Exploring spatial methods for prediction of traffic volumes

Georgios Sarlas, IVT, ETH – Zürich
Kay W. Axhausen, IVT, ETH – Zürich

Conference paper STRC 2016

STRC
16th Swiss Transport Research Conference
Monte Verità / Ascona, May 18-20, 2016

Exploring spatial methods for prediction of traffic volumes

Georgios Sarlas
IVT, ETH Zürich
ETH Hönggerberg,
CH-8093 Zürich

☎ +41-44-633 37 93

💻 georgios.sarlas@ivt.baug.ethz.ch

Kay W. Axhausen
IVT, ETH Zürich
ETH Hönggerberg,
CH-8093 Zürich

☎ +41-44-633 39 43

💻 axhausen@ivt.baug.ethz.ch

May 2016

Abstract

In the present paper a direct demand modelling approach for traffic volume prediction on a nationwide network is presented, exploring the ability of different spatial modelling alternatives to be applied for such purposes. A particular focus is on the identification of variables that can capture the interregional demand patterns, utilizing concepts from network theory. A new variable called accessibility-weighted centrality is introduced, constructed by applying a set of modifications on the stress centrality index, tailored for the task of the annual average daily traffic (AADT) prediction. The results exhibit clearly that the inclusion of network theory-based variables in the model formulation can lead to a significant enhancement on the predictive accuracy. In addition to the already tested models in the literature, two spatial simultaneous autoregressive models are estimated and it is shown that they have the potential to be applied both for interpolation and forecasting since their estimated parameters are unbiased and consistent. A comparison of the different estimated models to the output of a traditional four-step model is conducted to show to what extent direct demand models on nationwide scale can constitute a trustworthy alternative to more advanced, but definitely more data demanding and computationally burdensome models.

Keywords

traffic volume prediction – AADT – spatial regression – GWR – centrality – accessibility

1. Introduction

Many studies in the field of transport modelling have dealt with the issue of annual average daily traffic (AADT) prediction, developing different methodologies to tackle the problem. In general, two main streams of literature can be found. One that exploits different modelling techniques aiming at resolving the issues of spatial dependence and heterogeneity, while in the second stream the construction and the inclusion of more variables describing the demand patterns in models is investigated. The employed methodologies vary from the aspatial regression techniques to the statistical techniques accounting for the spatial effects. In particular, the later encompass two different approaches. The first one is utilizing a data-driven approach of spatial statistics called kriging, while the second one utilizes the geographically weighted regression (GWR) of the class of spatial econometric models. Nevertheless, the majority of the studies developed methodologies tailored for small, or medium, scale level of analysis in terms of network size, having mainly the purpose to interpolate AADT from known to unmeasured locations.

1.1 Literature review

Xia et al. (1999) developed a multiple regression model for estimating AADT on non-state roads of Florida and found that the most important contributing predictors are the roadway characteristics along with the area type, while socioeconomic variables were found to have an insignificant impact on AADT. Similarly, Mohamad et al. (1998) developed a multiple regression model for AADT prediction for county roads in Indiana, incorporating various demographic variables which were found to be significant. In a similar context, Desylas et al. (2003) developed a multiple regression analysis model for pedestrian flows.

The plausibility of applying the GWR model for estimating AADT was demonstrated in another study (Zhao and Park, 2004) and it was shown that it can lead to the enhancement of the prediction accuracy, compared to the aspatial ordinary linear regression. Eom at al. (2006) exploited ordinary kriging for interpolating AADT for non-freeway facilities in Wake County, North Carolina, and concluded that its predictive capability is much better than the ordinary regression models. Along the same line of thought, Wang and Kockelman (2009) applied kriging-based methods for AADT prediciton at unmeasured locations, making use of Texas highway count data, and highlighted further the capability of applying kriging for prediction purposes on a statewide network. Selby and Kockelman (2013) explored the application of two spatial methods for prediction of AADT on the same statewide network (universal kriging and GWR), and they concluded that both methods reduce predictions errors over aspatial regression

techniques whereas the predictive capabilities of kriging exceed those of GWR. Interestingly, the estimation of the kriging parameters taking into account network distances, instead of Euclidean, showed no enhanced performance.

Furthermore, Pulugurtha and Kusam (2012) developed Generalized Estimating Equations models to estimate AADT using integrated spatial data from multiple network buffer bandwidths. Spatial data included off-network characteristics such as demographic, socio-economic and land use characteristics, captured over multiple network buffer bandwidths around a link and integrated by the employment of distance decreasing weights. The methodology was applied on a city level (Charlotte, North Carolina). As a continuation of the previous study, Duddu and Pulugurtha (2013) exploited the application of the principle of demographic gravitation to estimate AADT based on land-use characteristics on the same network. A negative binomial model was estimated along with neural network models. Interestingly, the results obtained showed that the developed models gave significantly lower errors in comparison to outputs from traditional four-step method used by regional modellers.

In a recent study by Lowry (2014), a new method for interpolating AADT was presented, tailored for communities where attributes such as roadway characteristics, land-use etc., are uniform over space, and thus their inclusion in the model bears no explanatory power. The new method used novel explanatory variables that are derived through a modified form of stress centrality, a network analysis metric that quantifies the topological importance of a link in a network. The case study showed high quality results. The same methodology found application as well for estimating directional bicycle volumes (McDaniel et al., 2014).

1.2 Description of the framework of the paper

The objective of the current research is to develop a direct demand modelling approach for prediction of AADT on a nationwide network, a task which has not been addressed sufficiently in the existing literature. The particularity of the nationwide network level case stems from the inherent incapability of spatial densities of different socioeconomic data to capture the interregional demand patterns that occur on the links, since they fail to explain the high volume of interregional through traffic. Driven by this and building upon the work of Lowry (2014), we have expanded the stress centrality index to align with travel demand modelling aspects. In brief, the main advantage of this is that it can facilitate a quantification of the interregional demand patterns by associating the network structure with the travel accessibility concept. This allows to bring into the modelling formulation a way to capture both the spatial direction and

extent along with the trip attraction competition that govern the travel demand, allowing us to capture the demand capacity interaction at the core of transport modelling.

