Prediction of AADT on a nationwide network based on an accessibility-weighted centrality measure

Prediction of AADT on a nationwide network based on an accessibility-weighted centrality measure

Georgios Sarlas*
PhD Student
Institute for Transport Planning and Systems
ETH Zurich
HIL F 51.3, Stefano-Franscini-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-Franscini-Platz 5
8093, Zurich, Switzerland
Phone: +41 44 633 39 43
E-mail: axhausen@ivt.baug.ethz.ch

* Corresponding author

Total Word Count: **5348 (Text) + 7 (Figures/Tables) * 250 + References (500) = 7598 words**

Submitted for Publication and Presentation at the Transportation Research Board 95th Annual Meeting January 10-14, 2015, Washington, D.C.

ABSTRACT

In the present paper a direct demand modelling approach for AADT prediction on a nationwide network is presented. A particular focus if given on the construction of a variable that can capture the interregional demand patterns by taking into account the direction of potential interactions over space, called accessibility-weighted centrality, by applying different modifications on the stress centrality measure, tailored for the task of AADT prediction. It is exhibited that it can lead to a significant enhancement on the accuracy of the models. In addition to the already tested models in the literature, two SAR models are estimated and it is shown that GWR and kriging are more appropriate for interpolation purposes, while spatial error and OLS models might give slightly worse results but they have the potential to be applied both for interpolation and forecasting since their estimated parameters are unbiased and consistent. A comparison of models predictive accuracy to the output of a traditional four-step model is conducted to show that direct demand models on nationwide scale can constitute a trustworthy alternative to more advanced, but definitely more data demanding and computationally burdensome models.

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 various 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.

Literature review

Xia et al.*(1)* 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.*(2)* 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. *(3)* developed a multiple regression analysis model for pedestrian flows.

The plausibility of applying the GWR model for estimating AADT was demonstrated in another study *(4)* 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. *(5)* 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 *(6)* 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 *(7)* 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 *(8)* 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 *(9)*, the authors 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 modelers.

In a recent study by Lowry *(10)*, 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 *(11)*.

Description of the framework of the paper

The objective of the current paper is to develop a direct demand modelling approach for prediction of AADT on a nationwide network, which has not been addressed in the existing literature. The particularity of the nationwide network level case stems from the incapability of the spatial densities of different socioeconomic data to capture adequately the demand patterns that occur on the links, since they fail to bear explanatory power with respect to high volume of interregional through traffic. Naturally, the construction of a variable that can account for

interregional flows necessitates, taking into account the direction of potential interactions, allowing us to capture the demand capacity interaction at the core of transport modelling. More specifically, a variable called accessibility-weighted centrality measure is constructed, building upon the work of Lowry *(10)* who showed that the use of stress centrality constitutes a huge improvement over traditional direct-demand models, and modifying accordingly the stress centrality measure for the particular problem at hand.

In addition to the already tested models in the literature, the family of spatial simultaneous autoregressive (SAR) models is exploited with their capability to be applied for AADT prediction purposes. The advantage of such models is that they can resolve spatial dependence issues, offering a structural explanation of the AADT and since their estimated coefficients are unbiased and consistent, they can fulfill both interpolation and forecasting purposes which is important for 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 constructed variable and also on 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.

METHODOLOGY

Accessibility-weighted centrality measure

The construction of a new variable capturing the interregional demand patterns, taking into account the direction of potential interactions over space, is of high importance for the estimation of AADT models on a nationwide network. Making use of the graph 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' analysis. 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 *(12)* and is defined as the number of shortest paths connecting all pairs of nodes of the network that pass from 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 where 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 that 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 *(13)*. Two variations of distance decay function are checked to identify the one that fits the data better *(14)*. 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 by the origin node in total. 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.

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

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

$$\text{Travel Accessibility}_i = \sum_i^j \text{Employm}_j * f(\text{cost}_{ij}) \quad (4)$$

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

The parameters of the distance-decay function can be either estimated if data availability allows it, or taken from another study, if required.

