J Geograph Syst (2007) 9:397–417
DOI 10.1007/s10109-007-0050-4

ORIGINAL ARTICLE

Predicting road system speeds using spatial structure variables and network characteristics

Jeremy K. Hackney · Michael Bernard · Sumit Bindra · Kay W. Axhausen

Received: 1 December 2006 / Accepted: 23 July 2007 / Published online: 25 September 2007

Abstract Spatial regression is applied to GPS floating car measurements to build a predictive model of road system speed as a function of link type, time period, and spatial structure. The models correct for correlated spatial errors and autocorrelation of speeds. Correlation neighborhoods are based on either Euclidean or network distance. Econometric and statistical methods are used to choose the best model form and statistical neighborhood. Models of different types have different coefficient estimates and fit quality, which might affect inferences. Speed predictions are validated against a holdout sample to illustrate the usefulness of spatial regression in road system speed monitoring.

Keywords Road speed model · Spatial regression · Spatial structure variables · Network · Transportation planning

JEL R41 · C51 · C52 · C53 · C21

J. K. Hackney (✉) · M. Bernard · K. W. Axhausen
Institute for Transport Planning and Systems, ETH Hönggerberg,
HIL F51.3, 8093 Zurich, Switzerland
e-mail: hackney@ivt.baug.ethz.ch

M. Bernard
e-mail: bernard@ivt.baug.ethz.ch

K. W. Axhausen
e-mail: axhausen@ivt.baug.ethz.ch

S. Bindra
Connecticut Transportation Institute, University of Connecticut,
Castleman Building, Room 205, 261 Glenbrook Road, Unit 2037,
Storrs, CT 06269-2037, USA
e-mail: sumit.bindra@uconn.edu

1 Short and long term monitoring of speeds

Estimates of link speeds at detailed spatio-temporal resolution are valuable for a range of short- and long-term objectives: from improving navigation systems and optimizing system operations, to estimating user benefits and monitoring system performance, adjustment, and expansion in the context of land use and regional development.

Measurements of road system speed by themselves are insufficient for the task. The spatial coverage of fixed detectors is too limited to be useful for navigation or system-wide monitoring, while measurement of all links and times with trackable probe vehicles would be too expensive, due to equipment, logistical, and telecommunication costs.

System-wide link speeds can be inferred from microsimulations (Nagel et al. 2000, 2003; Bradley and Bowman 2006), dynamic assignment (Peeta and Ziliaskopoulous 2001), and integrated transportation-land use models (Wegener 2004; Hunt et al. 2005; Salvini and Miller 2005; Waddell et al. 2005). However, these models are time-consuming to develop, to adapt to specific areas, and to maintain. The chief cost and uncertainty in assignment is estimating and calibrating origin-destination (OD) matrices on a zone structure and in time, though correctly specified networks (capacity) and accurate speed/flow relationships are also prerequisites. Agent microsimulation requires error-free networks, detailed socio-economic databases, and the computational resources and skill to scale models of agent behavior to large populations. Integrated land use and traffic models are also major efforts in data assimilation and tuning of feedbacks between the two systems.

The method presented here uses a computationally intensive but simpler approach to utilize spatial linear regression on a sample of average link speeds to infer speeds on the entire network, as a function of time of day, road network topology, and population structure. It can be thought of as a highly detailed direct demand model. The regression yields two speed components: the first, the average road speed by road type for the time period, is a non-spatial quantity. Spatial variation is added to the link speed estimates in the second component via the spatially resolved explanatory variables. Spatially resolved road network densities represent the effect of road supply on speed: e.g. local route alternatives could have the effect of raising the speed of traffic locally, or higher road densities could be associated with areas of high demand for access. Spatial data on population and employment is taken to be indicative of the intensity of local activities, reflecting travel demand locally.

The necessary high-resolution data on land use and a road network topology are found in most planning agencies. Samples of link speeds can be based on many sources. GPS datasets are proliferating due to private investment in dynamic navigation systems. Or, such data can be collected quickly at reasonable cost, as reliable and affordable GPS-based measurement units and the required map matching software is readily available (Marchal et al. 2006).

A model combining these variables to explain link speeds is attractive for several reasons. First, it is a way to quickly extract more value from existing speed and land use or population data. Second, the tools and skills needed to build the model are ubiquitous. Third, it is spatially and temporally specific while avoiding the high

costs associated with the traditional models. It offers a structural explanation of the speed in a more direct way than assignment models, even if it is not able to capture all the details of flow patterns. In a longer-term monitoring context, for example, it would be possible to estimate the effects of land use changes on speed, without re-specifying an origin-destination matrix. Finally, the resulting link speed estimates can be combined for analyses of OD or system speeds, or used as initial values in more extensive models.

This article presents the estimation of the spatial regression models, with an assessment of predictive power and transfer error. The discussion centers on the treatment of the spatial error structures through appropriate model form. A brief discussion of spatial regression follows. The data is then described and the choice of the weighting approach explained. A weighted least square (WLS) model is estimated first. A set of spatial autocorrelation models using different neighborhood matrices are estimated so that several treatments for spatial autocorrelation can be compared in depth. The best spatial models are described and compared to the WLS. The article concludes with recommendations for further research and advice for the practical application of the approach. Follow up work will compare results with results of conventional monitoring methods.

