Localized speed prediction with the use of spatial simultaneous autoregressive models

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

Publication Date:
2015

Permanent Link:
https://doi.org/10.3929/ethz-b-000087025 →

Localized speed prediction with the use of spatial simultaneous autoregressive models

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

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

* Corresponding author

Total Word Count: **5.547 (Text) + 6 (Figures/Tables) * 250 = 7.047 equivalent words**

Transportation Research Board 94th Annual Meeting, January 11-15, 2015, Washington, D.C.

ABSTRACT

This paper examines how to employ spatial regression modelling as a direct demand modelling approach to provide localized speed estimates in a large scale network. In particular, four different spatial simultaneous autoregressive (SAR) models are estimated and compared to an ordinary linear regression in order to highlight and evaluate their capability of explaining transport related phenomena and resolving issues that arise from the underlying spatial dependence. A particular focus is given on the identification, the construction, and the selection of the spatial weighting matrices. We conclude that the spatial autocorrelation (SAC) model outperforms the other SAR models, resolving spatial dependence issues, and thus is the proposed one for speed prediction purposes.

INTRODUCTION

Travel demand models have increased their data demands massively both in scope and scale, and in addition their complexity has increased in a similar way. The obvious reluctance of the practice to adopt such advanced models, raises the concern that the gap between academia and practice has become wider than ever *(1)*. On the one hand, the increased data collection abilities of the field, along with the expected wave of "big data" might allow the (academic) field to continue on its current trajectory, but on the other hand the use and abuse of big data raises the danger of a sudden change in the course of public policy and the sudden lack of high quality alternatives to the existing state-of-the-art (i.e. academic) models. At this point, it is tempting to oppose this trend and explore the formulation of an alternative direct travel demand model structure that requires only aggregate and anonymous data. Nevertheless, the alternative modelling structure should be able to make statements about the speed and the traffic volume on a link level, the items that constitute the minimum requirements for the transport project appraisal. This paper is a step in that direction. The alternative it explores is the set of spatial econometrics techniques, and more specifically spatial autoregressive models. These allow to link the relevant variables describing the demand with the network characteristic giving us a way to capture the demand capacity interaction at the core of transport modeling.

Literature review and background

Spatial econometrics was popularized by Anselin *(2)*, and is defined as "the *domain that deals with the peculiarities caused by space in the statistical analysis of regional science models*. More specifically, these peculiarities are caused by the dependence and the heterogeneity of data in space (spatial effects). *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. Spatial heterogeneity is considered to be the lack of structural stability of the various phenomena over space, and also the lack of homogeneity of the spatial units of the observations."(2)*

Different modelling techniques have evolved over the years to account for the spatial effects. In the case of the spatial dependence, SAR models can account for it by the inclusion of relevant spatial autoregressive components. On the front of spatial heterogeneity, geographically weighted regression constitutes a technique which allows different relationships to exist in space, instead of a global relationship *(3)*.

A number of applications of spatial modelling techniques can be found in the urban analysis area. A comprehensive review of such applications is presented by Paez and Scott *(4)*. However, the presence of spatial effects constitutes a dimension which normally is neglected in the existing transport modelling approaches. Black *(5)* introduced and described the existence of autocorrelation among the variables in the context of networks, stating that *"spatial autocorrelation usually concerns itself with variable values at given locations being influenced by variable values at nearby or (contiguous) locations in a spatial context. Network autocorrelation concerns the dependence of variable values on given links to such values on other links to which it is connected in a network context."*

Wang et al. *(6)* review and assess the methodological issues that arise from the application of spatial models in transport**.** Nevertheless, there is a relatively limited number of applications employing spatial regression models for the explanation of how transport related phenomena, such as speed or flows, occur and evolve over the space. The correlation of speed observations is demonstrated by Bernard et al. *(7)*, pointing out the necessity of accounting for spatial dependency when it comes to the estimation of speed. Hackney et al. *(8)* demonstrate the plausibility of accounting for the spatial dependence in the estimation of speed where three SAR models were estimated and compared. Cheng et al. *(9)* examine the spatio-temporal dependence structure of road networks and they argue on the need of a dynamic spatial weight matrix for forecasting purposes on real-time data. Jenelius and Koutsopoulos *(10)* present a statistical network model for travel time estimation, allowing for correlation between travel times on different links based on a spatial moving average structure.

In a different context, Lopes et al. *(11)* examine the effect of the spatial dependence on the transportation demand models and specifically in the trip generation phase of the four step model. Selby and Kockelman *(12)* explore the application of two spatial methods for prediction of average daily traffic counts (universal kriging and GWR). Similarly, Zhao and Park applied GWR for the same purpose *(13)*. In another study *(14)*, a model to estimate average annual daily traffic (AADT) along a link using integrated spatial data from multiple network buffer bandwidths is presented. Spatial models have also been applied in the area of road crash models (e.g. *(15)*, *(16)*).