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 *(15)* 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 *(16)*. 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:

$$Y = \rho WY + \beta X + \varepsilon \quad (9)$$

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 *(17)*.

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

Where the notation $\beta_i(z)$ indicates that the parameter describes a relationship around location u and is specific to that location *(17)*.

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. In previous studies of AADT (e.g. *(7)*), 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.

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 accessibility-weighted

centrality measure 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 314 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.

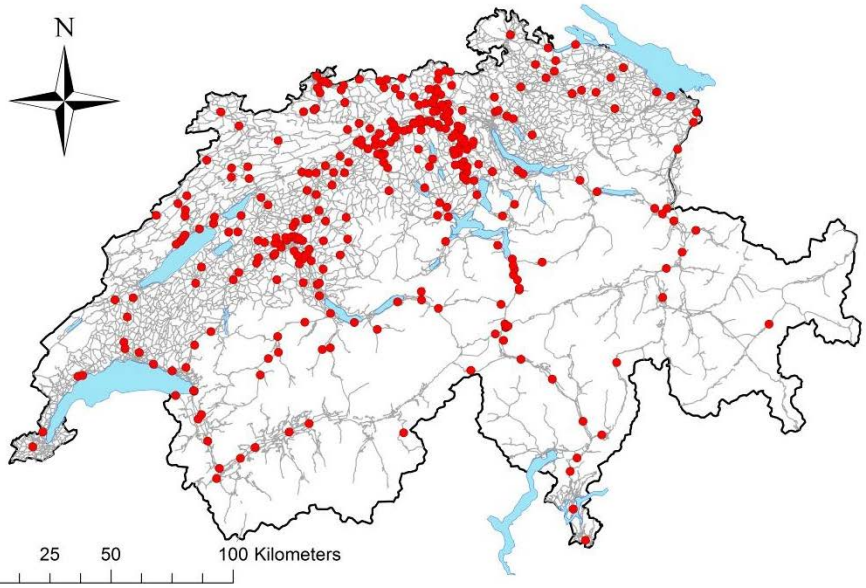

FIGURE 1 Case study network and count locations (source: ARE, National Transport Model, 2010)

Accessibility-weighted centrality measure

The first step is to proceed to the construction of the accessibility-weighted centrality measure for the study network, according to the defined methodology. As mentioned before, 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 *(14)*. 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 mainly contributing to the available AADT values 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 an existing Origin-Destination (O-D) matrix, that corresponds to the demand for trips among all the municipalities of Switzerland in year 2010 (ARE; National Transport Model, 2010). The 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).

Subsequently, the portion of total daily commuters/trips that lie within each bin of given length is calculated. The length of the bin is chosen to be three minutes and three kilometers respectively. These portions take values between 0 and 1 and they are referred to as the interaction intensity. A normalization of the aforementioned portions by the percentage of the first bin (maximum) follows to ensure that we have values covering the whole range of potential values. The next step is to quantify how interaction intensity decreases over space, which actually corresponds to the parameters of the distance decay functions. The nonlinear least-squares estimates of the parameters are calculated by following 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 cases of distance and travel time, and thus is the chosen one.

