Spatial modelling of origin-destination commuting flows in Switzerland

Thomas Schatzmann*
Institute for Transport Planning and Systems, ETH Zurich
HIL F 32.1, Stefano-Francini-Platz 5, 8093 Zurich, Switzerland
Tel: +41 44 633 30 88; Email: thomas.schatzmann@ivt.baug.ethz.ch

Georgios Sarlas
Institute for Transport Planning and Systems, ETH Zurich
HIL F 51.3, Stefano-Francini-Platz 5, 8093 Zurich, Switzerland
Tel: +41 44 633 37 99; Email: georgios.sarlas@ivt.baug.ethz.ch

Prof. Dr. Kay W. Axhausen
Institute for Transport Planning and Systems, ETH Zurich
HIL F 31.3, Stefano-Francini-Platz 5, 8093 Zurich, Switzerland
Tel: +41 44 633 39 43; Email: axhausen@ivt.baug.ethz.ch

*Corresponding author

Word count: 6,128 words text + 4 tables x 250 words (each) = 7,128 words

August 1, 2018
ABSTRACT

We present a direct demand modelling approach for origin-destination (OD) public transportation commuting flows between municipalities in Switzerland. The purpose is to improve the gravity modelling approach for OD flows by applying a spatial autoregressive regression model and testing different spatial weighting schemes. Besides the usual characteristics to explain commuting, we include a variable based on mean income differences to examine interregional demand patterns. In addition, we treat for the endogenous nature of the newly constructed variable and test its ability to serve as the basis for the construction of a spatial weight matrix, thus replacing the commonly used travel time / distance metric. We apply Ordinary Least Squares (OLS), Generalized Method of Moments (GMM) and Intrumental Variable (IV) estimators to obtain unbiased and consistent parameter estimates. We compare in-sample predictions of the models among each other and to the flows of the National transport model. We use data from the 2000 Federal Census and found significant spatial dependence in the residuals of the gravity model and thus the need for spatial regression models. We use a valid set of instruments to account for endogeneity and show that income differences are underestimated in the gravity and spatial models if assumed exogenous. Neighbouring municipalities affect flows under consideration positively at origins and negatively at destinations. Last, the spatial autoregressive models relying on a combination of origin- and destination-centric weight matrix outperform the gravity models in terms of the predictive accuracy when network and economic distance weights are used.

Keywords: commuting flows, spatial autoregressive regression, origin-destination flow modelling, instrumental variable, endogeneity, endogenous weight matrix
INTRODUCTION

From an economic perspective, commuting in the sense of travel-to-work trips can be seen as the consequence of a spatial discrepancy between work and living (1). If the spatial distribution of employment and that of workers is different, commuting will be the outcome of such a mismatch. It describes human interaction as a movement of people from a fixed origin to a destination that results from a previous decision (2). Distance decay and gravity forces are two important aspects explaining human interaction in terms of commuting. The demand for transport is derived from people’s activities. Ortuzar and Willumsen (3) specify: "It is the distribution of activities over space which makes for transport demand". Thus, different methods for treating distance and allocating origins and destinations over space and time are essential in transport analysis.

Transport demand modelling aims at replicating actual travel flows and is based on conventional data such as Federal Census and travel diary surveys. The classical four step transport model originated in the 1950s according to Ortuzar and Willumsen and generally tries to estimate the number of trips for different travel modes and routes taken between any pair of origin and destination zones in the study area. The modelling of origin-destination (OD) commuting flows corresponds to the trip distribution step which is most commonly addressed through the use and estimation of a gravity model, estimated by means of ordinary least squares estimators. However, the estimation techniques of such models require the independence of observations (flows) in order for the basic assumptions to hold, resulting in biased and inconsistent estimates if the conditions do not hold. A common misconception when dealing with spatial data is that independence holds and thus no further testing is done towards that direction. Previous research (4) has acknowledged the implications of that and has suggested a modelling approach for addressing the issue of spatial autocorrelation in the context of OD flows.

It remains an open research question with how to deal with potential endogeneity issues in the spatial autoregressive components of the model if they rely on economic characteristics rather than euclidean or network distance (5).

Objective

In this paper, we develop a direct demand model for public transport commuting that combines the first three steps of the classical transport model. The goal is to apply a spatial autoregressive regression model with different weighting schemes to the case of public transport commuting flows in Switzerland. We test whether the spatial weight matrix and/or any regressor is endogenous and provide a solution based on existing literature in such a case.

Literature review

Spatial interaction models

The well-known gravity model has probably emerged as the most frequently used spatial interaction model in migration, trade and transportation demand research. Gravity models incorporate explanatory variables that represent characteristics of both origin and destination zones as well as a function of distance between them. Wilson (6) proposes a gravity model based on a statistical equilibrium concept.

\[ T_{ij} = A_i B_j O_i D_j \exp(-\beta c_{ij}) \]

where

\[ A_i = \frac{1}{\sum_j B_j D_j \exp(-\beta c_{ij})} \quad \text{and} \quad B_j = \frac{1}{\sum_i A_i O_i \exp(-\beta c_{ij})} \]

Interaction between two zones i and j, \( T_{ij} \), is described by origin (\( O_i \)) and destination (\( D_j \)) characteristics, a decreasing function of distance or generalized cost of travel between those zones.
as well as two balancing factors $A_t$ and $B_t$. Wilson’s theorem provides that the frequency of trips by distance will be negative exponentially distributed so that the logarithmic weighting of distance required by the gravity model is derived for the first time (8). The notion that conventional spatial interaction models make use of distance functions to capture spatial dependence in interaction flows has been challenged in recent years since spatial autocorrelation (local distance) and distance decay (global distance) effects may be confounded (4, 17). Thus, equation (1) should include spatial linear operators applied on flows to account for spatial autocorrelation in its origin and destination geographic distributions.

$$(I - \rho W)T \quad \text{and} \quad T(I - \rho_d W)$$

where, in matrix notation, $I$ is an n-by-n identity matrix, $T$ is the n-by-n flows matrix, $W$ is a row-standardized geographic connectivity matrix, and $\rho$ and $\rho_d$ are the spatial autocorrelation parameters for the origin and destination geographic distributions. Griffith and Jones show the existence of statistically significant, positive spatial autocorrelation in gravity model parameters (9). Furthermore, they state that flows from an origin are "enhanced or diminished in accordance with the propensity of emissiveness of its neighbouring origin locations". Flows associated with a destination are "enhanced or diminished in accordance with the propensity of attractiveness of its neighbouring destination locations".

