by parents</td>
<td>-1.36** .255</td>
<td>-1.48** .295</td>
<td>-1.53** .425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main child-care by relatives/friends</td>
<td>-1.37** .268</td>
<td>-1.12** .300</td>
<td>-1.33** .463</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price (CC)</td>
<td>-.052** .007</td>
<td>.016* .008</td>
<td>-.003 .013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price (FH)</td>
<td>.013 .007</td>
<td>-.049** .007</td>
<td>.001 .012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price (BS)</td>
<td>.000 .004</td>
<td>.002 .004</td>
<td>-.035** .010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of children per person (CC)</td>
<td>-.158** .036</td>
<td>-.007 .034</td>
<td>.034 .060</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of children per person (FH)</td>
<td>-.023 .040</td>
<td>-.101* .046</td>
<td>.072 .093</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from CC (in 5 minutes)</td>
<td>-.313** .037</td>
<td>.072* .034</td>
<td>-.023 .077</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from FH (in 5 minutes)</td>
<td>.092** .035</td>
<td>-.364** .038</td>
<td>.026 .071</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FH is open at least 1 week-end-day</td>
<td>-.179 .094</td>
<td>.253** .099</td>
<td>-.370 .206</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC requires 1-month-ahead scheduling</td>
<td>-.154 .096</td>
<td>-.049 .099</td>
<td>-.046 .195</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FH requires 1-month-ahead scheduling</td>
<td>.136 .097</td>
<td>-.190 .107</td>
<td>-.007 .189</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>.152 .896</td>
<td>-.041 .902</td>
<td>-2.541 1.931</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

597 households, 2972 observations, Pseudo $R^2 = 0.141$, logL = -3036.4, Correct prediction: 54.8%
Alternatives: Child care center (CC); Family day care home (FH); Baby-sitter (BS); Private (comparison group)
* significant at .05; ** significant at .01
Table 5: Marginal effects and elasticities (at the sample mean) of significant variables (at p<.05)

<table>
<thead>
<tr>
<th></th>
<th>Child care center</th>
<th></th>
<th>Family day care home</th>
<th></th>
<th>Baby sitter</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Marginal effect</td>
<td>Elasticity</td>
<td>Marginal effect</td>
<td>Elasticity</td>
<td>Marginal effect</td>
<td>Elasticity</td>
</tr>
<tr>
<td>One parent’s nationality is not Swiss</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French/Italian speaking region</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) younger than 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) of age 5-12</td>
<td></td>
<td>-0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional child(ren) of age 13-18</td>
<td></td>
<td></td>
<td>-0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother’s age</td>
<td>0.01</td>
<td>1.32</td>
<td>0.004</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother has a university degree</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monthly income in 1000 CHF</td>
<td>0.02</td>
<td>0.56</td>
<td>0.01</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent is the father</td>
<td></td>
<td>-0.10</td>
<td>-0.08</td>
<td></td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td>Main child-care by parents</td>
<td>-0.16</td>
<td>-0.15</td>
<td>-0.03</td>
<td></td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>Main child-care by relatives/friends</td>
<td>-0.18</td>
<td>-0.10</td>
<td>-0.02</td>
<td></td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>Price (CC)</td>
<td>-0.01</td>
<td>-1.20</td>
<td>0.01</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price (FH)</td>
<td></td>
<td>-0.01</td>
<td>-0.98</td>
<td></td>
<td>-0.01</td>
<td>-2.58</td>
</tr>
<tr>
<td>Price (BS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.001</td>
<td>-2.58</td>
</tr>
<tr>
<td>Number of children per person (CC)</td>
<td>-0.03</td>
<td>-0.59</td>
<td></td>
<td></td>
<td>-0.02</td>
<td>-0.34</td>
</tr>
<tr>
<td>Number of children per person (FH)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.07</td>
<td>-0.94</td>
</tr>
<tr>
<td>Distance from CC (in 5 minutes)</td>
<td>-0.06</td>
<td>-0.74</td>
<td>0.03</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from FH (in 5 min.)</td>
<td>0.04</td>
<td>0.44</td>
<td>-0.07</td>
<td>-0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FH is open at least 1 week-end-day</td>
<td></td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Child care center (CC); family day care home (FH); baby-sitter (BS)
The results also suggest that with a given education and income, the parents’ working status does not have any significant direct effect on their choices. The working status represents the actual employment situation that can be considered as exogenous to the hypothetical choice of the child care mode. In 9.5% of the households in the sample the father has responded to the survey (see table 2). These cases generally correspond to the households in which the responsibility of child care is mainly with the father rather than the mother. The results suggest that these households are less likely to choose an external option (by about 10 and 8 percentage points respectively for child care center and family home options).

Our data on the couples’ working status show that in these households, it is relatively likely that the father does not work and the mother has a full-time job. This suggests that the father-respondent dummy might capture some of the effects of the couple’s working status. However, our preliminary regressions indicate that if this dummy is replaced with the father’s working status variables such as percentage of working hours or full-time job dummies, their effects are not significant. Therefore this dummy appears to capture some other household characteristics as well. In fact, the rather “unconventional” feature of these households, as the actual outcome of the households’ bargaining process about the organization of child care, could be linked to some unobserved characteristics of those families which also affect the choice for extra-familiar care.

The last two household characteristics in the model measure the access to a private care. As expected, households that have access to parental care (the main actual child care provided by one or both parents) are about 15% less likely to choose a child care center or a family day care home. Similarly, the households with access to day care provided by relatives or friends are less likely to choose a non-private care. However, the effect is much higher with respect to the child care center, suggesting that families with access to a private child-care are relatively more willing to substitute a non-parental care for a family day care home rather than a care
center. These families are also less likely to choose the baby-sitter option. It should be noted that given the rate of positive response to each outcome the effect of these two variables is quite substantial in all three outcomes.

Our additional regressions indicate if the nine choice attributes are excluded from the model, the pseudo R-square’s value falls to .060, suggesting that the choice attributes explain a large fraction of the variation in choice probabilities. The most important attribute is price. As seen in table 4, in the case of external care types (CC and FH) the coefficient of own price is more or less similar. In fact statistical tests show that these coefficients are not significantly different from each other. This also applies to the cross price effects, namely the effect of CC price on FH choice is the same as that of FH price on CC choice. This result suggests that regarding price changes, households have a similar response across these two options. The results (table 5) show that the own price elasticity of both external care types is about -1.0 to -1.2. This suggests that a price increase of 10% will reduce the demand by about 10 to 12 percent. The cross price elasticities are significantly lower. According to the estimations, the demand for child care centers and family day care homes respectively increase by about 5% to 8% if the price of the alternative external care increases by 10%.

The baby-sitter option’s own price elasticity is estimated at about 2.6, suggesting a very high sensitivity of demand to prices. The results also indicate that there is no significant price effect between the baby-sitter option and the two external services. This is also valid for all other attributes. Namely, none of the changes in attributes of any of the external care alternatives has any significant effect on the demand of baby-sitter care. This can be partly due to the very small positive response to this option.

The ratio of children per care-person does not have a significant ‘cross’ effect, that is changing this factor in one alternative does not change the demand for the other alternative. Here again the own elasticities are slightly but not significantly different across alternatives. The
results suggest that the probabilities decrease with less staff, the respective elasticities being 0.59 and 0.34 for the alternatives child care center and family day care home. The distance to the child care provider significantly affects the probabilities. The own distance elasticity is respectively –0.74 and –0.94 for child care center and family day care home choices and the cross distance elasticities are about 0.4 in both cases. The distance elasticities are not significantly different across the two alternatives.

If families day care homes function during the weekend, their demand will increase by about 6% on average, but it does not affect the child care center demand. The scheduling flexibility has no significant effect. However, if the family day care homes require one-month-ahead scheduling, their demand may slightly fall, while the demand for child care centers can rise by almost the same amount. This may suggest that families are willing to plan their child-care demand with a center but not with a family home.

6. Summary and conclusions

Although extra-familial child care services have gained importance in Switzerland, the provision of day care facilities has not been fully developed. Most families cannot afford or do not have access to private non subsidized day care facilities that are generally limited to large city centers.

In order to improve the provision of child care centers and family day care homes, the government has initiated an incentive program for start-up financing of child care services. An effective promotion of child care facilities by policy makers requires detailed information about the conditions under which parents are willing to use such services.

Due to the current limitations of the child care market in Switzerland, we have used a stated preferences approach in order to elicit the preferences of households regarding the type of provided care. By applying a choice experiment it was possible to identify the importance of
the characteristics of child care facilities. The choice experiment considered four modes of
day care: the child care center, the day care family home, the baby-sitter and the private solu-
tion. The characteristics considered in the choice experiment were the price, the distance from
home, the number of children per staff member, the opening hours and the flexibility to adapt
the time of care to different needs.

The sample consists of 597 families living in Switzerland covering households living in all
parts of Switzerland. The comparison of the actual choices and the hypothetical choices
shows a considerable increase in demand for child care centers and family day care homes.
This can be partly explained by the current lack of supply. The results suggest that the exist-
ing subsidy programs cannot fill the observed gap between the actual market conditions and
the potential demand of affordable services. Therefore, to the extent that the families’ prefer-
ences can be used as a guide, the governments could put greater efforts in promoting external
child care services. Actually, the promotion of this kind of public services can have positive
social and economic impacts (on the fertility rate, labour supply of women, and on the inte-
gration possibilities of disadvantaged children).

The results of a multinomial logit regression model indicate that many of the household char-
acteristics have a significant effect on the choice probabilities. In particular, mother’s age and
education have positive effects on the incidence of external care options especially child-care
centers. This result suggests that mothers with higher education and experience, who have
relatively good working possibilities, are more likely to benefit from external child care ser-
dvices. Immigrant families are relatively more likely to choose child-care centers. This group
could therefore be an appropriate target group for policies that aim at social integration.

The estimation results also suggest that the factors related to the parents’ current work status,
do not have a significant effect on families’ hypothetical choice of child care. Suggesting that
both working and non-working mothers are equally interested in child care services, this re-

sult could be interpreted as an indication that parents might change their work status if their child care possibilities change. This may be true in particular for all the mothers who in the current situation have to renounce to an employment because of the lack of affordable child care services.

Another interesting result is that except for the baby-sitter option, the household’s income has a relatively small effect in their choice of child-care. This can be explained by the fact that the experiment was based on prices set more or less proportional to the family’s income, as it is the case for the Switzerland’s dominantly subsidized child care services. The results also point to a considerable potential demand for child-care services among those households that use private solutions such as parents or relatives’ aid. Also the family’s cultural background appears to play a role in their choice: Families living in the French or Italian-speaking parts of Switzerland are more likely to choose external day care, whereas those living in German-speaking regions tend to prefer the private solution.

In addition to the socio-demographic variables, the choice attributes explain a large fraction of the variation in choice probabilities. The most important attribute is price with an elasticity of about -1 for child care center and family day care home and -2.6 for the baby-sitter option. The distance to the day care provider and the ratio of children per care-person have also a significant and negative effect on the demand. The cross elasticities between the two external care alternatives indicate that families to some extent, substitute the child-care center and family-home options.

In general, the results suggest an important potential demand for non-parental day care facilities in Switzerland. The demand for these institutions depends mainly on their characteristics with the affordability (price) and access (distance) being the most important factors. That is, the demand is sensitive to explicit costs (through prices) but also to travel costs (including time costs).
While the demand for external day care varies across different groups of population, we can observe that overall, there are no strong preferences for one specific child care mode. As a consequence, an effective promotion policy should promote diversified and flexible child care alternatives. In fact the optimal form of day care with maximum accessibility and minimum costs could vary across different communities. For instance, family day care homes can be considered as a more economical alternative to child care centers, in particular in rural regions where the number of children is not sufficient to justify the creation of a center. Given the relatively high price elasticity, price subsidies could be considered as an effective way for promoting extra-familial child care.

The results of this study provide an empirical rationale for the promotion policies such as the incentive programs financed by the Federal government for starting a child-care business. However, it is interesting to note that in spite of the important potential demand for such facilities, not all the available financial resources for the start-ups have been spent. One reason for this surprisingly low demand could lie in the long-term risks that a new facility might fail to reach the economic break-even point. In fact, the actual incentive programs mainly focus on the initial investment costs, which are only a component of overall costs of operating a child care business, mainly consisting of the annual costs of personnel. Therefore, in addition to a start-up financing of child care facilities, there is a need for a medium and long term commitment usually from local governments in order to allow a cost covering procedure for these providers. This commitment would be in line with the goal of improving the compatibility of the mother’s family and career in Switzerland.

Finally, the significant variation of demand across different population groups suggests that promotion policies would be more effective if they could target selected groups of families. This can be considered as a contradiction to an equitable provision of day care services. In order to solve this equity problem, governments could consider discount vouchers for families
rather than subsidizing the providers. This allows the families to choose their best alternative
day care while letting the market mechanism induce a sufficient provision adapted to the de-
mand variations across different regions.

7. References

ANDERSON PM., LEVINE P.B., 2000, Child Care and Mothers’ Employment Decisions, in

BATEMAN, I, … (ET AL), 2002, Economic valuation with stated preference techniques: a
manual, Edward Elgar Publishing Limited, Cheltenham, UK.


BIROL E., KAROUSAKIS K., KOUNDOURI P., 2005, ‘Using a choice experiment to esti-
mate the non-use values of wetlands: the case of Cheimaditida wetland in Greece’, Wa-
ter supply, 5(6): 125-133.

CHAMP P.A., BOYLE K.J., BROWN T.C., 2003, A Primer on Nonmarket Valuation, The
Economics of Non-Market Goods and Resources, Ed. Bateman I.J., Kluwe Academic

CHAPLIN D.D., ROBINS P.K., HOFFERTH S.L., WISSOKER D.A., FRONSTIN P., 1996,
The Price Elasticity of Child Care Demand: A Sensitivity Analysis. Unpublished manu-
script.

Queue for Childcare? Unpublished paper, Institute for the Study of Social Change, Uni-
versity College Dublin.

CLEVELAND G., GUNDERSON M., HYATT D., 1996, ‘Child Care Costs and the Em-
ployment Decision of Women: Canadian Evidence’, Canadian Journal of Economics,

CONNELLY R., KIMMEL J., 2003, ‘Marital Status and Full-Time/Part-Time Work Status in
Child Care Choices’, Applied Economics, 35(7), 761-77.

DEL BOCA D., LOCATELLI M., VURI D., 2004, Child Care Choices by Italian House-
holds, Discussion Paper Series, Institute for the Study of Labor, IZA DP No. 983.

GYLDMARK M., MORRISON G.C., 2001, ‘Demand for health care in Denmark: results of a
national sample survey using contingent valuation’, Social Science & Medicine, 53,
1023–1036.

New Jersey.

perior Alternative for Environmental Valuation’, Journal of Economic Surveys, Vol. 15,
No. 3, 435–462.


Willingness to Pay for Energy-Saving Measures in Residential Buildings*

Silvia BANFI♣, Mehdi FARSI♣, Massimo FILIPPINI♠♠, Martin JAKOB♦♦

♣ Department of Management Technology and Economics, ETH Zurich, Switzerland
♠ Department of Economics, University of Lugano, Lugano, Switzerland

Abstract

This paper uses a choice experiment to evaluate the consumers' willingness to pay for energy-saving measures in Switzerland's residential buildings. These measures include air renewal (ventilation) systems and insulation of windows and facades. Two groups of respondents consisting respectively of 163 apartment tenants and 142 house owners were asked to choose between their housing status quo and each one of the several hypothetical situations with different attributes and prices. The estimation method is based on a fixed-effects logit model. The results suggest that the benefits of the energy-saving attributes are significantly valued by the consumers. These benefits include both individual energy savings and environmental benefits as well as comfort benefits namely, thermal comfort, air quality and noise protection.

JEL classification: Q40, R21, C25, C93

Keywords: energy efficiency, energy saving, choice experiment, conjoint analysis, discrete choice, housing, fixed effect logit

* This paper is based on a research project sponsored by the Swiss Federal Office of Energy, the Federal Office of Housing, the House Owners Association of Switzerland, Canton of Basle and the electric utility of Zurich (EWZ), which is gratefully acknowledged. The authors would also like to thank Martin Baur and Walter Ott from Econcept AG, Zurich for their invaluable support and insightful discussions throughout this study.
1 Introduction

As is the case in most industrialized countries in temperate zones, residential buildings in Switzerland incur an important share of the end use energy consumption. Thus, improvements of energy efficiency in the building sector could have an important impact on the country’s total energy consumption and a considerable contribution in attaining the CO₂-emissions objectives for a sustainable development. The overall energy efficiency of a building is identified mainly by the insulation characteristics of the building envelope and the presence of an air renewal system.¹ Provided with an energy-efficient implementation, these measures yield two kinds of benefits: First, they reduce the energy consumption of the buildings hence costs, and secondly they generate comfort benefits namely, improved indoor air quality, thermal comfort and enhanced protection against external noise.