Moreover, spatial regression models have been applied in the modeling of hedonic housing residential rents. More specifically, Löchl and Axhausen *(17)* have modelled residential asking rents by applying SAR and GWR techniques in order to generate the required data for land use and transport simulation. In a similar way, Efthymiou and Antoniou *(18)* studied the direct and indirect effects of transportation infrastructure and policies on house prices and rents. In addition, spatial regression models have also found application in the problem of optimal site selection for facilities *(19)*.

Description of the framework of the paper

This work builds upon the work of Hackney et al. *(8)* and take it further by employing a larger network and utilizing a different source of data to enhance the understanding of the application of SAR models for speed predictions. In addition, the inclusion of additional variables in the model specification is explored while a particular focus is given to the construction of the weighting matrices and the identification of the optimum number of neighbors. Last, the results of four different SAR models are presented and evaluated in order to draw conclusions regarding the most appropriate model formulation to account properly for the spatial dependence of speeds.

METHODOLOGY

As mentioned above, the alternative of spatial econometrics techniques constitutes an option that is examined to be a direct demand modelling approach to provide localised speed predictions. Conceptually, it is arguable that a simplified approach cannot exhibit the predictive accuracy and the sensitivity of the existing approaches (4-step or agent-based models), however it cannot be overseen the fact that when it comes to the appraisal of public transport projects, as Flyvbjerg et al. *(20)* argue, the quality of the demand forecasts has not been improved over the years even though more complex and advanced models have been employed. The advantage of that choice, in comparison to the classical approaches, is that it is significantly less cumbersome to apply, less data demanding, and also offers a structural explanation of the observed transport phenomena such as speed and flows in a direct way. The underlying assumption and hypothesis is that by accounting properly for the impact of spatial dependence in the context of regression modelling, accurate demand forecasts can be provided.

Spatial regression models are defined as the use of regression models by accounting for the impact of spatial effects in their specification and estimation, avoiding the statistical problems such as unreliable statistical tests and biased and inconsistent estimated parameters. This is accomplished by incorporating in the model the information about the spatial structure of the data, in the form of a contiguity matrix. Spatial simultaneous autoregressive (SAR) models is a popular category of such models that they have been applied in many cases. As suggested by Ord *(21)*, their estimation should be conducted by means of maximum likelihood since the ordinary least square (OLS) estimation produces inconsistent estimates. The assumption of these models is that the response variable at each location is a combination of the explanatory variables at that location but also of the response of neighboring locations.

Three main types of SAR models can be found in the literature, each one having different characteristics based on their underlying assumptions about where the autoregressive process occurs (*(22)*, *(23)*). At first, the spatial error autoregressive model (SARerr) assumes that the spatial dependence is in the error term of the model, and thus the spatial autoregressive process is applied to it. As Elhorst states *(24)*, *" Interaction effects among the error terms are consistent with a situation where determinants of the dependent variable omitted from the model are spatially autocorrelated, or with a situation where unobserved shocks follow a spatial pattern".* The formulation of the model is:

$$Y = \beta X + u \ (1)$$

$$with \ \ u = \lambda W u + \varepsilon$$

where Y is a vector with N values of the dependent variable, β is a vector with the regression coefficients, X is a matrix with the independent variables, u the error term, λ the spatial autoregressive coefficient, W a matrix with the contiguity structure having dimensions N x N, and ε a vector of independent and identically distributed (iid) error terms.

The spatial lag autoregressive model (SARlag) assumes that the spatial dependence exists in the response variable (endogenous interaction effects), and applies the spatial autoregressive process to the response variable, treating it as a lagged variable. The formulation of the model is:

$$Y = \rho W Y + \beta X + \varepsilon \ (3)$$

where ρ is the spatial autocorrelation parameter, and WY is the term for the lagged variable.

The third type of autoregressive model, namely spatial lag of x (SLX), assumes that the spatial dependence exists in the independent variables (exogenous interaction effects), and applies the spatial autoregressive process to the independent variables. The formulation of the model is:

$$Y = \beta X + W X \gamma + \varepsilon \ (4)$$

Apart from the three main SAR models, where each one assumes different interaction effects, SAR models assuming more than one interaction effects can be formulated. More specifically, the spatial mixed autoregressive model (SARmix, also denoted as spatial Durbin model in some application (e.g. *(23)*) assumes that the spatial dependence exists in both the response and the independent variables. The formulation of the model is:

$$Y = \rho W Y + \beta X + W X \gamma + \varepsilon \ (5)$$

$$with \ \ \gamma = -\rho \beta$$

The spatial autocorrelation model (SAC) assumes that the spatial dependence exists both in the response variable and the error term. The formulation of the model in that case is:

$$Y = \rho WY + \beta X + u \ (6)$$

$$\text{with } u = \lambda Wu + \varepsilon$$

Adjacency matrices

A key aspect of the spatial regression models is to determine the spatial structure of the data. This is facilitated by the inclusion of a spatial weight matrix in the model specification. Thereupon, the spatial weight matrix incorporates in the model information about the extent of the neighborhood, the type of the adjacency, and the relative weight that should be assigned on the neighboring locations. In the transport network case, it specifies the expected direction and mechanism of influence.

