



Working Paper

## **Mean speed prediction with endogenous volume and spatial autocorrelation A Swiss case study**

**Author(s):**

Sarlas, Georgios; Axhausen, Kay W.

**Publication Date:**

2017-08

**Permanent Link:**

<https://doi.org/10.3929/ethz-b-000175412> →

**Rights / License:**

[In Copyright - Non-Commercial Use Permitted](#) →

This page was generated automatically upon download from the [ETH Zurich Research Collection](#). For more information please consult the [Terms of use](#).

1 **Mean speed prediction with endogenous volume and spatial autocorrelation: A Swiss**  
2 **case study**

3  
4  
5

6 **Georgios Sarlas\***

7 Institute for Transport Planning and Systems, ETH Zurich  
8 HIL F 51.3, Stefano-Frascini-Platz 5, 8093, Zurich, Switzerland  
9 Tel: +41 44 633 37 93; E-mail: [georgios.sarlas@ivt.baug.ethz.ch](mailto:georgios.sarlas@ivt.baug.ethz.ch)

10

11 **Kay W. Axhausen**

12 Institute for Transport Planning and Systems, ETH Zurich  
13 HIL F 31.3, Stefano-Frascini-Platz 5, 8093, Zurich, Switzerland  
14 Phone: +41 44 633 39 43; E-mail: [axhausen@ivt.baug.ethz.ch](mailto:axhausen@ivt.baug.ethz.ch)

15

16

17 Working paper 1275, IVT ETH Zurich

18

19

20

21

22

23 \* Corresponding author

24

25

26 **ABSTRACT**

27 In the present paper a modeling approach to address the issue of mean speed prediction on a  
28 large scale network is presented. Methodologically, we exploit the family of spatial regression  
29 models to treat both for the spatial dependence and the endogeneity aspect between speed and  
30 volume. The estimation of the model takes place by means of a 2-step instrumental variables-  
31 generalized method of moments estimator, allowing us to obtain consistent and unbiased  
32 parameter estimates.

33 An empirical case study is designed and conducted in order to model mean speed values on a  
34 national planning network in order to check the applicability and the predictive performance of  
35 the proposed modeling approach. A particular focus is given on the instrumentation of demand  
36 on truly exogenous and capable of capturing the interregional demand patterns variables.  
37 Moreover, different spatial weight matrices are tested thoroughly to conclude on a matrix  
38 identification based on free-flow travel time. Our findings suggest that the proposed model has  
39 the ability to provide accurate estimates, outperforming a much more complex and data  
40 demanding transport planning model, even though the superiority of such models is taken for  
41 granted in many cases. Last, the developed modeling approach coupled with the implied volume  
42 regression model can form a coherent direct demand modeling approach, suitable both for  
43 prediction and forecasting applications.

44

45

46

47 *Keywords:* Average speed, spatial regression, instrumental variables, endogeneity, AADT

48

## 49 INTRODUCTION

50 The provision of accurate link speed estimates constitutes a core task of transport modeling. A  
51 closer look at the nature of the task, points out that speed is essentially the outcome of the  
52 interaction between supply and demand. Driven by the inherent complexities of these two  
53 aspects, various modeling approaches have emerged over the years to tackle the problem.

54  
55 The most prominent approaches involve the use of simulation models, comprising a set of sub-  
56 models to quantify the different aspects of the transport system, and then through an iterative  
57 process facilitate the interaction between demand and supply until an equilibrium has been  
58 reached. Two broad categories of such models exist, depending on how they simulate the  
59 interactions within the system, involving and allowing different considerations and sub-model  
60 formulations. The first category, the macroscopic approach, focuses on the system as a whole  
61 and models its different components and their interactions in an aggregate way. The second  
62 category, the microscopic approach, considers the individual components of the system and  
63 models their behavior and interactions in a disaggregate way, making use of advanced statistical  
64 models.

65  
66 A further distinction of the transport simulation models can be made based on how the demand  
67 aspect is modeled. On the one hand, models focusing on the operational side of the system  
68 consider demand fixed and turn their focus on simulating how individuals move and interact  
69 (e.g. traffic simulation models). On the other hand, models focusing on planning purposes,  
70 model the demand aspect of the system under the assumption that demand is not fixed.  
71 Depending on how the generation of demand is formulated, a further distinction can be made  
72 between microscopic and macroscopic demand modeling approaches. A prominent example of  
73 the former one is the traditional 4-step model, while in the case of the latter are the agent-based  
74 models (e.g. (1)). Obviously, a microscopic demand model requires much more detailed data  
75 on a person level and the development and employment of many statistical sub-models,  
76 increasing considerably the computational effort to reach the equilibrium point.

77  
78 However, when it comes to the appraisal of public transport projects, as Flyvbjerg et al. (2)  
79 argue, the quality of the demand forecasts has not been improved over the years even though  
80 more complex and behaviorally sound models have been employed. In a similar line of thought,  
81 Dowling and Skabardonis (3) highlight the fact that large scale planning models, which are only  
82 calibrated against volume estimates, typically fail to provide reasonable speed estimates. This  
83 aspect has been systematically neglected in the literature, along with its implications. More  
84 specifically, travel time (calculated on the basis of the estimated speed values) is a core element  
85 when it comes to the system performance evaluation and the appraisal of new projects (e.g.  
86 estimation of travel time savings).

87  
88 Based on the above, formulating a direct demand modeling approach as an alternative seems  
89 appealing for a number of reasons. First, it can offer a structural explanation of the modeled  
90 phenomena at any location on the network in a direct and straightforward manner. Second, the  
91 data and computational requirements are considerably lower than the prevailing simulation  
92 approaches. Last, it can constitute a worthwhile alternative especially if its predictive accuracy  
93 is found to be acceptable. Nevertheless, the alternative modeling structure should be capable of  
94 making statements about the speed and the traffic volume on a link level of a regional planning  
95 scale network, the items that constitute the minimum requirements for the appraisal of transport  
96 projects.