The potential for energy efficiency improvements are not being sufficiently exploited. With a relatively long cycle of energy-relevant renovations in buildings (usually about 20 to 40 years), the Swiss building sector has still a very low usage of energy saving measures. Every year only one to two percent of the existing building envelopes undergo maintenance or renovation. In 30 to 50 percent of these cases the renovation measures include insulation with a reduction of the energy consumption by 50% to 70% and only a very small fraction opt for enhanced energy-efficiency measures that exploit the energy saving potential completely (see Jakob and Jochem, 2003). Houses with the latter measures satisfy the conditions set by Minergie label² reducing energy consumption by 70% to 85% for old buildings (constructed prior to the 1970s) or by 50% for today’s new buildings.

¹ Air renewal or ventilation systems have a reduced air exchange and provide the indoor spaces with fresh and filtered air (pre-heated by a heat-exchanger) without great heat losses through windows or traditional aeration systems. Not to confuse these systems, also known as “housing ventilation” or “comfort ventilation”, with conventional air conditioning used for cooling or moisturizing.

² Minergie is a quality label that combines high comfort of living and low energy consumption with a limited cost surplus of at most 10% of the construction price. Controlled air exchange requirement, is mostly met with a ventilation system. More information is available at www.minergie.com.
The Swiss and cantonal governments support the renovation or new investments in houses satisfying the Minergie requirements through subsidies and/or reduced interest rates. Although energy-efficient buildings are being promoted, a relatively small number of houses are constructed (5 to 10 percent of new single family houses and less than 5% of new apartment buildings) and hardly any are renovated according to Minergie guidelines.

In a recent study, Ott et al (2005) have identified legal and social factors as well as market structural barriers and lack of consciousness as the possible explanations of low usage of energy-saving systems for the case of the Swiss residential building sector. Moreover, as shown by Jakob (2006) depending on the adopted assumptions and especially for ventilation systems, the discounted value of long-term savings in energy costs could be insufficient to justify such investments.

In order to identify effective policy measures to induce more investment in buildings’ energy efficiency, it is important to have detailed information on the factors that influence the homeowners’ investment decision and on their willingness to pay for the resulting improvements. Similarly in rental buildings it is important to know how consumers value apartments in energy-efficient buildings. However, there are only a few studies that addressed the consumer’s valuation of energy-saving measures in buildings. One of the first studies is Cameron (1985) that analyzed the demand for energy-efficiency retrofits such as insulation and storm windows using the actual data collected by a national survey on energy consumption. Using a nested logit model that study shows a considerable sensitivity of demand to changes in investment costs, energy prices and income. In more recent literature, conjoint analysis was used in order to elicit the choice behavior of households for energy-saving measures. For instance, Poortinga et al. (2003) have focused on the characteristics of 23 energy-savings measures including insulation and energy-efficient heating systems in the Netherlands. The conjoint analysis was judged to be a useful method to examine the acceptability of these measures and identify the characteristics influencing the choices. A choice experiment was also carried out in Canada aiming at understanding the preferences of residential consumers when making investment decisions regarding heating system or a renovation that impacts the efficiency of home energy consumption (Sadler 2003). The renovation choice was estimated using a binomial logit model and
the heating system choice using a multinomial logit model. The results of that study suggest a high preference for the energy efficient renovations and highlight the effect of comfort in addition to the capital costs, the annual heating expenses and the subsidy regime.

This paper adopts a choice-experiment approach to analyze the willingness to pay (WTP) for energy-saving measures in residential buildings. The results provide the first WTP estimates based on choice experiments in the context of the Swiss housing sector. The analysis includes both renovation cases and new buildings. The decisions are related to purchasing single family houses as well as renting apartments. The estimation methodology is based on a binomial logit model with individual fixed effects. The results suggest that energy-saving measures are significantly valued by the consumers, which in some cases can counter the implementation and operation costs.

The rest of the paper is organized as follows: Section 2 describes the experiment design; Section 3 presents the theoretical framework and the econometric methodology. A description of the data and the regression sample is provided in section 4. The estimation results are presented in section 5. A summary of the main results and the conclusions are given in section 6.

2 Experiment design

The data needed for the econometric estimation of the choice behavior can basically be collected with two different methods: the revealed and the stated preference method. The first method is based on the observation of the actual choice decisions of households from a set of alternatives that are known to the econometrician whereas the second method is based on information extracted from interviews or choice experiments. Verhoef and Franses (2002) or Louviere et al. (2000) provide overviews of the advantages and drawbacks of the two methods.

The aim of this study is to estimate the marginal willingness to pay (WTP) for different energy-saving characteristics. In principle, both revealed and stated preference methods could be used for this purpose. However, the small share of buildings with enhanced energy efficiency standards makes the
use of a revealed preference method difficult. Moreover, it is generally difficult to obtain data on the available choice set from which the alternative has been chosen. For the above reasons we use a stated preference method with choice experiment, initially developed by Louviere and Hensher (1982). This approach has been used in other energy-related topics, for example in Bergmann et al. (2006).

Two samples of households respectively consisting of residents of single family houses and rental apartments have been presented with several choice sets and asked to choose the alternative they prefer the most. In our case, respondents were asked to choose between their actual situation and a hypothetical housing with different energy efficiency attributes and a different price, with all other characteristics remaining the same. The price is defined as the purchase price for houses and the monthly rent for apartments. The following attributes are included in the experiment: windows with different energy efficiency standards; façade with different levels of insulation and esthetics; presence of a ventilation system; and price. These attributes and the related categories are listed in Table 1.

The respondents were asked to imagine that their actual housing situation would be improved (downgraded) in terms of the mentioned attributes, with all other characteristics such as number and size of rooms, location etc. being constant. The respondents’ actual housing situation was chosen as a reference to reduce the hypothetical character of the survey (as compared to two hypothetical situations to choose from). The respondents already living in housing situations with a high energy efficiency standard were asked to imagine a decline in one or several of these features. The price levels were related to the actual residence of the respondents and were chosen within a reasonable range. Each respondent was asked to do several choice tasks.

The valuation of different housing attributes can be estimated by applying the hedonic pricing approach to market data.

To make the choice tasks as realistic as possible, the set of categories of the hypothetical housing situations was adapted to the present situation of respondents. For respondents living in new buildings only category 1 and 2 of both window and façade were included in the choice set.
Each choice task consisted of reading a card listing the characteristics of the actual situation and those of one alternative and choosing the one of the two that was preferred. The respondents were provided with descriptive information about the attributes, in particular the relatively new and not widely installed housing ventilation system. This description included information about the characteristics of the attributes and about their positive impacts on the energy efficiency of the building and the comfort benefits such as thermal comfort, air quality and noise protection (see Ott et al. 2006 for more details). The respondents were also informed of the energy cost-savings as well as the entailed environmental improvements. However, we have not provided quantitative information about the extent of these benefits particularly on the potential cost savings at the individual level. In fact, in most real cases when buying or renting a residence, individuals do not have such quantitative information. Moreover, these benefits particularly the savings in energy costs vary across offered alternatives and strongly depend on the actual situation, hence many unobserved factors which were difficult to assess. It is assumed that the respondents assess the trade-off between prices and the overall benefits from different housing attributes. Thus, the willingness to pay estimates include comfort benefits and cost-savings as well as the respondents’ potential valuation for environmental benefits.

To reflect the real-world choice situations (and to prevent strategic behavior) each of hypothetical alternative consisted of an upgrade in some attributes and/or a downgrade in some other attributes. This design was chosen to enrich the structure of the sample and was based on the assumption the respondents could answer differently depending on their personal experience about different energy efficiency attributes. In particular, we intentionally included respondents living in situations equipped with ventilation.5

5 Further details of the experiment are documented in Ott, Baur and Jakob (2006).
3 Model specification

With reference to the utility theory, the paper models the choice of respondents (apartment tenants, house buyers) for energy relevant characteristics of apartments and houses respectively. The underlying assumption is that households evaluate the characteristics of different housing alternatives and then choose the one which leads to the highest utility.\(^6\) We assume that the utility of living in energy efficient apartments or houses is a function of the price, the housing’s energy efficiency characteristics (for instance the characteristics of windows and façade and the presence of a ventilation system), the building location, household characteristics, and a random component that captures the influence of unobserved factors. The household characteristics can include income, education, environmental consciousness, as well as site-specific characteristics of the household’s actual residence. Indeed, according to the random utility theory, the utility of goods or services is considered to depend on observable (deterministic) components, including a vector of attributes \((X)\) and individual characteristics \((Z)\), and a stochastic element \(e\) (cf. Louviere et al. 2000). Thus, the utility function of a bundle of characteristics \(i\) for individual \(q\) at choice task \(j\) can be represented as:

\[
U_{qij} = V(X_{qij}, Z_q) + e_{qij}
\]  

(1)

where \(V\) is the deterministic part and \(e_{qij}\) the stochastic element. The deterministic variables that will be used in an empirical model are the housing attributes \((X_{qij})\) and the respondent’s characteristics \((Z_q)\). Assuming an extreme value distribution for the stochastic term \(e_{qij}\) in model (1), the probability of choosing alternative \(i\) out of a set of available alternatives \(A = \{1, 2, ..., K\}\) can be written in a logistic form as:

\[
P_{qij} = \frac{\exp(V_{qij})}{\sum_{k=1}^{K} \exp(V_{qkj})}
\]  

(2)

Expression (2) is the basic equation of a multinomial logit (cf. Greene, 2000 and Thomas, 2000). Utility function \(V\) is generally assumed to be linear in parameters. In our case, the number of

\(^6\) In other words it is assumed that households maximize their utility function of hedonic commodities that they produce from the housing services and goods (Thompson, 2002).
alternatives in each choice task is limited to two possibilities. Thus, the choice set for a given choice task \( j \) can be written as \( A = \{0, j\} \) with 0 indicating the status quo and \( j \) representing the offered alternative. The random utilities of the resulting binary logit model can be written as:

\[
U_{qj} = \beta X_{qj} + \alpha Z_q + e_{qj} \quad ; \quad U_{q0} = 0
\]  

(3)

where \( Z_q \) represent the household characteristics that do not vary across choice tasks, and \( X_{qj} \) is the characteristics of the alternative situation of choice task \( j \) for individual \( q \). \( \alpha \) and \( \beta \) are the vectors of model parameters. In a multinomial logit framework, the parameters associated with one of the outcomes are normalized to zero namely, \( U_{q0} = 0 \). Therefore, \( U_{qj} \) is the random utility of choosing the alternative situation over the status quo.

If all the relevant respondent’s characteristics (\( Z_q \)) are observed, the model given in equation (3) is a simple binomial logit. In general however, \( Z_q \) can include a host of parameters, many of which are not observed. In this case, this term can be considered as an individual fixed effect. The resulting model is a fixed effect binary logit model proposed by Chamberlain (1980) and can be written as:

\[
U_{qj} = \beta X_{qj} + u_q + e_{qj} \quad \text{with} \quad u_q = \alpha Z_q
\]  

(4)

It should be noted that because of the presence of fixed effects in the model, vector \( X_{qj} \) can be equivalently replaced by the \( X_{qj} - X_{q0} \), which measures the difference between the characteristics of the hypothetical alternative with the status quo. This implies that \( U_{qj} \) measures the net gained value through moving from actual situation (status quo) to a hypothetical status offered in choice task \( j \).

Given that the hypothetical alternatives may equally involve a better or worse situation regarding comfort, the individual specific term \( u_q \) can be interpreted as the (dis)utility of respondent \( q \) from changing their status quo.

Assuming a logistic distribution for the error term, the above model can be estimated by maximization of the conditional likelihood given the fixed effects (\( u_q \)).\(^7\) Chamberlain shows that for a consistent

\(^7\) See Hsiao (1986) and Greene (2003) for more details about the fixed-effects logit model and Ferrer-i-Carbonell (2004) for an application.
estimation, incidental parameters $u_q$ should be replaced by a minimum sufficient statistic namely, the
number of positive responses for a given individual. If we denote the individual $q$’s response for $J$
choice tasks by the sequence $(y_{q1}, y_{q2}, ..., y_{qJ})$, where $y_{qj} = 1$ if offer $j$ is chosen, and $y_{qj} = 0$ if
offer $j$ is not chosen, then the number of positive responses (accepted offers) for individual $q$ is
obtained by the sum $s_q = \sum_{j=1}^{J} y_{qj}$. The conditional probability can therefore be written as:

$$
\Pr\left( y_{q1}, y_{q2}, ..., y_{qJ} \mid \theta_q \right) = \frac{\exp \left( \sum_{j=1}^{J} y_{qj} X_{qj} \beta \right)}{\sum_{d_{q1},d_{q2},...} \exp \left( \sum_{j=1}^{J} d_{qj} X_{qj} \beta \right)}
$$

(5)

where $\Omega$ is the set of all the sequences $(d_{q1}, d_{q2}, ..., d_{qJ})$ in which the number of positive responses
is equal to that of the chosen sequence namely, $\sum_{j=1}^{J} d_{qj} = \sum_{j=1}^{J} y_{qj} = s_q$. Hence, the numerator represents
the odds of choosing the sequence $(y_{q1}, y_{q2}, ..., y_{qJ})$ and the denominator indicates the sum of the
probabilities of all possible outcomes that entail the same number of accepted offers.

The fixed effect logit model is estimated using the maximum likelihood estimation method. Once the
model parameters are estimated, the marginal rate of substitution between different attributes can be
calculated. If one of the attributes is a numéraire or a monetary variable like price ($p$) the marginal
willingness to pay for attribute $x$ can be derived as:

$$
WTP = \frac{\frac{\partial V}{\partial x}}{\frac{\partial V}{\partial p}}
$$

which is equivalent to the ratio of the corresponding coefficients in equation (3).

4 Data description

The data used in this paper were collected during Summer 2003 by telephone interviews in five
cantons covering a major part of German-speaking Switzerland. The experiments have been
performed on two separate samples for apartment buildings and single-family houses respectively. The first sample consists of tenants of rental flats whereas the second sample includes home-owners. Both samples have been selected from the households who have recently moved, thus have faced a housing choice decision within a few months before the experiment. The samples were stratified with the purpose of including a sufficient share of new and existing buildings, of standard and energy-efficient ones, of buildings with and without ventilation receptively. Both samples cover an important share of the German speaking part of Switzerland. The data sources from which respondents were randomly chosen were different for each stratum of the sample. These data sources include the list of labeled energy efficient houses published by the Minergie association (see footnote 2), a data base of a supplier of internet housing ads, and a database of another survey on buildings (Jakob and Jochem, 2003). Respondents’ names and phone numbers were matched to the buildings’ address using the public phone directory.

The telephone interviews have been conducted in two stages. In the first stage the respondents were recruited and basic information were collected to match the respondents to the different sample strata and to obtain information about their actual housing situation. The respondents were then provided with written information and the choice tasks. In the second stage, the choices and additional socio-economic information were collected by phone.

The first stage included 397 interviews for the rented flats and 402 interviews for single family houses, corresponding to a response rate of 36% and 41% respectively. The response rate of the second stage was 66% for the rented flats and 63% for the single family houses, resulting in overall response rates of 24% and 26% respectively. The response rate of the second stage is quite high and non-responses are mainly due to unavailability of some of the respondents. Thus the selection bias at this stage is relatively low. However, there could be a self-selection bias at the first stage of the survey. Persons interested in the subject (of energy efficiency and housing comfort) could be more likely to participate

---

8 In the original study (Ott, Baur and Jakob 2006) the buildings constructed after 1995 and those with energy-efficiency labels have been distinguished from other buildings.
in the choice experiment. With the available data we cannot identify the extent of potential selection biases due to unobserved differences between the participants and the Swiss population. Compared to the average values of the Swiss population the studied samples show a slight over-representation of high-income and a considerable over-representation of educated individuals (Ott, Bauer and Jakob, 2006). Assuming that relatively educated individuals have a higher than average valuation of comfort and energy-efficiency, such sample selection biases might result in an overestimation of WTP.

The resulting samples obtained from the survey include 264 tenants of apartment buildings and 253 single family house owners with a total of 3861 and 3458 observations (choice tasks) respectively. After excluding the choice tasks with dominated alternatives and also the respondents that have consistently preferred their status quo over all the offered alternatives⁹, the final regression samples consist of 163 tenants with 1928 observations and 142 house purchasers with 1685 observations.

The considerable rate of respondents always preferring their actual situation (101 out of 264 and 111 out of 253 respondents) may suggest that focusing on the remaining sample may create selection bias in the estimations. However, it should be noted that the experiment design is such that the alternative state does not necessarily have always higher attributes than the actual state. Therefore, the respondents who have never accepted any offer might rather have a relatively high disutility of change, or simply might have not examined all the offers. Therefore, to the extent that such disutilities are not correlated with the WTP, it is reasonable to assume that the WTP estimated from the regression sample are representative of the entire sample.

A descriptive summary of the sample used in the analysis is given in Table 2. The upper panel of the table lists the descriptive statistics of the respondents and the characteristics of their actual residence while the lower panel gives the attributes of the hypothetical alternatives offered in the experiment.

As seen in this table the share of apartments with installed housing ventilation systems is about 14 percent of the sample and that of single family houses is about 9 percent. These shares are slightly

⁹ The respondents that have not shown any variation in their choices cannot be included in a fixed effects logit model.
lower than the corresponding ones of the entire samples (about 20 percent of 264 tenants and 17 percent of 253 single family houses). This difference suggests that the respondents living with a ventilation system are relatively less likely to give up their present situation regardless of the offered price discounts.