**Spatial econometric models**

Spatial econometrics focus on spatial dependence/autocorrelation of observations which is determined by relative space or relative location in cross-sectional data sets. Thus, because of the multidirectional nature of dependence in space compared to one direction in time, econometric results from time series analysis do not carry over straightforward to spatial dependence in cross-sectional samples. A general spatial linear regression model for cross-sectional data is given by Anselin (10):

$$y = \rho W_1 y + X\beta + \epsilon, \quad \epsilon = \lambda W_2 \epsilon + \mu$$

with $\mu \sim N(0, \Omega)$ and $\Omega_{ii} = h_i(z\alpha), \ h_i > 0$. $\beta$ is a parameter associated with exogenous dependent variables $X$, $\rho$ is the coefficient of the spatially lagged dependent variable $y$ and $\lambda$ is the coefficient in a spatial autoregressive structure for the disturbance $\epsilon$ which is taken to be normally distributed with a general diagonal covariance matrix $\Omega$. $W_1$ and $W_2$ are either standardized or unstandardized spatial weight matrices. The notion of neighbourhood and nearest neighbour was introduced to see whether and if so, how strong other units in the spatial system influence the particular unit under consideration. Cliff and Ord (19) extend the concept of binary contiguity by Moran (20) and include a general measure of the potential interaction between two spatial units such as inverse distance or negative exponentials of it combined with shared border length. In general, euclidean distance measured between centroids need not be the most accurate measure for weight matrices. For instance in transportation networks like public transport or roads, travel time (network distance) reflects an access variable and is often more appropriate than simple euclidean distance due to increasing mobility and different speeds of traffic modes (11).

Different spatial model structures result when restricting parameters of the general model in equation (3) to zero, what is called filtering in the existing literature (10):

Spatial autor. model (SAR): $y = \rho W_1 y + X\beta + \epsilon$, \quad with $\alpha = 0 \& \lambda = 0$ (4)
Spatial error model (SEM): \( y = X\beta + (I - \lambda W_2)^{-1}\mu, \) with \( \alpha = 0 \) & \( \rho = 0 \)

SAR & SEM = SAC: \( y = \rho W_1 y + X\beta + (I - \lambda W_2)^{-1}\mu, \) with \( \alpha = 0 \)

Four additional models arise when spatial heterogeneity is incorporated by restricting \( h_i(z\alpha) \) to specific forms. The spatial durbin model (SDM), where also \( X \) is weighted, is not listed here. Note that we use model names from LeSage and Pace (12).

**Modelling origin-destination (OD) flows**

OD models involve \( n^2 \) OD pairs in contrast to typical spatial econometric models that contain samples of \( n \) regions with each region being an observation (4). They aim at explaining variation in the levels of flows between the \( n^2 \) OD pairs. It has remained difficult how to structure the connectivity of \( n^2 \) OD pairs, which is why they propose a spatial weight structure in way that is consistent with standard spatial autoregressive models. Interestingly, LeSage and Pace set up a commuting flow example to illustrate the justification for using spatial lags in OD models (13). They develop spatial model specifications for origin-destination flows that are based on a general spatial autoregressive model which takes into account origin \( (a) \), destination \( (b) \) and so-called origin-to-destination \( (c) \) based dependence:

\[
y = \underbrace{\rho_o W_0 y}_{(a)} + \underbrace{\rho_d W_d y}_{(b)} + \underbrace{\rho_w W_w y}_{(c)} + \alpha l_{n^2} + X_o \beta_o + X_d \beta_d + \gamma g + \varepsilon
\]  

(5)

As before, different models from non-spatial up to fully spatial dependence result from restricting parameters \( \rho_i, \ i = o, d, w \). \( X_o \) and \( X_d \) represent origin and destination characteristics whereas \( g \) reflects a distance function. LeSage and Thomas-Agnan (13) examine commuting flows for 60 regions in France and impressively point out that the aspatial model underestimates the total impact of increasing residents or jobs on commuting flows that arises because of network effects, also known as spatial spillovers. Recent work has included spatial dependence in models in various fields: Porojcan (14), Lee and Pace (15), Cushing and Poot (16), Griffith (17) and Bolduc et al. (18).

**Endogeneity**

Endogeneity is a severe issue in any model as it results in biased and inconsistent parameter estimates and makes inference invalid. Kelejian and Piras emphasize that weighting matrices are typically assumed exogenous in spatial econometric models (5). Already Anselin (10) stressed that when the spatial interaction phenomenon under consideration is determined by factors such as purely economic variables, spatial weights linked to the physical features of spatial units are less meaningful. As an important extension of earlier work by Kelejian and Prucha (21, 22), Drukker et al. (23) worked out a two-step generalized method of moments and instrumental variable estimator (2IV/GMM) which takes endogenous regressors and heteroscedastic innovations into account, in addition to a spatially lagged variable. Kelejian and Piras (5) specify a general spatial panel data model with a spatially lagged dependent variable in terms of an endogenous weight matrix that accounts for the influence of economic variables in neighbouring regions. In their empirical application they study the dynamic demand for cigarettes based and show that buyers of cigarettes make purchases in the neighbouring states when there is a price advantage in doing so. In a transportation study, Zhou et al. (24) apply a spatial autoregressive binary probit model to a firm relocation choice problem, where the matrix weights rely on geographic and economic distance, \( w_{ij} = d_{ij} \times |inc_i - inc_j| \), what potentially induces endogeneity issues. As a result, a higher level of autocorrelation and endogeneity lead to a better predictive accuracy of the model.
\textit{Estimation}