97 Driven by these considerations, the alternative of regression modeling emerges as the most  
98 apparent one, allowing to model the impact of different variables directly on the outcomes of  
99 interest. Previous attempts to model mean speed values (4, 5) have highlighted that speed  
100 observations are spatially dependent and that should be treated in order to obtain valid estimated  
101 parameters by employing spatial regression models. The issue of dependence was also  
102 acknowledged for the case of operating speed modeling in another study (6), whereas a plethora  
103 of studies on modeling the operating speed exist (an overview can be found in (7)). However,  
104 their scope differs substantially than the one of the current study, since their purpose is to  
105 evaluate the impact of design characteristics on speed. Interestingly, the presence of  
106 endogeneity issues was demonstrated in a number of studies (8, 9), accounting for the  
107 simultaneity between mean speed and speed deviations. The same issue was also acknowledged  
108 for the case of accident models, accounting for the simultaneity between speed and accident  
109 rates (e.g. (10, 11)). In another study, a simultaneous equation modeling approach was  
110 suggested to explore the relationships between mean speed, standard deviation of speed and  
111 work zone design characteristics (12). Another strand of literature is concerned with modeling  
112 of mean speed values for emission models (a review can be found in (13)).

113  
114 In conclusion, two main considerations should be made with regard to the application of linear  
115 regression techniques for speed and volume prediction purposes. First, the model should  
116 account for spatial dependence issues. Second, the presence of simultaneity between speed and  
117 volume should be addressed. Both of the issues, have the capacity of giving rise to invalid  
118 statistical testing, and inconsistent and biased estimates, if remain untreated.

## 119 **Description of the Framework**

120 This work builds upon the previous work undertaken by the authors on the mean speed and  
121 volume estimation with the use of spatial regression models (4, 14) and takes it further by  
122 accounting for the endogeneity aspects that govern the speed prediction in relation to the  
123 volume.  
124

## 125 **METHODOLOGY**

### 126 **General Overview**

127 As mentioned in the prior section, there are two considerations associated with the choice of a  
128 mean speed model. The first one relates to the endogenous character of volume in the speed  
129 model. More specifically, estimation by means of ordinary least-squares (OLS) constitute the  
130 standard for linear regression models. However, in the case of endogenous variable(s), the main  
131 OLS assumption of uncorrelated error terms with the independent variables is violated (15).  
132 This violation turns OLS to an inconsistent and biased estimator, and should be thoroughly  
133 tested and treated if present. This issue is dealt with by accounting for the endogeneity via  
134 utilizing instrumental variable(s) (IV), normally within a two-stage least-squares (2SLS)  
135 approach. Essentially, the endogenous variable(s) are replaced by the predicted one(s) from a  
136 set of variables (instruments). A strong prerequisite is that the instruments must be uncorrelated  
137 with the error term but substantially correlated with the endogenous variable(s). This estimation  
138 approach allows to obtain consistent and unbiased parameter estimates.  
139

140  
141 Secondly, the main implication of modeling data of a spatial nature is the existence of spatial  
142 dependence, thereby pointing to non-independent observations. As stated by Anselin (16), “*as*  
143 *spatial dependence, it can be considered to be the existence of a functional relationship between*  
144 *what happens at one point in space and what happens elsewhere*”. Thus implying that if the

145 correlation is not fully explained by the different variables included in the model specification,  
 146 the remaining correlation is “transmitted” to the residuals, leading to a violation of the  
 147 independent and identically distributed (*iid*) assumption of OLS. This violation of the *iid*  
 148 assumption leads to statistical problems such as unreliable statistical tests and biased and  
 149 inconsistent parameter estimates. Spatial simultaneous autoregressive (SAR) models constitute  
 150 a modeling medium allowing to treat for this issue in two ways, assuming different underlying  
 151 mechanisms that generate the spatial dependence. On the one hand, when a spatial variable has  
 152 been omitted from the model specification, the error terms tend to be spatially autocorrelated,  
 153 creating a need of an error term that inherently considers this (spatial error model). On the other  
 154 hand, when neighboring locations’ response variable has an indirect effect on the response at  
 155 location, then the inclusion of a spatially lagged dependent variable can mitigate the spatial  
 156 dependence issues, hence facilitate the estimation of explanatory variables’ direct effects on the  
 157 response variable (spatial lag model). A combined treatment of both aforementioned spatial  
 158 dependencies is also possible within a model formulation (spatial Durbin model). The  
 159 formulation of the first two SAR models is presented below.

160

161 Spatial error model:  $Y_i = \beta_k X_{ik} + u_i$ , with  $u_i = \lambda W u_{i-1} + \varepsilon_i$  (1)

162

163 where  $\lambda$  the spatial autoregressive coefficient,  $W$  the spatial weight matrix with dimensions  
 164  $N \times N$ ,  $u$  a vector of disturbances, and  $\varepsilon$  a vector of *iid* error terms (innovations).

165

166 Spatial lag model:  $Y_i = \rho W P_{i-1} + \beta_k X_{ik} + \varepsilon_i$  (2)

167

168 where  $\rho$  is a spatial autocorrelation parameter.

169

170 Spatial autocorrelation is normally measured in terms of the Moran’s  $I$  index which quantifies  
 171 the degree of autocorrelation on the residuals of a model (0 value indicates no autocorrelation,  
 172 while 1 or -1 perfect autocorrelation) (16). The spatial weight matrix  $W$  serves a two-fold  
 173 purpose. First, it specifies the neighborhood of each location, and second it assigns weights on  
 174 the neighboring locations on the basis of different schemes (e.g. binary, inverse distance  
 175 weighted etc.). Its determination takes place experimentally by identifying up to what spatial  
 176 extent there is statistically significant autocorrelation (a detailed discussion and illustration can  
 177 be found in (4)).