Regarding the energy efficiency attributes of the actual situation the sample can be described as follows: the most frequent type of windows is “Standard window” (67% of apartments, 80% of single-family houses) including coated glazing and sealing rubber. Only 13% of apartments and 9% of single-family houses (SFH) have enhanced windows (including coated triple glazing). 17% of the apartments and 9% of the SFHs have “old windows” (i.e. windows that were renovated before 1995 or not at all) including non coated double glazing and no sealing. A minor fraction of the buildings has still very old windows with only single glazing.

The two most frequent façade qualities in the samples are the standard insulation and the “old façade” (neither painted nor insulated the last few years) covering about one third each of them. More specifically, the shares of standard insulation are 34% (apartments) and 32% (SFH) and the “old façade” ones (nor painted or insulated the last few years) are 36% (apartments) and 31% (SFH).

In the final apartment sample, the number of valid observations (number of answered choice tasks per person) varies between 2 and 17 with an average of about 12 and a standard deviation of about 3.4. The number of accepted offers per person varies between 1 and 14 with an average of 3.4 accepted offers. The number of cards per person in the SFH sample varies between 7 and 18 with an average of about 14, from which 2.7 offers were accepted in average. The rental prices range between 430 and 4000 CHF/month and the standard deviation is 609 CHF/month. The purchase prices of the SFH range from CHF 100,000 to CHF 1.6 Million, with an average of CHF 659,000.

10 For the econometric estimation, the choice situations with dominated alternatives and undecided choice tasks were excluded from the sample. In all the remaining choice tasks, the price of the hypothetical alternative is higher (lower) than that of the actual situation if and only if the alternative offer provides a strict improvement (decline) in at least one of the attributes while other attributes remain at least (most) the same as in the actual state.
A descriptive summary of the characteristics of the hypothetical offers is given in the lower panel of Table 2. The sample of the choices can be described as a balanced sample in that there is a comparable share of old, standard and enhanced windows in the offered alternatives. This is also valid for the façade quality and the presence of a ventilation system. About 25% of the offers had very old windows. Rental prices of offers vary between 323 and 4600 CHF/month, with an average of 1509 CHF/month. In both samples the average price of offers is about the same as the average price of the actual situation, which is due to similar number of price increases and decreases. Despite the fact that the offers are balanced, only less than one third of the offers were accepted (29% in the apartment sample and 19% in the SFH sample). This result might suggest a significant disutility of change.

The explanatory variables include the price (monthly rent for apartments and purchase price for single family houses)\(^{11}\) and the energy efficiency attributes of the hypothetical offers. These attributes consist of three dummy variables for window attributes and three dummies for the facade characteristics with the standard (insulation) type being chosen as the omitted category in both cases and one dummy for ventilation system (see Table 1). An observation reported by Ott, Baur and Jakob (2006) is that the respondents who already have a given attribute in their households might attach a higher value to that attribute compared to other individuals. In order to control for potentially asymmetric choice behaviour,\(^{12}\) a dummy variable has been constructed to indicate the hypothetical offers with lower-than-actual prices, which entail a decline at least in some of the attributes while others being unchanged. The interaction of this dummy with price is included in the model.

\(^{11}\) Price variable is actually the difference between hypothetical and actual prices for each observation (choice task). Note that thanks to the fixed effects, it would not matter if the price levels of the hypothetical alternatives were used instead.

\(^{12}\) This asymmetric behavior, commonly referred to as the disparity between the willingness to accept (WTA) and the WTP, is usually observed in similar experiments and widely discussed in the literature (Horowitz and McConnell, 2002 and Sayman and Öncüler, 2005). This disparity has been explained by several factors including those related to the survey design and framing effects as well as economic and psychological factors. In our experiment the asymmetry might be exacerbated by the fact that some of the high-level attributes are completely new to the respondents and might be valued less than those already experienced.
Because of the fixed effects included in the model, the household characteristics can only be included through interaction terms. In a preliminary analysis several interaction terms between alternative attributes and household characteristics have been considered. Using several hypotheses we explored if households with different characteristics and socio-economic variables differ with respect to their valuation of energy efficiency attributes. For instance, we tested if households with smoking habits or with pets have a different valuation of ventilation systems and/or people living in noisy locations have a higher valuation of insulated windows. The results suggested that all the interaction terms were statistically insignificant at 10% significance level. We expected that household income might have an important effect. However, due to a relatively high share of missing values (about half of the sample) we could not include any income variable.

Therefore, in order to keep the model as parsimonious as possible and avoid unnecessary complication in the interpretation of the results, we decided to exclude such interaction terms from the model. The only exception is the different valuation of ventilation systems across new and old buildings. Our results suggest that the air renewal systems could be valued more in new buildings constructed after 1995 (less than 10-years-old). An interaction term is included in the model to account for such differences.

5 Results

The estimation results are shown in Table 3. The results regarding house purchasers and tenants show a very similar pattern. The coefficients of the price and of all energy-efficiency attributes have the

\[13 \text{ The details of these analyses are not included in this paper.} \]
expected sign and most of them are significantly different from zero at 5% significance level. Exceptions are the coefficients for enhanced windows and the interaction variable between ventilation system and new buildings for the rented flats. A significant difference in the price effects was found between price increases (price of the hypothetical alternative is higher than the price of the actual apartment) and price decreases.

Using equations (3) and (6) we can calculate the willingness to pay for each attribute, which is the ratio of the attribute’s coefficient and the price coefficient. The WTP results in Table 4 are expressed as a percentage of the reference purchase price for houses, and as a percentage of the reference rental price for flats. The average prices of both new and old buildings are used as reference. In new buildings the willingness to pay for enhanced façade insulation is about 3% whereas the ventilation system is valued at 4% to 12% of the reference price. In relative terms, house buyers and apartment tenants have a similar WTP for the case of new buildings. It is worth noting that the survey was conducted in Summer 2003 which was an extraordinary Summer with high temperature. This might explain the relatively high WTP for ventilation systems. Even though a comfort ventilation system as considered here is not designed for cooling, the respondents might have associated cooling with this system.

14 The WTA could be calculated similarly accounting for the interaction of price and the decreasing-price dummy. The estimation results suggest that the WTA/WTP ratio is 2.1 in the case of rental flats and 2.3 in the case of single family houses. This is consistent with the results reported in the literature (cf. Sayman and Öncüer, 2005). In the paper we focus on the WTP that has more importance from a policy point of view.

15 These prices are 650,000 and 686,000 CHF for new and existing single family houses respectively and 2030 and 1330 CHF/month for flats in new and in existing buildings respectively.
In the existing (not new) buildings we estimate the WTP for energy efficient façades and windows. Regarding the façade there is a WTP for insulation of 6% and 7% for SFH, whereas the estimated WTP for esthetic reasons is low (about 3%) and only for single family houses significant at the 10% level. In existing buildings, the willingness to pay is particularly high for window improvements. Indeed, the WTP for a standard insulation window as compared to an old window is 13% for tenants as well as for house purchasers. Note that today’s standard insulation windows are coated and have sealing rubber whereas old windows do not dispose of these properties. Coated windows have a higher surface temperature and sealing rubber protect from air infiltration and from external noise. Thus, such windows improve thermal comfort and comfort of living which might explain these relatively high WTP.

Comparing the results of windows and façades for old and new buildings shows that the marginal WTP for each further step of energy efficiency is decreasing. This result suggests that the “first” improvement provides a higher utility than that of an additional improvement.

The WTP for ventilation systems in old buildings is below that of new buildings. This could be explained by different preferences of residents living in old and new buildings or by the different reference price level. The respondents who live in new buildings might have a relatively high standard of living, thus higher WTP for comfort. Note that in the case of tenants, the willingness to pay in relative terms, is very similar across old and new buildings. That the willingness to pay for ventilation is different between persons living in new and old buildings could be interpreted as an income effect, since income of people living in new buildings is slightly higher than those living in not new buildings. Finally, it should be noted that the WTP values include both the willingness to pay for improved comfort, for increased energy efficiency i.e. reduced energy costs and eventually for environmental improvements.

The willingness to pay for energy-efficiency attributes can be compared with the capital costs of implementing such attributes. In Jakob et al (2006) some typical capital costs are given for the

((Insert Table 4 about here))
example of a typical flat of hundred square meters and for a typical single family house. For most of the considered attributes the monthly capital costs are significantly lower than the average willingness to pay of the sample as reported in Table 4.

That the willingness to pay exceeds the cost can be interpreted in different ways: On the one hand it could indicate that people actually desire enhanced efficiency but that the housing market has not yet reacted to this demand. On the other hand the values of the estimated willingness to pay could be overestimated.

The estimated values of WTP can be compared with the results obtained from hedonic pricing method (Ott et al, 2006). According to those results in the greater Zurich area the marginal value of Minergie label is about 7.5% of the rental price for new buildings and that of a renovated, insulated façade is about 8%. It is interesting to note that the estimated WTP values in this paper are comparable with these price effects obtained using data from revealed preferences through market prices.
6 Summary and Conclusion

This paper gives some insight into the willingness to pay for improvements in energy efficiency by studying stated choices of two samples of respondents respectively consisting of tenants of rental apartments and owners of single family houses. The considered energy saving measures include air renewal system and different energy efficiency standards of windows and façade. The data used for the econometric estimation were collected with a choice experiment. The respondents were presented with choice sets and asked to choose between their actual housing situation and a hypothetical one with different energy efficiency standards and a different price. The decision to use a stated preference method is supported by the fact that revealed reference data is only scarcely available since the market of energy efficient houses is still small. Further, this method made it possible to compare the willingness to pay of people who have already experienced the additional comfort benefits of energy-saving measures with those who do not have such information.

The econometric analysis of the data has been carried out using a fixed effect logit model. The coefficients of all attributes have the expected sign and most of them are significantly different from zero. The results show a significant willingness to pay (WTP) for energy efficiency attributes of rental apartments and of traded houses. The willingness to pay varies between 3% of the price for an enhanced insulated façade (in comparison to a standard insulation) and 8% to 13% of the price for a ventilation system in new buildings or insulated windows in old buildings (compared to old windows) respectively. Note that the interrelation of the WTP values for different attributes are quite plausible and the results reflect a decreasing marginal utility for increasing energy efficiency.

The WTP values presented in this paper could be an overestimation of the representative values in the Swiss population, due to possible overrepresentation of respondents with high education and/or income and the relatively high participation rate of environmentally conscious individuals. Moreover, an overestimation could result from the hypothetical choice situation, relying on individuals stating their behavioral intentions rather than on observable economic decisions.
The WTP is generally higher than the costs of implementing these attributes. Therefore it would be economically reasonable for owners and housing promoters to invest in energy saving measures. We assume that besides many legal, structural and socio-economic barriers the observed underinvestment is due to lack of information regarding the advantages of the efficiency measures and perhaps lack of methods to quantify these advantages in economic terms. Indeed house owners, architects, tenants and financial institutions have occasionally deplored this lack of economic foundation.

From a policy point of view, the government should be interested in reducing these barriers by supporting the communication and information for decision makers namely consumers, investors and financial institutions. A good example of this kind of promotion is given by advertising campaigns (so called “casa clima”) used by the government of the Italian province Alto Adige, or by information campaigns and subsidies applied to energy efficient buildings in Switzerland, namely Minergie guidelines that combines efficiency and comfort. The authors recommend that the WTP results presented in this paper could be included in these promotion campaigns. In addition to an enhancement in communication, the governments could grant additional financial resources needed by the house owners who want to invest in energy efficiency measures (less than 10% of total investment costs) to overcome financial barriers. Some Swiss financial institutions award credits with lower interest rates for Minergie labeled buildings. It should be considered that government intervention could speed up the cost reduction (learning curve) of measures improving energy efficiency in buildings.

Nonetheless the WTP values presented in this study should be considered with some caution. The results give a first estimate of the magnitude of benefits (willingness to pay) coming from energy-efficiency measures. Given the mostly lower costs of these measures, it may be possible by additional information of house owners, architects and tenants to increase significantly the share of energy efficient building.
7 References


Table 1 Categories of different attributes (in descending order) and price levels considered in the choice experiment

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>1. Enhanced insulation (triple glazing, double coated pane) *)</td>
</tr>
<tr>
<td></td>
<td>2. Standard insulation (coated, rubber)</td>
</tr>
<tr>
<td></td>
<td>3. Medium old (low insulation, not coated) **)</td>
</tr>
<tr>
<td></td>
<td>4. Very Old (single glazing) **)</td>
</tr>
<tr>
<td>Facade</td>
<td>1. Enhanced insulation *)</td>
</tr>
<tr>
<td></td>
<td>2. Standard insulation</td>
</tr>
<tr>
<td></td>
<td>3. No insulation but newly repainted **)</td>
</tr>
<tr>
<td></td>
<td>4. Old (not repainted) **)</td>
</tr>
<tr>
<td>Ventilation</td>
<td>1. With air renewal system</td>
</tr>
<tr>
<td></td>
<td>2. Without air renewal system</td>
</tr>
<tr>
<td>Price</td>
<td>In 5 levels: approximately -100, -50, 0, 50 and 100 CHF per month for rented apartments and -90,000, -45,000, 0 +45,000, +90,000 CHF per house, in addition to the actual price</td>
</tr>
</tbody>
</table>

*) applicable only to new buildings
**) applicable only to existing buildings
Table 2 Descriptive statistics

<table>
<thead>
<tr>
<th>Respondents and characteristics of their actual residence</th>
<th>Tenants</th>
<th></th>
<th></th>
<th></th>
<th>House buyers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of respondents</td>
<td>Sample Mean</td>
<td>Number of respondents</td>
<td>Sample Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of choice tasks per person</td>
<td>163</td>
<td>11.8 (3.4)</td>
<td>142</td>
<td>11.2 (2.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of accepted offers</td>
<td>163</td>
<td>3.40 (2.3)</td>
<td>142</td>
<td>2.68 (2.08)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price of actual situation</td>
<td>163</td>
<td>1550 ($^1$) (609)</td>
<td>142</td>
<td>659 ($^2$) (230)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhanced window in actual situation</td>
<td>163</td>
<td>0.135</td>
<td>142</td>
<td>0.092</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard insulated window in actual situation (*)</td>
<td>163</td>
<td>0.669</td>
<td>142</td>
<td>0.796</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium old window in actual situation</td>
<td>163</td>
<td>0.166</td>
<td>142</td>
<td>0.085</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very old window in actual situation</td>
<td>163</td>
<td>0.030 (2.1)</td>
<td>142</td>
<td>0.028 (2.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhanced facade insulation in actual situation</td>
<td>163</td>
<td>0.337</td>
<td>142</td>
<td>0.317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard facade insulation in actual situation (**)</td>
<td>163</td>
<td>0.117</td>
<td>142</td>
<td>0.162</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repainted facade in actual situation</td>
<td>163</td>
<td>0.356</td>
<td>142</td>
<td>0.317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old facade in actual situation</td>
<td>163</td>
<td>0.141</td>
<td>142</td>
<td>0.085</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old buildings (constructed before 1995)</td>
<td>163</td>
<td>0.650</td>
<td>142</td>
<td>0.549</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothetical offers</th>
<th>Number of offers</th>
<th>Sample Mean</th>
<th>Number of offers</th>
<th>Sample Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted offers (positive outcomes)</td>
<td>1928</td>
<td>0.288</td>
<td>1182</td>
<td>0.270</td>
</tr>
<tr>
<td>Price</td>
<td>1928</td>
<td>1509 ($^1$) (624)</td>
<td>1685</td>
<td>661 ($^2$) (242)</td>
</tr>
<tr>
<td>Enhanced window</td>
<td>1928</td>
<td>0.183</td>
<td>1685</td>
<td>0.188</td>
</tr>
<tr>
<td>Standard window (*)</td>
<td>1928</td>
<td>0.293</td>
<td>1685</td>
<td>0.256</td>
</tr>
<tr>
<td>Medium old window</td>
<td>1928</td>
<td>0.272</td>
<td>1685</td>
<td>0.292</td>
</tr>
<tr>
<td>Very old window</td>
<td>1928</td>
<td>0.252</td>
<td>1685</td>
<td>0.264</td>
</tr>
<tr>
<td>Enhanced facade</td>
<td>1928</td>
<td>0.172</td>
<td>1685</td>
<td>0.160</td>
</tr>
<tr>
<td>Standard facade insulation (**)</td>
<td>1928</td>
<td>0.401</td>
<td>1685</td>
<td>0.398</td>
</tr>
<tr>
<td>Repainted facade</td>
<td>1928</td>
<td>0.217</td>
<td>1685</td>
<td>0.216</td>
</tr>
<tr>
<td>Old facade</td>
<td>1928</td>
<td>0.210</td>
<td>1685</td>
<td>0.227</td>
</tr>
<tr>
<td>Ventilation</td>
<td>1928</td>
<td>0.661</td>
<td>1685</td>
<td>0.690</td>
</tr>
</tbody>
</table>