It has been common to estimate the parameters of spatial interaction models by linearising the equations in terms of their parameters. Taking the logarithm of both sides of a standard gravity model as in Sen and Smith (7) yields equation (5) without spatial lags. The parameters are estimated by means ordinary least squares (OLS). Commuting flow data typically belong to the class of count data due to its non-negative character. Flowerdew and Aitkin (26) treat spatial interaction as counts and propose to take the Poisson distribution as opposed to the Normal distribution, which seems reasonable as for example the number of commuters must be non-negative and that there is a constant probability of an individual to travel from zone i to j. Even though Poisson regression models seem to be appropriate to model commuting flows, the problem of over-dispersion can not be addressed due to the restrictive assumption of equidispersion of those models. In practice, over-dispersion of count data arises where the conditional variance is larger than the conditional mean of the dependent variable \( T_{ij} \). One way to correct for it is to assume a negative binomial or a zero-inflated negative binomial distribution of \( T_{ij} \), which is a generalisation of the Poisson distribution with an additional dispersion parameter in order to allow the conditional variance to exceed the conditional mean.

Since standard OLS estimation leads to an inconsistent estimator due to the spatial lag of the dependent variable that is correlated with the disturbance term, other methods of estimation have to be considered. The log-likelihood function for the model in equation (5) is the basis for both maximum likelihood (ML) and Bayesian estimation (BAYES), where standard algorithms are applied to calculate it. However, as the number of observations increases these algorithms become more difficult, because it requires large amounts of computer memory to calculate the log-determinants of \( n^2 \)-by-\( n^2 \) matrices. LeSage and Pace (4) found a solution to such problems by reducing the troublesome log-determinant calculation to one involving only traces of n-by-n matrices. An approach by Kelejian and Prucha proposes a generalized spatial two-stage least squares (GS2SLS) procedure to estimate the same model, resulting in a consistent and asymptotically normal estimator. It has the advantage of less specified distributional assumptions compared to the ML estimator. A year later, they derived a generalized moment estimator (GMM) in a similar set up. Furthermore, both procedures were shown to be feasible with large samples (22). In a follow-up paper the same authors introduce a new class of GMM estimators for the autoregressive parameter of a spatially autoregressive disturbance process allowing for innovations with unknown heteroskedasticity.

The problem of zero flows and excess zeros can be tackled differently. In the context of OD flow modelling, ML estimation is not appropriate for cases where a large number of zero flows exist. ML estimates need a normally distributed dependent variable or at least one that can be transformed to achieve normality. Lambert et al. (29) use a Poisson estimation approach for cross-sectional spatial autoregressive models in order to account for the limitation of the ML method. Ranjan and Tobias (30) treat zero flows using a Tobit model for censoring, which they extend to the case of spatial autoregressive interaction models. There is progress in spatial modelling of discrete and continuous dependent variables, but currently only for maximum likelihood and Bayesian approaches (31).

\textbf{METHOD}

We proceed in three main steps. Following the structure of the literature review, we first estimate a log-transformed gravity model by means of OLS. An important difference to classic interaction modelling arises in how a n-by-n flow matrix translates into a n\(^2\) vector of flows, which defines the OD model structure. We employ an origin-centric ordering, where the first n elements in the
stacked flow vector reflect flows from origin 1 to all n destinations. The last n elements of this vector represent flows from origin n to destinations 1 to n. We use Moran’s I test to show the existence and strength of spatially autocorrelated residuals of the model for both the origins and destinations. Since the former produce transport demand and the latter attract it, we create weighting matrices for each and the sum of them to capture two spatial effects: Origin- and destination based dependence. For each matrix, we determine the threshold up to what spatial extent there is statistically significant autocorrelation.

In a second step, we apply a spatial autoregressive model with two weighting schemes and rely on Kelejian and Prucha’s combined GMM and IV approach. The weights in the first scheme are based on travel time in form of a generalized travel cost between origin and destination. The second one is an expansion of the first, essentially weighting travel times by relative income difference between the municipalities of the commuting flow under consideration. Furthermore, we also use this newly constructed variable as a regressor in the model.

Since economic variables may induce endogeneity in the regressor and the spatial weight matrix, we check for its presence and treat for it in the gravity model as well as in the spatial lag models. In order to that, we use an IV approach for the gravity model and Drukker et al.’s four-step procedure to re-estimate the spatial regression models. Last, we evaluate and compare in-sample predictions of the models among each other and to flows predicted by the national transport model to be able to draw a solid conclusion with respect to the suitability of the presented method.

CASE STUDY

Set up

We designed a case study for public transport commuting flows in Switzerland to illustrate the concept of OD flow modelling, based on travel-to-work trip data from the 2000 Federal Census, the Lohnstrukturerhebung 2000 and own calculations. The data cover 2,896 Swiss municipalities and contain over 250,000 observations in their initial form. However, the given data set does not fill the whole flow matrix that contains $2,896^2 = 8,386,816$ flows. The flows represent entries in the OD flow matrix, where columns reflect origins and rows destinations. For the remaining OD pairs we assume zero-valued travel flows. The mode "public transport" is an aggregation of trips-to-work by train, tram, bus and combinations of those. An important aspect is the issue with how to deal with zero flows. A large fraction of zero-valued OD flows would definitely point towards a Poisson or a (zero-inflated) negative Binomial interaction model, as mentioned above. However, neither a Poisson nor a negative Binomial spatial autoregressive regression model for OD flows has been developed so far that is also able to treat for endogeneity. As explained below, we include income differences between Swiss municipalities as an explanatory variable in our models, since a higher income probably gives an incentive to commute. Therefore, we filter the initial flow matrix for inter-communal travel trips, income data available only in 1,595 communes and all zero flows, which gives a final sample size of 46,659 OD flows. Due to filtering, a municipality does not have to be an origin and a destination anymore. Clearly, this is a limitation of our modelling approach. Nevertheless, the findings can be of apparent value for future research.