In addition to the already tested models in the literature (ordinary least squares [OLS] model, negative binomial model, universal kriging, and GWR), the utilization of the family of spatial simultaneous autoregressive (SAR) models (Anselin, 1988a) is tested, in terms of its capability to be applied for AADT prediction purposes. The advantage of such models is that they can resolve spatial dependence issues, accounting for the spatial correlation, offering a structural explanation of the AADT and since their estimated coefficients are unbiased and consistent, they can be used for both interpolation and forecasting purposes, an important aspect for both policy evaluation and project appraisal purposes.

In summary, a set of different models is estimated and evaluated in order to draw sound conclusions on the newly employed variables and also on SAR models' capabilities to be employed for AADT prediction purposes and thus highlight in a quantifiable way their strengths and weaknesses. At last, a comparison of models predictive accuracy to the output of a traditional four-step model is conducted to show to what extent such models can constitute a trustworthy alternative to more advanced, but definitely more data demanding and computationally burdensome, models.

2. Methodology

2.1 Centrality indices

The construction of a new variable capturing the interregional demand patterns, taking into account the direction of potential interactions over space, is of central importance for the estimation of AADT models on a nationwide network. Making use of network theory, centrality is an index that aims to identify the most influential persons in the context of a social network. Different centrality indices have been introduced over the years, aiming at the identification and the quantification of the importance of a particular person in a social network. In general, centrality indices take into account the number of shortest paths that pass by a given link/node, either for given pairs of nodes, or for all pair of nodes within the network. In the case where a capacity constraint exists in the form of a particular weight/cost associated with each link/node, then this weight should be taken into account in the routing algorithm for the identification of the shortest paths.

Departing from the social sciences questions, centrality indices are meaningful for all networks' analyses. From this viewpoint, centrality indices are meaningful for the analysis of transport networks as well and can provide a quantifiable measure of the importance of links, taking into account the network structure and the cost of traversing each link (distance or time). In the case of transportation, networks correspond to directed networks, given the allowed and prohibited turning movements on its vertices (nodes), and are modelled as higher level networks in order to account for them. Stress centrality index was introduced by Shimbel (1953) and is defined as the number of shortest paths connecting all pairs of nodes of the network that pass via a link.

$$\text{Stress centrality}_e = \sum_{i,j \in V} \sigma_{ij}(e) \quad (1)$$

Where e is any link of the network, V the set of all nodes, $\sigma_{ij}$the shortest path from node i to node j, and $\sigma_{ij}(e)$ is equal to one if the link e is part of the shortest path connecting i and j nodes.

By definition, higher hierarchical links have high centrality values, while that might be the case as well for lower hierarchical links given the network structure. In the case of transport networks, the hierarchy is given by the functional class of the roads whereas their importance is normally matched by the number of trips using the given link. Naturally, two issues with respect to the application of the stress centrality index for transport networks come to the surface. First, the issue of travel demand since not all nodes are attracting or producing the same number of trips and thus this should be taken into account in the centrality formulation. Second, interaction between nodes tends to diminish and becomes very small as the distance between them increases, which should be accounted for in a modified stress centrality formulation.

Addressing the aforementioned issues takes place in three steps. At first, the issue of trip production and attraction is addressed by making the assumption that production is related to the economically active population in the vicinity of the origin node, and attraction at the employment positions at the destination node. Second, the interaction intensity between the nodes should be associated with a function that diminishes by network distance. The distance decay function embedded in the measure of travel accessibility is employed for this reason, since accessibility is a measure of how far people are willing, or able, to travel on the course of their daily life and quantifies how interaction opportunities decrease over the distance (Hansen, 1959). Two variations of distance decay function are tested to identify the one that fits the data better (Halás et al., 2014). The parameters of the distance-decay function can be either estimated, if data availability allows it, or taken from another study. Last, a restriction has to be imposed with respect to the direction of potential interactions by standardizing the accessible opportunities from each node to each node, by the total number of opportunities

accessible from the origin node. The incorporation of these changes in the stress centrality index and the derivation of the constructed index, called *accessibility-weighted centrality*, is presented below. It should be noted that the constructed variable mirrors to a great extent the first two steps of the traditional four-step model, however this is inevitable due to the nature of the relationships that we need to capture in the variable.

$$Accessibility - weighted\ centrality_e = \sum_{i,j \in V} \sigma_{ij}(e) \quad (2)$$

$$\sigma_{ij}(e) = \sum_{i,j \in V} Popul_i \frac{Employm_j * f(cost_{ij})}{Travel\ Accessibility_i} \quad (3)$$

$$\text{Travel Accessibility}_\text{i} = \sum_\text{i}^\text{j} \text{Employm}_\text{j} * \text{f}\left(\text{cost}_\text{ij}\right) \quad (4)$$

$$f\left(cost_{ij}\right) = \begin{cases} e^{\beta * cost_{ij}} \\ e^{\beta * cost_{ij}{}^a} \end{cases} \quad (5)$$

Where e is any link of the network, V the set of all nodes, $\sigma_{ij}$the shortest path from node i to node j, and $\sigma_{ij}(e)$ is equal to the sum total according to formula 3, if the link e is part of the shortest path connecting i and j nodes.

