Mean speed prediction with endogenous volume and spatial autocorrelation

A Swiss case study

Working Paper

Author(s):
Sarlas, Georgios; Axhausen, Kay W.

Publication date:
2017-08

Permanent link:
https://doi.org/10.3929/ethz-b-000175412

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 1275

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

Georgios Sarlas*
Institute for Transport Planning and Systems, ETH Zurich
HIL F 51.3, Stefano-Franscini-Platz 5, 8093, Zurich, Switzerland
Tel: +41 44 633 37 93; E-mail: georgios.sarlas@ivt.baug.ethz.ch

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

Working paper 1275, IVT ETH Zurich

* Corresponding author

ABSTRACT

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

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

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

INTRODUCTION

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

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

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

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

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

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

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

Description of the Framework

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

METHODOLOGY

General Overview

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

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

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

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

where λ the spatial autoregressive coefficient, W the spatial weight matrix with dimensions NxN, u a vector of disturbances, and ε a vector of iid error terms (innovations).

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

where ρ is a spatial autocorrelation parameter.

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

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

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

Modeling Approach

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

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

$$\sigma_{ij}(e) = \sum_{i,j \in V} Popul_i \frac{Employm_j * f\left(cost_{ij}^{car}\right)/(f(cost_{ij}^{car}) + f(cost_{ij}^{PuT}))}{Travel\ Accessibility_i} \quad (4)$$

$$Travel\ Accessibility_{\mathrm{i}} = \sum_{\mathrm{i}}^{\mathrm{j}} \mathrm{Employm_j} * \max\left\{\mathrm{f}\left(\mathrm{cost_{ij}^{car}}\right), f\left(cost_{ij}^{PuT}\right)\right\} \quad (5)$$

$$f\left(cost_{ij}^{mode}\right) = e^{\beta * cost_{ij}{}^{a}} \quad (6)$$

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

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

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

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

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

CASE STUDY

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

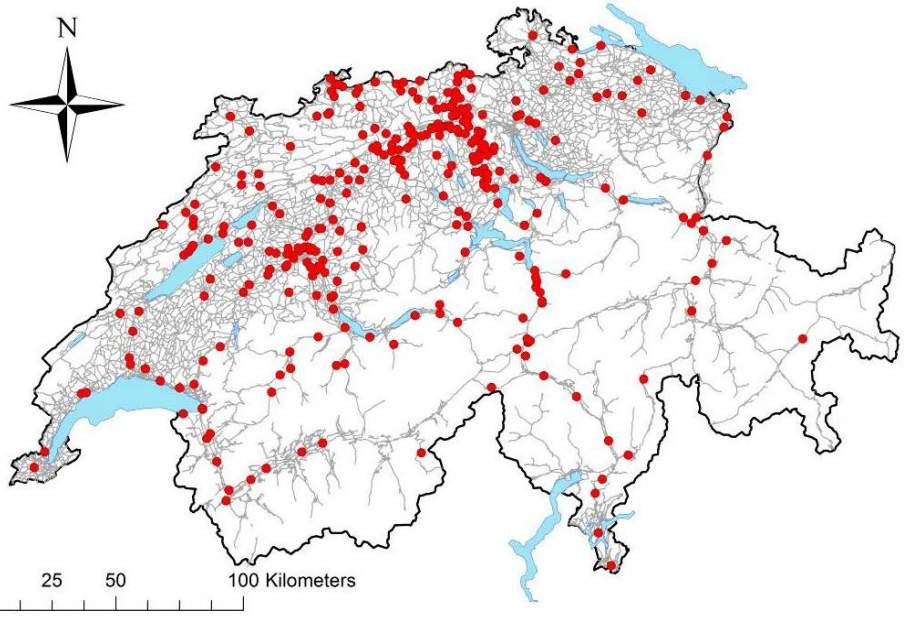


FIGURE 1 Case study network and count locations (based on: ARE, 2010)

[1] ARE; Swiss National Transport Model (2010): A 4-step model, implemented in VISUM.