CASE STUDY

In order to assess the plausibility of applying SAR models for localized speed prediction purposes, a large-scale case study is conducted. A part of the national network of Switzerland is selected, including the canton of Zurich and the neighboring cantons. In particular, the full road network of the North-East Switzerland is included in the chosen network. A navigational network is used, commercially available by Tom-Tom, including average daily speed estimations based on Tom-Tom GPS measurements for the majority of the links. In detail, the study network includes approximately 220.000 links (having excluded the secondary, or less important links) while the remaining links are classified based on five available types. In addition to the estimated speeds, the set speed limit is available. A map of the study network can be seen in Figure 1.

The average daily speed of a typical weekday is the dependent variable of interest for the regression. The regression yields two speed components; first, the average road speed which is a function of the speed limit, the link type, and the length, and constitutes a non-spatial quantity. Spatial variation is added to the link speed estimates in the second component through the spatially resolved explanatory variables. Spatially resolved road and public transport network densities represent the effect of road supply on speed. Spatial data on population and employment densities are taken to be indicative of the intensity of local activities, reflecting travel demand locally *(8)*.

FIGURE 1 Case study network

Spatially resolved variables

Apart from the network data that presented above, the spatial resolved variables constitute an important component of the regression model since they introduce variation in the estimated average values, as these result from the non-spatial component of the model. At first, the road and public transport densities are of apparent interest since they represent the effect of road supply and also the spatial competition between the private and public modes, especially in urban areas. The road density is estimated as the total length of links within a given area and it is calculated for different radii. The full navigational network is used for the density calculation. Besides the network's density, the density of ramp links is calculated as well as it is expected to have local impact on speed by given access to the motorway network. In the case of accounting for the impact of the public transport network on speed, it is less straightforward the way that a pertinent variable can be constructed. As an approximation, the density of public transport stops within a given area, is considered to be a good starting point.

Another source of spatially resolved variables corresponds to the demand impact on speed. More specifically, the socio-demographic data of interest are the population and the employment locations, aggregated per hectare available from the Swiss Federal Statistical Office (BFS: Bundesamt für Statistik). The population data were collected in the year 2011, while the employment data in the year of 2008. Given the disaggregate level of these data (hectar based), they are taken into account as densities over different radii. In addition to the normal densities, Gaussian kernel densities are calculated as well to account for the diminishing impact of the socio-demographic data over the space. At last, the spatially resolved variables need to be associated to the links of the network. Thereupon, each link of the network is associated with the hectare (cell) values of each spatial variable, closest to the upstream endpoint of the link.

MODEL ESTIMATION AND RESULTS

In this section the different regression models estimations are presented and compared to see, if they capture the impact of accounting properly for the spatial dependency of speeds. More specifically, a standard linear regression model is estimated in terms of ordinary least squares (OLS), while four SAR models are estimated subsequently. A comparison of the estimated models is conducted in order to shed some light on the plausibility of the SAR models to predict traffic related variables, as speed, and also to what extent they can accomplish that.

Linear regression model

At first, an OLS model is estimated to serve as the basis for testing the necessity of accounting properly for the spatial dependence (auto-correlation). It is expected that the OLS model is going to give rise to biased and inconsistent estimates and thus the resulted adjusted coefficient of determination will be inconsistent and not true.

In addition, OLS predicted values are going to be used for testing if spatial association exists in the residuals by estimating Moran's I measure. Depending on the results of the Moran's I, a justified explanation of whether or not the need to account for the spatial dependence properly arises. The independent variables that are included in the model are determined based on their predictive power and in accordance to the appropriate statistical tests, avoiding to give rise to multi-collinearity issues (none of the correlations is higher than 0.41). The summary statistics of the included independent variables are presented in Table 1, while the specification of the model and the estimated coefficients are presented in Table 2.