178

179 The prevailing estimation approach of SAR models is by means of maximum likelihood.  
 180 However, this entails a number of drawbacks, such as being computationally infeasible for large  
 181 samples, and most importantly lacking the ability to account for the presence of endogenous  
 182 regressor(s) and heteroscedastic disturbances. Regarding the former, Kelejian and Prucha (17)  
 183 suggested a generalized method of moments (GMM) estimator which can be seen as a major  
 184 breakthrough in the field of spatial econometrics, and has paved the way for addressing the  
 185 latter shortcoming as well. More specifically, in a follow-up paper (18), the same authors  
 186 developed a methodology for accounting for unknown forms of heteroscedasticity in  
 187 conjunction with an IV estimator for the parameters. Later on, Drukker et al. (19) extended their  
 188 work by developing a two-step generalized method of moments and instrumental variable  
 189 estimator (2IV/GMM), capable of accounting both for endogeneity and heteroscedastic  
 190 innovations, in addition to a spatially lagged variable. In summary, their estimator involves four  
 191 steps. Initially, a two stage least-squares (2SLS) approach is applied to obtain the starting values  
 192 of the parameters of the model (betas), similar to the traditional IV estimation approach. In the

193 next step, a GMM estimator is applied to obtain the value of the autoregressive coefficient  $\lambda$ .  
 194 The moment conditions are defined on the basis of conforming to the orthogonality assumption,  
 195 imposing the independence of the residuals with their first and second order neighbors'  
 196 counterparts. In the third step, a generalized spatial two-stage least-squares estimator is applied  
 197 on a Cochrane-Orcutt transformed model to obtain the new values of betas along with the  
 198 residuals. In the final step, the residuals from the previous step are utilized within a GMM  
 199 estimator to obtain the true value of  $\lambda$ , imposing the same moment conditions as before.

200

### 201 **Modeling Approach**

202 Previous attempts to model average speed (e.g. (4, 5)) have resorted to the use of proxy variables  
 203 for the traffic volume, operationalized in the form of spatial density values of various  
 204 sociodemographic variables (e.g. population, employment). Yet, and as identified in (14), such  
 205 variables fail to capture the directionality and the complexities of the interregional demand, and  
 206 thus can suffice only for small area cases. Based on this and given the objective of the current  
 207 study, we choose to instrument traffic volume on a set of variables capable of capturing the  
 208 interregional demand aspects. Based on our previous work on AADT prediction (14), we can  
 209 utilize the identified set of independent variables for the investigation of the choice of  
 210 instruments. In particular, the so-called constructed accessibility weighted centrality variable is  
 211 modified to account for the different mode shares, and the new formulation is presented below:

212

$$213 \text{Accessibility} - \text{weighted centrality}_e = \sum_{i,j \in V} \sigma_{ij}(e) \quad (3)$$

214

$$215 \sigma_{ij}(e) = \sum_{i,j \in V} \text{Popul}_i \frac{\text{Employ}_j * f(\text{cost}_{ij}^{\text{car}}) / (f(\text{cost}_{ij}^{\text{car}}) + f(\text{cost}_{ij}^{\text{PuT}}))}{\text{Travel Accessibility}_i} \quad (4)$$

216

$$217 \text{Travel Accessibility}_i = \sum_j \text{Employ}_j * \max\{f(\text{cost}_{ij}^{\text{car}}), f(\text{cost}_{ij}^{\text{PuT}})\} \quad (5)$$

218

$$219 f(\text{cost}_{ij}^{\text{mode}}) = e^{\beta * \text{cost}_{ij}^a} \quad (6)$$

220

221 The accessibility weighted centrality variable is calculated on a link level  $e$ , by summing up the  
 222 accessibility weighted shortest paths between all pairs of zones  $i$  and  $j$ , passing through that  
 223 link. In addition, travel accessibility of each zone  $i$  is defined as the sum of the maximum  
 224 intensity interaction between the two modes for each  $j$ , multiplied with the employment  
 225 opportunities at  $j$ . Additional information regarding the construction of the variable can be  
 226 found at (14).

227

228 In contrast to prior studies on the topic, we choose to follow a distinct approach concerning the  
 229 dependent variable. Specifically, and in line with the BPR functions' formulation, we specify  
 230 as dependent variable the mean travel time difference, defined as the difference between the  
 231 free-flow travel time and the mean travel time. The apparent advantage of employing such a  
 232 formulation is that it can better capture the relation between volume and speed. Furthermore,  
 233 this way the problem can be transformed into a linear model, overcoming the non-linearity  
 234 issues present. This choice allows us also to make use of the 2IV/GMM estimator, in order to  
 235 account and treat both for endogeneity and spatial dependence issues. The model formulation  
 236 is presented below:

237

$$238 \quad tt = t_0 + t_0^\alpha \frac{AADT^\beta}{Capacity^\gamma} \quad (7)$$

$$239 \quad \log(tt - t_0) = \log\left(t_0^\alpha * \frac{AADT^\beta}{Capacity^\gamma}\right) = \alpha \log(t_0) + \beta \log(AADT) - \gamma \log(Capacity) + u \quad (8)$$