2.2 Modelling approaches

In order to test the predictive accuracy of models for AADT prediction, the application of different models is examined. In particular, the classical ordinary least square (OLS) model constitutes the starting point due to its simplicity, where the dependent variable Y is described by a linear function of independent variables X with the parameters β being the least squares estimates. One of the main assumptions of the model requires that the error should be spherical, meaning that they should be homoscedastic and not auto-correlated.

$$Y = \beta X + \varepsilon \quad (6)$$

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 and ε a vector of error terms.

However, the application of the OLS estimator for the statistical analysis of spatial data results to residuals that are not independent, but spatially correlated, leading to the violation of the assumptions of the OLS estimator.

Spatial econometrics was popularized by Anselin (1988a) and are defined as the use of regression models by accounting for the impact of spatial effects (spatial dependence and heterogeneity) in their specification and estimation, avoiding the statistical problems such as unreliable statistical tests and biased and inconsistent estimated parameters. This is facilitated by the inclusion of a spatial weight matrix (W) in the model specification that incorporates 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.

In the case of the spatial dependence, SAR models can account for it by the inclusion of relevant spatial autoregressive components (Kissling and Carl, 2007). In particular, the spatial error model assumes that the spatial dependence exists in the error term of the model, and thus the spatial autoregressive process is applied to it.

$$Y = \beta X + u \quad (7)$$

$$with \;\; u = \lambda W u + \varepsilon \quad (8)$$

where 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 model assumes that the spatial dependence exists in the response variable and applies the spatial autoregressive process to the response variable, treating it as a lagged variable. The formulation of the model is:

$$Y = \rho W Y + \beta X + \varepsilon \; (9)$$

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

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, and provides localized estimates of the coefficients (Charlton and Fotheringham, 2009).

$$Y(z) = \beta_i(z) X + u \quad (10)$$

Where the notation $\beta_i(z)$ indicates that the parameter describes a relationship around location u and is specific to that location (Charlton and Fotheringham, 2009).

Kriging is a geostatistical technique used for interpolation purposes. In the case of ordinary kriging, the assumption is that the unobserved value is decomposed into two terms, the local

trend βX, and the error terms which are spatially correlated and their variance is assumed to follow a semivariogram relation $\gamma(h_{ij})$, as a function of the distance h between the points (detailed information can be found at Oliver and Webster (1990)). In line with previous studies of AADT prediction (e.g. Selby and Kockelman (2013)), three semivariogram functions are evaluated:

Exponential: $\gamma(h_{ij}; c_0, c_e, a_s) = c_0 + c_e\left(1 - e^{-\frac{h_{ij}}{a_s^2}}\right)$ (11)

Gaussian: $\gamma(h_{ij}; c_0, c_e, a_s) = c_0 + c_e\left(\frac{1.5h_{ij}}{a_s} - 0.5\left(\frac{h_{ij}}{a_s}\right)^3\right)$ (12)

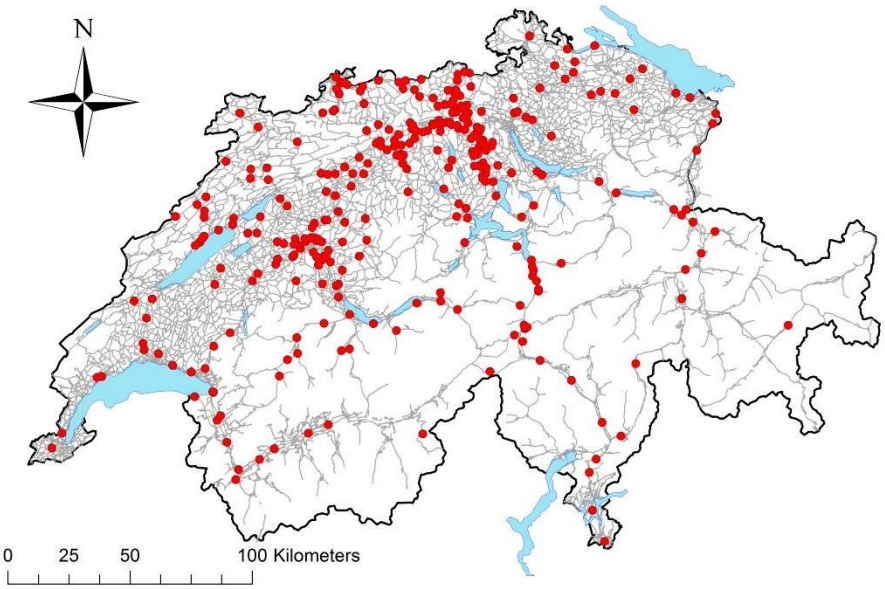
Spherical: $\gamma(h_{ij}; c_0, c_e, a_s) = c_0 + c_e\left(1 - e^{-\frac{h_{ij}}{a_s}}\right)$ (13)

Last, the negative binomial regression is widely used along with the Poisson regression, for the modelling of count data, accounting properly for their non-negative nature.

3. Case study

In order to assess the plausibility of applying a direct demand modelling approach for prediction of AADT on a nationwide network, and evaluate the capability of the centrality indices to enhance the predictive accuracy of such models, a case study is designed and conducted. More specifically, the network of Switzerland is employed as the study network (ARE; National Transport Model, 2010), where the Federal Roads Office collects count data at various locations of the network and calculates AADT values. As the basis year, the year 2010 is chosen in order to be comparable with the output of the latest version of the National Transport Model. In particular, for the basis year AADT data on 398 links exist which are used for the model estimation as dependent values. A map of the study network along with the spatial distribution of the count locations can be seen in Figure 1.

Figure 1 Case study network and count locations



Source: ARE, National Transport Model, 2010.

3.1 Centrality indices

3.1.1 Stress centrality

The first centrality measure that is of interest for evaluation mainly due to its simplicity, is the stress centrality as defined in formula 1. The number of shortest paths connecting all pairs of nodes of the network for each link is a variable that can be constructed with a relative ease, making use of existing routines (e.g. igraph package for R (Csárdi and Nepusz, 2006)).