TABLE 1 Summary statistics of independent variables

Variables	Units	Mean	Median	St.Dev.
Speed-limit	km/ hour	52.930	50.000	13.28
Highways: Constant	dummy	0.027	-	-
Trunk roads: Constant	dummy	0.006	-	-
Collector roads: Constant	dummy	0.008	-	-
Distributor roads: Constant	dummy	0.392	-	-
Urban roads: Constant	dummy	0.567	-	-
Road curveness	degrees	0.048	0.000	0.186
Distributor: Public transp. stops density, r=0.5km	stops per sq. km	3.242	2.546	3.566
Urban: Public transp. stops density, r=0.2km	stops per sq. km	5.170	0.000	7.370
Highways: logarithm of population, r=5km	residents per sq. km	683.394	401.605	820.818
Trunk roads: logarithm of population, r=2km	residents per sq. km	955.504	468.234	1365.736
Collector roads: logarithm of employment positions, r=2km, kernel weighted	residents per sq. km	726.682	194.704	1798.524
Distributor roads: logarithm of employment positions, r=1km, kernel weighted	residents per sq. km	927.006	244.653	2408.971
Urban roads: logarithm of employment positions, r=0.5km, kernel weighted	residents per sq. km	1114.429	279.858	3003.314
Urban roads: ramps' density, r=1km	meters per sq. km	0.128	0.000	0.316
Distributor roads: road density, r=500 m	meters per sq. km	16.769	15.913	7.670
Urban roads: road density, r=100 m	meters per sq. km	28.841	27.858	11.215
Highways with length less than 0.1 km	dummy	0.441	-	-
Trunk roads with length less than 0.1 km	dummy	0.641	-	-
Collector roads with length less than 0.1 km	dummy	0.762	-	-
Distributor roads with length less than 0.1 km	dummy	0.739	-	-
Urban roads with length less than 0.1 km	dummy	0.705	-	-
Highways with length between 0.1 km and 0.2 km	dummy	0.169	-	-
Trunk roads with length between 0.1 km and 0.2 km	dummy	0.103	-	-
Collector roads with length between 0.1 km and 0.2 km	dummy	0.064	-	-
Distributor roads with length between 0.1 km and 0.2 km	dummy	0.067	-	-
Urban roads with length between 0.1 km and 0.2 km	dummy	0.082	-	-
Number of observations	220599			

TABLE 2 Estimated OLS coefficients

Y = Average daily speed Explanatory variables	Coeff.	Std. Error	t value	signif.
Speed-limit	0.472	0.003	185.229	***
Highways: Constant	73.801	0.945	78.063	***
Trunk roads: Constant	52.559	1.181	44.508	***
Collector roads: Constant	54.919	1.002	54.836	***
Distributor roads: Constant	45.655	0.235	193.923	***
Urban roads: Constant	34.975	0.199	176.196	***
Road curveness	-10.420	0.109	-95.393	***
Distributor: PuT stops density, r=0.5km	-0.339	0.011	-31.605	***
Urban: PuT stops density, r=0.2km	-0.149	0.004	-34.735	***
Highways: ln(population), r=5km	-3.795	0.134	-24.443	***
Trunk roads: ln(population), r=2km	-2.939	0.162	-18.165	***
Collector roads: ln(employment), r=2km,kernel weighted	-3.529	0.127	-27.780	***
Distributor roads: ln(employment), r=1km, kernel weighted	-1.822	0.030	-61.485	***
Urban roads: ln(employment), r=0.5km, kernel weighted	-0.937	0.017	-56.251	***
Urban roads: Ramps' density, r=1km	-0.666	0.119	-5.612	***
Distributor roads: Road density, r=500 m	-0.263	0.006	-42.270	***
Urban roads: Road density, r=100 m	-0.152	0.003	-54.576	***
Highways: Length < 0.1 km, Dummy	-3.897	0.282	-13.799	***
Trunk roads: Length < 0.1 km, Dummy	-10.902	0.733	-14.867	***
Collector roads: Length < 0.1 km, Dummy	-11.611	0.798	-14.550	***
Distributor roads: Length < 0.1 km, Dummy	-6.439	0.114	-56.347	***
Urban roads: Length < 0.1 km, Dummy	-4.648	0.090	-51.616	***
Highways:0.2 km > Length > 0.1 km, Dummy	-2.812	0.339	-8.289	***
Trunk roads: 0.2 km > Length > 0.1 km, Dummy	-4.222	0.868	-4.863	***
Collector roads: 0.2 km > Length > 0.1 km, Dummy	-6.447	0.955	-6.753	***
Distributor roads: 0.2 km > Length > 0.1 km, Dummy	-0.263	0.006	-42.270	***
Urban roads: 0.2 km > Length > 0.1 km, Dummy	-1.989	0.103	-19.343	***
adjusted R-square		0.964		
Log-likelihood		-807851.8		
AIC		1615760		
Observations		220599		
Signifance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

As it can be seen, the adjusted R square is extremely high while the estimated parameters are all statistically significant. Employment locations and population densities are not used at the same time due to their high correlation. Notably, a differentiation of the employed densities radius for different links' types is found to be more appropriate and thus chosen, instead of a fixed radius density for all links' types. This finding exhibits that depending on the type of the link, the impact of spatial resolved variables on speed is not homogeneous, indicating a rather localized impact in the case of lower link types. Ramps' density variables have a negative impact on speed which can be explained by the fact that the higher the density of ramps, more vehicles are diffused in the adjacent local road network, leading to higher traffic loads. The negative sign of the line density is not according to our expectations since it would be more reasonable to assume that the higher the local supply of roads, the more alternative routes exists, and thus lower congestion occurs. However, the sign of line density exhibits the opposite which reveals that locally, the higher number of roads corresponds to more intersections where the lower classified links often have to yield priority. The impact of short length variables differs among the link types and shows that the short length has an impact on the speed, possibly because of the close proximity of intersections.