240

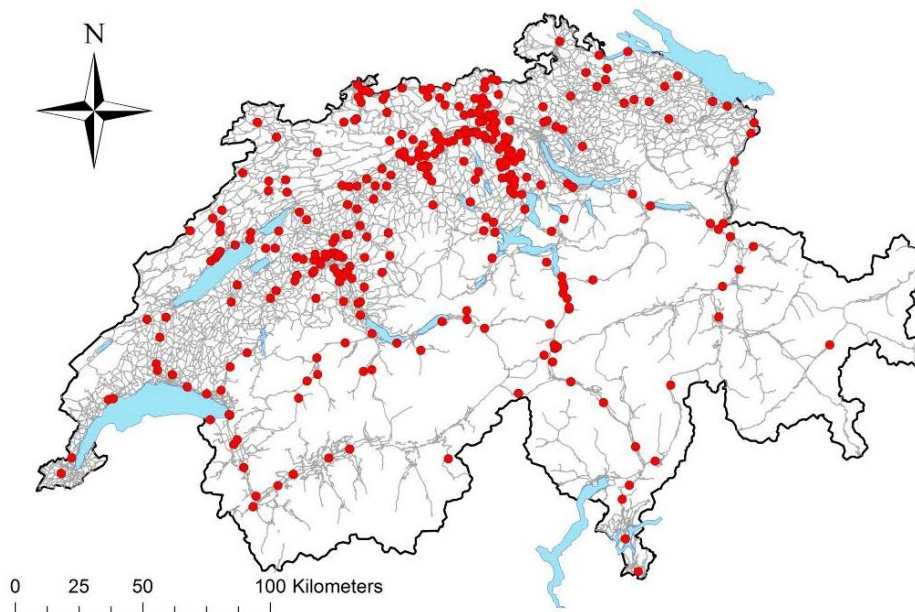
241 Conceptually, the various forms of BPR functions attempt to model the interaction between  
 242 demand and supply on a link level by incorporating various congestion functions. The majority  
 243 of those quantify the congestion as a function of the volume to capacity ratio. Capacity values  
 244 are normally calculated on the basis of standardized values coming from the Highway Capacity  
 245 Manual (7). In our case, we prefer to employ instead a set of typically explanatory variables as  
 246 proxy ones for the capacity instead, say number of lanes, legal speed limit etc. Consequently,  
 247 the need of a priori estimating capacity values diminishes. Especially, if we take into account  
 248 that for the case of national planning models, which involve some degree of abstraction and a  
 249 link might actually correspond to a set of links in reality, determining a single capacity value  
 250 for a set of links with varying attributes can be challenging and potentially troublesome.

251

## 252 CASE STUDY

253 An empirical case study is designed and conducted in order to model mean speed values in a  
 254 set up that resembles a national planning model's configuration. Particularly, the network of  
 255 Switzerland is exploited as the study network in the form that is present in a state-of-the-practice  
 256 national transport model<sup>1</sup>. The network consists of approximately 60'000 directed links, while  
 257 the links are classified into four hierarchical types, namely highway, major road, rural main  
 258 road, and urban arterial road. Two independent sources of volume and speed data are utilized  
 259 to facilitate the construction of the observations' database. At first, traffic volume observations  
 260 are obtained from the Federal Roads Office where they collect count data at various locations  
 261 on the network and calculate annual average daily traffic (AADT) values. In summary, for the  
 262 year 2010, data on 420 locations are retrieved. Subsequently, the count locations are matched  
 263 to the network employed. An overview of the network along with the spatial distribution of the  
 264 count locations is shown in Figure 1.

265



266

267 **FIGURE 1 Case study network and count locations (based on: ARE, 2010)**

<sup>1</sup> ARE; Swiss National Transport Model (2010): A 4-step model, implemented in VISUM.



## 268 **Network Matching**

269 On the speed front, we utilize a commercial floating car data source that has emerged in the last  
270 years. Tom-Tom provides historical travel time databases for the entire network of Switzerland,  
271 including daily mean speed estimates. The acquired data correspond also to the basis year,  
272 however they are reported on a navigational network, which is considerably more detailed than  
273 the study network (1.5 million links). Naturally, the need for matching the two network arises  
274 in order to be able to integrate the historical travel time observations to the study network. The  
275 network matching is facilitated by developing an automated procedure that incorporates an  
276 adaptive radius search, operationalizing an edge matching approach. In brief, the matching is  
277 established based on two assumptions. First, the nodes of the study network should correspond,  
278 or at least be nearby to the actual nodes. Second, the links in the study network can be perceived  
279 as paths, meaning that their reported length should correspond to the total length of the links  
280 composing the path. At the first step, a circle with a 50 meters radius is drawn around each node  
281 of the study network in order to identify the Tom-Tom nodes that lie in the encompassed area  
282 (matched nodes). In the second step, for each pair of starting and ending nodes of the study  
283 network's links, the shortest path between all pairs of matched nodes are identified.  
284 Subsequently, the path with the lowest absolute deviation from the study network's link length  
285 is chosen as the most probable one. However, if the deviation is higher than a threshold value  
286 (set to 2%), or a path is not identified at all, then the radius increases and the procedure starts  
287 anew. The radius increase happens in increments of 50 meters and it continues up to a maximum  
288 distance of 300 meters, allowing for a maximum number of 6 iterations.

289  
290 Nonetheless, the incompatibility of the two networks surfaces in various instances (e.g.  
291 presence of a node where no actual intersection exists), giving rise to erroneous matching. As  
292 a remedy, the matched paths are checked visually to conclude on whether or not a correct  
293 identification is achieved. For the given problem at hand involving 831 links, the accuracy of  
294 the developed matching routine is found to be close to 70%. For the remaining cases, the  
295 matching is conducted manually to ensure that no systematic error is introduced due to  
296 mismatching. Interestingly, a comparison of the common attributes between the two networks  
297 reveals that in many cases attributes such as free flow speed, speed limit, number of lanes, are  
298 not aligned. This is anticipated to a greater extent, since in less detailed networks a single link  
299 can correspond to a number of links with varying attributes. To mitigate the impact of a wrong  
300 specification on the attributes side, the length weighted attributes of the links forming each path  
301 are adopted as attributes of the study network's links. An exception is made on the length  
302 attribute which we assume that is correctly specified in the study network. In the case of free-  
303 flow travel time, the mean relative difference between the two is found to be 13.30%, with a  
304 standard deviation of 19.70%, supporting the argument that less detailed networks have a higher  
305 degree of abstraction on their attributes specification.