3.1.2 Accessibility-weighted centrality measure

The construction of the accessibility-weighted centrality measure for the study network is conducted according to the previously defined methodology. In particular, the new measure includes a distance decay function which serves the purpose of capturing the diminishing intensity interactions over distance and two variations of distance decay function are checked to identify the one that fits better the data, in line with a previous study (Halás et al., 2014).

Obviously, different parameters are associated with different trip purposes; e.g. people are willing to travel shorter distances for shopping activities than for commuting to work. In our case, the interregional commuting to work trips are the ones contributing to the available AADT values the most and thus the estimated parameters should correspond to this trip purpose.

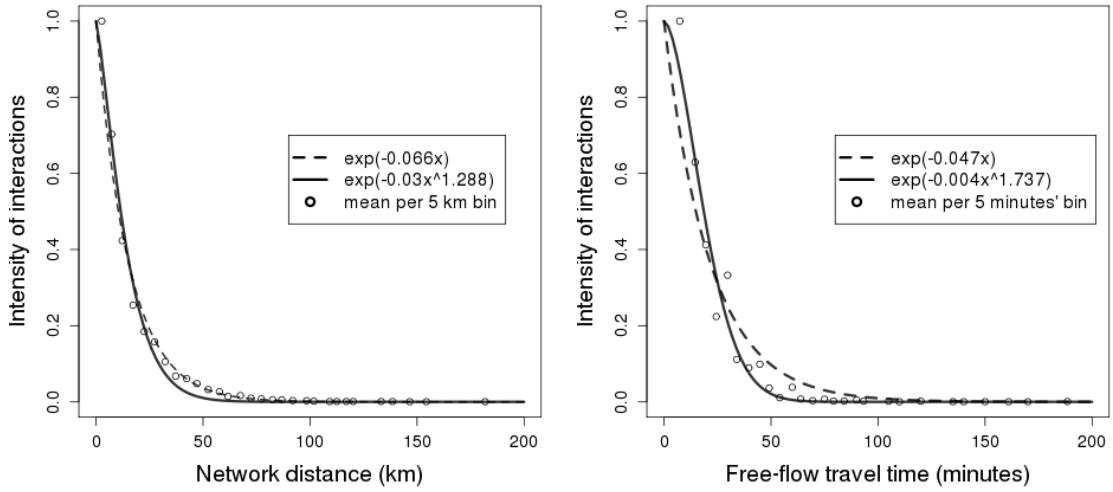
In order to facilitate the estimation of the parameters of these two functions, we make use of the 2010 Microcensus data, where the residential and employment location of the participants is reported and subsequently it is matched to a municipal level. The associated travel cost among all municipalities is calculated by identifying their shortest paths on the employed weighted directed network, both in terms of distance and travel time (free-flow travel time).

The nonlinear least-squares estimates of the parameters are calculated using the Gauss-Newton algorithm. The estimated parameters and the shape of the distance decay functions are presented in Figure 2, where the function with the two parameters is found to fit better to the data, for both distance and travel time, and thus is the chosen one. Alternatively, these parameters could be taken from previous studies as long as the associated cost metric is consistent with the one of the case study to avoid giving rise to inconsistencies that can lead to erroneous results.

The next step is to define the origin and the destination nodes of the network that their shortest paths are accounted in the calculation of the centrality measure. Given the interregional character of the trips, a convenient choice is to employ a zonal level according to the administrative level of municipalities. In this case, a node close to the centroid of each zone serves as the origin and destination node for the trips of each zone, associating on it the population and the employment positions of each zone. The advantage of that choice is the availability of socioeconomic data aggregated on this level while the methodology can be easily applied if more disaggregated data (e.g. on a hectar level) exist along with the identification of different population and employment clusters, which can then replace the employed zonal analysis level.

Finally, the calculation of the accessibility-weighted centrality value takes place for the subset of links with count data, for both metric costs of network distance and travel time. For computational reasons, given the finding that zones with distances more than 60 kilometers or minutes between them (Figure 2) have an interaction intensity close to zero, we restrict the time/distance window around each link to these values. Essentially that means that only the shortest paths among the origins and destinations within a radius of 60 kilometers or 60 minutes around each link are found and taken into account.

Figure 2 Estimated parameters of the accessibility distance decay functions



3.2 Independent variables

In essence, the regression yields two components; one that captures the impact of supply on AADT, and one that captures the impact of demand allowing to model their interaction. On the supply side, variables describing the road capacity are put to use. More specifically, the functional class of the road and the number of lanes are the chosen explanatory variables. On the demand side, a set of variables is tested thoroughly in order to capture to the greatest possible extent the demand patterns. These variables correspond to the spatial densities of socioeconomic variables for various radii, stress centrality indices and the constructed accessibility-weighted centrality measure. Additional spatial variation is added on the demand side by the inclusion of the public transport network density in the vicinity of each road (density of public transport stops within 5 km radius), as indicative of the intensity of local activities, and thus of local demand. The summary statistics of the included variables are presented in Table 1. As it can been, in conjunction with the box-plot in Figure 3, the newly constructed variable has a similar magnitude as the AADT while their correlation is close to 0.75, providing evidence that the new variable has the capability of reproducing satisfyingly the variation of demand over space.