Estimation of spatial regression models

The first key aspect before proceeding to the estimation of the spatial regression models is to examine the existence of spatial dependence in the data and thus justify if the need for the estimation of spatial regression models arises. This is accomplished by calculating the Moran's I measure which is a measure of the spatial correlation of the OLS error terms. However, in order to facilitate this calculation, the spatial structure of data should be defined beforehand, in the form of an adjacency matrix.

The inclusion of the adjacency in the model specification incorporates information in the model about the extent of the neighborhood, the type of the adjacency, and the relative weight that should be assigned on the neighboring locations.In order to identify the optimum spatial matrix for the problem at hand, the impact of different adjacency matrices type is assessed. Consequently, three different adjacency matrices schemes are constructed and tested thoroughly; one that identifies all the k-nearest neighbors based on the Euclidean distance, one that identifies the k-th order nearest neighbors in terms of network distance, and last, one where only the k-th order straight movements are included in the adjacency matrix.

For all three schemes, two variations of the adjacency matrix are examined to conclude if the assigned weight should be uniform for all observations (denoted as normal), or calculated based on a weighting function aiming in capturing the diminishing dependence of links over the distance or order (distance and order decay). In the first case, the weight is defined based on the inverse distance of each link from the midpoint of the base link. In the case of the other two weighting schemes, the weight is assigned as the inverse of the order of connectivity (k). Moreover, in all tested cases, each row of the weighing matrices is standardized to one.

The spatial autocorrelation of the OLS residuals is estimated in terms of the Moran's I measure, for the different developed spatial weighting schemes, and it is found to lie within the range of 0.5 to 0.7, being statistically significant. This finding reveals the presence of strong spatial autocorrelation in the residuals of the OLS models, that should be treated through the estimation of SAR models.

Subsequently, four SAR models are estimated for different number of neighbors and weighting schemes, namely the SARerror, the SARlag, the SAC, and the Durbin model. The optimum number of neighbors for each model is identified on the basis of minimizing the Akaike Criterion (AIC).

The estimation of the SAR models and the construction of the weighting matrices was conducted in R *(25)*, making use of the package "spdep"*(26)*. It should be noted that to facilitate computationally the estimation of the SAR models, the LU method for the decomposition of sparse matrix is used *(27)*. The analytical results are presented in Table 3.

At first, the optimum number of neighbors for each scheme and variation is found to be the same for all the estimated models. More specifically, for the case of the k-nearest adjacency matrix, the optimum number of neighbors is equal to six for the normal weighting scheme, while for the distance decay weighting scheme it is eight. In the case of the other two weighting schemes, the optimum number of k-th order neighbors is found to be one. An interesting finding is that the application of an order decay function produces better results compared to the normal weighting function, a trend which is not present in the case of the k-nearest weighting scheme. Nevertheless, it can be concluded that the third weighting scheme, including the 1st order straight links, gives the best results and hence is the one employed for the comparison and the evaluation of the SAR models that follows.

TABLE 3 Measures of quality of fit for SAR models for different weight matrices

AIC		SARerror		SARlag		SAC		Durbin	
Weighting scheme	k-nearest	normal	decay	normal	decay	normal	decay	normal	decay
k-nearest	3	1474896	1498912	1507270	1525942	1474050	1492775	1460516	1484429
	4	1459532	1479500	1497931	1512246	1459090	1473969	1446664	1465823
	5	1458598	1473010	1498341	1507921	1458480	1469365	1446866	1460196
	6	**1454827**	1467068	**1496637**	1504080	**1454806**	1464987	**1444357**	1455146
	7	1457784	1465396	1499480	1503302	1457630	1464304	1448472	1454361
	8	1459152	**1463702**	1501433	**1502640**	1458607	**1463252**	1450904	**1453379**
	9	1463834	1463964	1505937	1503471	1463135	1463809	1456580	1454368
	10	1466634	1463821	1508881	1503907	1465394	1463811	1460209	1454815
	k-th order	normal	decay	normal	decay	normal	decay	normal	decay
network	1	**1426432**	**1426432**	**1459870**	**1459870**	**1425438**	**1425438**	**1414796**	**1414796**
	2	1463463	1443308	1498131	1481536	1459081	1437002	1456881	1436600
	3	1489760	1456989	1523922	1497499	1485152	1448503	1485180	1452220
	4	1508012	1466917	1541482	1509124	1503793	1457342	1504598	1463234
	5	1521512	1474590	1554187	1518153	1517834	1464451	1518795	1471537
	k-th order	normal	decay	normal	decay	normal	decay	normal	decay
straight network	1	**1410453**	**1410453**	**1466226**	**1466226**	**1388647**	**1388647**	**1390509**	**1390509**
	2	1454948	1434346	1489788	1480813	1424690	1404782	1433716	1415466
	3	1481920	1447590	1505595	1490135	1449457	1415911	1459457	1429576
	4	1499128	1455559	1516544	1496516	1466190	1423346	1475865	1438366
	5	1510986	1460843	1524538	1501180	1478142	1428652	1487132	1444354