Spatial regression models saw widespread application after Anselin's (1988) description of the method. The broad application of spatial analysis in the transportation and urban planning context is reviewed by Miller (1999) and Páez and Scott (2004). Primarily, one is concerned with the explicit accounting for the spatial assumptions inherent in aggregation, such as defining travel zones, and for correlations in statistical models of spatially interacting processes like network flows and competing or complementary land uses. Recent applications in the field of transportation have been made in long-term land use models by Páez and Suzuki (2001) and Zhou and Kockelman (2005), trip generation by Páez et al. (2007), and in short-term location choice by Zhao and Bhat (2002), Bhat and Guo (2004), and Guo and Bhat (2007). The work of Steenberghen et al. (2004) and Black and Thomas (1998) are examples of spatial autocorrelation analyses of incidents on networks. Spatial autoregression is used by Bolduc et al. (1992, 1989) to model network flows between origin and destination zones, and by Kim and Niemeier (2001) to estimate mobile-source emissions along roadways.

2 Spatial analysis

A priori, one would expect spatial structure and traffic volumes to be spatially correlated. Spatial dependence is explicit for many activities that generate traffic because they are located (strategically or otherwise) according to spatially interacting activities, and in such a way as to optimize access to roadways. Similarly, traffic speed on a section of the network is influenced by the traffic or by signalization ahead, or else by correlation caused by adherence to speed limits. One would therefore expect that explaining link speeds with structural variables would yield spatially correlated residuals. If not treated, this will result in biased and inconsistent parameter estimates that cannot reliably be used for inference (LeSage 2000).

An ordinary or weighted least squares (OLS, WLS) model can be corrected for spatial correlations by adding information about the neighborhood (see Anselin 1988; LeSage 2000, whose terminology is used here). The spatial correlation is derived either from the regression residuals or the values of the independent variable in some set of (not necessarily spatially) neighboring observations in the dataset. A neighborhood weighting matrix W ($n \times n$) is employed to introduce the information into the equations predicting each of the n locations. Each row sum of the W-matrix is normalized to one. In contrast to most applications of spatial regression, the definition of the neighborhood is not obvious in the case of models including both spatially fixed quantities and quantities derived from flows in time, like road speeds, and will be discussed in detail.

The spatial lag, or spatial autoregressive model (SAR), is a linear regression of a dependent variable y on independent variables X that includes a term for the spatial dependence of the observations in X. The procedure is analogous to *detrending* a correlated time series:

$$y = \rho W_a y + \beta X + \varepsilon \tag{1}$$

with

$$\varepsilon \sim N(0, \sigma). \tag{2}$$

The spatial error model (SEM) corrects for the spatial correlation of the error terms, and is analogous to stationary correlated errors in time series data:

$$y = \beta X + u \tag{3}$$

with

$$u = \lambda W_e u + \varepsilon, \quad \varepsilon \sim N(0, \sigma). \tag{4}$$

The general spatial autoregressive model with a correlated error term (SAC) includes both the spatial lag term and the correlation of the error terms:

$$y = \rho W_a y + \beta X + u \tag{5}$$

with

$$u = \lambda W_e u + \varepsilon, \quad \varepsilon \sim N(0, \sigma). \tag{6}$$

The parameter ρ in the SAR and SAC models represents the additional influence of neighboring observed values on the dependent variable. In the SEM and SAC models, the parameter λ corrects for spatially correlated errors.

3 Dataset

The regressions use a dataset of 3 weeks' continuous daytime floating car measurements on a sample of zone-to-zone legs within the Canton Zurich that were chosen to represent the daily average demand matrix. The GPS data is cleaned

of driver errors and personal stops (as logged onboard by the drivers), and matched topologically to a network model (Marchal et al. 2006). The network model's directed links are categorized by link type, according to infrastructure features and speed limit (Public Works Office of the Canton Zurich 2002): highways, trunk roads, collector roads, distributor roads, and other roads. The matching yields 52,000 speed observations on 3,680 directed links, where the link speed is defined as the link length divided by the travel time over the link (exit time–entry time), in km/h. The average number of measurements per link is 13, and the median is 11.

The observations are averaged by link number and time period for consistency with the practice of the Canton's planning office. The four periods are weekday peak (6:30–8:30 a.m. and 4:30–6:30 p.m.), weekday shoulder (8:30 a.m.–4:30 p.m. and 6:30–8:30 p.m.), weekday off-peak (8:30 p.m.–6:00 a.m.), and Saturday. Thus, each observation used in the regression represents the average speed on a directed link during one time period. The mean speed, standard deviation, and number of observations for the resulting 10,506 observations by road type are: highways (86.1, 14.0, 1,509), trunk roads (37.0, 13.3, 5,855), collector roads (46.9, 8.9, 1,393), distributor roads (25.5, 8.4, 1,195), and other roads (24.9, 4.5, 554).

The dataset is partitioned into an estimation sample and a validation sample that is used to determine goodness-of-fit for forecasting, and to quantify the predictive quality of the model (see Fig. 1). The sampling is constrained by the requirement that the network of links be contiguous. The two samples were chosen to represent the two urban centers in the region in order to include similar land uses, densities, and network characteristics. The estimation sample is the Zurich metropolitan region of 9,297 observations. The validation sample is the Winterthur metropolitan region consisting of 1,209 observations.

The higher speed links have higher variance independent of sample size, indicating a heteroscedastic dependent variable. Indeed the OLS residuals are also heteroscedastic. Because a single error term (one regression equation) for all road types is desired for later spatial treatments, weighted least squares is indicated. The procedure groups appropriate observations of the heteroscedastic variable and divides the OLS equation by the group-specific residual variances (Maddala 2001). Here, the framework of the problem provides convenient groups based on road type. Heteroscedastic residuals are no longer detected after dividing the OLS equation by the residual variance according to road type. The WLS parameters have the same units as in the OLS, and can be used directly to calculate link speed predictions.