Table 1 Summary statistics of variables

Variables	Unit	Mean	Median	St. Dev.
AADT (before transformation)	Vehicles	14370	8668	14146
Freeway - Highway	Dummy	184	-	-
Major road	Dummy	136	-	-
Rural major road	Dummy	78	-	-
One-lane road	Dummy	231	-	-
Two-lane road	Dummy	147	-	-
Three-lane road	Dummy	20	-	-
Population density (kernel weighted): 10 km	Residents/ sq. km	571	327	626
Population density (kernel weighted): 20 km	Residents/ sq. km	369	303	323
Stress centrality	Importance	$8.30*10^6$	$2.21*10^6$	$13.17*10^6$
Accessibility-weighted centrality	Accessible empl. opportunities	22350	9646	28651
Public transp. density: 5km radius	Stops/ sq. km	1.33	0.89	1.28

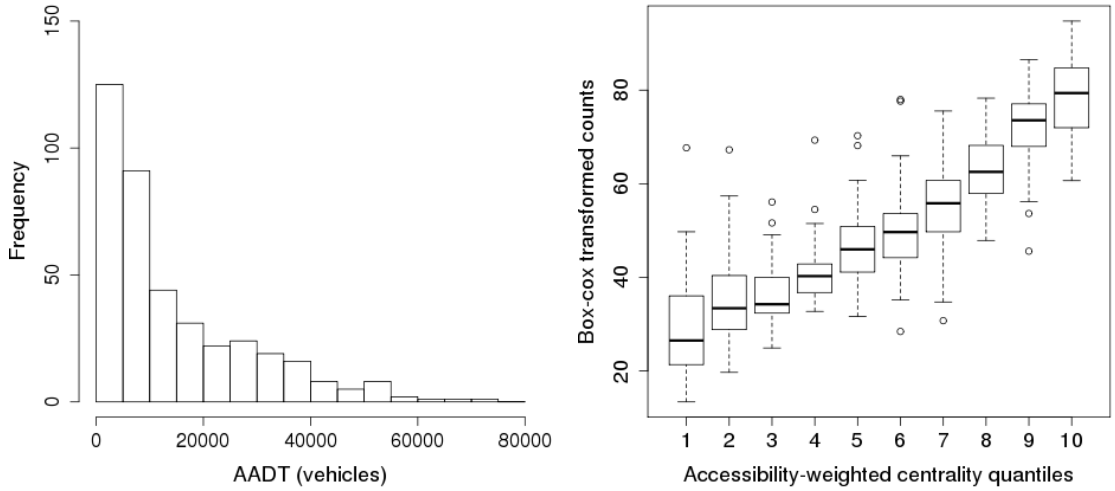
3.3 AADT transformation

The particularity of using count data as the dependent variable in the context of linear regression models, stems from their non-negative character which can lead to a number of shortcomings (Winkelmann, 2008). In this case, models accounting for it should be employed such as Poisson or negative binomial regression models, or the dependent variable should be transformed to conform to the assumptions of normality and/ or homoscedasticity of variance (Osborne, 2010). Based on that, the Box-Cox transformation (Box and Cox 1964) is applied on the AADT data in order to allow the estimation of linear regression models. The transformation form is presented below while the identified ξ value for the AADT data is found to be equal to 0.222.

$$Y_{tr} = \begin{cases} \frac{Y^{\xi} - 1}{\lambda}, \xi \neq 0 \\ lnY, \xi = 0 \end{cases} \quad (14)$$

Given the high correlation of the centrality variable with the AADT, we choose to apply an identical Box-Cox transformation to it in order to maintain their strong linear relation in the model. The histogram of the AADT values before the transformation is presented in Figure 3 (left side), while on the right side the box-plot of the transformed centrality quantiles are plotted against the transformed AADT values to show their strong linear correlation.

Figure 3 Histogram of AADT and box-plot of accessibility-weighted centrality quantiles with respect to AADT



It should be noted that the involved data processing, models estimation, and network processing are undertaken with the statistical programming language R (R Development core team, 2011), making use of different available packages (*igraph* (Csárdi and Nepusz, 2006); *spdep* (Bivand et al., 2005); *gstat* (Pebesma, 2004)).

4. Model estimation - Results

In this section, a set of different models is estimated and evaluated in order to draw safe conclusions on both the newly constructed variable and also on models' capabilities. In addition to models already tested in the literature, the family of spatial simultaneous autoregressive (SAR) models is tested as well. An assessment of models predictive accuracy and comparison to the output of a traditional four-step model is conducted to show to what extent such models can constitute a trustworthy alternative.

Three variations of OLS models are estimated serving a twofold purpose. At first, to examine the ability of different sets of variables to capture long-distance trips that occur on a nationwide network within a direct demand model formulation and thus draw conclusions concerning this aspect. The particularity of the long-distance trips is that spatial density variables fail to capture due to their inherent incapability to take into account the directionality and the mechanism that governs the demand. More specifically, the first model includes the spatial density of population in a 20 kilometres radius to resemble the travel demand patterns in a medium scale. The second model, includes in addition the stress centrality variable where the importance of the links is quantified. Last, in the third OLS model the aforementioned variables are replaced with the accessibility weighted centrality variable which simultaneously quantifies both the network structure and the directionality and magnitude of travel demand. Furthermore, the spatial density of population in a shorter radius than before (10 kilometres) is included as well to capture more localized demand patterns that the constructed variable fails to capture sufficiently.

Secondly, to serve as the comparison benchmark and also for examining the existence of spatial autocorrelation in the residuals and thus justify if the need for the estimation of spatial regression models arises. The spatial autocorrelation is calculated in terms of the Moran's I measure which shows that there is statistically significant autocorrelation of 0.21. The implication of this, as mentioned before, is that the estimates are biased and inconsistent since more (or less) explanatory power is attributed to them than it should. The estimated coefficients for the different OLS models are presented in Table 2. In addition, the models have been tested for heteroscedasticity by making use of the appropriate tests (Breusch and Pagan, 1979; Goldfeld and Quandt, 1965) and no strong indication of it was found.