RESULTS AND DISCUSSION

In Table 4, the estimated coefficients, along with the relevant goodness of fit measurements, can be seen. In summary, the coefficients of the OLS model differ significantly from the corresponding ones of the SAR models, reflecting that in the absence of accounting properly for the spatial dependence, the estimated coefficients are inconsistent and biased since more (or less) explanatory power is attributed to them. SAR models are significantly better than the OLS one, all of them having smaller values (in absolute terms) of both the AIC and the Log-likelihood measure.

It should be noted that the formulation of the model remains the same in the different model estimations in purpose, in order to allow a comparison of all models in terms of identifying the impacts that the four different SAR models have both on the estimated coefficients and on the results.

In addition, the in-sample predictive power of each model is calculated, in order to facilitate their comparison and draw some conclusions also with respect to their ability to make accurate predictions. The predictive accuracy, in terms of predicted values that are within different specified ranges, is presented in Table 5. In summary, OLS model performs relatively bad since less than 40% of the predictions fall within a range of 10%. On the other hand, the predictive accuracy of the SAR models is much better and it is clearly reflected that accounting for the spatial dependence, in addition to the structural variables, can lead to significantly improved predictions. The summary statistics of the percent error term also provide support to this argument.

Between the first two SAR models, clearly the SARerror model is better than the SARlag model in terms of AIC, indicating that accounting for the spatial dependence in the error terms of the model is more important than accounting for the spatial dependence in the response variable. Nevertheless, the SAC model gives the best results and improves further the results of SARerror model which is logical since both of the models account for the spatial dependence in the error terms, while the slight improvement in terms of AIC and predictive power can be attributed to the additional accounted spatial interaction between the dependent variables.

TABLE 4 Estimated coefficients for the different spatial models

Y = Average daily speed Explanatory variables	SAR error coeff.	SAR lag coeff.	SAC coeff.	Durbin coeff.	 lag. Coef.
Speed-limit	0.254***	0.272***	0.26***	0.267***	-0.161***
Highways: Constant	96.456***	38.421***	83.897***	93.021***	-76.444***
Trunk roads: Constant	56.704***	26.84***	51.514***	53.107***	-40.41***
Collector roads: Constant	54.042***	30.047***	51.287***	52.499***	-39.803***
Distributor roads: Constant	38.941***	24.363***	38.95***	36.618***	-25.288***
Urban roads: Constant	30.332***	20.189***	30.428***	29.003***	-19.623***
Road curveness	-3.592***	-4.248***	-3.597***	-4.147***	1.477***
Distributor: PuT stops density, r=0.5km	-0.083***	-0.186***	-0.143***	-0.079***	-0.007***
Urban: PuT stops density, r=0.2km	-0.095***	-0.073***	-0.094***	-0.087***	0.051***
Highways: ln(population), r=5km	-7.978***	-2.073***	-5.962***	-7.776***	7.026***
Trunk roads: ln(population), r=2km	-3.602***	-1.497***	-3.15***	-3.21***	2.58***
Collector roads: ln(employment), r=2km,kernel weighted	-3.429***	-2.04***	-3.452***	-3.25***	2.625***
Distributor roads: ln(employment), r=1km, kernel weighted	-1.081***	-0.881***	-1.244***	-1.009***	0.635***
Urban roads: ln(employment), r=0.5km, kernel weighted	-0.501***	-0.404***	-0.554***	-0.477***	0.302***
Urban roads: Ramps' density, r=1km	0.346*	-0.054	-0.049	0.543***	-0.51***
Distributor roads: Road density, r=500 m	-0.271***	-0.133***	-0.256***	-0.225***	0.165***
Urban roads: Road density, r=100 m	-0.112***	-0.093***	-0.115***	-0.117***	0.058***
Highways: Length < 0.1 km, Dummy	-0.713***	-1.723***	-0.859***	-1.315***	-0.23*
Trunk roads: Length < 0.1 km, Dummy	-2.064***	-4.967***	-2.368***	-3.554***	-0.177*
Collector roads: Length < 0.1 km, Dummy	-3.109***	-5.915***	-3.218***	-4.912***	0.336**
Distributor roads: Length < 0.1 km, Dummy	-2.645***	-4.147***	-2.786***	-3.573***	0.913***
Urban roads: Length < 0.1 km, Dummy	-3.622***	-4.127***	-3.823***	-3.994***	2.293***
Highways:0.2 km > Length > 0.1 km, Dummy	-0.725***	-0.797***	-0.769***	-0.843***	0.298***
Trunk roads: 0.2 km > Length > 0.1 km, Dummy	-1.632***	-2.64**	-1.835***	-2.168***	0.618*
Collector roads: 0.2 km > Length > 0.1 km, Dummy	-3.047***	-3.148***	-3.027***	-3.039***	2.377***
Distributor roads: 0.2 km > Length > 0.1 km, Dummy	-1.931***	-2.285***	-2.009***	-2.108***	1.287***
Urban roads: 0.2 km > Length > 0.1 km, Dummy	-2.474***	-2.258***	-2.56***	-2.477***	1.884***
Lamda	*0.928****	-	0.742***		
Rho	-	0.459***	0.215***	0.722***	
Log-likelihood	*-705197*	*-733084*	*-694294*	*-695199*	
AIC	*1410453*	*1466226*	*1388647*	*1390509*	
Residuals spatial autocorrelation	*0.013****	*0.342****	*-0.034*	*0.101****	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