306

## 307 **Explanatory Variables**

308 Having as an objective to model mean speed values on a nationwide network, a set of  
309 explanatory variables need to be included in the model specification, either directly if they  
310 comply with the exogeneity assumption, or indirectly as instruments of the traffic volume.  
311 Before proceeding further, a closer look at the phenomenon we are aiming to model can provide  
312 valuable insights. Essentially, mean speed is the outcome of the interaction between two  
313 interrelated mechanisms, supply and demand.

314

315 On the supply side, link characteristics associated with the design and the operation aspects are  
 316 the main determinants of the link's capacity. Therefore, variables such as the free-flow speed  
 317 and the number of lanes determine to a large extent the capacity. Other variables such as the  
 318 link's hierarchy type (e.g. highway, etc.) have been also found in the literature to affect the  
 319 capacity, however the inclusion of such variables can lead to multi-collinearity problems since  
 320 they are highly correlated with free flow speed. Variables such as curvedness, and link type  
 321 (tunnel or not) are expected to have a two-way impact on the capacity. At first, indirectly  
 322 through affecting the free flow speed values, and directly either by affecting the driving  
 323 behavior (e.g. more alert drivers), or by the existence of driving restrictions (e.g. prohibited  
 324 overtaking).

325  
 326 On the demand side, variables such as the population density and the potential number of  
 327 persons passing through (accessibility weighted centrality) can clearly be identified as the main  
 328 determinants of the travel demand. Variables associated with the network design (e.g. stress  
 329 centrality) are also expected to exert some influence on the demand. It should be noted that  
 330 stress centrality is defined as the number of shortest paths connecting all pairs of nodes of the  
 331 network that pass through a link. Variables such as the link type can be considered as not  
 332 directly related to the demand, however they bear on the ability of capturing the character of  
 333 the surrounding area, and thus of different demand aspects. Last, the presence of public  
 334 transport stops in the vicinity of a link can be viewed as an economic activity indicator, capable  
 335 of generating additional demand.

336  
 337 Clearly, supply and demand are interrelated. Nevertheless, in the case of a given national  
 338 network we can assume that the interaction between demand and supply on a link level, is not  
 339 affecting the demand. This assumption is tested in the following section. The summary statistics  
 340 of the different employed variables are presented in Table 1.

341  
 342 **TABLE 1 Summary statistics for variables**

343

Variable	Mean	Stand. Deviation
Travel time difference [sec.]	5.71	5.81
Free-flow travel time [sec.]	120.58	97.94
Free-flow speed [km/h]	78.14	24.60
Free-flow speed > 90 km/h [dummy]	0.41	-
One-lane road [dummy]	0.67	-
Mean curvedness [degrees]	0.04	0.06
Tunnel percentage [%]	0.13	0.24
AADT [vehicles]	13968.40	13906.08
<i>AADT Instruments</i>		
Population density (kernel, R=10km) [residents/ sq. km]	593.45	647.28
Freeway [dummy]	0.44	-
Rural main road [dummy]	0.19	-
Main road [dummy] * Public Transp. stops within 2 km radius	6.36	15.35
Accessibility-weighted centrality [persons]	4736.64	6291.38
Stress centrality [crossings]	7927228	13033574

344  
 345 The involved data processing, network matching, and model estimation is undertaken with the  
 346 statistical programming language R (20), making use of additional packages (21, 22).

## 347 **RESULTS – DISCUSSION**

348 Having identified the set of potential variables, we proceed to the model estimation. In total,  
349 our sample consists of 420 links. It should be mentioned that for each count location with  
350 bidirectional traffic, only one of the two directions is randomly chosen and included in the  
351 sample. This choice is made because the available AADT data are reported per location, and  
352 not per link. In the absence of information regarding the shares per direction, we choose to  
353 assume that AADT is equal on both directions.

354

### 355 **Model Estimation**

356 At first, an IV model is estimated by means of a 2SLS estimator to account for the endogeneity  
357 of AADT. The estimated model corresponds to the one presented in formula 8, where the  
358 dependent and the endogenous variable are in a logarithmic form, while the capacity is replaced  
359 by a number of proxy variables. The 2SLS model serves as the benchmark model for checking  
360 for the presence of spatial autocorrelation, hence drawing conclusions on the need to utilize the  
361 aforementioned 2IV/GMM estimator.

362

363 The variables presented in Table 1 are chosen as instruments, whereas a bigger set of variables  
364 was thoroughly tested as well on their ability to serve as instruments. A number of statistical  
365 tests is performed in order to conclude on the presence of endogeneity, and on the ability of the  
366 instruments to comply with the prerequisites. At first, a weak instruments test is performed  
367 through the formation of an F-test on the instruments. More specifically, the null hypothesis of  
368 weak instruments is rejected with a lower than 0.1% p-value. The presence of endogeneity is  
369 checked with the Wu-Hausman test (23) and the null hypothesis of no endogeneity is rejected  
370 at the 5% level. Last, the validity of the instruments is tested with the Sargan test (24). The null  
371 hypothesis of the instruments validity (exogenous) fails to be rejected at any of the examined  
372 levels. In summary, the performed statistical endogeneity tests demonstrate clearly that AADT  
373 is indeed endogenous, while the chosen instruments are found to be statistically valid. The  
374 existence of multi-collinearity issues is checked with variance inflation factors, and found not  
375 be the case. In addition, it is worthwhile to put into perspective that the adjusted R square of  
376 the corresponding AADT model with respect to the employed set of instruments is 0.829,  
377 demonstrating that the chosen instruments are strong predictors of AADT. Last, the presence  
378 of simultaneity bias between speed and AADT is tested by formulating an AADT model and  
379 instrumenting the speed. The results validate our hypothesis that for the current setting there is  
380 only a one-way endogeneity issue. Nonetheless, if that was not the case the estimation of a  
381 structural equation model with spatial considerations would constitute the appropriate way to  
382 tackle the problem.