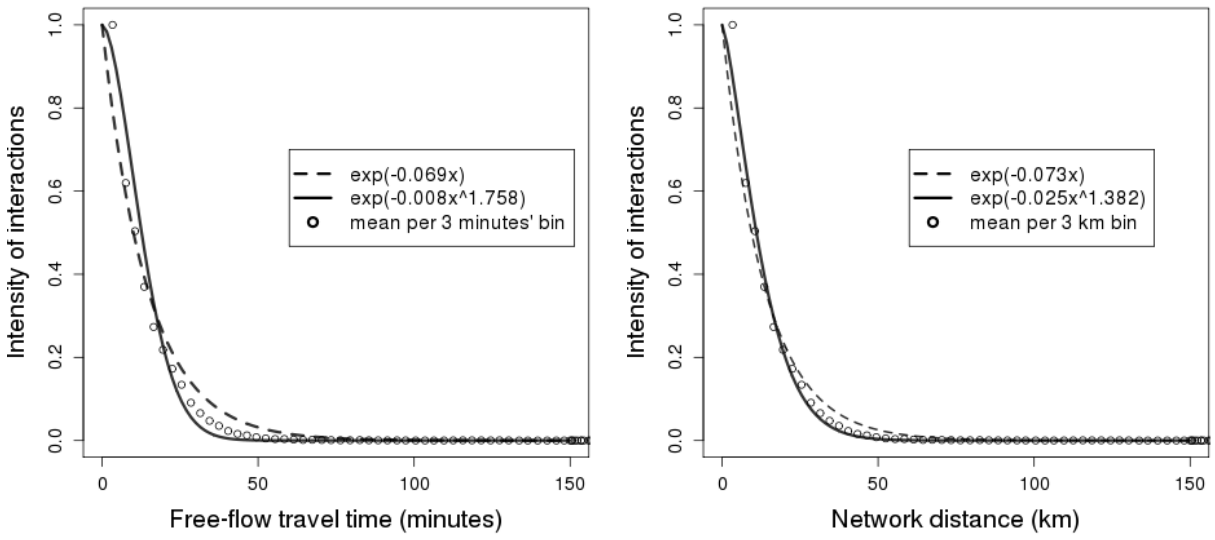

FIGURE 2 Estimated parameters of the distance decay functions

Alternatively, these parameters could be taken from previous studies as long as the employed cost metric is consistent with the one of the case study.

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 all the links where count data exists 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 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.

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, the number of lanes, and the speed limit are the chosen explanatory variables. On the demand side, the constructed accessibility-weighted centrality measure is introduced for incorporating information about the magnitude and the direction of the spatial interactions, serving as an approximation of spatial flows. 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 new variable has a similar magnitude as the AADT while their correlation is close to 0.75, taking it as 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	9834	5851	10399
Collector road	Dummy	34	-	-
Alpine road	Dummy	21	-	-
Rural major road	Dummy	28	-	-
Major road	Dummy	112	-	-
Freeway-Highway	Dummy	119	-	-
Two-lane road	Dummy	92	-	-
Three-lane road	Dummy	13	-	-
Free-flow speed	km/hr	85.75	80	22.56
Accessibility-weighted centrality	Empl. opportunities	9846	5400	11737
Public transp. density: 5km radius	stops/ sq. km	1.1	0.82	0.86

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 *(18)*. In this case, either 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 *(19)*. Based on that, the Box-Cox transformation*(20)* 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 0.414, indicating a transformation somewhere between the square and the third root.

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

Given the high correlation of the centrality variable with the AADT, we choose to apply an identical Box-Cox transformation to it, 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 to show their clear linear correlation with the transformed AADT values.