The set of spatial variables detailing the spatial population structure and the structure of road network was constructed and intersected with the network links using geographic information system software. The variables available at hectare grid resolution are the *population*, *employment opportunities*, and *employed persons* from the national census (Swiss Federal Statistical Office 2001). *Employed persons* and *population* are nearly perfectly correlated. Population, instead of the number of employed people, is included in the models because it is more likely to be a variable available to planners. These densities were weighted with a kernel density function over radii R of 1, 3, and 5 km to create different variables that capture the effect of the spatial structure with increasing distance from a link (see Fig. 2). The kernel density estimators take the form:

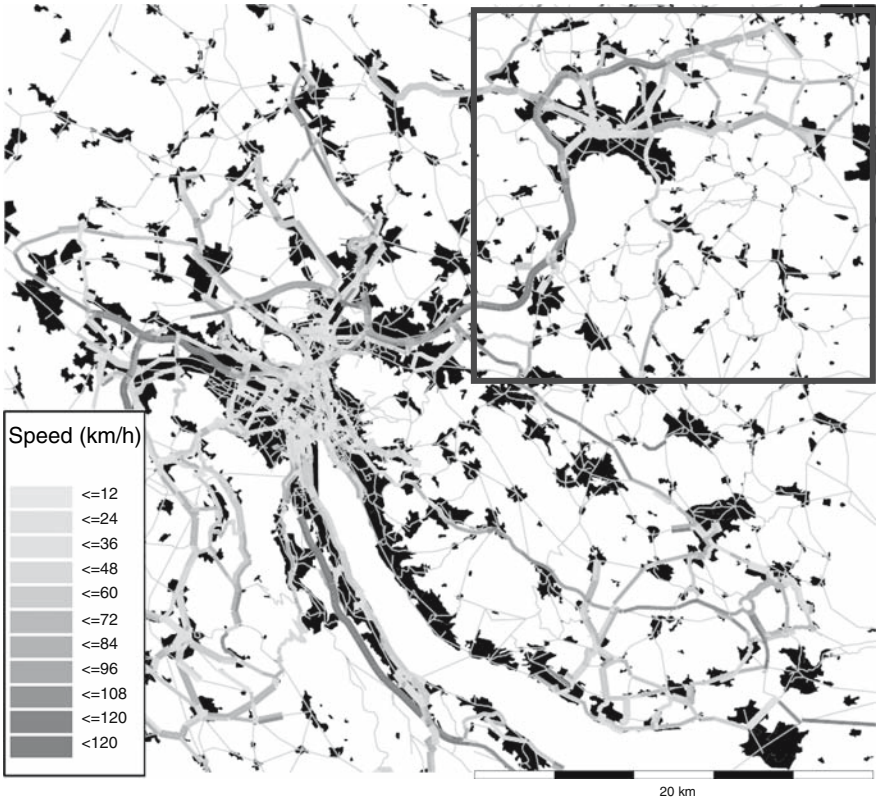

Fig. 1 Average measured road speeds (all times) matched to the Canton network model. The city of Zurich (fit sample) lies around a lake. The holdout sample centered on Winterthur is marked by the *box*. The black areas are developed. The finely drawn links were not measured. Discontinuities in measurements are caused by disturbances to the GPS signal

$$\lambda(s) = n^{-1}b^{-2}\Sigma\kappa[(s - s_i)/b] \tag{7}$$

where s_1, s_2,…, s_n are the variable values in the n hectares within the region R, b is the bandwidth of 70.72 m (half the diagonal of the hectare) and κ is the Gaussian spatial probability density function.

The length of road by type per hectare (*road density*) and the number of *highway access points* (on/off ramps) per hectare were calculated with a high-resolution GIS network model of the Canton that was intersected with the hectares (Navteq Corporation 2004). The *road density*, in units of meters per hectare, and the number of *highway access points* per hectare, are indicators of the local routing alternatives and the number of intersections near a link which could influence speed on the link by way of flow volume, signalization, or flow continuity. They are also indicators of land use, but the correlation with these variables is sufficiently low as to not cause concern for the regression. The *road densities* are not kernel-weighted,