383

384 The estimated parameters are in line with our expectations about how the different variables  
385 affect the travel time difference. Apart from the one-lane dummy variable, all the rest are found  
386 to be statistically significant at different levels. However, its inclusion in the model  
387 specification is justified on the basis of improving the overall goodness of fit. Furthermore, 11  
388 observations with high leverage are identified and excluded from the dataset based on Cook's  
389 distance. The estimates of the 2SLS with IV model is presented in Table 2. The detailed results  
390 for the log(AADT) prediction are omitted, but are available on request.

391

392 **TABLE 2 Models Estimates**

Dependent variable=log( $\Delta$ TT)	IV (2SLS)			IV with spatial error (2IV/GMM)		
	Estimate	Sign.	Std. Error	Estimate	Sign.	Std. Error
log(free-flow travel time)	1.02	***	(0.040)	1.04	***	(0.040)
log(free-flow speed)	-1.29	***	(0.128)	-1.31	***	(0.147)
Free-flow speed>90km/h (dummy)	-0.27	**	(0.095)	-0.21	*	(0.085)
One-lane road (dummy)	0.12		(0.085)	0.13		(0.097)
Mean road curvature	-1.56	**	(0.509)	-1.44	*	(0.577)
Tunnel percentage	-0.38	**	(0.134)	-0.35	.	(0.179)
log(AADT)	0.27	***	(0.047)	0.26	***	(0.058)
Lamda	-			0.80	***	(0.167)
Adjusted R-squared:			0.865			-
Breusch-Pagan test:	27.36	***				-
Moran's I measure:	0.09	***		-0.02		
<i>Endogeneity diagnostics</i>						
Weak Instrument (Ho=weak instr.)	114.86	***		114.86	***	
Wu-Hausman test (Ho=no endog.)	4.23	*		4.23	*	
Sargan test (Ho= valid instruments)	3.09			1.91		

*Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1, number of observations:409*

393

394 In order to check for the presence of spatial autocorrelation on the residuals of the 2SLS model,  
 395 and justify the choice to proceed to the estimation of the 2IV/GMM model, Moran's *I* measure  
 396 is utilized. More specifically, different spatial matrices variants are constructed and tested,  
 397 based on Euclidean and network distances. In the case of the latter, two distance metrics are  
 398 employed, the free-flow travel time and the network traveled distance. In particular, for the  
 399 Euclidean and the network distance, the Moran's *I* measure demonstrates that the  
 400 autocorrelation exists up to a radius of 20 and 30 kilometers respectively. In the case of the  
 401 network time, autocorrelation remains significant up to a radius of 25 minutes of free-flow  
 402 travel time. Finally, the last part of the construction of the spatial weight matrices concerns the  
 403 determination of the weights that should be assigned to the neighboring locations. Making use  
 404 again of the Moran's *I* measure, we conclude that an inverse distance metric is the most  
 405 appropriate to capture the spatial structure. Moreover, in order to avoid having misspecification  
 406 issues as those highlighted in (18), a so-called min-max normalization of the weights is applied.  
 407 Among the tested spatial weighting schemes, the one based on the free-flow travel time is  
 408 concluded to be the most pertinent one. The calculated Moran's *I* measure for this spatial matrix  
 409 indicates that spatial autocorrelation is statistically significant with a value of 0.09 (Table 2).  
 410 Thus, we should account for the spatial dependence in the model formulation by using the  
 411 2IV/GMM estimator.

412

413 Initially, a model with the spatial Durbin formulation is estimated. However, the spatial  
 414 autocorrelation parameter is found to be statistically insignificant while this is not the case for  
 415 the spatial autoregressive parameter. This finding indicates that a spatial error formulation  
 416 should be adopted, dropping the spatially lagged dependent variable. In addition, it points  
 417 towards the case of omitted spatial variable(s) as the underlying source of dependence. The new  
 418 parameter estimates differ slightly in comparison to the previous ones. The results of the spatial  
 419 error model with endogenous AADT, and heteroscedasticity robust standard errors, are

420 presented in Table 2. On the endogeneity front, the Sargan test has to be modified to take into  
 421 account the innovations of the spatial error model. The results of the new version of the Sargan  
 422 test validate our prior results on the exogeneity of the chosen instruments.

423

### 424 **Predictive Performance**

425 Finally, the predictive performance of the estimated models versus the Swiss national transport  
 426 model (4-step model), is evaluated in order to allow us drawing some solid conclusions.  
 427 Thereof, the predictive performance is evaluated in terms of different accuracy measures such  
 428 as the mean absolute error, allowing a quantification and comparison to take place.