In summary, the OLS coefficients of the functional class variables have the expected order of magnitude, while the impact of the number of lanes and the functional class is in line with expectations. The demand relevant variables, have positive impact and they are statistically significant. It should be mentioned that they centrality value with the distance decay function as a relationship of the travel time distance is found to be slightly more statistically significant, and thus the one employed. The OLS model with the accessibility-weighted centrality variable has the highest fit among the models, in terms of adjusted R-squared and Akaike Information Criterion.

Table 2 Estimated coefficients for the different OLS models

	OLS		OLS stress centr.		OLS acc.weighted	
Indep. Variables	Estimate	Sign.	Estimate	Sign.	Estimate	Sign.
Intercept	21.35	***	8.97	***	13.83	***
Major road	-5.24	***	-3.37	***	-3.657	***
Rural major road	-6.59	***	-4.44	***	-4.55	***
Two-lane road	5.81	***	4.93	***	3.719	***
Three-lane road	10.88	***	8.97	***	6.917	***
Ln Population density: 10 km	-		-		1.65	***
Ln Population density: 20 km	1.71	***	1.27	***	-	
Ln Public transp. density: 5km	1.23	***	1.81	***	-	
Acc.weighted centrality (box-cox)	-		-		0.233	***
Ln Stress centrality	-		0.98	***	-	
Adjusted R	*0.839*		*0.857*		*0.875*	
Akaike Inf. criterion	*2108*		*2061*		*2006*	
Moran's I (network distance)	*0.13*	***	*0.19*	***	*0.2*	***
No. of observations						398

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Based on the Moran's I measure results, the estimation of spatial error and lag models necessitates in order to account for the autocorrelation issues. Driven by this, three spatial weight matrices are constructed based on Euclidean distance, and network cost, both in terms of time and distance, in order to evaluate the direction that correlation occurs. The identification of the spatial extent of autocorrelation in the OLS residuals is used as an indicator to define the extent of the neighborhood. In particular, for the Euclidean and the network distance, the Moran's I measure exhibits that the autocorrelation exists up to a radius of 20 and 30 kilometers respectively. In the case of network time, the autocorrelation remains significant up to a radius of 25 minutes of free-flow travel time. The last part of the construction of the spatial weight matrices is to determine the weight that should be assigned to each neighboring location. Based on the Moran's I measure, we conclude that the inverse distance metric along with a normalization of the sum of the weights of the neighboring locations to one, is the more appropriate to capture the spatial structure. Making use of the robust form of the Lagrange Multiplier diagnostics for spatial dependence (Anselin, 1988b), we conclude that the spatial dependence exists in the error term, hence the spatial error model is the appropriate model. Nevertheless, the spatial lag model is evaluated as well to test its predictive accuracy. The estimated coefficients for the spatial regression models are presented in Table 3.

Table 3 Estimated coefficients for SAR models

	Sp. Error netw. dist.		Sp. Lag netw. dist.	
Indep. Variables	Estimate	Sign.	Estimate	Sign.
Intercept	14.10	***	12.28	***
Major road	-3.68	***	-3.71	***
Rural major road	-4.73	***	-4.45	***
Two-lane road	3.06	***	3.50	***
Three-lane road	5.98	***	6.74	***
Ln Population density: 10 km	1.50	***	1.34	***
Acc.weighted centrality (box-cox)	0.27	***	0.23	***
lamda	0.51	***	-	
rho	-		0.11	**
Akaike Inf. criterion	*1963*		*2000*	
Moran's I measure	*0.01*		*0.15*	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

The same patterns as in the OLS model can be observed in the estimated coefficients of the spatial models, with the spatial autoregressive and autocorrelation parameters found to be statistically significant. In terms of goodness-of-fit measures, the Akaike information criterion shows that the spatial error model outperforms both the OLS and the spatial lag model.

The next model estimated corresponds to the GWR, which aims to resolve spatial heterogeneity issues and it is calculated by taking into account an adaptive bandwidth. The corresponding results are presented in Table 4.

Table 4 Estimated coefficients for GWR model

Indep. Variables	Min.	1st Quantile	Median	3rd Quantile	Max.
Intercept	-1.04	9.26	12.51	15.22	30.46
Major road	-11.02	-5.57	-3.84	-1.88	3.77
Rural major road	-14.95	-7.29	-4.32	-1.93	3.44
Two-lane road	-3.25	0.37	2.90	5.07	9.68
Three-lane road	-0.65	2.31	5.21	9.89	15.42
Ln Population density: 10 km	-0.51	1.22	1.70	2.12	3.09
Acc.weighted centrality (box-cox)	0.06	0.23	0.29	0.35	0.56
Local R square	0.748	0.891	0.929	0.941	0.9725

Interestingly, the statistics of the constructed centrality variable's coefficient show that it has relatively low variation over space, providing further evidence on its ability to approximate interregional demand patterns.

The negative binomial regression results are not reported, but the estimates exhibit the same patterns as in the OLS model. In this particular case, the untransformed AADT and centrality variables are employed.

4.1 Evaluation of predictive accuracy of models

The developed models are evaluated in terms of their predictive accuracy, both for in-sample and out-of-sample. For the out-of-sample, an 80% share of the count locations are randomly chosen and used for the estimation of the model while the remaining 20% is used for the validation part. Given the relatively low number of observations, the out-of-sample predictive accuracy of the model exhibits variation. In order to account for it, a number of 100 replications is performed to draw safe conclusions and the corresponding mean values are reported.

The following five accuracy measures are calculated in order to allow the evaluation to take place. Mean percentage error (MPE) and mean absolute percentage error (MAPE) are easily interpretable measures, having the main disadvantage though that they are influenced by outliers. Symmetric mean absolute percentage error (SMAPE) is a similar measure which has the advantage that it corrects for outlier's influence. Median absolute percentage error (MdAPE) has the advantage that it is not influenced by outliers and can provide an overview of the distribution of the errors in conjunction with MPE. Mean squared error (MSE) because of the quadratic term is influenced heavily by the outliers. An overview of the employed accuracy measures is given by Makridakis and Hibon (1995), where they conclude that for forecasting purposes MSE and SMAPE are the preferable measures. It should be noted that AADT predicted values are back-transformed before the calculation of the measures. The formulas of the accuracy measures are given below with $\widehat{Y}_i$ the predicted value, while the results are reported in Table 5.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{Y}_i - Y_i}{Y_i} * 100 \quad (15)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{Y}_i - Y_i}{Y_i}\right| * 100 \quad (16)$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{\frac{Y_i + \widehat{Y}_i}{2}}\right| * 100 \quad (17)$$

$$MdAPE = median\left(\left|\frac{\widehat{Y}_i - Y_i}{Y_i}\right| * 100\right) \quad (18)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad (19)$$

A comparison of the accuracy measures reveals similar patterns for both in-sample and out-of-sample. In particular, among the variations of SAR models, the ones that employ a spatial matrix based on network distances metrics yield slightly better results, highlighting the importance of using network over Euclidean distances. Among the kriging models, all of them yield similar results although it can be concluded that the one with the spherical semivariogram has slightly better accuracy.

The negative binomial model yields the results with lower predictive accuracy, providing support to the argument of the necessity of transforming the dependent variable that does not conform to the assumptions of normality.

Among the estimated models, GWR has the highest in-sample and out-of-sample accuracy. In terms of SMAPE, all models besides negative binomial regression yield similar out-of-sample results. Moreover, taking into account the fact that GWR and kriging models are aimed for interpolation purposes, it can be concluded that the spatial error model gives similar results, while having the advantage that it can be applied for forecasting purposes since its parameters are unbiased and consistent. Interestingly, OLS out-of-sample accuracy is slightly better than spatial error model, which is not the case in-sample.

A comparison with the Swiss national model's, which corresponds to the state-of-practice four-step model used for AADT estimation, exhibits that the national model outperforms the estimated direct demand models in terms of predictive accuracy. In summary, national transport model has higher accuracy than the other models but at the same it has to be pointed out that it has been calibrated against the count data and it requires much more data and complicated models. In addition, a potential source of introduced bias might have resulted from not accounting for international commuters which can lead to underestimation of AADT close to the borders, an aspect which is taken into account in the national model.

Table 5 In-sample and out-of-sample predictive accuracy of estimated models

	Model	MdAPE	MPE	MAPE	MSE	SMAPE
In-sample predictive accuracy	OLS	27.62	16.02	43.41	4.19E+07	0.088
	OLS stress centr.	24.73	14.11	39.51	3.57E+07	0.082
	OLS acc. weighted	24.08	11.79	35.88	3.43E+07	0.076
	Negative binomial	26.89	23.48	44.57	5.10E+07	0.084
	Sp. error: Eucl. distance	21.39	11.41	34.56	2.95E+07	0.073
	Sp. error: Netw. distance	20.38	10.57	33.04	2.72E+07	0.070
	Sp. error: Netw. fftt	20.43	10.63	33.28	2.75E+07	0.070
	Sp. lag: Netw. distance	21.29	11.51	35.21	3.30E+07	0.075
	GWR	16.92	7.43	25.65	1.85E+07	0.056
	National model (4-step)	4.78	5.73	14.65	3.85E+06	0.031
Out-of-sample predictive accuracy	OLS	28.50	15.98	44.07	4.53E+07	0.090
	OLS stress centr.	25.88	13.99	40.34	3.83E+07	0.084
	OLS acc. weighted	25.95	13.26	39.31	3.81E+07	0.082
	Negative binomial	27.05	23.89	45.62	4.46E+07	0.086
	Sp. error: Eucl. distance	25.41	13.67	39.31	3.79E+07	0.082
	Sp. error: Netw. distance	25.60	13.76	39.28	3.79E+07	0.082
	Sp. error: Netw. fftt	25.55	13.35	39.16	3.79E+07	0.082
	Sp. lag Netw. distance	26.34	13.40	39.60	3.83E+07	0.083
	Kriging: Spherical	25.34	12.84	38.24	3.56E+07	0.080
	Kriging: Gaussian	25.47	12.86	38.26	3.56E+07	0.080
	Kriging: Exponential	25.95	13.26	39.31	3.81E+07	0.082
	GWR	25.52	9.68	36.86	3.60E+07	0.080
	National model (4-step)	4.82	5.84	14.39	3.66E+06	0.030

Attempting a comparison with the results of a similar scale study (Selby and Kockelman, 2013) where kriging models were estimated and the MAPE was calculated to be close to 60%. The difference in the magnitude of the accuracy can be attributed to a great extent to the inclusion of the centrality measures. In the case of the study conducted by Lowry for a community network though, the reported MdAPE values of 28%, are slightly larger but of similar magnitude with our results.

5. Conclusions

In the present paper a direct demand modelling approach for AADT prediction on a nationwide network is presented. It is exhibited that the inclusion of network theory-based variables in the model formulation can lead to a significant enhancement on the predictive accuracy. In addition, a methodology for expanding the stress centrality index to align with travel demand

modelling aspects is presented and evaluated, providing some concrete evidence in favour of it.

In addition to the already tested models in the literature, it is shown that while GWR has the highest predictive accuracy its underlying assumptions make it more appropriate for interpolation purposes. In contrast, spatial error and OLS models have the potential to be applied for forecasting purposes as well since they are estimated parameters are unbiased and consistent. Given this consideration, spatial error model and OLS can be used within a structural equation framework to make statements about the speed and the AADT on a link level, accounting for both their well-known interdependencies and the spatial autocorrelation (Sarlas and Axhausen, 2015a). These two constitute the minimum requirements for the transport project appraisal.

At last, a comparison of models predictive accuracy to the output of a traditional four-step model is conducted to show that direct demand models can constitute a trustworthy alternative to more advanced, but definitely more data demanding and computationally burdensome models. Conceptually, it is arguable that a simplified approach cannot exhibit the predictive accuracy and the sensitivity of the existing approaches (four-step or agent-based models). However, the higher sensitivity might allow to address more issues, but then raises the issue if the forecast is better, as there are more independent variables to forecast/fix. Furthermore, it cannot be overlooked that when it comes to the appraisal of public transport projects, as Flyvbjerg et al. (Flyvbjerg et al., 2005) 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 developed methodology can be easily applied to different scales of network, where a finer zonal analysis level and the identification of clusters of trip production and attraction can be used. Moreover, it requires only publicly available socioeconomic data and can utilize different available networks (e.g. Open street map).

Acknowledgements

This paper incorporates incremental improvements in a previous working paper of the authors (Sarlas and Axhausen, 2015b), and it is based on an ongoing research project funded by the Swiss National Science Foundation entitled *"Models without (personal) data?"*.

6. References

Anselin, L. (1988a) *Spatial Econometrics: Methods and Models*, Springer, Dordrecht, Netherlands.

Anselin, L. (1988b) Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity, *Geographical Analysis*, **20** (1), 1–17.

Bivand, R. et al. (2011) spdep: Spatial dependence: weighting schemes, statistics and models.

Box, G.E.P. and D.R. Cox (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B (Methodological)*, **26** (2), 211–252.

Breusch, T.S. and A.R. Pagan (1979), A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica*, **47**, 1287–1294.

Charlton, M. and S. Fotheringham (2009) Geographically Weighted Regression, *White paper*, National Centre for Geocomputation, National University of Ireland Maynooth.

Csárdi, G. and T. Nepusz (2006) The igraph software package for complex network research, *InterJournal Complex Systems*, **1695**.

Desyllas, J. et al. (2003) Pedestrian demand modelling of large cities: an applied example from London, Center for Advanced Spatial Analysis, University College London.

Duddu, V. and S. Pulugurtha (2013) Principle of demographic gravitation to estimate annual average daily traffic: Comparison of statistical and neural network models, *Journal of Transportation Engineering*, **139** (6), 585–595.

Eom, J. et al. (2006) Improving the Prediction of Annual Average Daily Traffic for Non-freeway Facilities by Applying a Spatial Statistical Method, *Transportation Research Record: Journal of the Transportation Research Board*, **1968** (1968), 20–29.

Flyvbjerg, B et al. (2005) How (In) accurate are demand forecasts in public works projects, *Journal of the American planning association*, **71**.

Goldfeld, S.M. and R.E. Quandt (1965), Some Tests for Homoskedasticity, *Journal of the American Statistical Association*, **60**, 539–547.

Halás, M., P. Klapka and P. Kladivo (2014) Distance-decay functions for daily travel-to-work flows, *Journal of Transport Geography*, **35**, 107–119.

Hansen, W.G. (1959) How Accessibility Shapes Land Use, *Journal of the American Institute of Planners*, **25** (2), 73–76.

Shimbel, A. (1953) Structural parameters of communication networks. *Bull. Math. Biophys*, **15** (4), 501–507.

Kissling, W.D. and G. Carl (2007) Spatial autocorrelation and the selection of simultaneous autoregressive models, *Global Ecology and Biogeography*, **17** (1), 59–71.

Lowry, M. (2014) Spatial interpolation of traffic counts based on origin–destination centrality, *Journal of Transport* Geography, **36**, 98–105.

Makridakis, S. and M. Hibon (1995) Evaluating accuracy (or error) measures. *Insead*, 1–41.

McDaniel, S., M. Lowry and M. Dixon (2014) Using Origin-Destination Centrality to Estimate Directional Bicycle Volumes, *Transportation Research Record: Journal of the Transportation Research Board*, **2430**, 12–19.

Mohamad, D. et al. (1998) Annual Average Daily Traffic Prediction Model for County Roads, *Transportation Research Record: Journal of the Transportation Research Board*, **1617**, 69–77.

Oliver, M. and R. Webster (1990) Kriging: a method of interpolation for geographical information systems, *International journal of geographical information systems*, **4** (3), 313–332.

Osborne, J.W. (2010) Improving your data transformations : Applying the Box-Cox transformation, *Practical Assessment, Research & Evaluation*, **15** (12), 1–9.

Pebesma, E.J. (2004) Multivariable geostatistics in S: The gstat package, *Computers and Geosciences*, **30** (7), 683–691.

Pulugurtha, S. and P. Kusam (2012) Modeling Annual Average Daily Traffic with Integrated Spatial Data from Multiple Network Buffer Bandwidths, *Transportation Research Record: Journal of the Transportation Research Board*, **2291**, 53–60.

Sarlas, G. and K.W. Axhausen (2015a) Localized speed prediction with the use of spatial simultaneous autoregressive models, *Paper presented at the 94th Annual Transportation Research Board Meeting*, Washington D.C., January 2015.

Sarlas, G. and K.W. Axhausen (2015b) Prediction of AADT on a nationwide network level based on an accessibility weighted centrality measure, *Arbeitsberichte Verkehrs- und Raumplanung*, **1094,** IVT, ETH Zürich, Zürich.

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

Wang, X. and K.M. Kockelman (2009) Forecasting Network Data, *Transportation Research Record: Journal of the Transportation Research Board*, **2105**, 100–108.

Winkelmann, R. (2008). Econometric analysis of count data, *Springer*.

Xia, Q. et al. (1999) Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County, *Transportation Research Record: Journal of the Transportation Research Board*, **1660**, 32–40.

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