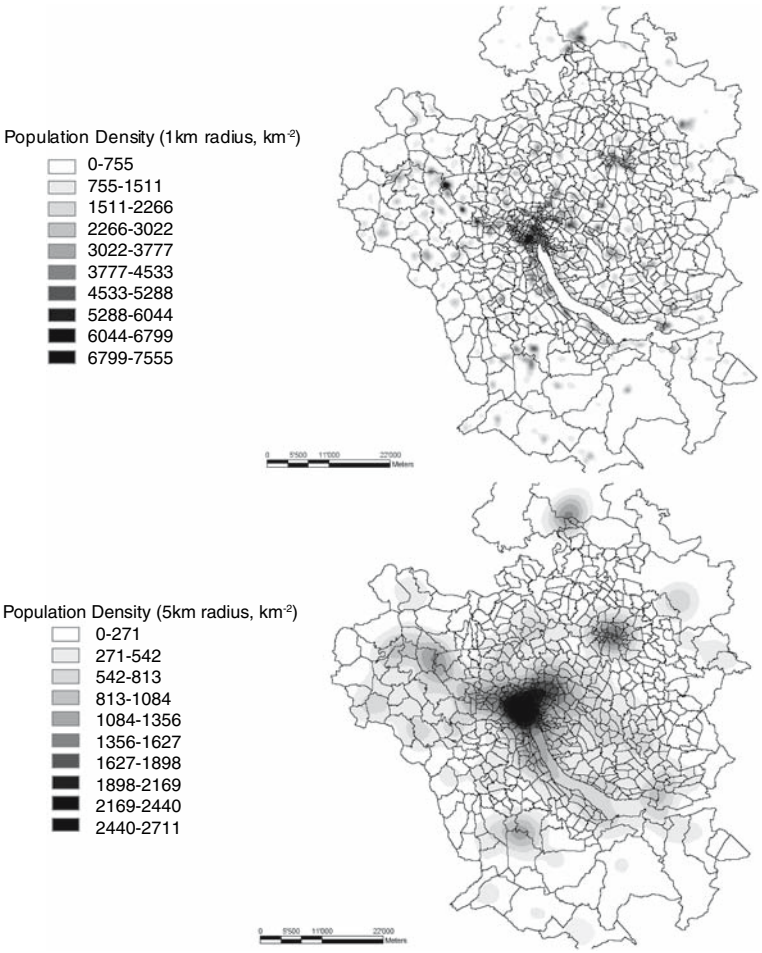

Fig. 2 Spatial patterns of population density in Canton Zurich with kernel radii 1 and 5 km

corresponding to the assumption that their effect on speed is localized. The *highway access points* are kernel weighted in the same manner as the population variables.

The regressions in this paper associate each network link with the spatial hectare value closest to the downstream endpoint of the link. The upstream link endpoint could just as well have been used. An endpoint is chosen for association with spatial data for two reasons. First, only the geographic position of the endpoints of links is known for certain. The routing of the link is not geographically accurate; as a link is really only a pointer between nodes for assignment models, the real path of the road might intersect hectares other than those along a straight line between nodes. Second, the links have differing lengths for reasons other than local land use (like road type designation). In general, the shorter links correspond to dense portions of Zurich with lower-speed roads, and longer links are highways and overland routes.

The mean, median, and mode of link length are 465, 280, and 100, respectively. Thus, most links span several hectares. Computing an average value of the hectare variables along each link or between upstream and downstream endpoints would aggregate the spatial data on arbitrary and variable distance scales (link length), resulting in a kind of modifiable areal unit problem on a link rather than zonal basis. Using network nodes avoids this statistical inconsistency.

4 WLS results

The OLS regression is estimated using the program SPSS with a stepwise estimation/validation procedure that adds and eliminates variables by seeking marginal improvements in the F statistic, retaining only coefficients significant at the 5% level. The method is robust against overfitting but is insensitive to correlated independent variables which would invalidate the standard errors of the estimates. Specific combinations of network and structure variables were chosen for the stepwise regression based on their qualitative meaning in explaining speeds, their correlation with speed, and a low correlation with each other. The logarithm of the structural variables fits the relationship better and correlates stronger with speeds. Finally, only the combinations of variables with the lowest Variance Inflation Factor (Maddala 2001) were used, to minimize correlation of the variables with the regression residuals.

Often there is little difference in fit quality or parameter statistics across different combinations of structural variables. Among those with the best statistics, the model with the most plausible qualitative explanation was retained for the final form of the WLS.

The WLS is estimated, like the following spatial regressions, using the econometrics library in Matlab (LeSage 2005). It uses dummies for road type and time of day to capture assumed independent effects on average speeds (thus this is not a temporal model). The variables used and the estimated parameters are in Table 2. Variables were kept if they were significant at $\alpha = 5\%$ or if they served illustrative purposes for the effects of the spatial correlation treatments. The adjusted R^2 for 9,297 observations and 34 variables is 0.657.

The average speeds (dummy coefficients) correspond to the relative hierarchy in the Canton's road system, the travel period, and the speed limits on the different road types. Speeds are highest on Saturdays and during shoulder/off peak periods for all road types. During peak periods, speeds on the Highways and major Trunk Roads are strongly reduced. The variation across time period is less pronounced on secondary road types, reflecting consistency of flow, traffic control, etc.

The parameters for the kernel density-smoothed spatial variables *employment opportunities*, *population*, and number of *highway access points* have negative sign consistent with expectations: speeds decrease with increasing activity densities. The radii of maximum effect are slightly different for the different road types. Highway speeds are more strongly associated with job density at a wide radius of 5 km, and with highway access density locally at a radius of 1 km (this is nearly the average distance between highway on- and off-ramps). Speed on lower ranked roads is

associated more with the local *employment* density (1 km) and the *population* density in a 5 km radius.

The parameters for the *road density* are all positive except for the Urban Collector Roads parameter, which is insignificant, and the Urban Distributor Roads parameter, which is strongly negative. The magnitude of each influence at average road densities is 2–8 km/h. The interpretation is that the presence of higher-speed roads near a link is an indicator of land use dedicated to traffic throughput to destinations not directly involved with the immediate hectare, with higher-speed flow as a result. The presence of lower-speed (urban) roads, as an indicator of land use requiring high accessibility to a local origin or destination, would be expected to be associated with lower speeds on the link (and perhaps correlated with land uses).

5 Neighborhood matrices

The clear correlation of speed observations demonstrated by Bernard et al. (2006) supports the discussion above that spatial correlations should be expected a priori in the WLS residuals. This section describes the rationale for two alternative approaches to defining distance and neighborhood. One can measure distance between a pair of links either along the shortest network path between them, or as Euclidean ("planar spatial", Okabe et al. 2006) distance by the midpoints of the links.