429

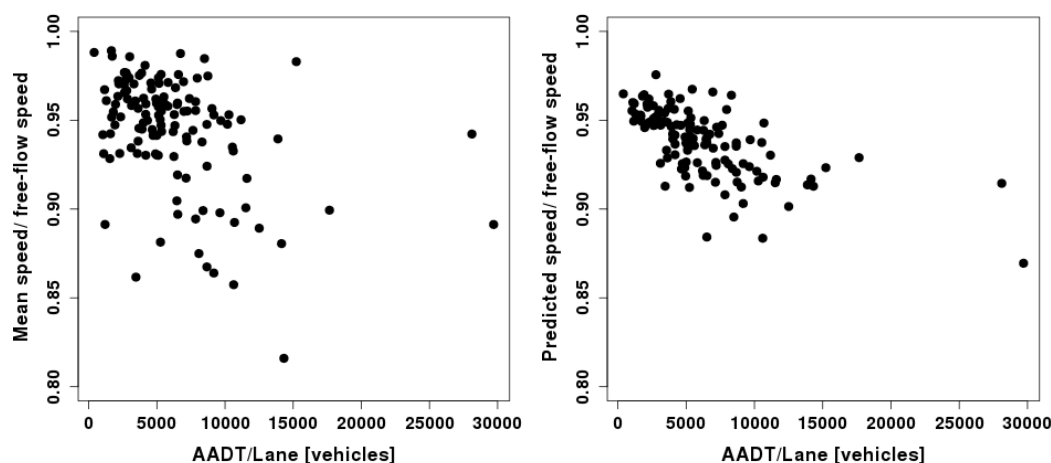
430 Given the identified differences on the travel times between the two networks, the choice of a  
 431 different metric is more appropriate. More specifically, instead of measuring the accuracy of  
 432 predicting the actual travel time differences, we measure the accuracy of predicting the relative  
 433 congestion, defined as the relative decrease on the free-flow speed values. It should be noted  
 434 that given the log transformation of the dependent variable, when back transforming to the  
 435 original scale we account for the fact that the model predicts the geometric mean instead of the  
 436 arithmetic one, in the way suggested by Wooldridge (15). Omitting this correction will give rise  
 437 to systematic underestimation problems. Two variations of the different accuracy measures are  
 438 calculated, a simple one and an AADT-weighted one that allows us to generalize our findings  
 439 on a national level (e.g. whole population). The calculated predictive accuracy measures are  
 440 presented in Table 3. The results demonstrate that the estimated statistical models outperform  
 441 substantially the 4-step model, while between the two models the spatial one exhibits slightly  
 442 better predictive accuracy. Interestingly, in the case of the AADT-weighted measures the mean  
 443 error of the regression models is around 32%, while the 4-step model yields a value of higher  
 444 than 100%. In the case of the symmetric mean absolute error, a measure which is less influenced  
 445 by the presence of outliers, the magnitude of the measure is substantially lower, but still the  
 446 statistical models outperform the 4-step model by more than a factor of two. However, one has  
 447 to remember that the 4-step model was calibrated against volumes and not against speeds.

448 **TABLE 3 Models Predictive Accuracy for Relative Speed Reductions**

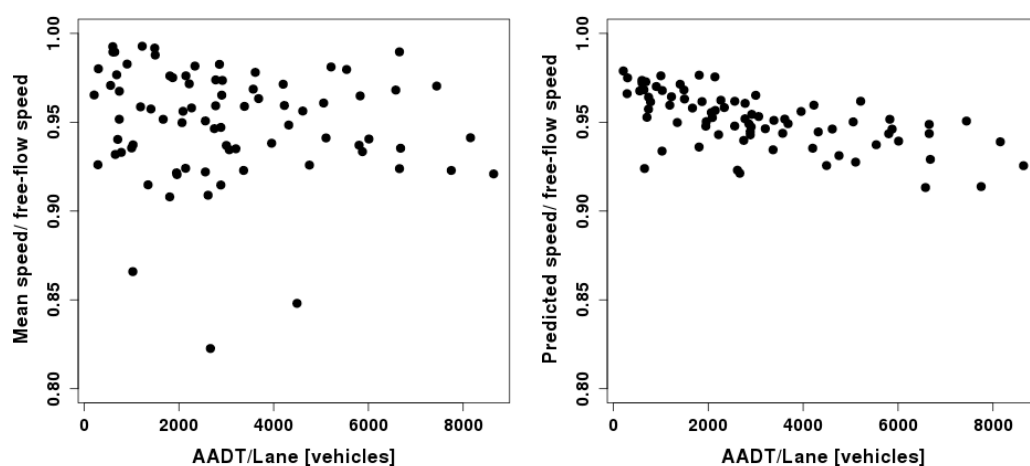
Model	Median Abs Error[%]	Mean Abs. Error[%]	Mean Error[%]	Symmetric Mean Abs. Error[%]
2SLS	38.60	63.78	39.82	10.95
2SLS: AADT weighted	41.25	63.67	32.31	11.60
2IV/GMM	37.12	63.08	38.43	10.92
2IV/GMM: AADT weighted	42.28	63.22	31.21	11.61
4-step model	89.21	133.15	41.04	27.06
4-step model: AADT weighted	100.66	167.02	117.63	23.27

449

450 At last, a visual evaluation of the predictive accuracy of the estimated spatial regression model  
 451 (2IV/GMM) is attempted in the following plots (Figure 2). More specifically, the ratios of mean  
 452 speed to free flow speed are plotted versus the AADT values per lane (as a simplified capacity  
 453 proxy) to visualize the predictive ability of the estimated model. In general, it appears that the  
 454 model has the ability to reproduce relatively well the relationship between the demand and  
 455 congestion.



(a) Main roads



(b) Rural main roads

456  
457  
458  
459  
460 **FIGURE 2 Actual speed ratios (left) versus the predicted ones (right) for link types of main road**  
461 **(a) and rural main road (b).**

## 462 CONCLUSIONS

463  
464 In the present paper a methodology to estimate mean speed values on a large scale network was  
465 presented, treating both the endogeneity and spatial dependence aspects. A particular focus was  
466 given on the instrumentation of demand to allow obtaining consistent and unbiased parameter  
467 estimates, thus making the model capable both for prediction and forecasting applications. Our  
468 findings suggest that a correctly specified statistical model has the ability to provide accurate  
469 estimates, outperforming a much more complex and data demanding transport planning model,  
470 even though the superiority of such models is taken for granted in many cases. Nonetheless, the  
471 low predictive accuracy of the 4-step model is alarming and it raises some well-founded  
472 concerns regarding the reasonableness of the speed predictions of such models, an aspect which  
473 is normally neglected in the calibration processes. Taking into account that such models  
474 normally constitute the medium for the evaluation of different policies and projects on a  
475 national level, the implications of unreasonable speed predictions can be rather huge.