All variables except prices are dummy variables. Standard deviations for prices are given in parentheses. (*) Reference Category for windows (**) Reference category for facade ($^1$) Monthly rent in Swiss Francs. ($^2$) Purchase prices in thousand Swiss Francs.
### Table 3  Estimation results of the logit model with individual fixed effects

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Rented flats in apartment buildings</th>
<th></th>
<th></th>
<th>Purchase of single family houses</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>Std. Err.</td>
<td>Sig.</td>
<td>Coeff.</td>
<td>Std. Err.</td>
<td>Sig.</td>
</tr>
<tr>
<td>Price</td>
<td>-0.0089</td>
<td>0.0009</td>
<td>***</td>
<td>-0.0229</td>
<td>0.0033</td>
<td>***</td>
</tr>
<tr>
<td>Price * dummy decreasing price</td>
<td>0.0047</td>
<td>0.0014</td>
<td>***</td>
<td>0.013</td>
<td>0.0055</td>
<td>**</td>
</tr>
<tr>
<td>Enhanced insulated window</td>
<td>0.15</td>
<td>0.21</td>
<td>n.s.</td>
<td>0.14</td>
<td>0.23</td>
<td>n.s.</td>
</tr>
<tr>
<td>Enhanced facade insulation</td>
<td>0.50</td>
<td>0.20</td>
<td>**</td>
<td>0.51</td>
<td>0.23</td>
<td>**</td>
</tr>
<tr>
<td>Housing ventilation system</td>
<td>0.90</td>
<td>0.17</td>
<td>***</td>
<td>0.54</td>
<td>0.21</td>
<td>***</td>
</tr>
<tr>
<td>Housing ventilation system * new building</td>
<td>0.46</td>
<td>0.32</td>
<td>n.s.</td>
<td>1.33</td>
<td>0.38</td>
<td>***</td>
</tr>
<tr>
<td>Medium old windows</td>
<td>-1.49</td>
<td>0.22</td>
<td>***</td>
<td>-1.95</td>
<td>0.24</td>
<td>***</td>
</tr>
<tr>
<td>Very old windows</td>
<td>-2.68</td>
<td>0.25</td>
<td>***</td>
<td>-3.08</td>
<td>0.29</td>
<td>***</td>
</tr>
<tr>
<td>Painted facade</td>
<td>-0.73</td>
<td>0.22</td>
<td>***</td>
<td>-0.97</td>
<td>0.25</td>
<td>***</td>
</tr>
<tr>
<td>Unpainted facade</td>
<td>-1.10</td>
<td>0.22</td>
<td>***</td>
<td>-1.48</td>
<td>0.25</td>
<td>***</td>
</tr>
</tbody>
</table>

| | | | | | | |
| No. of persons | 157 | | | 142 | | |
| No. of observations (choice tasks) | 1928 | | | 1685 | | |
| Log likelihood | -540.44 | | | -435.12 | | |
| Pseudo R² | 0.318 | | | 0.298 | | |

---

1) Prices are expressed in CHF/month for rented flats and in thousand CHF for single family houses  
2) Reference category: new standard insulated windows  
3) Reference category: standard insulated facade  
Sig. = Significance level: *** 0.01, ** 0.05, * 0.1, n.s. = not significantly different from 0 at 10% significance level

---

25
Table 4 Marginal willingness to pay derived from discrete choice models, expressed as % of rental price (flats) and purchase price (single family houses) respectively 15

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Rented flats in multi-family houses</th>
<th></th>
<th></th>
<th></th>
<th>Purchase of single family houses</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WTP</td>
<td>Sig.</td>
<td>95%-Interval</td>
<td>WTP</td>
<td>Sig.</td>
<td>95%-Interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhanced insulated window</td>
<td>1%</td>
<td>n.s.</td>
<td>-1% 3%</td>
<td>1%</td>
<td>n.s.</td>
<td>-2% 4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as compared to standard insulated windows)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhanced facade insulation</td>
<td>3%</td>
<td>*</td>
<td>1% 5%</td>
<td>3%</td>
<td>**</td>
<td>0% 6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(As compared to standard insulation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housing ventilation system (new buildings)</td>
<td>8%</td>
<td>***</td>
<td>4% 11%</td>
<td>12%</td>
<td>***</td>
<td>6% 17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housing ventilation system (existing buildings)</td>
<td>8%</td>
<td>***</td>
<td>4% 11%</td>
<td>4%</td>
<td>**</td>
<td>1% 7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New windows</td>
<td>13%</td>
<td>***</td>
<td>8% 17%</td>
<td>13%</td>
<td>***</td>
<td>9% 18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as compared to medium old ones)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium old windows</td>
<td>10%</td>
<td>***</td>
<td>6% 14%</td>
<td>8%</td>
<td>***</td>
<td>4% 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as compared very old ones)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard facade insulation</td>
<td>6%</td>
<td>**</td>
<td>3% 10%</td>
<td>7%</td>
<td>***</td>
<td>3% 10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as compared to facade painting)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facade painting</td>
<td>3%</td>
<td>n.s.</td>
<td>-1% 7%</td>
<td>3%</td>
<td>*</td>
<td>0% 7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as compared to old unpainted facade)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

WTP = Willingness To Pay, expressed as % of rental price (flats) and purchase price (single family houses) respectively
Sig. = Significance level: *** 0.01, ** 0.05, * 0.1, n.s. = not significantly different from 0 at 10% significance level

26
Abstract

This paper applies an ordered discrete choice framework to model fuel choices and patterns of cooking fuel use in urban Indian households. The choices considered are for three main cooking fuels: firewood, kerosene and LPG (liquid petroleum gas). The models, estimated using a large microeconomic dataset, show a reasonably good performance in the prediction of households’ primary and secondary fuel choices. This suggests that ordered models can be used to analyze multiple fuel use patterns in the Indian context. The results show that lack of sufficient income is one of the main factors that retard households from using cleaner fuels, which usually also require the purchase of relatively expensive equipments. The results also indicate that households are sensitive to LPG prices. In addition to income and price, several socio-demographic factors such as education and sex of the head of the household are also found to be important in determining household fuel choice.
SUMMARY

Promotion of efficient use of energy and mitigation of the adverse environmental and health effects associated with the use of biomass fuels is an important policy issue in India. Despite a major shift from biomass fuels to commercial fuels, a considerable number of Indian households still use inefficient fuels such as firewood. This is not limited to rural areas where the choices might be constrained by lack of access. The use of LPG (liquid petroleum gas) is common in urban areas, but the use of less efficient fuels like kerosene and wood is still widespread. Analyzing the effect of different factors such as prices and income on the household’s fuel choice has received increasing attention in the energy literature. However, only a few studies have used micro-level data to undertake an empirical analysis. This paper analyzes patterns of cooking fuels used in urban Indian households using a large microeconomic dataset. Cooking fuels account for the bulk of residential energy consumption in India. We focus on the three main cooking fuels: LPG, kerosene and firewood. Based on the observed choices and also the comfort and environmental differences across these fuels, it is argued that there is an implicit ordering in the households’ preferences, suggesting that the transition between fuels can be simulated as a ‘ladder’ from firewood at the bottom to LPG at the top. Moreover, most households use multiple fuels. This suggests that switching from one fuel to another is at best partial and gradual. The correct prediction of households’ switching behavior, therefore, requires an adequate modeling of the preference ordering and the secondary fuel choices. Also, given the considerable amount of fuel consumed as secondary fuels, the prediction of such choices is especially important. This paper applies a discrete choice model that assumes a uniform ordering across different households. The results indicate that in addition to income, there are several socio-demographic factors such as education and sex of the head of the household, which are important in determining household fuel choice. While prices of kerosene and wood do not show a considerable effect on fuel choices, LPG price has an important effect particularly for low and moderate-income households. However, given that most low income households are wood users and that household income has a considerable effect on the choice of fuels, policies that promote rebates on the purchase of LPG stoves and easier access to credit or purchase on installment plans for such equipment might be an effective way of promoting the adoption of cleaner fuels. Finally, this analysis highlights the important effect of education and empowerment of women in development policies aimed at promoting efficient energy use.
I. Introduction

For a number of developing countries, including India, issues relating to energy choice and household energy transitions are important from a policy standpoint. Efforts at encouraging and facilitating households to make substitutions that will result in more efficient energy use and less adverse environmental, social and health impacts are advocated in many of these countries. But the effective design of public policy in this area requires, as a first step, research and analysis of the factors that affect energy choices and consumption patterns in rural and urban areas of such countries. In rural areas, choices are constrained not only by low incomes, but also by the lack of access to more commercial fuels and markets for energy using equipments and appliances. Often, the choice of fuel is determined more by local availability and transaction and opportunity costs involved in gathering the fuel (mostly wood, dung and other biomass) rather than by household budget constraints, prices and costs. Modeling choices in such circumstances is complicated and often there is little data available on proximity to supply of biomass, opportunity costs or time needed for collection.

In contrast to rural households, urban ones often have a wider choice and greater availability and accessibility to modern commercial fuels, electricity, and energy using end-use equipment and appliances, and therefore, greater potential for fuel switching. The rapid growth of urban areas in developing countries has been accompanied by a huge surge in the demand for household fuels and electricity. In India, the share of urban population increased from 17.3% in 1951 to about 28% in 2001 and is projected to rise to about 41% by 2030 (UN, 2003). Changing urban lifestyles have important implications for the quantum and pattern of energy use in households residing in these areas and suggest various avenues for policy relevant research. In addition, an understanding of factors affecting fuel choices in urban households might also provide insights into how rural households might behave if supply of commercial fuels and access to markets were not constrained in these areas.

In India, household energy is required to meet the needs for cooking and water heating and for lighting and powering electrical equipment and appliances. However, the bulk of energy used in Indian households even today is for cooking. Therefore, an understanding of cooking energy consumption patterns is particularly important.

Despite a major shift away from the use of biomass fuels towards commercial fossil fuels and electricity over the last two decades in urban areas, there are still many poor Indian households who rely on firewood as their primary source of cooking energy. As income increases households tend to switch from firewood to kerosene and then LPG (liquid petroleum gas). However, all households do not necessarily switch completely or, in other words, terminate the use of one fuel when taking up the use of another.

LPG, when compared to kerosene or firewood, has clear health, environmental and efficiency benefits. Of course, choice is constrained by cost as well and it is not fuel costs alone that matter, but also the start-up costs of connections, equipment and

---

1 Indoor heating is limited to a short season in the northern areas that face relatively cool temperatures.
2 About 90% of the total residential energy consumption in India was reported to be for cooking by Natarajan (1985).
stoves. Some recent studies that have compared total costs of different cooking fuels in India (WB, 2003; Reddy, 2003; Gupta and Ravindranath, 1997) find that in some cases the option of purchased firewood can be even more expensive than LPG particularly, when the efficiency of use of the different fuels is taken into account. However, for most poor households, the capital costs associated with the use of LPG are still a large hindrance to wider adoption of this fuel for cooking. In addition, a number of factors other than the cost affect the choice of fuels used by the household. The energy ladder hypothesis, that has traditionally been used to describe household fuel switching strategies, prescribes income to be the sole factor. However, as will be shown later in the paper, there are, in fact, several other household characteristics that affect choice.  

In this paper, we analyze cooking fuel choices in urban households of India. For this purpose we use a microeconomic data set, which is derived from the Indian Household Consumer Expenditure Survey conducted by the National Sample Survey Organisation (NSSO, 2002). Fuel choice is modeled empirically using a discrete choice framework and the substitution relationships between fuels examined. The impact of income and prices on fuel choice are examined. The analysis also aims to identify whether and to what extent other socio-demographic variables determine fuel choice.

The rest of the paper is structured as follows. Section II includes a brief review of the literature. Section III describes the data and presents some descriptive statistics. Section IV presents the model, and section V discusses the results. Finally section VI concludes with a brief discussion of some of the main policy implications of the results.

II. Literature Review

Several studies that try to understand household energy use patterns in developing countries can be found in the literature. However, those that try to quantify patterns in household energy transitions and the underlying causal factors, or factors affecting fuel choice decisions using disaggregate household data are more limited. Recently, renewed focus on such studies has been stimulated by growing concerns about the health impacts of indoor air pollution associated with the burning of unprocessed biofuels like wood and dung in inefficient cooking stoves. Amongst studies on household fuel choices for developing countries, we can distinguish broadly between two types, those that use simple descriptive statistics and others that have employed econometric methods to analyze fuel choice.

The traditional view on fuel switching in the household sector of developing countries has been that households gradually ascend an “energy ladder” and that there is a simple linear progression from relatively inefficient fuels and energy end-use equipment to more efficient fuels, electricity and equipment, with increasing income levels and urbanization (Leach, 1992; Sathaye and Tyler, 1991; Smith et al., 1994; Reddy and Reddy, 1994). In general, much of the literature points to income being an important factor influencing energy choice. However, while income is important, in as

\[1\text{ For a discussion of the energy ladder hypothesis see Leach, 1992; Sathaye and Tyler, 1991; Smith et al., 1994; Reddy and Reddy, 1994; Barnes and Qian, 1992; Leach and Gowen, 1987.}\]
far as it increases the options available to a household, what in fact actually motivates households to switch between different fuels and triggers energy transitions is a much more complex interplay of factors. Recent literature on household energy use in developing countries also supports the view that in fact the picture drawn by the energy ladder theory is too simplistic and that there are many factors that determine fuel choice (Davis, 1998; Masera et al., 2000; Barnett, 2000). An early study by Hosier and Dowd, (1987) for household fuel choice in Zimbabwe using a multinomial logit model shows that although economic factors do affect fuel choice, a large number of other factors are also important. In addition, much of the recent literature bears out that fuel switching is often not complete and is in fact, a gradual process with many households often using multiple fuels. The reasons for multiple fuel use are varied and not dependent on economic factors alone, although the affordability or cost of the energy service also has an important bearing on the household’s choice. In some cases, households choose to use more than one fuel because they want to increase the security of supply. In other cases, the choice might be dependent on cultural, social or taste preferences.

Other recent work in this field include a study for Bolivia (Israel, 2002) that examines whether fixed costs associated with switching to LPG act as a barrier, how income growth effects fuelwood use and whether female earned income influences fuel choice. The study concludes that reducing the fixed costs associated with a switch to cleaner fuels like LPG and increasing income earning opportunities for women can go a long way in encouraging households to shift away from the use of fuelwood. A multicountry study by the World Bank (Heltberg, 2003, 2004) has also examined the factors affecting a switch from solid (traditional) fuels to non-solid (modern) fuels and the role of electrification in facilitating such a switch. In another study, Heltberg (2005) analyzes the factors determining fuel choice in Guatemalan households. Another recent study by Chaudhari and Pfaff, (2003) estimates Engel curves for traditional (dirty) and modern (cleaner) fuels using household survey data from Pakistan and concludes that there is evidence of a U-shaped relationship between indoor air quality and income akin to the EKC as households transition from traditional fuels to modern fuels with increases in household income. Evidence on the nature of household energy transitions in Africa includes studies by Campbell et al., (2003); Davis, (1998); Ezzati and Kammen, (2002); Hosier and Kipondya, (1993).

Evidence from empirical studies on the patterns of household energy use in India includes WB, 1999; WB, 2002; Alam et al., 1998. Viswanathan and Kavi Kumar, (2005) analyze fuel consumption patterns across rural and urban households in India by examining data on the share of expenditures for different fuels. However, prior empirical research using a discrete choice framework for households in India is limited to only two studies. The first of these studies is Reddy, (1995) that looks at energy carrier choices for a sample of households residing in the city of Bangalore. He employs a series of binomial logit models to determine the choice between each pair of energy carriers, to explain the shifts in and the pattern of consumption of different fuels used for cooking and water heating. Results of the study confirm the hypothesis that households ascend an energy ladder and the choice is largely determined by income. However, factors such as family size and occupation of the head of the household are also seen to play a role in fuel selection.

More recently, Gangopadhyay et al., (2003) employ a multinomial logit framework to represent household fuel choice separately for rural and urban Indian households. They also employ data from the NSSO household expenditure survey,
which we use in this paper. However, they model household decisions concerning the choice of both cooking and lighting fuels together and therefore consider a choice set that consists of all the key alternatives of different energy carrier combinations used by households. The objective of that study was to evaluate the effectiveness of the existing price subsidies in facilitating a shift to the cleaner and more efficient fuels – kerosene and LPG. Their results indicate that the existing subsidies are fiscally unsustainable and also of little help in meeting social policy objectives as they are seriously misdirected and favor the rich disproportionately.

Given the limited area and country specific empirical evidence that is available on this topic, this research aims to augment the knowledge in this field. The present paper differs from the previous studies described above in three important regards. First, we analyze choices only in urban households, as we believe an analysis of choice of household fuels within rural areas would require additional information on nearness of source of biomass or time required for collection. Secondly, the analysis focuses on cooking fuels, which still comprise the largest part of household energy needs in India, and are quite separate and disparate from the energy needs for either lighting or powering appliances. Finally, we assume that there is a natural order of progression in terms of the choice of fuels based on their efficiency, ease of use, and cleanliness and therefore employ an ordered discrete choice framework to model fuel choice.