The first measure is spatially inhomogeneous and not symmetric, due to, for example, one-way streets or limited access roads. The explanatory hypothesis is that the flows along the path create the correlations. The second measure is spatially symmetric. Here, the explanatory hypothesis is that the abutting land uses and their travel generate the correlations.

While it is quite useful to assume spatially symmetric error correlations for regressions of geographically fixed variables like land rents, there are good reasons to expect the residual correlations of a traffic speed regression to be stronger on networks than symmetrically distributed in space. First, spatially proximate road links might only connect with each other at a distant part of the network, so (contemporaneous) traffic loads on proximate links might not be related except by the type and intensity of local land use. This would weaken a spatial model's ability to discern between spatial error and autocorrelation terms. One example is the oncoming traffic lane: Travel demand is strongly directional at peak periods, so opposite lanes may carry much different flows, in which case the correlation of the speed variances in opposing directions will not be strong. A second reason that the error correlation structure for traffic is not likely to be spatially symmetric is the temporal dependence of a traffic state: events that occur upstream in the traffic flow cannot have relevance to concurrent events downstream. While upstream events may indeed be correlated to the speed on the link, it only makes causal sense to model correlation from links downstream in the flow.

The Euclidean set of nearest neighbors is constructed for link i by searching outward in all directions from the midpoint of i for the midpoints of the N nearest

links, where their Euclidean distance is the measure of nearness. The method *nnw* in the Matlab spatial econometrics toolbox is used (LeSage 2005).

The network neighborhood of link i is the set of downstream links within a given network distance D, in this case defined as the number of road intersections (nodes with three or more edges, Balmer et al. 2005). The distance between road intersections is different than link length referred to earlier. While links are abstract in the network model, the locations of intersections are accurately depicted in space and strongly associated with the dynamic between the accessibility to land enabled by an intersection and the intensity of land use. The network is searched from i in the direction of link flow, including all branches of links encountered, up to D downstream intersections (see Fig. 3). The number of nearest neighbor links will vary according to how many links join at each intersection. The oncoming lane is only reachable by a U-turn and has a distance of at least one intersection.

Speeds and residuals are assumed to be independent across the four time periods used. Thus, if links i and j are within distance D on the network or within N nearest neighbors in space, they are only considered neighbors if there is a speed observation for both i and j during the same time period.

6 Spatial analysis results

Spatial regressions are indicated if analysis shows that the least squares residuals are correlated across the neighborhood matrix. Fit statistics (e.g. Moran's I or Lagrange Multiplier Statistic for SAR models) are desirable indicators of residual spatial errors. But their calculation requires inversion of the n^2 neighborhood matrix. Four GB of computer memory were not sufficient to calculate fit statistics for this dataset. In order to identify spatially correlated residuals, it is computationally cheaper to

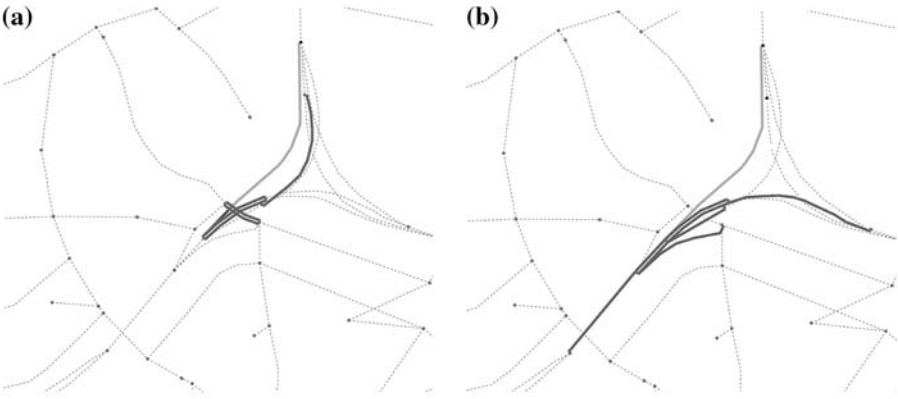

Fig. 3 Two link neighborhoods. Link i is *solid grey*, the neighbors are *solid black*, and non-neighboring links are *dotted lines*: **a** spatially symmetric nearest neighbors by Euclidean distance (five neighbors), **b** neighbors within a network distance of two intersections (also five neighbors)

estimate the full regressions and to compare the significance of the estimated correlation parameter and the log-likelihoods. The regressions can be calculated using sparse W matrices which save computer memory (LeSage 2005).

The SAR and SEM models explain the correlated spatial model variance differently. The SEM model assumes a common but unidentified spatial process which affects all of the variables associated by the W matrix. A significant parameter indicates missing spatial variables (Bivand 1998). Examples are areas where older architecture or topography forces roads to be narrower and more curvy, areas where fog or ice was present (endemic to the study area in November), or the specific composition and distribution of structural variables within a hectare that impact on travel speed differently, such as whether the employment opportunities are associated with a large shopping mall versus offices.

The SAR model should be investigated if a process can be assumed which would lead to spatially autocorrelated dependent variables. In this case it is an attempt to explain directly the speed on a link as a function of the speed of downstream traffic or signalization, as effects spill over from one road segment to the next along the path of influence in the W matrix. The SAR model must still be tested for spatially correlated residuals and corrected if necessary. The determination of the best spatial model using both an autoregressive and a spatial error term is described later.

Determining the relevant correlation neighborhood is discussed in Griffith (1996), and Stetzer (1982) summarizes experience with weighting versus neighborhood area. In this case, the resources were available to estimate models with a range of neighborhood matrices and to work with those models with the statistical best fits. Fit and maximum likelihood estimation statistics of the WLS and of spatial regressions using the first eight network and 16 Euclidean orders of neighborhood matrices are shown in Table 1.