The Swiss network

A presentation of the resulting commuting flows is given in Figure 1 where flows emanate from the centroid of each municipality. Higher flow values correspond to a thicker representation of the linkages and only flows bigger than the median are shown. This figure clearly shows dense linear features emanating among larger cities in Switzerland, which also hints at the monocentric nature of employment in the area of big cities and towns.
Examining travel-to-work distances within the public transportation flow network reveals that even after filtering for zero flows the distribution of distances is heavily right skewed (Figure 2). Apparently, low flows in the initial data set are filtered leading to higher median values for flows in the first five deciles, which shows the importance of big cities in Switzerland.

**FIGURE 1** Public transportation commuting flows in Switzerland

**FIGURE 2** Distributions of network distances and flows (in logs) before and after filtering
**The model variables**

Modelling commuting behaviour requires a set of relevant explanatory variables that capture origin and destination characteristics along with mechanisms that generate and attract the trips among them. The dependent variable, inter-communal travel flows, is regressed on several independent variables obtained or derived from the Federal Census 2000, the Swiss national transport model and the Institute for Transport Planning and Systems (IVT) of ETH Zurich. We use the following variables in our framework, which are also common in explaining public transport demand in the literature mentioned above (Table 1).

Network distance is reported as travel time in minutes between municipalities. It resembles a generalised cost of travelling with public transport and incorporates not only the raw travel time, but also the waiting time at stations and the number of transfers on travel-to-work trips. Hence, network distance reflects the structure of public transportation and affects the dependent variable negatively. Income is an important variable for transport demand as the difference of income between destinations and origins could be a reason to commute. We construct a relative measure based on mean wages per municipality. Refer to Sarlas et al. (27) for the derivation of the wages on municipality level.

\[
Relative \text{ income difference} (inc) = \frac{income_{destination} - income_{origin}}{income_{origin}}
\]  

In general, one would expect that relative income differences have positive influence on flows. The positive sample mean supports our expectation. The question about the influence of income is, whether it has a direct impact or not. To start, we assume that income directly influences commuting flows exogenously without any other confounding effects. Job and population accessibility by public transportation are measures of available job positions and population in surrounding municipalities of origins and destinations (equation (7)). The parameters of the distance decay functions are taken from Sarlas and Axhausen (28).

\[
Job \text{ accessibility}_i = \sum_j jobs_i \times \exp(\beta cost_{ij}) \\
Population \text{ accessibility}_i = \sum_j Population_i \times \exp(\beta cost_{ij})
\]  

Because they capture how municipalities generally compete against each other in terms of available population and jobs, both measures should have a negative impact on the flow under consideration, either at an origin or destination level. We expect the total number of jobs at destinations and population at origins to influence OD flows positively, because they normally attract and produce trips in transportation networks. The positive influence on commuting flows should also hold for the share of economically active people per municipality (workers) at both origins and destinations. It makes sense to expect that bigger shares of jobs in the third sector at destination municipalities have a positive impact on OD flows, since it is the largest sector and many firms in the service sector are located in Switzerland’s big cities. However, we assume the share of jobs in the service sector at origins to affect flows negatively, since also job possibilities at origins in the service sector gives incentives not to commute.
TABLE 1 Summary statistics for the model variables

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commuting flow</td>
<td>average daily flows</td>
<td>11.1</td>
<td>79.1</td>
<td>1</td>
<td>5,698</td>
</tr>
<tr>
<td>Network distance</td>
<td>minutes</td>
<td>74.1</td>
<td>42.8</td>
<td>5.8</td>
<td>730.3</td>
</tr>
<tr>
<td>Relative income difference</td>
<td>CHF (in 1,000)</td>
<td>0.068</td>
<td>0.235</td>
<td>-0.643</td>
<td>1.744</td>
</tr>
<tr>
<td>Jobs (d)</td>
<td>Number of jobs</td>
<td>19,919</td>
<td>55,560</td>
<td>11</td>
<td>341,213</td>
</tr>
<tr>
<td>Job accessibility (d)</td>
<td>Number of accessible jobs</td>
<td>154,286</td>
<td>131,066</td>
<td>35.366</td>
<td>567,509</td>
</tr>
<tr>
<td>Workers (d)</td>
<td>Share of workers in population</td>
<td>0.528</td>
<td>0.037</td>
<td>0.277</td>
<td>0.726</td>
</tr>
<tr>
<td>Jobs 3rd sector (d)</td>
<td>Share of jobs in the 3rd sector</td>
<td>0.652</td>
<td>0.174</td>
<td>0.044</td>
<td>0.726</td>
</tr>
<tr>
<td>Population (o)</td>
<td>Number of jobs</td>
<td>13,662</td>
<td>40,809</td>
<td>2,094</td>
<td>363,273</td>
</tr>
<tr>
<td>Population accessibility (o)</td>
<td>Number of accessible jobs</td>
<td>217,619</td>
<td>183,387</td>
<td>93.783</td>
<td>1,064,884</td>
</tr>
<tr>
<td>Workers (o)</td>
<td>Share of workers in population</td>
<td>0.524</td>
<td>0.036</td>
<td>0.277</td>
<td>0.726</td>
</tr>
<tr>
<td>Jobs 3rd sector (o)</td>
<td>Share of jobs in the 3rd sector</td>
<td>0.595</td>
<td>0.176</td>
<td>0.044</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: 
N = 46,659; (o),(d) = at origin, at destination municipalities