III. Data Source and Descriptive Statistics

The household micro budget data used in this study is from the household expenditure survey Round 55 covering the period July 1999 to June 2000 conducted by the National Sample Survey Organization (NSSO), a part of the department of statistics of the Indian government (NSSO, 2002). We selected the 1999-00 cross-section data to analyze fuel choices because it is the most recent quinquennial round of the survey available. The survey collects information on quantity consumed and value of household consumption for a wide variety of consumer goods and services. In addition, data on a host of other socio-economic and infrastructural variables is collected via the survey. The data is collected from a large nation-wide sample of households living in both rural and urban areas using the interview method. For the analysis presented in this paper, we make use of data only from the urban sample and the quantity and expenditure data for fuels/energy on a 30-day recall basis.

For the urban sector, the complete sample from Round 55 consists of 48,924 households representing 51.4 million households and a total urban population of approximately 314 million people. The information on cooking energy consumption is available for 46,918 households. Data pertaining to a few observations where there were missing or extreme values were excluded. We also excluded all observations where the household had no cooking arrangement or “other fuels”, that is, fuels other than LPG, kerosene or firewood were used as a cooking fuel. This comprised about

---

4 For details regarding the sampling methodology refer to NSSO, (2002).
5 The official definition of urban areas is based on number of criteria including “(a) the population of the place should be greater than 5000; (b) a density of not less than 400 persons per square km.; (c) three-fourths of the male workers are engaged in non-agricultural pursuits” (GoI, 2001).
11% of the total urban sample. The final analysis was conducted using a sample of 41,593 household level observations.

Amongst urban households in India, the main cooking fuels in use are firewood (often commercially bought), kerosene and LPG. The data indicate that in 1999-00, 30% of urban households still used firewood as a cooking fuel, while the percentage using kerosene was about 70% and about 50% used LPG. As different fuels vary in their efficiency, the main cooking fuel is defined as the fuel that provides the highest share of total useful cooking energy used by the household. This does not necessarily correspond with the reported primary cooking fuel in the questionnaire. The rates for converting to useful energy for LPG, kerosene and wood are calculated by assuming specific average levels of efficiency in the use of these fuels for cooking. The reason for using useful energy as the basis for the analysis is that households in fact do not demand energy in itself, but in fact demand services such as a hot cooked meal that energy helps provide. While ideally, one would like to capture demand at the level of energy services, this is not possible and thus useful energy proves to be the best approximation to the level of energy services.

Both the choice of household cooking fuel and the amount consumed are related to the income (proxied by the per capita expenditure level) and the size of the household. The relationship between the primary fuel choice and income level is illustrated in Figure 1. This figure shows that as income increases the likelihood of choosing wood drops while that of LPG rises. As for kerosene, the likelihood first rises at low incomes, peaks at the third decile and then declines. This suggests that while a higher income is likely to be associated with a switch from wood to kerosene/LPG among moderate income groups, in high-income groups the likely effect is a switch from wood/kerosene to LPG.

The data also show that about 54% of households use two or more fuels. The incidence of single-fuel use is about 50% among LPG users and as high as 74% among households using kerosene. However, this is not the case for firewood users, most of whom use kerosene as well. Finally, there are few households that use a combination of LPG and firewood or who use all three fuels. Thus, we see that multiple fuel use is more frequent in poorer households that are more dependent on less efficient biomass fuels (see Pachauri et al., 2004 for more on energy poverty in India).

Figure 2 plots total useful energy use for cooking across income decile groups. One observes an increase in the amount of energy use with income, but leveling off among the highest income deciles. The share of different fuels used varies significantly across deciles with a larger share of firewood and kerosene among lower deciles and a predominate share of LPG among higher income groups. The observed pattern seems to suggest that among households in the lower income groups, fuels are used more as compliments and there is a greater degree of fuel stacking in evidence. It

---

6 These observations mainly consist of 1,768 households with no cooking arrangement, 2,087 using coal and 877 using dung cake as their main cooking fuel and 542 households that use LPG, kerosene or wood as their main fuel but use other fuels as well.

7 Refer to Pachauri and Spreng, 2004 and Pachauri et al., 2004 for a description of how useful energy is calculated for households using the survey data.

8 The values used in this paper are 276 kJ/liter for LPG, 148.5 kJ/liter for Kerosene and 21 kJ/kg for wood.
is only among those in the higher decile groups that some cases of complete fuel switching occur, with one fuel dominating total energy needs for cooking, and additional fuels possibly only used occasionally as back-up.

**Insert Figure 2:**

The distribution of households by their main fuel choice is given in Table 1. This table shows that for the majority of the urban LPG using households, more than two third (69%) of cooking energy needs is met from LPG. Kerosene and firewood are used as the main energy source in a considerable number of urban households (26 and 21% of the sample respectively). Even in households that mainly use LPG, the share of kerosene is, on average, about 7% of total cooking energy. In the case of LPG and kerosene users, the median share of secondary fuels drops to zero suggesting that the supplemental fuel might be used only as a back up. Table 1 also shows the average share of kerosene purchased from the private market as opposed to the subsidized public distribution system. These numbers show that households that use kerosene as their primary fuel purchase more than half of their fuel from the market, whereas the majority of those who use kerosene as a secondary fuel tend to purchase the subsidized kerosene. This suggests that both market and subsidized prices may affect the choice probabilities.

**Insert Table 1:**

The above descriptive analysis suggests that the observed patterns in the data are consistent in part with the “energy ladder” theory. In other words, there is a clear order in the distribution of energy shares by the primary fuel (see Table 1). Firewood and LPG at the two extremes are more likely to be used with kerosene in the middle, than with each other. Moreover, at the bottom of the ladder, households are more likely to use two fuels. In contrast, at the top of the ladder (LPG), single fuel choices are more likely. The econometric model used in this paper is in line with the ordered preferences observed in the data.

Table 2 presents the descriptive statistics of the household characteristic variables included in the model specification. As seen in the previous discussion, household income (proxied by the household’s per capita monthly expenditure) has a considerable effect on the fuel choice. Dummy variables for the level of education of the head of the household, occupation, female headed households, season, and geographic location (state dummies and a dummy for households residing in metropolitan areas) are included in the model in addition to variables relating to household size, fuel prices, and age of the head of the household. While the dataset includes a wide variety of information on household level characteristics, expenditure and consumption information, one area where the data are lacking is regarding independent and reliable information on fuel prices. For this reason, fuel prices are calculated as the median value of individual prices for each one of the 78 regions (sub-states) in the sample. Unit or average values calculated by dividing expenditures on each fuel type by the corresponding quantities for each household are used as a proxy for individual prices. By including dummies relating to state regions and seasons, we hope to capture some of the unexplained variation on account of the lack of direct price data in addition to any spatial and geo-climatic differences that

---

9 The extremely low age of the household head in a few observations (Table 2) is due to an Indian tradition that specifies that in the absence of a father the eldest male son is considered the head of the family. Excluding these observations does not change our estimation results significantly.

10 See the appendix for sample means of state dummies.
might exist. Finally, the table also includes descriptive statistics relating to the number of LPG dealers per 100,000 households. This variable is included in the model so as to capture differences on account of LPG availability and accessibility at the state level.

**Insert Table 2:**

### IV. Model and Estimation Methods

As discussed in the previous sections, the observed patterns in the data suggest that the fuel choice in urban households is consistent with an ordered discrete choice framework. These models such as ordered logit and probit are often used for ordered categorical response variables that represent groups of continuous variables, such as income groups. However, the application of these models can be extended to categorical variables that have an “assessed” order, such as “the extent of pain relief after treatment” (Anderson, 1984). These variables are referred to as assessed, ordered variables. In many of these response variables, the ordering is not obvious at first sight. We contend that the cooking fuel type in an Indian household can be considered as an ordered variable, in that the three fuel types can be clearly ordered in terms of efficiency, comfort and ease of use.

In this paper we report results of the estimation of an ordered probit model (see Green, 2003; and Wooldridge, 2002 for more details). In this model it is assumed that the individual’s choices are based on a latent variable, which can be considered as a measure of random utility. This latent variable is defined as a linear function of explanatory variables:

\[ y_i^* = X_i \beta + Z_i \gamma + \varepsilon_i, \]

where \( X_i \) is the vector of alternative fuel prices faced by household \( i \); \( Z_i \) is the vector of household characteristics; \( \beta \) and \( \gamma \) are the parameter vectors to be estimated; and \( \varepsilon_i \) is an iid stochastic error term that represents the unobserved heterogeneity. The probability of choosing alternative \( j \) is defined as:

\[ \Pr(y_j = j) = \Pr(k_{j-1} < y_i^* \leq k_j) ; \quad -\infty = k_0 < k_1 < \ldots < k_j = +\infty, j \in \{1, 2, \ldots, J\}, \]

where \( k_j \)'s are the threshold parameters.

The error term \( \varepsilon_i \) is assumed to follow a normal distribution with mean zero and variance \( \sigma^2 \). In this model, the probability of choice \( j \) can be written as:

\[ \Pr(y_j = j) = \Phi \left( \frac{-k_j + X_i \beta + Z_i \gamma}{\sigma} \right) - \Phi \left( \frac{-k_{j-1} + X_i \beta + Z_i \gamma}{\sigma} \right), \]

where \( \Phi \) is the CDF of a standard normal variable.\(^{11}\) The model in Equation (3) can be estimated using maximum likelihood estimation method. As seen in Equation (2), the choice probabilities are assumed to be a function of a continuous latent variable \( (y^*) \) that can be considered as the household’s “energy status” or the position of the household on the energy ladder.

\(^{11}\) An ordered logit model was also estimated. The results (available upon request from the authors) are generally similar to those of the ordered probit model.
The model described above requires that the alternatives are ordered, namely \( j = 1, 2, 3 \) correspond to firewood, kerosene and LPG respectively. This assumption implies that households are more likely to substitute two fuels that are adjacent on the specified ordering. For instance, if a wood user is to choose another alternative, kerosene is more likely to be chosen as opposed to LPG. In order to explore if such an assumption is realistic, we also considered two non-ordered discrete choice models namely, a multinomial probit model in line with Geweke et al., (1994) and a multinomial logit specification.\(^{12}\) Comparing the prediction results between ordered and non-ordered models indicates a slightly better prediction rate for primary fuel choice in non-ordered models. However, when we consider the secondary fuel choices, the situation gets reversed with the ordered probit model making correct predictions in about 63% of the cases, compared to 51% for the non-ordered models.\(^{13}\) The results also suggest that all models especially the ordered ones are weak in predicting the primary fuel for kerosene users. However, when we consider both primary and secondary choices among multiple fuel users, the ordered models have a clear advantage. Given that fuel switching is a transitional and gradual behavior, correct prediction of the preferences of multiple fuel users is necessary for understanding the substitution possibilities of single fuel users as well.

As for the estimated marginal effects, the results for ordered and non-ordered models are comparable for almost all the socio-economic factors and the LPG prices.\(^{14}\) Overall, these comparisons show that the ordered models have a better performance especially considering that these models include only half as many parameters as the non-ordered models. Therefore, we retain the ordering assumption and focus on the ordered probit model in the rest of the paper.

V. Results

The maximum likelihood results from the estimations of the ordered probit model described in section IV are presented in Table 3. These results indicate that most of the explanatory variables included in the model have significant effects and show the expected signs. The results clearly show that there are a number of factors, other than income, influencing the choice of household cooking fuels in urban India. The coefficients listed in Table 3 can be interpreted as the effects on the households’ energy status that is the position of the household on the energy transition line (ladder). As expected, income and education have a positive and significant effect.

Table 3 also indicates that LPG and kerosene prices have negative effects suggesting that higher prices can result in a lower energy status, with LPG price having the greatest and kerosene price having the least effect. This can be explained by the combination of an effect on purchasing power (income effect) and a substitution effect. While the income effect for a price increase is always downward (away from LPG), the substitution effect is upward for wood prices, downward for LPG and ambiguous for kerosene prices. Therefore, the resulting effect is relatively

\(^{12}\) The estimation results are not included in the paper but are available upon request.

\(^{13}\) The predicted primary and secondary fuels are defined as the alternatives that have the highest and second-highest probabilities.

\(^{14}\) There however, is a significant difference regarding the effects of kerosene and wood prices with non-ordered models suggesting a counter-intuitive effect for kerosene prices on the probabilities.
high for LPG prices, whereas for wood prices the two effects cancel out hence resulting in an insignificant overall effect. The negative effect for kerosene prices suggest that the resulting effect of a price increase is toward inferior fuels and away from LPG. Similarly, one can conclude that a decrease in kerosene prices might be effective in raising energy status and encouraging adoption of cleaner fuels. The price of wood is statistically insignificant for fuel switching. This might be explained by the fact that wood-users are mainly the low-income people who cannot afford other alternative fuels that are considerably more expensive especially in terms of fixed costs of appliances.

The size of the household and the age of the head of the household have a positive effect on the probability of choosing cleaner fuels, as does the household being headed by a female. Living in larger cities or metros also increases the probability of choosing cleaner fuels, as does having more LPG distributors and hence easier accessibility. The seasonal dummies have no significant effects, suggesting that urban households do not significantly change their cooking energy choices across different seasons. A number of state dummies are also included in the model and the coefficients on these are mostly significant, suggesting that there are differences in the choice behavior of households living in different regions of the country (see Table A1 in the appendix for the estimated effects of the state dummy variables). The rates of correct prediction of the household’s main fuel are given at the bottom of Table 3.

**Insert Table 3:**

In order to better understand the nature of the substitution patterns between the three main cooking fuels amongst different households, the marginal effects of the significant variables at the sample means are also calculated and presented in Table 4. The numbers in this table show the effect of a one-unit change in a given explanatory variable (or a switch in the case of dummy variables) on the probability of choosing each one of the three fuels. As all the continuous variables are in logarithms, the corresponding marginal effects can be interpreted as the effect of a relative change, thus can be used for a direct comparison of the magnitude of different effects. The first observation from these results is that among the continuous explanatory variables, LPG price and household income have the most important effects and among the dummy variables, those associated with the household head’s education have the greatest effects.

The household head being illiterate or only having primary education increases the probability of choosing firewood or kerosene as a cooking fuel, whereas those households where the head has a higher level of education are more likely to use LPG. For instance, households with illiterate heads are on average about 22% more likely than those with a secondary school education (base category) to use wood and about 34% less likely to use LPG. These results also indicate that a 10% increase in income will raise the share of LPG users by 4.7% while decreasing the share of wood and kerosene users by 2.6 and 2.1% respectively.

The results in Table 4 also suggest that higher LPG prices are associated with a significant negative shift away from LPG, as one might expect. According to the model results, a 10% decrease in LPG price, for instance will increase the average share of LPG users by about 7% while decreasing the share of wood and kerosene users by 3.1 and 3.9% respectively. The effect of kerosene price is much smaller and more ambiguous. For instance, a 10% increase in the kerosene market price will decrease the share of LPG users by about 0.8% while increasing the users of kerosene
and wood by about 0.4%. Such a positive effect on wood demand can be explained by the behavior of moderate and low-income households in substituting kerosene with wood. Whereas, the positive effect on kerosene demand and the negative effect on LPG demand can be explained by the ambiguous effect of kerosene prices on households that use LPG and kerosene together. These households, accounting for about 24% of the sample, when faced with high kerosene prices, in some cases might tend to substitute kerosene with LPG but in other cases are subject to an income effect resulting from higher kerosene prices, which might push them towards inferior fuels and consuming less energy. The results suggest that the latter effect dominates the former. Hence, an increase in kerosene price could cause a higher share of kerosene consumption for these households.

Insert Table 4:

In addition, the marginal effects for the variables household income and price of LPG calculated for different income tiles of the population are listed in Table 5. Differences in the marginal effects of the price and expenditure variables are evident for households belonging to different income tile groups. As expected, the marginal effect of income on the probability of using wood is much greater for low-income households. For instance a 10% increase in household income decreases the probability of using wood by .03 to .04 in low-income households, while the decrease is only about .003 to .007 for high-income households (see upper panel of Table5). A contrasting pattern can be observed for the use of LPG that is subject to a relatively low marginal effect for income among high-income households.

Similar patterns can be seen regarding the effects of LPG prices among different income groups. As the lower panel of Table 5 indicates, higher LPG prices push households away from LPG use, but this effect is greater for moderate and median income groups and relatively low for high-income households. For instance a 10% increase in LPG prices can decrease the LPG share by 7% among first quartile income group but only 5% among the third quartile income group. Another interesting pattern observed in Table 5 is that being in a lower income group increases the probability of choosing wood over kerosene when LPG price increases. Conversely, when facing higher LPG prices, the moderate and high-income households are more likely to substitute LPG with kerosene than with wood.

Insert Table 5:

VI. Conclusions

The paper provides results of the estimation of an ordered discrete choice model on fuel choices and patterns of cooking fuels in urban Indian households using a large database consisting of 46,918 observations. The analysis is used to determine the responsiveness of fuel choices to own price, income, price of alternate fuels and variables relating to socio-demographic and geographic characteristics of households.

From a methodological point of view, this paper differs from previous literature in that we assume that there is a natural order of progression in terms of the choice of fuels based on their efficiency, ease of use, and cleanliness and therefore, we employ an ordered discrete choice framework to model fuel choice. Our analysis shows that in the Indian context, such ordered models can be as useful and instructive
as non-ordered multinomial models. Our work suggests that ordered models that have fewer parameters and are easier to interpret, provide a better performance in predicting the choice of multiple fuel users.