The speed (v) and residual (r) correlation (ρ_v and ρ_r) are calculated using neighboring pairs of values, i.e. for all non-zero elements of the adjacency matrix m_{ij} (=0 if w_{ij} = 0 and 1 otherwise). As this matrix is not necessarily symmetric, the mean values and standard deviation have to be calculated separately for the first and second elements of the pair (indicated by '1' and '2' in the expression). Equation (8) shows the calculation for a generic variable x which is to be replaced with v for speed or r for the residuals, respectively:

$$\rho_x = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}\left(x_j - \bar{x}_1\right)\left(x_j - \bar{x}_2\right)}{\sigma_{x1}\,\sigma_{x2}} \tag{8}$$

with

$$\overline{x_1} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}x_i}{M},\ \overline{x_2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}x_j}{M}, \tag{9}$$

$$\sigma_{x1} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}\left(x_i - \overline{x_1}\right)^2}{M-1}},\ \text{and}\ \sigma_{x2} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}\left(x_j - \overline{x_2}\right)^2}{M-1}}, \tag{10}$$

Table 1 Measures of quality of fit for spatial regressions for different W matrices

Number of neighbors in W	Speed correlation ρ_v	WLS residual correlation ρ_r	SAR adj. R^2	SAR log lik.	SEM adj. R^2	SEM log lik.
WLS result						
0	na	na	0.6569	na	0.6569	na
Euclidean nearest neighbor matrix						
1	0.59	0.25	0.6626	−22,746	0.6744	−22,836
2	0.52	0.22	0.6606	−22,704	0.6816	−22,762
3	0.48	0.20	0.6603	−22,678	0.6846	−22,724
4	0.46	0.19	0.6606	−22,661	0.6860	−22,705
5	0.44	0.17	0.6614	−22,661	0.6857	−22,703
6	0.42	0.16	0.6616	−22,666	0.6849	−22,710
7	0.41	0.15	0.6615	−22,669	0.6853	−22,702
8	0.39	0.14	0.6626	−22,672	0.6840	−22,716
9	0.38	0.14	0.6627	−22,669	0.6845	−22,709
10	0.37	0.13	0.6626	−22,669	0.6845	−22,707
11	0.36	0.13	0.6625	−22,677	0.6837	−22,715
12	0.35	0.12	0.6624	−22,689	0.6831	−22,725
13	0.35	0.11	0.6624	−22,702	0.6817	−22,739
14	0.34	0.11	0.6623	−22,713	0.6809	−22,745
15	0.33	0.10	0.6624	−22,727	0.6797	−22,759
16	0.32	0.10	0.6622	−22,738	0.6787	−22,768
Network matrix: average number of neighbors within the ((*n*) nearest intersections)						
2 (1)	0.56	0.23	0.6673	−22,602	0.6834	−22,748
6 (2)	0.46	0.15	0.6710	−22,531	0.6882	−22,686
12 (3)	0.40	0.11	0.6711	−22,547	0.6911	−22,656
21 (4)	0.35	0.09	0.6714	−22,589	0.6915	−22,645
31 (5)	0.32	0.07	0.6712	−22,634	0.6881	−22,672
45 (6)	0.28	0.05	0.6702	−22,679	0.6855	−22,698
60 (7)	0.25	0.04	0.6689	−22,727	0.6813	−22,739
78 (8)	0.23	0.04	0.6676	−22,764	0.6773	−22,780

where N is the number of observations, and M is the number of nonzero entries in the $N \times N$ adjacency matrix.

The adjusted R^2 as well as log-likelihoods of all the spatial models are higher than for the WLS, indicating that the SEM and the SAR models fit the data slightly better than WLS. However the spatial coefficients, ρ or λ, of all the spatial models are highly significant, meaning that the WLS results are biased and inconsistent due to the uncorrected spatial correlations.

The best models chosen for illustration purposes are based on the statistics in Table 1. Because the iterative solution to the spatial regression maximizes the log-likelihood, the models with the highest log-likelihood are chosen as best fits.

Though the highest residual correlation occurs as expected between nearest neighbors (e.g. Tobler 1970), the best fits are usually achieved with more neighbors than one. Also, the network distance W matrices fit the data better than the Euclidean nearest neighborhoods.

In the SEM model, one would exclude the density of Urban Distributor roads from the regression on the basis of its t-statistic, leaving all road densities with positive and significant coefficients which are slightly smaller than in the WLS (see Table 2). The other coefficients also change only slightly relative to the WLS.

The coefficient λ shifts explanatory power from structure variables to the neighborhood context of the link and resolves the problems of residual correlations that confound inference. The best-fit spatial error models result by using either the seven nearest Euclidean neighbors, or a network distance of four intersections (on average, 21 neighbors). This indicates that persistence in speed variations is stronger along the network paths than across space. λ is 0.63 with the network neighborhood and 0.37 with the Euclidean-distance based neighborhood (see Table 2), meaning that random error correlations in the network neighborhood contribute nearly twice as much to speeds as in a spatial neighborhood on the Euclidean plane. The dummy variables on trunk roads with the network neighborhood is 20% smaller relative to the Euclidean distance model. Evidently, the unobserved characteristics of the spatial and network neighbors are quite different for this class of road, resulting in different effects on the regression errors.