On the other hand, Durbin model gives slightly worse results in terms of AIC and similar predictive results. Attempting a closer look at the estimated coefficients of the Durbin model, the majority of the lagged variables' coefficients have opposite sign, compared to the variables' coefficients at the response location, which matches our expectations due to the formulation of the model. However, the magnitude of the spatial autocorrelation parameter (rho) indicates that alarmingly high weight is attributed to the dependent variable of the neighbor(s), which as mentioned earlier is endogenous interaction and thus can raise concerns, while the impact of the independent variables at the response location is outweighed significantly. In addition, the nature of the included variables in the model specification, especially of the socio-demographic ones, can give rise to multi-collinearity issues since they are not truly independent in space. Moreover, a substantial and statistically significant spatial autocorrelation remains in the residuals of the SARlag and the Durbin model, that can be taken as indicative of biased coefficients' estimation. On the other hand, the SARerror model has relatively low remaining spatial autocorrelation (0.01), while the SAC model has statistically insignificant remaining spatial autocorrelation which shows that the residuals have spatial randomness.

TABLE 5 Predictive accuracy of estimated models in terms of predicted speeds within specified range of actual speeds, and summary statistics of their errors

Model	2% range	5% range	10% range	15% range	20% range	30% range	SDE	ME
OLS	8.01%	20.35%	39.86%	57.07%	70.36%	84.69%	27.25%	-5.13%
SARerror	21.25%	47.20%	69.89%	81.07%	87.21%	93.68%	16.81%	-2.05%
SARlag	14.57%	35.27%	61.09%	75.31%	82.82%	90.88%	19.33%	-2.58%
Durbin	20.63%	46.19%	70.04%	81.18%	87.39%	93.95%	16.81%	-2.05%
SAC	21.09%	47.26%	70.99%	81.92%	87.84%	94.05%	17.04%	-1.92%

*SD=standard deviation of error; ME=mean error

CONCLUSIONS

In the present paper a methodology for predicting localized speed estimates, for large scale networks, is presented. The alternative of SAR models as direct demand model is examined and evaluated. In summary, the presented results of the SAR models can be considered that they highlight the impact of accounting properly for the spatial dependence of transport related data and it can be concluded that the SAC model is the most appropriate among the SAR models for speed prediction purposes, outperforming the simple OLS model, both in terms of measures of quality, and in terms of predictive ability. Furthermore, the inconsistency of the estimated parameters of the OLS model comes to the surface and provides arguments in favor of accounting for the spatial dependence.

The formulation of different adjacency matrices is checked, based on different hypotheses, to conclude to the most appropriate one for the given situation which is based on the network connectivity. The process of identifying the optimum adjacency matrix is presented analytically. Conclusively, the SAR models can be applied for the prediction of localized speeds and can be considered as a worthy alternative for such purposes, especially given the limited data requirements and computational effort that they require. At the same time, they can enhance the understanding of transport related phenomena in a direct and structural way, and can point directions for further improvements in other existing modelling techniques.

FURTHER RESEARCH

An apparent extension of the methodology presented is to apply spatial models for the estimation of AADT and thus formulate a coherent framework for direct demand modelling that can provide the required answers for project appraisal purposes. A comparison to the widely applied models (4-step and agent based model) but computationally burdensome, can quantify the advantages and disadvantages of each model choice in terms of required effort, data demand, accuracy and their capability to support the decision making process.

By definition, the employed method of estimating speeds cannot capture the variability of speed during the day like dynamic traffic assignment models can. However it is of great interest to compare the results with the respective results from static assignment models, which are widely used for transportation planning purposes. Still, it would be possible to estimate time-of-day specific models, when the data are available.

Furthermore, the effectiveness of a number of policies might not be able to be studied in the context of spatial modeling (e.g. congestion pricing), at least not in a direct way but perhaps through the incorporation of other spatially resolved variables that there are strong evidences to be the causal outcome of such policies (e.g. reduction of car ownership).

Nevertheless, the value of the research towards this direction is not only as a competing alternative to the existing methods but it can also serve to point directions regarding the importance of accounting properly for the spatial effects and thus can provide insight on how the existing approaches can be improved. In line with this, their potential to contribute to the further improvement of the existing models needs to be investigated and evaluated (e.g. facilitate a quicker convergence by setting the initial values in iterative processes).