The descriptive analysis and the econometric results reported in the paper suggest that the observed patterns in the data are consistent with the stylized “energy ladder” theory. In other words, there is an order in the distribution of energy shares by the primary fuel that depends to a large extent on the level of income of the household. Firewood and LPG at the two extremes are more likely to be used with kerosene in the middle, than with each other. However, the results also show that in addition to income, there are several socio-demographic factors such as education and sex of the head of the household, which are important in determining the choice of fuels in urban Indian households. Our results thus corroborate that of other recent studies (Heltberg, 2005; Masera et al., 2000) that suggest that fuel choice is not determined purely by economic factors and that a more general interpretation of the energy ladder theory is needed.

Overall, fuel choice decisions in urban Indian households appear to be flexible and dynamic with many households maintaining the ability to use two or more different fuels for cooking at any given point in time. Our results seem to suggest several reasons why households shift to the use of modern fuels. In urban areas, where firewood is often bought and opportunity costs for collecting wood are high, economic considerations and availability are crucially important in determining fuel choices. Higher incomes increase the ability of households to afford both the equipment and fuel costs of modern fuels like LPG, which are also more widely available in urban areas. Better education increases the awareness of households of the negative health impacts associated with the use of firewood and also the advantages of modern fuel use, in terms of efficiency and convenience. In larger cities and areas where modern fuel supplies are more regularly and reliably distributed, households are more likely to choose modern fuels and less likely to require back-up or supplemental use of other fuels. In addition, households where women are more empowered are less likely to use less efficient wood. Other reasons, such as tastes, customs and status, may also influence fuel choice and require further investigation.

From a policy point of view, our results suggest that in order to encourage households to make fuel substitutions that will result in more efficient energy use and less adverse environmental, social and health impacts, a subsidization of LPG gas provision, a promotion of higher levels of education, greater empowerment of women and a promotion of general economic development could be effective instruments. Given the high fiscal costs associated with LPG fuel price subsidies, however, it may be more sustainable to promote policies that promote rebates on the purchase of LPG stoves and easier access to credit or purchase on installment plans for the equipment needed to use cleaner fuels such as LPG in a more targeted fashion. In addition, since multiple fuels are more likely to be used by the poor and the share of secondary fuels in total cooking fuel consumption is higher for households in lower income decile groups, a LPG fuel subsidy policy is likely to benefit richer rather than poorer households and may not result in a complete transition away from the use of inferior fuels like wood and kerosene. As the results of the analysis presented in this paper highlight several other variables in addition to fuel price as affecting fuel choice, this points also to the importance of exploring other policy options than pricing alone.
References


Figure 1. Main cooking fuel by income

Fraction of households using the fuel as their primary cooking fuel

Decile of household monthly expenditure per person

LPG
Kerosene
Firewood
Figure 2. Total cooking energy by Income (41,593 households)

Decile of household monthly expenditure per person

Total useful energy (1000 MJ)

- Firewood
- Kerosene
- LPG
<table>
<thead>
<tr>
<th>Primary fuel used for cooking</th>
<th>Firewood</th>
<th>Kerosene</th>
<th>LPG</th>
<th>Fraction of households</th>
<th>Average share of kerosene purchased in the market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firewood</td>
<td>76.2% (78%)</td>
<td>23.0% (22%)</td>
<td>0.8% (0)</td>
<td>20.9%</td>
<td>31.2% (0)</td>
</tr>
<tr>
<td>Kerosene</td>
<td>7.1% (0)</td>
<td>91.4% (100%)</td>
<td>1.5% (0)</td>
<td>25.5%</td>
<td>55.2% (56%)</td>
</tr>
<tr>
<td>LPG</td>
<td>1.5% (0)</td>
<td>7.2% (0)</td>
<td>91.3% (100%)</td>
<td>53.6%</td>
<td>28.6% (0)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>18.5% (0)</td>
<td>32.0% (16%)</td>
<td>49.5% (69%)</td>
<td>100%</td>
<td>38.9% (0)</td>
</tr>
</tbody>
</table>

- Median shares are given in parantheses.
Table 2. Descriptive statistics (41,593 urban households)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPG price (Rps/liter) *</td>
<td>11.808</td>
<td>0.610</td>
<td>10.56</td>
<td>13.33</td>
</tr>
<tr>
<td>Kerosene market price (Rps/liter) *</td>
<td>9.145</td>
<td>2.139</td>
<td>4.80</td>
<td>13.00</td>
</tr>
<tr>
<td>Kero. price in public system (Rps/liter) *</td>
<td>3.218</td>
<td>0.383</td>
<td>2.70</td>
<td>5.00</td>
</tr>
<tr>
<td>Firewood market price (Rps/kg) *</td>
<td>1.448</td>
<td>0.465</td>
<td>0.67</td>
<td>3.50</td>
</tr>
<tr>
<td>No. of LPG distributors per 100,000 HHs **</td>
<td>5.159</td>
<td>3.407</td>
<td>1.25</td>
<td>14.53</td>
</tr>
<tr>
<td>Household monthly income (Rps)</td>
<td>4232.1</td>
<td>3136.2</td>
<td>108</td>
<td>68805</td>
</tr>
<tr>
<td>HH monthly expenditure per person (Rps)</td>
<td>1020.4</td>
<td>796.6</td>
<td>18</td>
<td>35612</td>
</tr>
<tr>
<td>Age of the HH head</td>
<td>44.83</td>
<td>13.32</td>
<td>5</td>
<td>98</td>
</tr>
<tr>
<td>Number of persons in the HH</td>
<td>4.711</td>
<td>2.387</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>HHs with a single member</td>
<td>0.063</td>
<td>0.243</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HHs with a female head</td>
<td>0.104</td>
<td>0.305</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Main HH income from casual labor</td>
<td>0.122</td>
<td>0.327</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HH head illiterate</td>
<td>0.178</td>
<td>0.382</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HH head's education primary school or lower</td>
<td>0.218</td>
<td>0.413</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HH head has a university education</td>
<td>0.190</td>
<td>0.392</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HH residence in a metropolitan area ***</td>
<td>0.214</td>
<td>0.410</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Interview was held in Monsoon</td>
<td>0.249</td>
<td>0.433</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Interview was held in Winter</td>
<td>0.248</td>
<td>0.432</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

* Median prices at the district level (78 districts).
** Calculated at the state level (32 states).
*** Cities with more than a million habitants.
Table 3. Regression results

<table>
<thead>
<tr>
<th>Alternatives in ascending order: Firewood, Kerosene, LPG</th>
<th>Coeff.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>In (LPG price)</td>
<td>-1.780 **</td>
<td>0.228</td>
</tr>
<tr>
<td>In (Kerosene market price)</td>
<td>-0.203 **</td>
<td>0.050</td>
</tr>
<tr>
<td>In (Kero. price in public system)</td>
<td>-0.201 *</td>
<td>0.087</td>
</tr>
<tr>
<td>In (Firewood market price)</td>
<td>0.010</td>
<td>0.045</td>
</tr>
<tr>
<td>In (# of LPG distributors per 100,000 HHs)</td>
<td>0.071</td>
<td>0.036</td>
</tr>
<tr>
<td>In (HH monthly expenditure per person)</td>
<td>1.182 **</td>
<td>0.017</td>
</tr>
<tr>
<td>In (Age of the HH head)</td>
<td>0.519 **</td>
<td>0.023</td>
</tr>
<tr>
<td>In (Number of persons in the HH)</td>
<td>0.424 **</td>
<td>0.018</td>
</tr>
<tr>
<td>HHs with a single member</td>
<td>-0.388 **</td>
<td>0.034</td>
</tr>
<tr>
<td>HHs with a female head</td>
<td>0.302 **</td>
<td>0.022</td>
</tr>
<tr>
<td>Main HH income from casual labor</td>
<td>-0.438 **</td>
<td>0.020</td>
</tr>
<tr>
<td>HH head illiterate</td>
<td>-0.899 **</td>
<td>0.019</td>
</tr>
<tr>
<td>HH head's education primary school or less</td>
<td>-0.537 **</td>
<td>0.016</td>
</tr>
<tr>
<td>HH head has a university education</td>
<td>0.622 **</td>
<td>0.024</td>
</tr>
<tr>
<td>HH residence in a metropolitan area</td>
<td>0.187 **</td>
<td>0.018</td>
</tr>
<tr>
<td>Interview was held in Monsoon</td>
<td>-0.006</td>
<td>0.016</td>
</tr>
<tr>
<td>Interview was held in Winter</td>
<td>-0.025</td>
<td>0.016</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-29721.2</td>
<td></td>
</tr>
<tr>
<td>Pseudo R-squared</td>
<td>0.2923</td>
<td></td>
</tr>
</tbody>
</table>

Percentage of correct prediction of chosen fuels:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary fuel for all the sample (41,593 households):</td>
<td>67.07%</td>
</tr>
<tr>
<td>Both 1st and 2nd fuels for multiple-fuel users (22,264 households):</td>
<td>62.57%</td>
</tr>
</tbody>
</table>

* significant at .05; ** significant at .01; State dummies (18 groups) are included in the model (see Table A1).
Table 4. Marginal effects at the sample mean

<table>
<thead>
<tr>
<th></th>
<th>Wood</th>
<th>Kero.</th>
<th>LPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln (LPG price)</td>
<td>0.310</td>
<td>0.391</td>
<td>-0.701</td>
</tr>
<tr>
<td>ln (Kerosene market price)</td>
<td>0.035</td>
<td>0.045</td>
<td>-0.080</td>
</tr>
<tr>
<td>ln (Kero. price in public system)</td>
<td>0.035</td>
<td>0.044</td>
<td>-0.079</td>
</tr>
<tr>
<td>ln (HH monthly expenditure per person)</td>
<td>-0.206</td>
<td>-0.260</td>
<td>0.466</td>
</tr>
<tr>
<td>ln (Age of the HH head)</td>
<td>-0.090</td>
<td>-0.114</td>
<td>0.204</td>
</tr>
<tr>
<td>ln (Number of persons in the HH)</td>
<td>-0.074</td>
<td>-0.093</td>
<td>0.167</td>
</tr>
<tr>
<td>HHs with a single member</td>
<td>0.083</td>
<td>0.071</td>
<td>-0.154</td>
</tr>
<tr>
<td>HHs with a female head</td>
<td>-0.045</td>
<td>-0.070</td>
<td>0.115</td>
</tr>
<tr>
<td>Main HH income from casual labor</td>
<td>0.093</td>
<td>0.080</td>
<td>-0.173</td>
</tr>
<tr>
<td>HH head illiterate</td>
<td>0.218</td>
<td>0.127</td>
<td>-0.344</td>
</tr>
<tr>
<td>HH head's education primary school or less</td>
<td>0.113</td>
<td>0.099</td>
<td>-0.212</td>
</tr>
<tr>
<td>HH head has a university education</td>
<td>-0.084</td>
<td>-0.145</td>
<td>0.230</td>
</tr>
<tr>
<td>HH residence in a metropolitan area</td>
<td>-0.030</td>
<td>-0.042</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Only the statistically significant effects are listed. For dummy variables the effects are obtained from probability differences.
Table 5. Marginal price and income effects at the sample median by income category

<table>
<thead>
<tr>
<th>Alternative:</th>
<th>Wood</th>
<th>Kero.</th>
<th>LPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>In (HH monthly expenditure per person)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH expenditure per person:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 percentile</td>
<td>-0.394</td>
<td>-0.015</td>
<td>0.410</td>
</tr>
<tr>
<td>25 percentile</td>
<td>-0.295</td>
<td>-0.171</td>
<td>0.466</td>
</tr>
<tr>
<td>Median</td>
<td>-0.168</td>
<td>-0.282</td>
<td>0.450</td>
</tr>
<tr>
<td>75 percentile</td>
<td>-0.072</td>
<td>-0.267</td>
<td>0.339</td>
</tr>
<tr>
<td>90 percentile</td>
<td>-0.027</td>
<td>-0.185</td>
<td>0.211</td>
</tr>
<tr>
<td>In (LPG price)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH expenditure per person:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 percentile</td>
<td>0.594</td>
<td>0.023</td>
<td>-0.617</td>
</tr>
<tr>
<td>25 percentile</td>
<td>0.444</td>
<td>0.257</td>
<td>-0.701</td>
</tr>
<tr>
<td>Median</td>
<td>0.253</td>
<td>0.424</td>
<td>-0.677</td>
</tr>
<tr>
<td>75 percentile</td>
<td>0.108</td>
<td>0.402</td>
<td>-0.510</td>
</tr>
<tr>
<td>90 percentile</td>
<td>0.040</td>
<td>0.278</td>
<td>-0.318</td>
</tr>
</tbody>
</table>
## Appendix.

Table A1. Regression coefficients and sample means for state dummies

<table>
<thead>
<tr>
<th>State dummy</th>
<th>Sample Mean</th>
<th>Coeff.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.0833</td>
<td>0.180 **</td>
<td>0.040</td>
</tr>
<tr>
<td>ARP,ASM,MPR,MEG, MIZ,NGL,SKM,TRI</td>
<td>0.0838</td>
<td>-0.302 **</td>
<td>0.043</td>
</tr>
<tr>
<td>BHR</td>
<td>0.0298</td>
<td>-0.065</td>
<td>0.064</td>
</tr>
<tr>
<td>GOA, D&amp;D, A&amp;N Islands, LKS, D&amp;N Hoveli</td>
<td>0.0217</td>
<td>-0.282 **</td>
<td>0.050</td>
</tr>
<tr>
<td>GUJ</td>
<td>0.0606</td>
<td>0.368 **</td>
<td>0.041</td>
</tr>
<tr>
<td>HAR, PUN</td>
<td>0.0564</td>
<td>0.404 **</td>
<td>0.040</td>
</tr>
<tr>
<td>HP, J&amp;K</td>
<td>0.0203</td>
<td>0.304 **</td>
<td>0.052</td>
</tr>
<tr>
<td>KAR</td>
<td>0.0530</td>
<td>-0.160 **</td>
<td>0.044</td>
</tr>
<tr>
<td>KER</td>
<td>0.0451</td>
<td>-0.939 **</td>
<td>0.050</td>
</tr>
<tr>
<td>MP</td>
<td>0.0655</td>
<td>-0.126 **</td>
<td>0.040</td>
</tr>
<tr>
<td>ORS</td>
<td>0.0188</td>
<td>-0.378 **</td>
<td>0.065</td>
</tr>
<tr>
<td>RAJ</td>
<td>0.0440</td>
<td>-0.117 *</td>
<td>0.048</td>
</tr>
<tr>
<td>TN, PON</td>
<td>0.0994</td>
<td>-0.136 **</td>
<td>0.038</td>
</tr>
<tr>
<td>UP</td>
<td>0.0925</td>
<td>-0.062</td>
<td>0.036</td>
</tr>
<tr>
<td>WB</td>
<td>0.0505</td>
<td>0.088</td>
<td>0.054</td>
</tr>
<tr>
<td>CHD</td>
<td>0.0173</td>
<td>0.456 **</td>
<td>0.082</td>
</tr>
<tr>
<td>DEL</td>
<td>0.0242</td>
<td>0.200 **</td>
<td>0.073</td>
</tr>
</tbody>
</table>

The omitted state: MHR; * significant at .05; ** significant at .01.
Mehdi Farsi, Lara Gitto

A statistical analysis of pain relief after surgical operations

Quaderno N. 06-01

Decanato della Facoltà di Scienze economiche
Via G. Buffi, 13 CH-6900 Lugano
A statistical analysis of pain relief after surgical operations

Mehdi Farsi†, Lara Gitto‡

† Department of Management Technology and Economics, ETH Zurich, and Department of Economics, University of Lugano, Switzerland

‡ CEIS Sanità, University of Rome Tor Vergata and University of Messina, Italy.

Correspondence: Lara Gitto, Department of Economics, University of Messina, Piazza Pugliatti 1, 98100 Messina, Italy. Email: Lara.Gitto@unime.it. Phone: +39090719202. Fax: +39090719202

Acknowledgment: We would like to thank Dr. Letterio Guglielmo and Dr. Nicola Torina from Ospedale Buccheri La Ferla, Palermo, Italy, for providing the data and their support. We are responsible for remaining errors and omissions.

Abstract

Adequate and timely application of analgesics after surgical operations is important from both clinical and economic perspectives. Administering pain relief measures requires information about development of post-operative pain and the effect of analgesics. Such information can be obtained from studying patients’ perception of pain in different periods after the operation. This paper applies an ordered response model to a sample of patients undergoing orthopedic surgeries. All studied patients have been received at least one analgesics. The patients’ subjective pain levels have been recorded for several intervals up to 24 hours after their respective operations. The adopted statistical model accounts for the unobserved heterogeneity among patients through random coefficients. Such heterogeneity could be due to differences in patients’ subjective measure of pain as well as their health status and sensitivity to pain. The analysis indicates that post-operative pain gradually increases over time but with a slightly diminishing rate. The results suggest that analgesics are quite effective in containing the development of pain. However, the analgesic effects manifest gradually, at a rate which is more or less similar to that of post-operative pain. This result implies that the optimal time of administering analgesics is immediately after the operation, suggesting preemptive analgesics could be most effective.