The gravity model

As mentioned above, we start with a logged least-squares gravity model for OD commuting flows,

\[
\log(y) = \alpha \log(l_{n,2}) + \beta_o \log(X_o) + \beta_d \log(X_d) + \gamma_o X_o + \gamma_d X_d + \delta (inc) + \theta \log(g) + \epsilon \tag{8}
\]

where \(X_o, X_d\) represent origin and destination characteristics, \(g\) denotes the network distance and \(inc\) stands for the relative income difference between destination and origin municipalities. Estimation results for the gravity model are shown in Table 2. The associated adjusted R-squared of 51.7% shows that the OLS model explains a bit more than half of the variation in the flows. The residuals of the gravity model are almost normally distributed, yet they exhibit a slightly right skewed distribution and feature a higher kurtosis (Figure 3). Furthermore, heteroskedasticity robust standard errors are calculated and presented to account for potential non-constant variance in the residuals. We checked the variance inflation factors, which we do not show here, for all independent variables and found no multi-collinearity issues.

(a) Distribution of the model residuals
(b) Boxplot of the model residuals

FIGURE 3 Gravity model diagnostics
All parameters estimates are highly significant and show the expected signs. The network distance decay parameter (-1.532) is within the expected range for commuting patterns and in accordance with previous studies. Besides the share of workers at destinations, all other explanatory variables have a much weaker impact on the dependent variable. This finding is in line with the expectations of existing literature (12, 25). Income differences between destinations and origins have a significant and positive effect on travel-to-work trips and read as an elasticity, since relative differences are used. Note that all estimates read as elasticities. Among the destination characteristics, an increase in the share of workers (share of economically active population per municipality) and the number of jobs yield the biggest influence on travel-to-work trips (1.332 and 0.476), whereas an increase in the accessibility of jobs in neighbouring municipalities has the strongest negative influence on commuting (-0.153). A higher accessibility of jobs by public transport results in less transport demand in the destination of the OD flow under consideration and thus act as a kind of competition variable. The share of jobs in the third sector also has a positive impact on flows. Regarding the origin-specific variables, the parameters for population and the share of workers show the strongest positive impact (0.420 and 0.669) on commuting flows. An increase in both variables is positively related to travel demand, leading to higher flows away from origins. Interestingly, workers at origins influence flows only half as strong as at destinations. If more people in the neighbouring municipalities are available by means of public transportation, this has a negative and effect on travel-to-work trips (-0.211). The share of jobs in the service sector affects commuting flows negatively, which supports our view that the service sector is an important factor for any origin in the context of commuting.

OLS assumes independent observations. In the context of OD commuting flows, this assumes that the use of a network distance variable should eradicate the spatial dependence among the sample OD pairs, what is not the case in our setting as “flows associated with a destination are enhanced or diminished in accordance with the propensity of attractiveness of its neighbouring destination locations” (9). The same principle holds for flows from origins. Hence, residuals of gravity models indicate the presence of untreated spatial effects (8). By applying Moran’s I test (20), which in our framework weight the mean residuals by network and economic distance (see next subsection), we find that they indeed exhibit spatial dependence and thus justify the need for spatial models. That is, the mean residuals of either origins or destinations are positively correlated with their spatially lagged disturbances. In addition, spatial autocorrelation for origins and destinations is significant up to a radius of 100 minutes of travel time.

The spatial autoregressive model

We use the spatial autoregressive model (SAR) in log form,

\[
\log(y) = \alpha \log(l_n) + \rho_i W_i \log(y) + \beta_o \log(X_o) + \beta_d \log(X_d) + \gamma o X_o + \gamma d X_d + \delta (inc) + \theta \log(g) + \varepsilon, \quad \text{with } i = o, d, b. \tag{9}
\]

where in our case the weights \(w_{ij}\) for the weight matrix \(W_i\) are defined as

\[
\text{Network distance weights (netwdist)} = w_{ij} = \frac{1}{\text{travel time}_{ij}}
\]

\[
\text{Economic distance weights (ecodist)} = w_{ij} = \left(\frac{\text{travel time}_{ij}}{\exp((inc_j - inc_i)/inc_i)}\right)^{-1} \tag{10}
\]
**TABLE 2 Coefficient estimates for the gravity and SAR model(s)**

<table>
<thead>
<tr>
<th>Dependent variable: log(Commuting flows)</th>
<th>(I) Gravity model (OLS)</th>
<th>(II) SAR (GMM) (o)</th>
<th>(III) SAR (GMM) (d)</th>
<th>(IV) SAR (GMM) (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.837 ***</td>
<td>2.740 ***</td>
<td>2.767 ***</td>
<td>3.057 ***</td>
</tr>
<tr>
<td>log(Network distance)</td>
<td>-1.532 ***</td>
<td>-1.501 ***</td>
<td>-1.507 ***</td>
<td>-1.603 ***</td>
</tr>
<tr>
<td>Relative income difference</td>
<td>0.105 ***</td>
<td>0.101 ***</td>
<td>0.103 ***</td>
<td>0.110 ***</td>
</tr>
<tr>
<td>log(Jobs) (d)</td>
<td>0.476 ***</td>
<td>0.457 ***</td>
<td>0.458 ***</td>
<td>0.469 ***</td>
</tr>
<tr>
<td>log(Job accessibility) (d)</td>
<td>-0.153 ***</td>
<td>-0.151 ***</td>
<td>-0.154 ***</td>
<td>-0.166 ***</td>
</tr>
<tr>
<td>Workers (d)</td>
<td>1.332 ***</td>
<td>1.325 ***</td>
<td>1.321 ***</td>
<td>1.349 ***</td>
</tr>
<tr>
<td>Jobs 3rd sector (d)</td>
<td>0.301 ***</td>
<td>0.291 ***</td>
<td>0.290 ***</td>
<td>0.293 ***</td>
</tr>
<tr>
<td>log(Population) (o)</td>
<td>0.420 ***</td>
<td>0.420 ***</td>
<td>0.420 ***</td>
<td>0.482 ***</td>
</tr>
<tr>
<td>log(Population access.) (o)</td>
<td>-0.211 ***</td>
<td>-0.206 ***</td>
<td>-0.204 ***</td>
<td>-0.212 ***</td>
</tr>
<tr>
<td>Workers (o)</td>
<td>0.669 ***</td>
<td>0.636 ***</td>
<td>0.657 ***</td>
<td>0.792 ***</td>
</tr>
<tr>
<td>Jobs 3rd sector (o)</td>
<td>-0.129 ***</td>
<td>-0.118 ***</td>
<td>-0.122 ***</td>
<td>-0.129 ***</td>
</tr>
<tr>
<td>rho</td>
<td>0.057 ***</td>
<td>0.054 **</td>
<td>-0.270 ***</td>
<td>-0.398 ***</td>
</tr>
</tbody>
</table>