476  
477 In addition, the developed modeling approach if coupled with an AADT regression model (e.g.  
478 the one implied by the chosen instruments), it forms a coherent direct demand modeling

479 approach which makes use only of aggregated data. A direct demand modeling approach has  
 480 the apparent advantage that it can be set up within a short time frame with very low associated  
 481 computational, maintenance, and monetary costs, while it can still provide the required answers  
 482 for a number of transport planning problems. Last, the developed approach can supplement the  
 483 existing simulation approaches to improve their predictions and overall their function.

484

#### 485 **ACKNOWLEDGEMENTS**

486 This paper is based on an ongoing research project funded by the Swiss National Science Foundation  
 487 entitled “Models without (personal) data?” (Project number 144134).

488

#### 489 **REFERENCES**

- 490 1. Horni, A., K. Nagel, and K. W. Axhausen. *The Multi-Agent Transport Simulation MATSim*.  
 491 Ubiquity Press, London, 2016.
- 492 2. Flyvbjerg, B., M. K. Skamris Holm, and S. L. Buhl. How (In) accurate are demand forecasts in  
 493 public works projects. *Journal of the American planning association*, Vol. 71, 2005.
- 494 3. Dowling, R., and A. Skabardonis. Improving Average Travel Speeds Estimated by Planning  
 495 Models. *Transportation Research Record*, 1993, pp. 68–74.
- 496 4. Sarlás, G., and K. W. Axhausen. Localized speed prediction with the use of spatial simultaneous  
 497 autoregressive models. *Paper presented at the 94th Annual Transportation Research Board*  
 498 *Meeting, Washington D.C*, 2015.
- 499 5. Hackney, J. K., M. Bernard, S. Bindra, and K. W. Axhausen. Predicting road system speeds  
 500 using spatial structure variables and network characteristics. *Journal of Geographical Systems*,  
 501 Vol. 9, No. 4, 2007, pp. 397–417.
- 502 6. Park, Y. J., and F. F. Saccomanno. Evaluating speed consistency between successive elements  
 503 of a two-lane rural highway. *Transportation Research Part A: Policy and Practice*, Vol. 40, No.  
 504 5, 2006, pp. 375–385.
- 505 7. *HCM 2010: Highway Capacity Manual, Washington, D.C.: Transportation Research Board,*  
 506 *2010*.
- 507 8. Shankar, V., and F. Mannering. Modeling the endogeneity of lane-mean speeds and lane-speed  
 508 deviations: a structural equations approach. *Transportation Research Part A: Policy and*  
 509 *Practice*, Vol. 32, No. 5, 1998, pp. 311–322.
- 510 9. Himes, S. C., E. T. Donnell, and D. Ph. Speed Prediction Models for Multilane Highways:  
 511 Simultaneous Equations Approach. *Journal of Transportation Engineering*, Vol. 136, 2010, pp.  
 512 855–862.
- 513 10. Quddus, M. Exploring the Relationship Between Average Speed, Speed Variation, and Accident  
 514 Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*,  
 515 Vol. 5, No. 1, 2013, pp. 27–45.
- 516 11. Cheng, W., J.-H. Wang, G. Bryden, X. Ye, and X. Jia. An Examination of the Endogeneity of  
 517 Speed Limits and Accident Counts in Crash Models. *Journal of Transportation Safety &*  
 518 *Security*, Vol. 5, No. 4, 2013, p. pp 314-326.
- 519 12. Porter, R. J., and J. S. Wood. Exploring Endogeneity of Macroscopic Speed Parameters:  
 520 Empirical Study During Low Volume Conditions in Construction Work Zones. *Transportation*  
 521 *Letters*, Vol. 5, No. 1, 2013, pp. 27–37.
- 522 13. Boulter, P., I. McCrae, and T. Barlow. *A review of instantaneous emission models for road*  
 523 *vehicles*. TRL Limited. 2007.
- 524 14. Sarlás, G., and K. W. Axhausen. Prediction of AADT on a nationwide network based on an  
 525 accessibility- weighted centrality measure. *Arbeitsberichte Verkehrs- und Raumplanung*, No.  
 526 1094, IVT, ETH Zurich, Zurich., 2015.
- 527 15. Wooldridge, J. M. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.
- 528 16. Anselin, L. *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business  
 529 Media, 2013..
- 530 17. Kelejian, H. H., and I. R. Prucha. A Generalized Moments Estimator for the Autoregressive  
 531 Parameter in a Spatial Model. *International Economic Review*, Vol. 40, No. 2, 1999, pp. 509–  
 532 533.
- 533 18. Kelejian, H. H., and I. R. Prucha. Specification and estimation of spatial autoregressive models

- 534 with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, Vol. 157, No. 1,  
535 2010, pp. 53–67.
- 536 19. Drukker, D. M., P. Egger, and I. R. Prucha. On Two-Step Estimation of a Spatial Autoregressive  
537 Model with Autoregressive Disturbances and Endogenous Regressors. *Econometric Reviews*,  
538 Vol. 32, No. 5–6, 2012, pp. 686–733.
- 539 20. Team, R. C. R: A language and environment for statistical computing. *Vienna, Austria: R*  
540 *Foundation for Statistical Computing*, 2014, pp. 3–36.
- 541 21. Csárdi, G., and T. Nepusz. The igraph software package for complex network research.  
542 *InterJournal Complex Systems*, Vol. 1695, 2006, p. 1695.
- 543 22. Piras, G. sphet: Spatial Models with Heteroskedastic Innovations in R. *Journal of Statistical*  
544 *Software*, Vol. 35, No. 1, 2010.
- 545 23. Greene, W. H. *Econometric Analysis*. Pearson, 2007.
- 546 24. Sargan, J. D. The Estimation of Economic Relationships using Instrumental Variables.  
547 *Econometrica*, Vol. 26, No. 3, 1958, p. 393.  
548