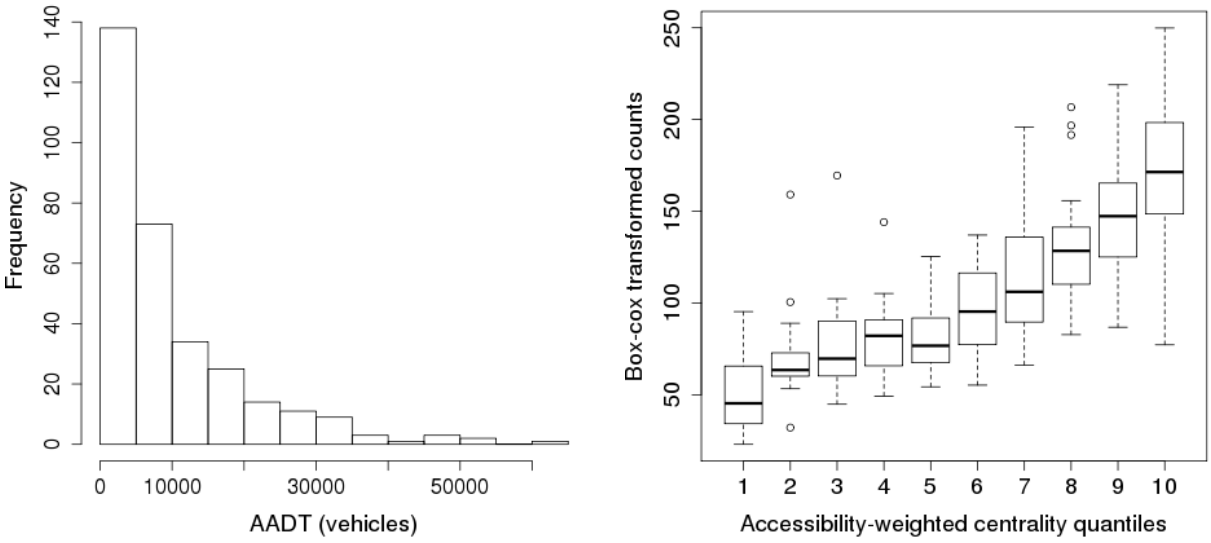

FIGURE 3 Histogram of the AADT data and box-plot of the centrality with respect to AADT

It should be noted that the involved data processing, models estimation, and network processing are undertaken with the statistical programming language R *(21)*, making use of different available packages *(22–24)*.

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 the already tested models 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.

At first, an OLS model is estimated to serve as the basis for the comparison 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 and more (or less) explanatory power is attributed to them than it should.

Therefore, 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 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. Making use again of 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. The estimated coefficients for the OLS and the spatial regression models are presented in Table 2.

TABLE 2 Estimated coefficients for the different models

	OLS		Sp. error model		Sp. lag model	
Indep. Variables	Estimate	Sign.	Estimate	Sign.	Estimate	Sign.
Collector road	23.50	***	22.17	***	9.42	
Alpine road	26.59	***	29.92	***	15.88	.
Rural major road	39.59	***	37.79	***	24.98	***
Major road	41.05	***	39.91	***	25.43	***
Freeway-Highway	53.43	***	53.36	***	39.84	***
Two-lane road	28.35	***	26.49	***	24.93	***
Three-lane road	76.20	***	73.65	***	74.20	***
Free-flow speed	0.16	.	0.18	*	0.14	.
Acc.-weighted centrality	0.26	***	0.25	***	0.24	***
Public transp. density: 5km	7.16	***	6.89	***	5.69	***
lamda	-		0.55	***	-	
rho	-		-		0.20	***
Adjusted R	*0.978*		-		-	
Akaike Inf. criterion	*2663*		*2625*		*2641*	
Moran's I measure	*0.21****		*-0.006*		*0.13*	***
No. of observations						*314*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*						

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 free-flow speed is in line with expectations. The demand relevant variables, public transport density and accessibility-weighted centrality, 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 same pattern 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

366 measures, the Akaike information criterion shows that the spatial error model is the best one
367 among the three.

368 The next estimated model is GWR, which aims to resolve spatial heterogeneity issues and it is
369 calculated by taking into account an adaptive bandwidth. The results are reported in Table 3.

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

374 **TABLE 2 Estimated coefficients for GWR model**

Variable	Min.	1st Quant.	Median	3rd Quant.	Max.
Collector road	2.94	21.28	27.46	27.45	35.26
Alpine road	0.59	23.92	29.73	29.59	37.03
Rural major road	16.43	33.75	42.43	41.64	50.13
Major road	14.54	37.16	44.27	44.76	54.77
Freeway-Highway	23.10	45.31	58.42	58.48	69.13
Two-lane road	18.57	24.55	30.60	30.45	34.86
Three-lane road	53.07	67.87	76.61	76.16	84.54
Free-flow speed	-0.31	-0.02	0.12	0.12	0.28
Acc.-weighted centrality	0.08	0.23	0.26	0.27	0.31
Public transp. density: 5km	-3.25	4.67	6.60	6.34	9.31
Local R square	0.86	0.90	0.92	0.92	0.94

375

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

379 **Evaluation of predictive accuracy of models**

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

386 The following five accuracy measures are calculated in order to allow the evaluation to take
387 place. Mean percentage error (MPE) and mean absolute percentage error (MAPE) are easily
388 interpretable measures, having the main disadvantage though that they are influenced by outliers.
389 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 *(25)*, where they conclude that for forecasting purposes MSE and SMAPE are found to be the more preferable measures. It should be noted that AADT predicted values are reversely 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 4.