HC robust std. errors | yes | yes | yes | yes | yes | yes | yes
Adj. R² | 0.517
Pseudo adj. R² | 0.523 | 0.523 | 0.513 | 0.513 | 0.504 | 0.498

Note: N=46,659; (o)/(d) = at origin/at destination municipalities; *p<0.1; **p<0.05; ***p<0.01
Economic distance weights are essentially travel times weighted by the exponential of relative income differences. For instance, a positive difference resulting from a higher income in destinations than origins for a given OD dyad lowers travel times, implying a higher weight overall because of taking the inverse. These weights are now assigned to neighbouring origins of an OD pair in the case of an origin-centric weight matrix, essentially weighting the corresponding commuting flows from neighbours of an origin to a specific destination ($W_o$). The same principle holds for the case of destination-centric spatial weight matrices ($W_d$). A third weight matrix sums the origin and destination based weight matrix and accounts for both effects ($W_b$). Furthermore, we row-normalize the weight matrices.

Estimation results for the SAR models are shown in Table 2. In the transition to the spatial models (II), (III) and (IV), all parameter estimates remain statistically significant, relatively stable in magnitude and keep their signs. We emphasize that the magnitude of the coefficients from least-squares and spatial models can not be compared due to the fact that the least-squares coefficient for any explanatory variable $x$ represents $\frac{\partial y}{\partial x}$, whereas one from the spatial lag model does not because of possible feedback structures induced by the spatial weighting. LeSage and Thomas-Agnan (13) propose scalar summary measures of these impacts that average over changes applied to a single independent variable for all regions. We do not calculate those summary measures here, but it is important to mention the limited interpretation. The autoregressive parameter is of particular interest. It is statistically significant in all models. For origin-centric weight matrices ($W_o$) based on network and economic distance, rho is positive and (0.057 and 0.054). Thus, neighbours of an origin influence the particular OD flow under consideration positively. In the case of destination-centric weightig ($W_d$), rho is negative for both weights which basically indicates that destinations are competing for commuters. It shows an even stronger negative effect for economic distance weights (-0.398). When we account for both effects ($W_b$), rho remains negative what suggests that destination dependence is stronger than origin dependence. In addition, the spatial models yield similar goodness-of-fit measures as the gravity model. Since the Akaike information criterion cannot be calculated for GMM models, a pseudo adj. R² is reported. Note that these values must be treated with caution, as they are not equivalent to OLS-based R² measures.

### Treating for endogeneity in the gravity and spatial model(s)

Endogeneity is a severe problem and leads to biased and inconsistent estimates, making inference invalid. In this framework, the mean income as an economic characteristic of origins/destinations and as part of the spatial weight matrix is used to explain variation in commuting flows. Due to the economic nature of income it may be that there is an omitted variable bias, causing the disturbances to be correlated with the regressor in the case of the OLS gravity model. This fact violates the conditional mean assumption, what means that it is not possible to distinguish the influence of and between each variable in the model. To account for endogeneity in the gravity model we employ an Instrumental variable (IV) approach in order to get a consistent, but biased, and less efficient estimator compared to OLS. It is difficult to think of income to be exogenous in the case of commuting. First, there may be other (omitted) variables explaining variation in travel-to-work trips that are correlated with income - taxes at municipality level for example. Second, it is difficult to assume no interaction with other variables in the model, because of strong interrelations of transportation, human settlement, urban agglomeration and economic activities concentrated in cities, the gravity model should be tested for endogeneity. Usually, family background, workforce variables or characteristics of job positions are used as instruments for income. Sarlas et al. (27) found evidence for the positive impact of job characteristics on mean salaries. We use six instruments that relate to different requirements or characteristics of job
positions: High and low managerial skills, tertiary or minimal education as well as high and medium qualification demands.

Coefficient estimates for the IV gravity model (V) are shown in Table 3. The IV diagnostics confirm that we have a valid IV model since we do not have weak instruments (F-test), i.e. no weak first stage relationship, we rely on inconsistent OLS estimates suggesting that endogeneity is present (Wu-Hausman test) and we use valid instruments (Sargan test). Except for the income variable, the parameter estimates differ only little compared to the OLS gravity model. Interestingly, the estimate of relative income difference has changed significantly in the IV framework, yielding stronger and positive impact on commuting flows (0.173 versus 0.105). The model fit statistic has not changed compared to the adjusted R-squared of the OLS model.