Keywords: post-operative pain; analgesics; ordered logit; random coefficients

1. Introduction

The International Association for the Study of Pain, (IASP, 1979) defines pain as “an unpleasant sensorial and emotional experience associated to present or potential disadvantages for patients”. Pain is not only considered a physiological phenomenon, such as a symptom of a latent or incipient disease, or a consequence of an intervention, but as a pathological state that is likely to evolve to an illness. Neglecting pain might lead to complications in the long run (Dworkin, 1997), a lower quality of life and a longer stay at hospital (Berry and Dahl, 2000). Therefore, tolerating pain in any number of patients could result in considerable economic losses through additional hospitalization costs as well as lower productivity for the patients at least in the short run. The treatment of postsurgical pain has also proved effective in improving the quality of care: patients who have undergone analgesic therapy are reported to experience a better surgical follow up (Berry and Dahl, 2000).
National health organizations throughout the world recognize pain relief as a primordial medical objective – both ethically as well as from a cultural point of view (Zborowski, 1982). Several prevalence studies carried out for the United States, Great Britain and Holland during the 80s and 90s, report that the number of patients suffering from post-surgical pain has grown considerably from 45% up to 79% in all the countries where the surveys have been performed (for a brief review on this topic see Donovan et al., 1987, Visentin et al., 2005). Similar trends have been reported for Italy, where the steps taken against post-surgical pain have been characterized by inefficiencies (cf. Notaro et al. 2001, Lattuada et al. 2004, and Visentin et al. 2005). For instance, Italy has historically one of the lowest scores in Europe in consumption of opiates, an indicator of pain relief performance established by the World Health Organization1.

The level of pain is often considered as an indicator of the quality of care. Every time a patient reports a positive level of pain it could imply that pain is not controlled efficiently and that the clinical procedures should be revised. In this perspective, the “pain-free hospital” project (ospedale senza dolore) carried out in Italy, among other European countries, provides guidelines that could control the level of pain in the majority of patients complaining of post-surgical pain (Visentin et al., 2005, have estimated the project “pain-free hospital” should achieve the objective of lowering or eliminating pain for 90% of cases. That means that a great number of patients still suffer from a pain that could be avoided through the implementing of such guidelines. However, as Lattuada et al. (2004) reported for a region in Italy, implementing such projects and achieving the stated targets could be difficult in many cases.

Studying the development of post-surgical pain and the effect of analgesics could help improve the administration of analgesics and achieve a better allocation of resources used for pain relief. For instance knowing the evolution of pain, one can better identify the optimal timing of analgesics, or based on the effect of gender, age, type of surgical intervention, one can identify cases that are more sensitive and thus need more help. This study is aimed at modeling the probability of experiencing different levels of post-operative pain as a function of the time after surgery, the time after analgesic therapy, and several patient characteristics. The pain level is modeled using a discrete ordered response model namely, ordered logit. The unobserved heterogeneity among patients is accounted for by random coefficients. The studied sample includes 49 inpatient orthopedic surgeries. The patients were observed in a hospital (Buccheri La Ferla) in Palermo (Sicily, Italy), in 2002 and 2003. The analysis is expected to cast light on the impact of individual characteristics and analgesic administrations on the probability of post-surgical pain.

---

1 Italy’s poor performance could be explained by the laws that have limited for a long time, the use of narcotics and, at the same time, precluded the use of opiates in clinical practice.
To make such an analysis possible, it is necessary to select appropriate pain assessment tools that could identify the crucial variable that is the level of pain across different cases. Secondly, observations should be frequent in order to monitor every variation in the level of reported pain. In this study observations have been performed over three-hour intervals during the first 24 hours following the operation. The pain level is recorded in five levels from no pain to extremely painful (intolerable level of pain).

The estimation results suggest that the adopted model can be used for selecting an appropriate strategy to control post-surgical pain. In particular, the studied data indicate that time is an important factor in the evolution of pain as well as in the manifestation of analgesic effects. The estimated effect of patient characteristics such as age and gender are consistent with other studies, and can be used for a better allocation of resources among different cases. The model used in this paper can also be used to predict the duration of post-operative pain. To our knowledge, this question has not been addressed in the previous literature thus far. Although the hospital that is subject of this study has not joined “pain-free hospital” project, the conclusions of this paper and the adopted methodology can be easily related to other medical units involved in that project or other similar programs.

The rest of the paper is organized as follows. The next section briefly describes the tools employed to assess pain and the methodology applied here. Section 3 presents the data and estimation results. Section 4 concludes the paper with a discussion of the main results and recommendations for further research.

2. Measurement of post-operative pain

Post-surgical pain has been only recently recognized as an outcome measure of care quality. In a seminal paper in 1973, Marks and Sachar first noticed how pain was under-treated. They found a large discrepancy between the amount of analgesics ordered and the amount actually administered to surgical patients, which resulted in significant unrelied pain. Other studies, carried out after more than 10 years, documented a high incidence of uncontrolled and severe pain in hospitals (Donovan et al., 1987; Oden, 1989). The necessity to develop a precise strategy for pain management has been put in evidence later on, in a work by Kuhn et al. (1990), and later stressed in several studies examined by systematic reviews such as Kitson (1993) and Brown (2004).

Post-surgical pain management is complex because of considerable variations among patients regarding pain perception and the intensity and duration of pain as well as patients’ characteristics such as type of disease, age and psychological state and differences in nature of surgeries, in
analgesic techniques, and in the actual procedures used in managing pain (Lynch et al., 1997). The need to account for clinical and social factors together with factors due to patients’ characteristics has been stressed in a study by Short and Kluger (1998) about post-anaesthetic outcomes. Measures to control pain are based on pre-intervention anesthetics, together with the administration of analgesics after surgery, as soon as pain arises. There are many types of pain medication including opiates or narcotics, local anesthetics, anti-inflammatory medications, and many different delivery methods such as oral, intravenous and epidural administration. Depending on the procedure and clinical situation, a single medication or a combination, and one or more modalities of delivery may be used.  

Together with efficiency aspects for physicians and nurses in providing post-surgical assistance, medical literature has underlined qualitative aspects of pain management. Moreover, there is compelling evidence that unrelieved acute pain is directly linked to subsequent long-term pain problems. Unrelieved acute pain complicates recovery and could lead to more complications, longer hospital stays, greater disability and potentially long-term pain (Watt-Watson et al., 1999; Berry and Dahl, 2000; MacLellan, 2004). There is some evidence that extreme suffering from pain could weaken the patient’s immune system, while the risk of addiction to pain medication has been shown to be extremely low in patients who received medications for post-operative pain.

Inadequate management of pain not only decreases the quality of life but creates a financial burden on the health care system. A few studies carried out in American hospitals concluded that unrelieved pain could cost millions of dollars annually as a result of longer hospital stays, re-hospitalizations, and visits to outpatient clinics and emergency rooms (Grant et al., 1995; Sheehan et al., 1996). Zimberg (2003) describes some other economic advantages associated to pain management such as saving the additional cost of disability programs because of patients who are unable to work due to pain. The financial burden of analgesics is marginal compared to such considerable benefits for patients and medical units. Chauvin (1998) puts in evidence how there is actually a limited resort to analgesics, in spite of a minimal financial burden and considerable benefits for patients and medical units. Development of an effective pain management plan requires monitoring of the types of analgesic used, pain management regime selected, and knowledge of patients’ attitudes. In order to implement such plans, nurses need a minimum level of expertise in pain management and should be able to provide patients with means for describing and assessing their pain.

---

2 As stressed in the AHCPR (1992), an integrated approach to pain management includes cognitive behavioral interventions, systematic administration of appropriate medications, education of both staff and patients, as well as routine assessment of pain. All factors are recommended in order to make progress in providing pain management of a high standard.
Difficulties encountered in measuring the pain and its treatment have been emphasized in medical literature. An extensive review is that of Coll et al. (2004). Other studies have highlighted the necessity to interpret patients’ needs, in order to correctly assess pain and avoid discrepancies between intensity of pain reported by patients and that reported by physicians and nurses. For instance, Manias et al. (2002) report that the nurses’ assessment of post-surgical pain could differ from patients’ own assessment. It has been observed that nurses tend to over-evaluate pain when it is high and to under-evaluate when it is moderate: such a contrasting evaluation might be due to nurses’ personal experiences or to their perceptions about patients’ pain (Harmer and Davies, 1998). Pain assessment tools need to distinguish the true suffering patients from patients who are simulating (Torina, 2005). The questionnaires should not be too complicated to fill or misleading, allowing all patients to report correctly their level of pain. Hernandez-Quevedo et al. (2004) describe problems that might occur in reporting bias. Discrepancies across observations might depend either on variables as patients’ different income, age, education and personal experience of illness: this means that different groups tend to assess pain within their own specific context, using different reference points when they are responding to the same question. Hence, there could be any index shifts (that occur if the shape of the distribution of health, that is self-assessed remain the same, but there is a change in its location such that a parallel shift in all of the reporting thresholds for particular subgroups of the population is observed) or cut-point shifts (that implies that there is a change in the relative positions of the reporting thresholds). In order to overcome such problems, the authors suggest to apply models such as pooled ordered probit or random effects ordered probit. The literature stresses how pain can only be assessed on an individual basis using self-report pain and external indicators (Brown, 2000). Unfortunately, external indicators are frequently dismissed, as nurses and physicians rely heavily on validated pain assessment scales to assist in the monitoring of the patient’s pain.

A wide variety of pain scales have been used in the literature. Assessment scales mostly used include visual analogue scale (VAS) as in Gudex et al. (1996) and Fletcher et al. (1995),3 verbal rating scale (VRS) as in Bucknall and Manias (2001), verbal assessment scale (Downie et al., 1978), and finally face expression scale (Torina, 2005), the latter being particularly useful for children. The main limitation of all these scales is in the fact that they offer a unilateral approach to pain assessment, by omitting important factors, including those that may exacerbate or reduce pain and/or cognitive and behavioral changes. These factors include changes in sleeping or eating patterns, increased frustration, agitated or aggressive behaviors or withdrawal from family and

---

3 Fletcher et al. (1995) assessed post-surgical pain in a sample of 60 patients who underwent orthopaedic surgery. That study analyses the impact of the timing of certain analgesics, stressing the benefits of analgesics prior to the operation as opposed to post-surgical administration. In our data, the patients did not have any pre-surgical analgesics.
friends or avoidance of activity. However, thanks to their simplicity, the assessment scales are widely used in medical literature.

In this study a VRS with five levels has been used to measure pain. In this method, the patient is offered a set of adjectives, from which she is asked to choose the one that best describes the intensity of her pain. The main weakness of this scale is that it could be affected by the differences in patients’ age, language, educational and cognitive status (Bucknall and Manias, 2001). However, VRS measures provide a quick and simple method of pain assessment for acute pain. As we see later, the adopted model in this paper can take into account differences among patients through random coefficients. Nurses and physicians have recorded the assessment scales with regular observations in three-hour intervals during the 24 hours following the surgery, resulting in 8 observations for each patient. Five pain categories from “no pain” to “unbearable pain” have been considered. The observation sheets also include patient’s age, gender, time of the administered analgesics and a measure of patient’s satisfaction of the quality of care received during the surgical follow up. It should be noted that the analgesics administration during the observation period has not been random. Rather, it has been decided based upon the complexity of the surgical intervention and also the level of pain declared by the patient after the surgery. Therefore, it can be expected that more severe cases are more likely to have earlier analgesics and perhaps repeated treatments.

3. Model specification and estimation methods

Ordered discrete choice models such as ordered logit and probit have been often used for ordered categorical response variables that represent groups of continuous variables, such as income groups. The application of these models can be extended to categorical variables that have an “assessed” order, such as “the extent of pain relief after treatment” (Anderson, 1984). These variables are referred to as assessed, ordered variables. In this paper we use an ordered logit model as explained in Greene (2003) and Wooldridge (2002). In this model it is assumed that the individual choices are based on a latent variable, which can be considered as a measure of the individual’s random utility. This latent variable is defined as a function of explanatory variables. In the context of this study, the latent variable \( y_{it}^* \) is defined as the pain level of patient \( i \) at period \( t \). The latent variable \( y_{it}^* \) is assumed to be a continuous additive function of a vector of time-variant factors denoted by

---

4 Recently, this kind of models have been used in a RAND study assessing patients’ use and preferences for information about the technical and interpersonal quality of care delivered by individual physicians (Center of Excellence for the Study of Healthcare Provider Behavior, 2005). See also McKelvey and Zavoina (1975) for an earlier application. To our knowledge, there are few studies applying these models in the context of post-operation pain.
\( X_{it} \), and of a vector of patient’s characteristics represented by \( Z_{it} \). The vector \( X_{it} \) includes, for instance, the time period after the operation and the number of hours after the application of analgesics.

Considering the additive stochastic term the latent pain level can be written as:

\[
y^*_it = \alpha + X_{it} \beta + Z_{i}\gamma + \varepsilon_{it},
\]

where subscripts \( i \) and \( t \) respectively represent a given patient and the number of hours after her operation; \( \alpha, \beta \), and \( \gamma \) are the parameter vectors to be estimated; and \( \varepsilon_{it} \) is an iid stochastic error term that represents the unobserved factors. We assume that patients translate their continuous pain level (latent to us) to a finite number of pain levels (\( J \)) asked in the survey. The probability of choosing pain level \( j \) is defined as:

\[
Pr(y_{it} = j) = Pr(\mu_{j-1} < y^*_it \leq \mu_j) ; \quad -\infty = \mu_0 < \mu_1 < \ldots < \mu_J = +\infty, j \in \{1, 2, \ldots, J\},
\]

where \( y_{it} \) is the discrete response variable, that is patient \( i \)'s pain level, \( t \) hours after her operation; and \( \mu_j \)'s are the threshold parameters. Assuming a logistic probability distribution for the error term \( \varepsilon_{it} \), the above probability can be written as:

\[
Pr(y_{it} = j) = \frac{1}{1 + \exp(-\mu_j + X_{it} \beta + Z_{i}\gamma)} - \frac{1}{1 + \exp(-\mu_{j-1} + X_{it} \beta + Z_{i}\gamma)},
\]

that can be estimated using the maximum likelihood estimation method.

In the above model it is assumed that all individual patients use a similar measure of pain. This is a restrictive assumption in that individuals differ with respect to their sensitivity facing pain. Moreover, a given patient’s perception of pain could vary depending on her expectation regarding the seriousness of her operation. For instance a patient with a complex operation could complain less than a similar patient who had a relatively simple operation but who suffers the same level of pain. This restrictive assumption can be partly relaxed by considering that the intercept parameter \( \alpha \) varies across different individuals. Moreover, the effect of independent variables \( X_{it} \) could also vary across patients. For instance, the development of pain after the operation and the effect of analgesics depend on the type and complexity of the operation as well as other unobserved patient characteristics. Such unobserved heterogeneity can be partly accounted by randomizing coefficient vector \( \beta \) across patients.

Assuming that \( \alpha \) and \( \beta \) are normal random variables across patients, the model in Equation (1) can be written as:
As seen in Equation (2), the probability of choosing a certain pain category is assumed to be a function of a continuous latent variable ($y^*_i$) that can be considered as the patient’s pain level. Ordered logit model is a proportional odds model in the sense that the odds ratio of switching from an alternative to the next one is invariant to the alternative. Namely, the probability ratio
\[
\frac{\Pr(y^*_i > y)}{\Pr(y^*_i \leq y)} = \exp(\alpha_i + X_i \beta_i + Z_i \gamma),
\]
is shown to be equal to: $\exp(\alpha_i + X_i \beta_i + Z_i \gamma)$, thus not a function of the actual pain category $j$. This assumption implies that for any given patient, as long as the independent variables remain unchanged, the probability of moving up one level in the pain categories does not depend on the actual pain category.

In this paper, we consider both the simple model with constant parameters as in (1) and the alternative model with random coefficients as expressed in (4). The patient characteristics included in the model ($Z_i$) are gender, age with six cohort dummies, a dummy variable indicating whether more than one analgesic have been used, and a dummy variable controlling for the quality of care (poor or acceptable) as perceived by the patient.

\[y^*_i = \alpha_i + X_i \beta_i + Z_i \gamma + \epsilon_w\]

\[\alpha_i \sim N(\alpha, \sigma^2_{\alpha})\]

\[\beta_i \sim N(\beta, \sigma^2_{\beta})\]

and equation (3) is changed by substituting $\alpha$ and $\beta$ with $\alpha_i$ and $\beta_i$ respectively, resulting in:
\[
\Pr(y_w = j) = \frac{1}{1 + \exp(-\mu_j + \alpha_i + X_i \beta_i + Z_i \gamma)} - \frac{1}{1 + \exp(-\mu_j + \alpha_i + X_i \beta_i + Z_i \gamma)}.
\]
The time-variant explanatory variables ($X_{it}$) include the number of hours after the operation and the number of hours elapsed after the analgesic as well as their respective squares. Considering this specification, the adopted model is based on the assumption that after controlling for observed patient characteristics and stochastic errors, the evolution of pain level ($y^*$) follows a quadratic function of the following form:

$$y^*_t = y^*_0 + \theta_1 t + \theta_2 t^2 + \omega_1 s + \omega_2 s^2,$$

where $\theta_1, \theta_2, \omega_1$ and $\omega_2$ are the parameters to estimate; $y^*_t$ and $y^*_0$ are respectively the pain level at $t$ and the initial pain level at $t=0$ (right after the operation); and $s$ is the time elapsed after the analgesic (or after the most recent analgesic for patients with several analgesics). Note that in the random coefficient model explained above, the pain function (6) is assumed to be patient-specific, that is parameters $\theta_1, \theta_2, \omega_1$ and $\omega_2$ vary across individual cases.