Last, the simplicity of spatial models can allow them to be incorporated within the land use transport interaction models framework and it needs to be examined how they can contribute in the advance of such models. An initial idea of the potential advantages of such integration is given by Zeiler et al. *(28)*.

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. Wachs, M., L.L. Cove, T.B. Deen, G.B. Dresser, R.W. Eash, R.A. Johnston, E.J. Miller, M.R. Morris, R.H. Pratt, C.L. Purvis, G. Rousseau, M.L. Tischer, R.E. Walker, J.M Williams. Metropolitan Travel Forecasting: Current Practice and Future Direction, *Special Report*, *Transportation Research Board,* No. 288, 2007.

2. Anselin L. Spatial econometrics: methods and models, *Kluwer Academic, Dordrecht*, 2008.

3. Brunsdon, C., S. Fotheringham, and M. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, Vol. 28, No. 4, 1996.

4. Páez, A., and D. M. Scott. Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal*, Vol. 61, No. 1, 2004, pp. 53–67.

5. Black, W. Network autocorrelation in transport network and flow systems. *Geographical Analysis*, Vol. 24, No. 3, 1992.

6. Wang, C., M. Quddus, and T. Ryley. Spatial models in transport: a review and assessment of methodological issues. *In: Proceedings of the 91st Annual Meeting of the Transportation Research Board*, 2012.

7. Bernard, M., J. Hackney, and K. Axhausen. Correlation of link travel speeds. *6th Swiss Transport Research* Conference, 2006.

8. 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, Sep. 2007, pp. 397–417.

9. Cheng, T., J. Haworth, and J. Wang. Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, Vol. 14, No. 4, Apr. 2011, pp. 389–413.

10. Jenelius, E., and H. N. Koutsopoulos. Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, Vol. 53, Jul. 2013, pp. 64–81.

11. Lopes, S., N. Brondino, and A. da Silva. GIS-Based Analytical Tools for Transport Planning: Spatial Regression Models for Transportation Demand Forecast. *ISPRS International Journal of Geo-Information*, Vol. 3, No. 2, Apr. 2014, pp. 565–583.

12. Selby, B., and K. M. Kockelman. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, Vol. 29, May 2013, pp. 24–32.

13. Zhao, F., and N. Park. Using geographically weighted regression models to estimate annual average daily traffic. *Transportation Research Record: Journal of the Transportation Research Board*,, Vol. 1879, 2004, pp. 99–107.

14. Pulugurtha, S., and P. Kusam. Modeling AADT using integrated spatial data from multiple network buffer bandwidths. *In: Proceedings of the 91st Annual Meeting of the Transportation Research Board*, 2012.

15. Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, Vol. 97, No. 1, Jan. 2006, pp. 246–273.

16. Aguero-Valverde, J., and P. P. Jovanis. Spatial Correlation in Multilevel Crash Frequency Models. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2165, Dec. 2010, pp. 21–32.

17. Löchl, M., and K. W. Axhausen. Modelling hedonic residential rents for land use and transport simulation while considering spatial effects. *Journal of Transport and Land Use*, Vol. 3, No. 2, Sep. 2010, pp. 39–63.

18. Efthymiou, D., and C. Antoniou. How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece. *Transportation Research Part A: Policy and Practice*, Vol. 52, Jun. 2013, pp. 1–22.

19. Efthymiou, D., C. Antoniou, and Y. Tyrinopoulos. Spatially Aware Model for Optimal Site Selection. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2276, Dec. 2012, pp. 146–155.

20. 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.

21. Ord, J. K. Estimation Methods for Models of Spatial Interaction*, Journal of the American Statistical Association*, Vol. 70, pp. 120-126, 1975.

22. Kissling, W. D., and G. Carl. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, Vol. 17, No. 1, Jun. 2007, pp. 59–71.

23. LeSage, J., and R. Pace. Spatial and spatiotemporal econometrics. *Advances in Econometrics*, Vol. 18, No. 04, 2004, pp. 1–32.

24. Elhorst, J. Spatial econometrics: from cross-sectional data to spatial panels. *Springer*, 2014.

25. R Development Core Team. R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, 2011.

26. Bivand, R., L. Anselin, O. Berke, A. Bernat, M. Carvalho, Y. Chun, C. F. Dormann et al. spdep: Spatial dependence: weighting schemes, statistics and models, 2011.

27. LeSage, J. and R. K. Pace. Introduction to Spatial Econometrics. *CRC Press, Taylor and Francis Group*, 2009.

28. Zeiler A., G. Sarlas, M. Kuliowsky, B.R. Bodenmann, B. Sanchez, J. Bode, P. Furtak, K.W. Axhausen. FaLC Transport Simulation Module: How accurate can a simplified transport model be?, *14th Swiss Transport Research Conference*, 2014.