Clearly, endogeneity affects spatial models in a more complex way than gravity models, since we do not only include an endogenous regressor, but also endogenous weight matrices. Thus, the spatial model estimates in Table 2 are biased and inconsistent. Since we have found valid instruments for the income difference between origin and destination municipalities, it is possible to use the predicted and thus corrected income values (see the first equation below) of the first stage in IV in order to construct the spatial weights.

\[
\left(\frac{\text{inc}_j - \text{inc}_i}{\text{inc}_i}\right) = \text{all instruments} + \text{all exogenous variables} + \epsilon 
\]

\[\text{Economic distance weights (ecodist)} = w_{ij} = \left(\frac{\text{travel time}_{ij}}{\exp\left(\frac{(\text{inc}_j - \text{inc}_i)/\text{inc}_i}{\text{inc}_i}\right)}\right)^{-1}\]

By directly including predicted values of income in the construction of the spatial weight matrix, we can account for previously endogenous elements. We then use Drukker et al.’s 4-step estimation method (23) to get consistent coefficients.

The results for the SAR models that account for endogeneity are presented in Table 3. All parameter estimates for model (VI), (VII) and (VIII) remain highly statistically significant. When we compare the coefficient estimates for relative income differences among all models, we see that especially for origin-centric weighting \(W_o\) they are substantially underestimated for both network and economic distance weights when we do not treat for endogeneity. This finding makes sense and shows that differences in income between municipalities have a higher impact for origin-centric weighting compared to one relying on destinations. Overall, we find that when we account for endogeneity, income differences have a higher impact on flows. Moreover, they are stronger for economic distance weights in all weighting schemes. The estimates for rho do not follow the same pattern. We observe big differences in their coefficients again for economic distance weights in the origin- and destination-centric weighting scheme (0.195 in model (VI) versus 0.054 in model (II) and -0.836 in model (VII) versus -0.398 in model (III)). All other estimates have expected signs and do not vary much throughout model (VI)-(VIII).
TABLE 3 Coefficient estimates for the gravity and SAR model(s), accounting for endogeneity

<table>
<thead>
<tr>
<th>Dependent variable: log(Commuting flows)</th>
<th>Gravity model (IV/2SLS)</th>
<th>SAR (2IV/GMM) (o)</th>
<th>SAR (2IV/GMM) (d)</th>
<th>SAR (2IV/GMM) (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>netwdist</td>
<td>ecodist</td>
<td>netwdist</td>
<td>ecodist</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.835 ***</td>
<td></td>
<td>2.745 ***</td>
<td></td>
</tr>
<tr>
<td>log(Network distance)</td>
<td>-1.533 ***</td>
<td></td>
<td>-1.506 ***</td>
<td></td>
</tr>
<tr>
<td>Relative income difference</td>
<td>0.173 ***</td>
<td></td>
<td>0.251 ***</td>
<td></td>
</tr>
<tr>
<td>log(Jobs) (d)</td>
<td>0.474 ***</td>
<td></td>
<td>0.453 ***</td>
<td></td>
</tr>
<tr>
<td>log(Job accessibility) (d)</td>
<td>-0.156 ***</td>
<td></td>
<td>-0.158 ***</td>
<td></td>
</tr>
<tr>
<td>Workers (d)</td>
<td>1.346 ***</td>
<td></td>
<td>1.356 ***</td>
<td></td>
</tr>
<tr>
<td>Jobs 3rd sector (d)</td>
<td>0.300 ***</td>
<td></td>
<td>0.291 ***</td>
<td></td>
</tr>
<tr>
<td>log(Population) (o)</td>
<td>0.423 ***</td>
<td></td>
<td>0.428 ***</td>
<td></td>
</tr>
<tr>
<td>log(Population access.) (o)</td>
<td>-0.209 ***</td>
<td></td>
<td>-0.201 ***</td>
<td></td>
</tr>
<tr>
<td>Workers (o)</td>
<td>0.665 ***</td>
<td></td>
<td>0.629 ***</td>
<td></td>
</tr>
<tr>
<td>Jobs 3rd sector (o)</td>
<td>-0.130 ***</td>
<td></td>
<td>-0.122 ***</td>
<td></td>
</tr>
<tr>
<td>rho</td>
<td>0.053 ***</td>
<td></td>
<td>0.195 ***</td>
<td></td>
</tr>
<tr>
<td>HC robust std. errors</td>
<td>yes</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.517</td>
<td></td>
<td>0.523</td>
<td></td>
</tr>
<tr>
<td>Pseudo adj. R²</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV diagnostics:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak instrument test</td>
<td>1226.462 ***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu-Hausman endog. test</td>
<td>5.413 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sargan instr. validity test</td>
<td>4.098</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N=46,648; (o)/(d) = at origin/at destination municipalities; *p<0.1; **p<0.05; ***p<0.01
Evaluation of predictive accuracy of models

We evaluate the accuracy of our model predictions by two adjusted measures: The root weighted mean squared percentage error (RWMSPE) and the weighted mean absolute percentage error (WMAPE). Percentage errors have the advantage of being scale-independent and are frequently used to compare the prediction performance across different models. MAPE measures have the main disadvantage that they are influenced by outliers. RMSE is often preferred to the MSE as it is on the same scale as the data. Instead of using measures based on symmetric median values, which are more robust to outliers and therefore smaller than those based on mean values, we use a weighted mean approach that weights predictions with the corresponding size of the real flow according to their importance in the Swiss commuting flow network.

It seems that outliers, possibly leading to the high percentage errors we see here, still heavily influence WMAPE measures. We therefore focus on RWMSPE measures. Among all estimated models, the SAR model (IV) with the combined weight matrix (b) using economic distance weights has the highest in-sample accuracy. The origin-centric SAR models relying on network distance weights, both destination-centric SAR models based on economic distance weights together with the gravity models yield the worst accuracies. However, all four spatial autoregressive models based on the sum of the origin- and destination-centric weight matrix show lower RWMSPE values than both gravity models. The models accounting for endogeneity do only slightly better for origin- and destination-centric weighting matrices, while they do slightly worse for their combined version. If we compare the performance of all our estimated models to the National transport model from 2000, which is the state-the-art four step model used to estimate flows, only the previously mentioned SAR (b) model (IV) provides more accurate predictions.