4. Data

The questionnaire for monitoring pain includes the following information: name, age, sex, type of surgery, level of pain at the end of the operation, level of pain in the hours following the surgery, quality of care with optional comments. The questionnaire has been filled by physicians (anesthesiologists) and nurses. Data have been collected in years 2002 and 2003, in the surgical division of orthopedics department at “Buccheri La Ferla” hospital in Palermo, Italy.

Patients who have been administered a continuous analgesic such as epidural or femoral, have been excluded. Hence, the patients considered for the final estimations have been treated with one of the following analgesics: Diclofenac; Meperidina; Paracetamolo; Tavor; Toradol (tromethamine ketorolac). Such analgesics are very commonly used in the treatment of post-surgical pain, do not have severe contraindications and are immediately available from the pharmacy at the hospital.

The type of analgesics is constant for each given case and has been selected based on the complexity and type of the operation as well as the patient’s individual medical data. We assume that these analgesics have a similar appeasement effect. Given that different analgesic types might have completely different time-effects, the above assumption may be restrictive. However, since the random coefficients vary across individual patients, they could capture these differences to the extent that the time effects can be modeled with the same functional form for all the above analgesic types.

The final sample consists of 49 orthopedic cases including 24 male and 25 female patients with an overall average age of 59.6 years old. The patients’ age varies from 15 to 86 years old for men and 22 to 89 for women with a median value of 67 for both male and female patients.
Several factors might influence intensity, characteristics and length of post-surgical pain. One can consider the complexity of surgical intervention, its duration, patients’ psycho-physiological conditions and the quality of preparation. In spite of a low level of pain immediately after the operation, pain could develop with greater intensity in the following hours.

The pain level is defined in five categories labeled from 0 (no pain) to 4 (intolerable pain). Collection of data has been done through questionnaires. Patients have been interviewed by an anesthetist together with a nurse. They have been told about the studies that would have been carried out using the information provided, stressing the relevance that this could have for quality of care. Patients have been asked about how they would have categorized the level of pain and the quality of care received. While the latter has been classified into three levels (poor, good and very good), level of pain has been categorized into five levels: Level 0 corresponds to no pain at all, at rest or in movement; level 1 indicates no pain at rest, but slight pain while moving; level 2 represents slight pain at rest and high pain in movement; level 3 indicates high pain at rest and intolerable pain when moving; and finally, level 4 means intolerable pain even at rest.

Table 1 summarizes reported values for pain as monitored at three-hour intervals after the operation. As this table indicates, most patients do not declare any pain immediately after the surgery, which is an expected outcome because of the effects of anesthesia. Table 1 also lists the number patients that had analgesics. These numbers show that in most of the cases, patients do not receive analgesics within a several hours after their operation. In total 60 analgesics have been administered to the patients in the sample, most of which have been performed about 12 to 18 hours after the operation.

As it is seen in the table, the maximum value of pain is often limited to level 2 that is, slight pain at rest and high pain at movement. The peak in the level of pain (increase in the number of patients declaring a level of pain higher than 1) arises in the time interval between 6 to 12 hours after the surgery. Number of patients declaring a level of pain more than 2 increases from 10 to 15 in these time intervals. Pain rises again for some patients 18 hours after the surgery (4 patients declare to feel intolerable pain (level 4) and ask for further analgesics. In some of these cases the previous analgesic might have had no effect or its effect might have vanished. Overall, the observation of maximum pain level is quite rare. To avoid the problem of small sample, these values, that are limited to 2.6 percent of the observations, are stacked to the next lower level (level 3).

Patients’ subjective assessment about the quality of care is described in three categories: poor, good and very good. However, most patients have assessed a good quality. Only five patients reported a poor quality and only 2 assessed a very good quality. We constructed a new dummy variable
(POOR) that takes 1 if the care is evaluated as bad and 0 otherwise. As mentioned earlier, the following patient characteristics are also included in the model: dummy variable MALE for patient’s sex, age dummies in 6 categories, and a dummy variable (REPEAT) that takes 1 if the patient has received more than one analgesics during the sample period and 0 otherwise. T_SURG is the number of hours after the operation and T_ANAL is the number of hours after the administration of the most recent analgesic. A descriptive summary of the variables included in the analysis is provided in Table 2.

[Tables 2 here]

5. Estimation results

The estimation results are listed in Table 3. Comparing the estimated coefficients across the two models indicate significant differences between corresponding coefficients, suggesting that ignoring the unobserved heterogeneity across patients could create significant biases in the estimated parameters. As it can be seen in the tables, in the mixed models with random parameters, the standard deviation of virtually all the random coefficients is statistically significant. Moreover, the significant variations in the coefficients of the time variables suggest that the development of pain and the effect of analgesics vary depending on the case.

[Table 3 here]

As Table 3 shows, the results suggest that male patients show lower pain levels compared to female cases. This result is consistent with the findings of medical literature, according to which the way men and women experience pain is different. More importantly, the positive effect of time after surgery suggests that the pain increases with time at least within the 24 hours period recorded in the data. The negative coefficient of the square term indicates however, that post-operative pain increases at a decreasing rate. On the other hand, the negative and significant effect of time after the analgesics suggests that these interventions are effective in lowering pain, but that their effects appear gradually. Here, the second order effect is positive, suggesting that the appeasing effect diminishes with time.

The main result of this analysis is that time is an important factor in pain relief. Table 3 shows that the coefficient of T_ANAL is lower in absolute value than that of T_SURG, whereas the coefficients of the corresponding square terms are more or less similar in absolute values. This

---

7 Some of the differences between men and women in responding to pain and analgesics are documented in Levine et al. (1998).
suggests that the analgesics will be most effective if they are administered as early as possible that is, immediately after the operation or possibly even a few hours prior to the operation.

Age variables show a peculiar effect on pain level, which cannot be explained clearly. For instance, while groups of 40-49, 50-59 and those older than 80 years old feel more pain than patients younger than 40 years old, patients aged between 60 and 80 develop more pain than the youngest group but less pain than the middle age patients. These results could suggest that age variables actually capture other unobserved differences across patients. Given the limited number of patients in the sample and the relatively high variation within each age group, the present data cannot provide reliable information about the age effects. The results also indicate a positive effect for repeated analgesics but this effect is not statistically significant in eth random-coefficient model. The positive effect in the first model can be explained by the higher severity of the case where repeated analgesics were applied, which is partly captured by the random coefficients in the second model.

The results in Table 3 also suggest that subjective quality has a negative impact on pain level: however, it is a tautological result that patients who reported a poor quality of care must have experienced a higher pain. It is also possible that in the patient’s assessment about care that is the psychological factors linked to nurses’ and physicians’ assistance have conditioned the perception of pain.

The estimated marginal effects at the sample mean are provided in Table 4. In general, these estimates suggest that the marginal effect of virtually all the variables has a lower magnitude for high levels of pain. This might suggest that in lower pain levels, a larger fraction of variation in pain can be explained by the explanatory variables. Particularly, this can be considered as suggestive evidence that analgesics can be more effective when pain is moderate.

4. Conclusions

This study has been aimed at modeling the development of post-surgical pain for a sample of 49 orthopedic patients who underwent surgery in the period 2002-2003. The probability of patients to experiment a certain level of pain has been estimated using an ordered discrete choice model. Two alternative models have been considered: a simple ordered logit model and a mixed ordered logit model with random parameters. The latter model considers some of the unobserved factors, especially those related to differences in pain perception among patients. The evolution of pain has been considered with two main time variables measuring respectively the time after the operation and the hours elapsed after the administration of the analgesics.
The results indicate that the perception of pain and the analgesic effects might vary across individual cases. This paper shows that such unobserved heterogeneity can be partly taken into account using random coefficient models. The present analysis also highlights the effect of patient characteristics such as gender.

The results suggest a significant appeasing effect for analgesics. The analysis shows that time has a crucial effect in the development of the post-operative pain as well as the analgesic effects: both these effects appear gradually but at a decreasing rate. The results also suggest that pain can be lowered considerably or, in some cases, prevented altogether if the analgesics are applied in a timely manner. The results of this paper suggest that analgesics are most effective if they are administered as early as possible after the operation.

The present paper is a contribution in investigating the development of post-surgical pain. Moreover, information obtained by this study can be useful for physicians and nurses in developing an effective strategy in the management of pain. As a result, there could be an improvement of care, savings in resources and a better health outcome.

References


Table 1 – Distribution of patients by pain level and by time after surgery

<table>
<thead>
<tr>
<th>Level of pain</th>
<th>3 h</th>
<th>6 h</th>
<th>9 h</th>
<th>12 h</th>
<th>15 h</th>
<th>18 h</th>
<th>21 h</th>
<th>24 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>39</td>
<td>38</td>
<td>28</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>11</td>
<td>13</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total number of patients</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Patients that had an analgesic</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>27</td>
<td>8</td>
<td>5</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2 – Descriptive statistics (392 observations from 49 patients)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEVEL OF PAIN</td>
<td>0.717</td>
<td>0.97</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>T_SURG</td>
<td>13.50</td>
<td>6.88</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>T_ANAL</td>
<td>6.29</td>
<td>4.41</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>REPEAT</td>
<td>0.163</td>
<td>0.37</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MALE</td>
<td>0.490</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>59.6</td>
<td>20.5</td>
<td>15</td>
<td>89</td>
</tr>
<tr>
<td>AGE 30_39</td>
<td>0.082</td>
<td>0.27</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE 40_49</td>
<td>0.122</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE 50_59</td>
<td>0.102</td>
<td>0.30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE 60_69</td>
<td>0.143</td>
<td>0.35</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE 70_79</td>
<td>0.245</td>
<td>0.43</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE &gt; 80</td>
<td>0.184</td>
<td>0.38</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>POOR</td>
<td>0.102</td>
<td>0.30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Model with constant parameters</td>
<td>Model with random parameters</td>
<td>Coefficient Mean</td>
<td>Coefficient's Std. Dev.</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>T_SURG</td>
<td>0.489** (0.088)</td>
<td>0.643** (0.0704)</td>
<td></td>
<td>0.0956** (0.0077)</td>
</tr>
<tr>
<td>T2_SURG</td>
<td>-0.0109** (0.0031)</td>
<td>-0.0182** (0.0024)</td>
<td></td>
<td>0.0006 (0.0003)</td>
</tr>
<tr>
<td>T_ANAL</td>
<td>-0.473** (0.117)</td>
<td>-0.468** (0.0896)</td>
<td></td>
<td>0.2029** (0.017)</td>
</tr>
<tr>
<td>T2_ANAL</td>
<td>0.0182** (0.0052)</td>
<td>0.0147** (0.0040)</td>
<td></td>
<td>0.00007 (0.0008)</td>
</tr>
<tr>
<td>REPEAT</td>
<td>0.903** (0.315)</td>
<td>0.0188 (0.258)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MALE</td>
<td>-0.839** (0.225)</td>
<td>-1.264** (0.194)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE 30_39</td>
<td>0.334 (0.577)</td>
<td>1.209* (0.497)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE 40_49</td>
<td>1.748** (0.492)</td>
<td>4.174** (0.457)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE 50_59</td>
<td>1.913** (0.511)</td>
<td>4.168** (0.470)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE 60_69</td>
<td>0.969* (0.499)</td>
<td>1.115* (0.445)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE 70_79</td>
<td>0.937* (0.450)</td>
<td>2.079** (0.397)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE &gt; 80</td>
<td>1.806** (0.4549)</td>
<td>3.266** (0.407)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POOR</td>
<td>2.077** (0.400)</td>
<td>2.545** (0.320)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSTANT</td>
<td>-3.556** (0.707)</td>
<td>-5.522** (0.591)</td>
<td>1.129** (0.102)</td>
<td></td>
</tr>
<tr>
<td>Threshold parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>μ₁</td>
<td>1.682** (.141)</td>
<td>2.635** (.128)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>μ₂</td>
<td>3.014** (.220)</td>
<td>4.329** (.199)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at 5%; ** Significant at 1%.
Standard errors are given in brackets.
The dependent variable is the level of pain in four categories from zero to 3.
Table 4 – Marginal effects estimated at the sample mean

<table>
<thead>
<tr>
<th>Simple ordered logit model</th>
<th>Y=0</th>
<th>Y=1</th>
<th>Y=2</th>
<th>Y=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_SURG</td>
<td>-.1200</td>
<td>.0667</td>
<td>.0362</td>
<td>.0170</td>
</tr>
<tr>
<td>T2_SURG</td>
<td>.0027</td>
<td>-.0015</td>
<td>-.0008</td>
<td>-.0004</td>
</tr>
<tr>
<td>T_ANAL</td>
<td>.1161</td>
<td>-.0646</td>
<td>-.0351</td>
<td>-.0165</td>
</tr>
<tr>
<td>T2_ANAL</td>
<td>-.0045</td>
<td>.0025</td>
<td>.0013</td>
<td>.0006</td>
</tr>
<tr>
<td>REPEAT</td>
<td>-.2219</td>
<td>.0988</td>
<td>.0805</td>
<td>.0426</td>
</tr>
<tr>
<td>MALE</td>
<td>.2030</td>
<td>-.1112</td>
<td>-.0621</td>
<td>-.0297</td>
</tr>
<tr>
<td>AGE 30_39</td>
<td>-.0830</td>
<td>.0426</td>
<td>.0271</td>
<td>.0133</td>
</tr>
<tr>
<td>AGE 40_49</td>
<td>-.3983</td>
<td>.1040</td>
<td>.1757</td>
<td>.1187</td>
</tr>
<tr>
<td>AGE 50_59</td>
<td>-.4237</td>
<td>.0865</td>
<td>.1944</td>
<td>.1428</td>
</tr>
<tr>
<td>AGE 60_69</td>
<td>-.2373</td>
<td>.1014</td>
<td>.0882</td>
<td>.0476</td>
</tr>
<tr>
<td>AGE 70_79</td>
<td>-.2299</td>
<td>.1076</td>
<td>.0806</td>
<td>.0417</td>
</tr>
<tr>
<td>AGE &gt; 80</td>
<td>-.4153</td>
<td>.1247</td>
<td>.1761</td>
<td>.1145</td>
</tr>
<tr>
<td>POOR</td>
<td>-.4494</td>
<td>.0739</td>
<td>.2101</td>
<td>.1654</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random coefficient ordered logit model</th>
<th>Y=0</th>
<th>Y=1</th>
<th>Y=2</th>
<th>Y=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_SURG</td>
<td>-.1439</td>
<td>.1219</td>
<td>.0176</td>
<td>.0043</td>
</tr>
<tr>
<td>T2_SURG</td>
<td>.0036</td>
<td>-.0031</td>
<td>-.0004</td>
<td>-.0001</td>
</tr>
<tr>
<td>T_ANAL</td>
<td>.1048</td>
<td>-.0889</td>
<td>-.0128</td>
<td>-.0031</td>
</tr>
<tr>
<td>T2_ANAL</td>
<td>-.0033</td>
<td>.0028</td>
<td>.0004</td>
<td>.0001</td>
</tr>
<tr>
<td>REPEAT</td>
<td>-.0042</td>
<td>.0036</td>
<td>.0005</td>
<td>.0001</td>
</tr>
<tr>
<td>MALE</td>
<td>.2755</td>
<td>-.2307</td>
<td>-.0360</td>
<td>-.0088</td>
</tr>
<tr>
<td>AGE 30_39</td>
<td>-.2915</td>
<td>.2236</td>
<td>.0539</td>
<td>.0140</td>
</tr>
<tr>
<td>AGE 40_49</td>
<td>-.7177</td>
<td>.1511</td>
<td>.3627</td>
<td>.2038</td>
</tr>
<tr>
<td>AGE 50_59</td>
<td>-.7054</td>
<td>.1216</td>
<td>.3669</td>
<td>.2169</td>
</tr>
<tr>
<td>AGE 60_69</td>
<td>-.2672</td>
<td>.2105</td>
<td>.0452</td>
<td>.0115</td>
</tr>
<tr>
<td>AGE 70_79</td>
<td>-.4757</td>
<td>.3476</td>
<td>.1008</td>
<td>.0273</td>
</tr>
<tr>
<td>AGE &gt; 80</td>
<td>-.6613</td>
<td>.3359</td>
<td>.2407</td>
<td>.0846</td>
</tr>
<tr>
<td>POOR</td>
<td>-.5513</td>
<td>.3141</td>
<td>.1803</td>
<td>.0569</td>
</tr>
</tbody>
</table>