Network Matching

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

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

Explanatory Variables

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

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

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

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

TABLE 1 Summary statistics for variables

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>90km/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
AADT Instruments		
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

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

RESULTS – DISCUSSION

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

Model Estimation

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

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

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

TABLE 2 Models Estimates

	IV (2SLS)			IV with spatial error (2IV/GMM)		
Dependent variable=log(ΔTT)	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		
Endogeneity diagnostics						
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*

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

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

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

Predictive Performance

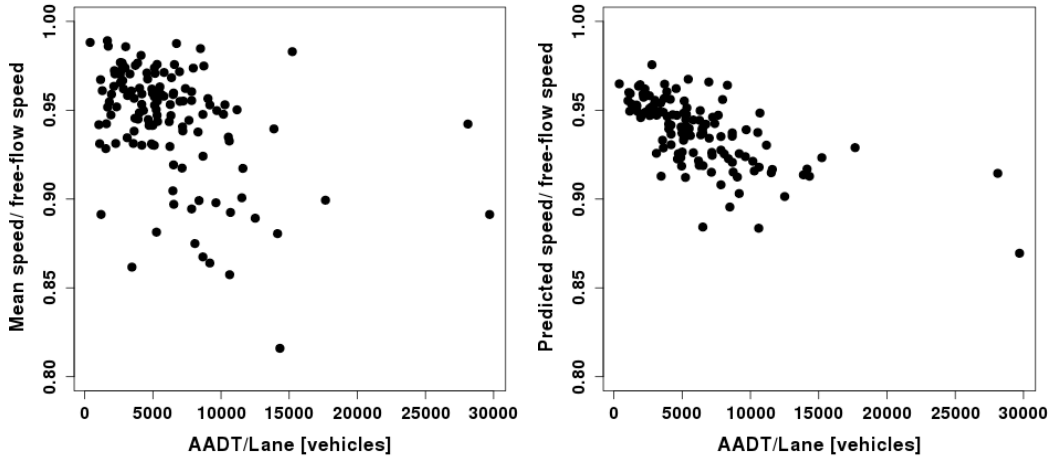
Finally, the predictive performance of the estimated models versus the Swiss national transport model (4-step model), is evaluated in order to allow us drawing some solid conclusions. Thereof, the predictive performance is evaluated in terms of different accuracy measures such as the mean absolute error,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

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

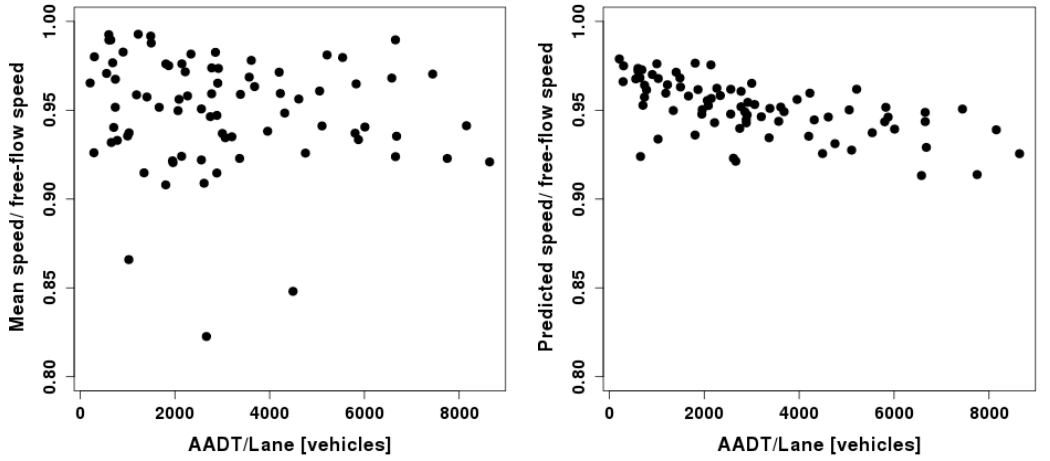
TABLE 3 Models Predictive Accuracy for Relative Speed Reductions

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

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



(a) Main roads



(b) Rural main roads

FIGURE 2 Actual speed ratios (left) versus the predicted ones (right) for link types of main road (a) and rural main road (b).

CONCLUSIONS

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

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

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

ACKNOWLEDGEMENTS

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

REFERENCES

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

with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, Vol. 157, No. 1, 2010, pp. 53–67.

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