We provide a relatively simple and direct approach to predict commuting flows and get similar accuracies in terms of RWMSPE compared to the complex 4-step procedure of the National transport model, which we see as a clear advantage of our approach.

### TABLE 4 In-sample predictions

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Model</th>
<th>RWMSPE</th>
<th>WMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td>National transport model</td>
<td>44.68</td>
<td>176.24</td>
</tr>
<tr>
<td>(V)</td>
<td>OLS gravity model</td>
<td>47.94</td>
<td>200.79</td>
</tr>
<tr>
<td>(V)</td>
<td>IV gravity model</td>
<td>47.92</td>
<td>201.23</td>
</tr>
<tr>
<td>(II)</td>
<td>SAR (o): Network distance</td>
<td>49.00</td>
<td>208.80</td>
</tr>
<tr>
<td>(II)</td>
<td>SAR (o): Economic distance</td>
<td>46.55</td>
<td>192.30</td>
</tr>
<tr>
<td>(III)</td>
<td>SAR (d): Network distance</td>
<td>44.80</td>
<td>181.43</td>
</tr>
<tr>
<td>(III)</td>
<td>SAR (d): Economic distance</td>
<td>48.98</td>
<td>208.24</td>
</tr>
<tr>
<td>(IV)</td>
<td>SAR (b): Network distance</td>
<td>46.17</td>
<td>187.99</td>
</tr>
<tr>
<td>(IV)</td>
<td>SAR (b): Economic distance</td>
<td>43.55</td>
<td>171.06</td>
</tr>
<tr>
<td>(VI)</td>
<td>Accounting for endogeneity:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(VI)</td>
<td>SAR (o): Network distance</td>
<td>48.85</td>
<td>208.86</td>
</tr>
<tr>
<td>(VI)</td>
<td>SAR (o): Economic distance</td>
<td>46.53</td>
<td>192.36</td>
</tr>
<tr>
<td>(VII)</td>
<td>SAR (d): Network distance</td>
<td>44.76</td>
<td>181.64</td>
</tr>
<tr>
<td>(VII)</td>
<td>SAR (d): Economic distance</td>
<td>48.39</td>
<td>212.43</td>
</tr>
<tr>
<td>(VIII)</td>
<td>SAR (b): Network distance</td>
<td>45.12</td>
<td>186.58</td>
</tr>
<tr>
<td>(VIII)</td>
<td>SAR (b): Economic distance</td>
<td>47.10</td>
<td>191.93</td>
</tr>
</tbody>
</table>
CONCLUSION

In this paper we implemented a direct demand model for OD public transport commuting flows in Switzerland. It is based on data from the 2000 Federal Census, the Lohnstrukturerhebung 2000 and the National transport model 2000. Further variables are based on calculations by the Institute for Transport Planning and Systems of ETH Zurich.

Methodologically, we employed a three-step process to examine the problem of spatial dependence in OD commuting flows when network and economic distance are used as underlying impedance function. The starting point was a simple OLS gravity model relying on independent observations, which we then replaced by spatial autoregressive models that are based on two weighting schemes (network and economic distance) in order to account for untreated spatial dependence in the residuals of the OLS gravity model. We used origin- and destination-centric weight matrices and a combination of both to account for both origin and destination effects. In the last step, we checked if endogeneity is present in the OLS gravity model by applying an IV regression approach using valid instruments for relative income differences. Subsequently, we used a combination of a GMM and IV procedure according to Drukker et al. to re-estimate the spatial regression models in order to get consistent coefficient estimates. Last, we compare in-sample predictions of our models by root weighted mean squared percentage error measures.

We applied a filter method due to a large fraction of zero-valued flows and income data available for only 1595 municipalities, which gave a final sample of 46,659 observations. Neither a Poisson nor a negative Binomial estimation approach for OD flow modelling has been developed so far to treat for endogeneity in the same framework. Clearly, this a limitation of our work. Nevertheless, it provides important insights into modelling OD commuting flows when an economic variable is used as a regressor as well as in the spatial weight matrix.

The estimates of the OLS gravity model were in line with expectations of existing literature concerning its statistical and economical importance. We used Moran’s I test to show that the residuals of the gravity model contain patterns of remaining autocorrelation up to a radius of 100 minutes of travel time. We estimated spatial autoregressive models where we used network and economic distance weighting schemes and were able to show that neighbouring municipalities at origins have a positive and influence on OD commuting flows under consideration. At destinations they have a larger and negative effect. For the combination of the origin- and destination-centric weight matrix, neighbours still have a negative but slightly lower influence on flows, which indicates that municipalities are generally competing for commuters. The remaining explanatory variables remained stable across all models in sign and magnitude. We used different job requirements such as education levels, managerial duties and qualification demands as instruments to show that income indeed is endogenous in our setting. We were able to show that relative income differences were significantly underestimated in the OLS gravity and SAR model once we account for endogeneity, with the biggest difference in the case of economic distance weights. Last, we compared the predictive accuracies of our models among each other and to those of the National transport model from 2000. Four SAR models that use a combined weight matrix outperformed the gravity models with respect to the root weighted mean squared percentage error (RWMSPE) measure, whereas only the one of those that relies on economic distance weights gave more accurate predictions than the National transport model.
AUTHOR CONTRIBUTION STATEMENT
The authors confirm contribution to the paper as follows: Study conception and design, data processing, analysis and interpretation of results, and manuscript preparation: Thomas Schatzmann, Georgios Sarlas. Interpretation of results and manuscript reviewing: Kay W. Axhausen. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES