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

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

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

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

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

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 one that employs a spatial matrix based on the free-flow time distance gives slightly better results. Among the kriging models, it can be concluded that the one with the exponential semivariogram has better accuracy.

The negative binomial model yields the results with the 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.

TABLE 3 In-sample and out-of-sample predictive accuracy of estimated models

	Model	MdAPE	MPE	MAPE	MSE	SMAPE
In-sample predictive accuracy	OLS	25.27	12.98	35.38	1.38E+07	7.59
	Negative binomial	23.25	24.61	45.01	2.54E+07	8.35
	Sp. error: Eucl. distance	22.10	12.17	33.96	1.18E+07	7.35
	Sp. error: Netw. distance	21.41	11.95	33.56	1.14E+07	7.26
	Sp. error: Netw. fftt	21.67	11.91	33.44	1.13E+07	7.24
	Sp. lag: Eucl. distance	23.87	12.05	33.95	1.33E+07	7.36
	Sp. lag: Netw. distance	24.18	12.05	34.09	1.31E+07	7.36
	Sp. lag: Netw. fftt	24.23	11.95	34.09	1.29E+07	7.37
	GWR	19.83	8.32	27.86	7.69E+06	6.22
	National model (4-step)	17.45	18.12	19.79	1.29E+07	4.36
out-of-sample predictive accuracy	OLS	26.13	14.58	37.68	1.56E+07	7.97
	Negative binomial	26.04	27.32	48.40	4.31E+07	8.83
	Sp. error: Eucl. distance	26.12	15.14	38.35	1.55E+07	8.03
	Sp. error: Netw. distance	26.06	15.20	38.38	1.56E+07	8.04
	Sp. error: Netw. fftt	26.13	14.36	38.23	1.57E+07	8.06
	Sp. lag: Eucl. distance	26.23	14.32	37.05	1.56E+07	7.87
	Sp. lag: Netw. distance	26.26	13.93	37.11	1.59E+07	7.90
	Sp. lag: Netw. fftt	26.27	13.52	37.00	1.58E+07	7.92
	Kriging: Spherical	24.58	12.66	35.29	1.44E+07	7.66
	Kriging: Gaussian	25.19	12.79	35.68	1.53E+07	7.75
	Kriging: Exponential	24.54	12.48	35.11	1.39E+07	7.63
	GWR	24.80	13.10	36.50	1.45E+07	7.84
	National model (4-step)	17.74	18.06	19.86	1.27E+07	4.37

Among the estimated models, GWR has the highest in-sample accuracy while kriging has the highest out-of-sample. However, in terms of MdAPE the out-of-sample difference between kriging and SAR models is 1.5%, while in terms of MAPE is almost 3 percent. 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.

Attempting a comparison with the Swiss national model's accuracy, which corresponds to the state-of-practice four-step model used for AADT estimation, the national model outperforms the estimated models. However, it has significantly higher MPE than the other models, which is of similar magnitude as MAPE, revealing that it systematically overestimates the AADT. In general, national transport model has higher accuracy than the other models but at the same it has to be pointed out that its higher MPE value raises some concerns, given the much more data and

424 complicated models it employs. In addition, a potential source of introduced might have resulted
425 from not accounting for international commuters which can lead to underestimation of AADT
426 close to the borders.

427 Attempting a comparison with the results of a study of a similar scale *(7)* where kriging models
428 were estimated and the MAPE was calculated to be close to 60%, the difference in the magnitude
429 of the accuracy can be attributed to a great extent to the inclusion of the centrality measure. In the
430 case of the study conducted by Lowry for a community network, the reported MdAPE values of
431 28%, are slightly larger but of similar magnitude with our results.

432 **CONCLUSIONS**

433 In the present paper a direct demand modelling approach for AADT prediction on a nationwide
434 network is presented. It is exhibited that the construction of a variable that can account for
435 interregional flows, such as the accessibility-weighted centrality measure, can lead to a
436 significant enhancement on the accuracy of the models. In addition to the already tested models
437 in the literature, the spatial error model is estimated and it is shown that GWR and kriging
438 models are more appropriate for interpolation purposes while spatial error and OLS models have
439 the potential to be applied for forecasting purposes as well since they are estimated parameters
440 are unbiased and consistent. Under this consideration, spatial error model and OLS can be used
441 within a structural equation framework to make statements about the speed and the AADT on a
442 link level, accounting for both their well-known interdependencies and the spatial autocorrelation
443 *(26)*. These two constitute the minimum requirements for the transport project appraisal.

444 At last, a comparison of models predictive accuracy to the output of a traditional four-step model
445 is conducted to show that direct demand models can constitute a trustworthy alternative to more
446 advanced, but definitely more data demanding and computationally burdensome models.
447 Conceptually, it is arguable that a simplified approach cannot exhibit the predictive accuracy and
448 the sensitivity of the existing approaches (four-step or agent-based models). However, the higher
449 sensitivity might allow to address more issues, but then raises the issue if the forecast is better, as
450 there are more independent variables to forecast/fix. Furthermore, it cannot be overseen that
451 when it comes to the appraisal of public transport projects, as Flyvbjerg et al. *(27)* argue, the
452 quality of the demand forecasts has not been improved over the years even though more complex
453 and advanced models have been employed.

454 The developed methodology can be easily applied to different scales of network, where a finer
455 zonal analysis level and the identification of clusters of trip production and attraction can be used.
456 Moreover, it requires only publicly available socioeconomic data and can utilize different
457 available networks (e.g. Open street map).

ACKNOWLEDGEMENTS

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

REFERENCES

1. Xia, Q., F. Zhao, Z. Chen, L. Shen, and D. Ospina. Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1660, Jan. 1999, pp. 32–40.

2. Mohamad, D., K. Sinha, T. Kuczek, and C. Scholer. Annual Average Daily Traffic Prediction Model for County Roads. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1617, Jan. 1998, pp. 69–77.

3. Desyllas, J., E. Duxbury, J. Ward, and A. Smith. Pedestrian demand modelling of large cities: an applied example from London. 2003.

4. 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.

5. Eom, J., M. Park, T.-Y. Heo, and L. Huntsinger. Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. *Transportation Research Record*, Vol. 1968, No. 1968, 2006, pp. 20–29.

6. Wang, X., and K. M. Kockelman. Forecasting Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2105, No. -1, Dec. 2009, pp. 100–108.

7. 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.

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

9. Duddu, V., and S. Pulugurtha. Principle of demographic gravitation to estimate annual average daily traffic: Comparison of statistical and neural network models. *Journal of Transportation Engineering*, No. June, 2013, pp. 585–595.

10. Lowry, M. Spatial interpolation of traffic counts based on origin–destination centrality. *Journal of Transport Geography*, Vol. 36, Apr. 2014, pp. 98–105.

11. McDaniel, S. Using Origin-Destination Centrality to Estimate Directional Bicycle Volumes. *assets.conferencespot.org*, 2014, pp. 1–16.

12. Shimbel, A. Structural parameters of communication networks. *Bull. Math. Biophys*, Vol. 15,No. 4, 1953, pp. 501–507.

13. Hansen, W. G. How Accessibility Shapes Land Use. *Journal of the American Institute of Planners*, Vol. 25, No. 2, May 1959, pp. 73–76.

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

15. Anselin, L. *Spatial Econometrics: Methods and Models*. Springer Netherlands, Dordrecht, 1988.

16. 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.

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

18. Winkelmann, R. *Econometric analysis of count data*. 2008.

19. Osborne, J. W. Improving your data transformations : Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, Vol. 15, No. 12, 2010, pp. 1–9.

20. Box, G. E. P., and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 26, No. 2, 1964, pp. 211–252.

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

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

23. 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.

24. Pebesma, E. J. Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, Vol. 30, No. 7, 2004, pp. 683–691.

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

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

27. 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.