The SAR model corrects for the spatial autocorrelation of the speeds. Though the autocorrelation parameters were significant for all neighborhood matrices tested in Table 1, like the WLS, the residuals remain correlated. The parameter estimates may thus be incorrect and are not shown in Table 2 for this reason. The best fit is obtained by using the four nearest Euclidean neighbors or a network distance of two intersections (on average six neighbors). The autoregressive parameters, ρ, are similar whether the Euclidean or network neighborhood is used, though statistically distinct (0.30 and 0.25). Both SAR models result in qualitatively similar differences in the fitted parameters relative to the WLS, which are also reflected in the SAC results.

The general spatial regression (SAC) requires the use of two neighborhood matrices: one for spatial autoregression and one for correlated spatial errors. It is not certain that the best SAR model will result in the best SAR–SAC model with the addition of a spatial error correction term. Therefore, models were estimated using combinations of network and Euclidean neighborhood matrices. An additional hybrid model using the logical combination of Euclidean neighbors for spatial error effects and network neighbors for speed autocorrelations was also estimated. The log-likelihoods are shown in Fig. 4.

The highest log-likelihoods are found with similar neighborhood matrices in the three model types. Using either network or Euclidean matrices improves the fit in the SAC model beyond the underlying SAR or SEM models, and the hybrid model results are similar to the network neighborhood results (see Table 2). The best fits are obtained with the 11 and 3 nearest Euclidean neighbors, the nearest network neighbors within three and one intersections (12.5 and 2 neighbors, on average), and with three intersections (12.5 neighbors) for the autoregression and four spatial

Table 2 Estimated model parameters for the subset of link speeds in Zurich (N = 9,297)

Model	WLS		SEM Eucl.		SAC Eucl.		SEM Net.		SAC Net.	
Variable	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t
HighwaysPk	148.1	24.2	150.3	21.8	118.5	20.4	150.2	15.7	117.3	39.3
HighwaysShoulder	137.4	22.4	139.8	20.3	109.9	18.7	139.1	14.6	109.1	30.1
HighwaysOffPk	146.3	24.0	148.3	21.7	117.7	20.3	147.6	15.6	116.1	37.3
HighwaysSa	148.4	23.8	149.0	21.1	117.3	19.9	148.4	15.2	117.0	39.5
Trunk roadsPk	119.7	36.6	123.3	30.0	86.3	37.9	112.6	16.1	84.5	7.1
Trunk roadsShoulder	113.5	34.6	117.3	28.5	81.8	34.1	107.1	15.3	80.4	17.5
Trunk roadsOffPk	115.5	35.4	119.1	29.1	83.2	35.6	108.7	15.6	81.6	17.2
Trunk roadsSa	118.8	35.6	122.7	29.2	85.9	35.4	113.0	15.6	83.9	17.3
Collector roadsPk	110.4	22.9	114.1	20.3	88.8	18.4	108.1	16.8	81.5	45.4
Collector roadsShoulder	104.1	21.5	107.4	19.0	83.8	17.2	101.9	15.8	77.2	30.9
Collector roadsOffPk	107.1	22.4	110.2	19.8	85.9	17.9	104.5	16.4	79.0	38.0
Collector roadsSa	105.3	21.4	111.0	19.4	85.7	17.3	105.5	15.9	78.0	32.0
Distributor roadsPk	119.0	13.4	113.5	11.8	82.3	9.7	108.9	9.8	86.4	12.9
Distributor roadsShoulder	114.1	12.9	109.3	11.4	79.1	9.3	104.7	9.5	83.8	12.2
Distributor roadsOffPk	115.4	13.1	110.9	11.6	80.6	9.5	106.1	9.6	84.5	12.4
Distributor roadsSa	119.6	13.2	113.8	11.5	82.8	9.5	109.1	9.6	86.9	12.6
Other roadsPk	75.5	10.6	73.7	10.3	54.6	7.8	75.0	10.4	52.7	9.1
Other roadsShoulder	72.1	10.2	70.0	9.9	51.3	7.4	71.8	10.0	50.6	8.6
Other roadsOffPk	74.1	10.6	72.0	10.3	53.6	7.9	72.9	10.3	52.1	9.1
Other roadsSa	77.5	10.2	76.8	10.1	57.4	7.8	79.3	10.3	55.7	8.8
Highways* Highway ramps, r = 1 km	−2.2	−7.5	−1.7	−5.3	−2.2	−7.6	−1.7	−5.2	−1.7	−5.8
Highways* LN(Jobs, r = 5 km)	−9.8	−6.9	−10.9	−6.9	−6.9	−4.8	−11.8	−5.4	−8.7	−6.1
Trunk roads* LN(Jobs, r = 1 km)	−7.0	−15.0	−5.9	−11.0	−4.6	−9.4	−6.0	−10.8	−4.4	−12.4

Table 2 continued

Model	WLS		SEM Eucl.		SAC Eucl.		SEM Net.		SAC Net.	
Variable	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t
Trunk roads* LN(Pop, r = 5 km)	−5.1	−8.5	−6.3	−8.8	−4.0	−6.4	−4.7	−4.3	−4.2	−7.7
Collector roads* LN(Jobs, r = 1 km)	−3.0	−3.9	−3.2	−4.2	−2.3	−3.2	−3.9	−5.3	−3.5	−4.9
Collector roads* LN(Pop, r = 3 km)	−6.1	−6.3	−6.4	−6.1	−5.5	−5.5	−5.1	−4.5	−3.8	−4.3
Distributor roads* LN(Jobs, r = 1 km)	−4.8	−5.1	−5.3	−5.2	−3.6	−3.8	−5.9	−5.8	−3.7	−4.0
Distributor roads* LN(Pop, r = 5 km)	−7.3	−5.2	−6.2	−4.0	−4.7	−3.3	−4.9	−2.9	−5.3	−3.9
Other roads * LN(Jobs, r = 1 km)	−7.1	−5.3	−6.6	−4.9	−5.2	−3.9	−6.7	−5.0	−5.3	−4.2
Density highways (m/m^2)	484.5	12.6	416.0	10.2	374.4	9.6	259.4	6.0	301.3	9.1
Density trunk roads (m/m^2)	165.4	6.0	159.4	5.3	135.5	4.8	195.7	6.5	136.1	5.1
Density ramps (m/m^2)	240.7	5.8	229.1	5.4	125.5	3.1	143.0	3.5	125.4	3.4
Density urban collector roads (m/m^2)	2.7	**0.1	26.4	**1.1	29.7	**1.3	57.2	2.5	41.5	*1.9
Den. urban distributor roads (m/m^2)	−44.2	−2.3	−20.1	**1.0	−8.2	**0.4	−0.6	**0.0	−9.3	**0.5
ρ	–		–	32.3	0.24	21.0	–	26.0	0.30	7.0
λ	–		0.37	–	0.17	19.9	0.63	15.7	0.14	31.9
Adjusted R^2	0.657		0.685		0.691		0.692		0.702	
Log-likelihood (×10^4)	–		−2.270		−1.726		−2.644		−1.712	

Probability of rejecting H_0: * means $5\% \leq P < 10\%$; ** means $P \geq 10\%$; others: $P < 5\%$

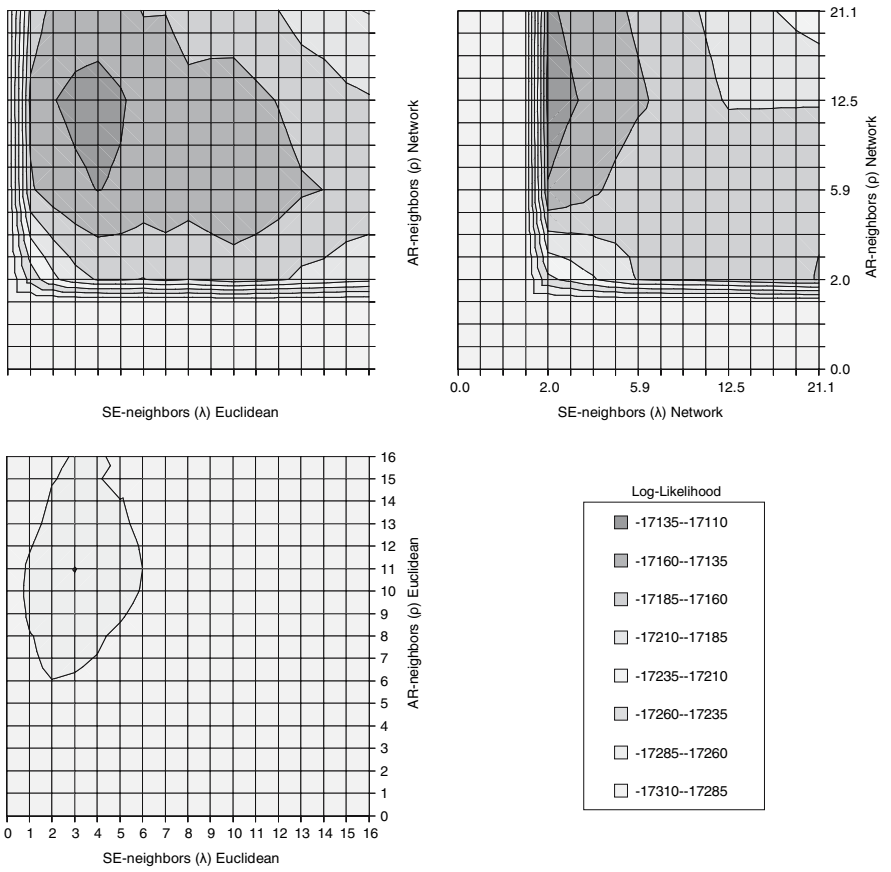

Fig. 4 Contours of the log-likelihood surface of the SAC model for different link neighborhood matrices: axes are the number of neighbors in the autoregressive (SAR) versus spatial error (SEM) matrix. Note the nonlinear scale for network neighborhoods

neighbors for spatial error in the hybrid model. The spatial error and autocorrelation coefficients of the models are highly significant.

All formulations of the SAC use more neighbors for autocorrelations and fewer for the spatial error correlation. The influence of autocorrelation is approximately double that of the spatial residual correlation for the network and hybrid models and 40% larger in the Euclidean model. These models therefore emphasize the persistence of speeds in traffic flows more than unobserved spatial influences. The hybrid model has very similar fit statistics and coefficients to the network model, an indication that (network) autocorrelation is the dominant process and that residual spatial correlations can be treated with either neighborhood matrix.

The SAC parameter estimates are rather similar to the SAR estimates. The link type and time dummies are much lower than for WLS and SEM. The difference is made up by the contribution of the speed on neighboring links. All contributions from urban roads are found to be insignificant. The parameters of spatial structure

indicators and road densities have less influence on speed when autocorrelation is explicitly modeled, meaning that autocorrelation explains spatial variations across road type.

Both SAC and SEM correct the problem of correlated residuals, but they offer different explanations of the processes causing spatial speed variations. The SAC, however, results in better fit statistics. Inferences made without accounting for spatial correlation would overemphasize the importance of structure variables and even ascribe significance to variables that have no explanatory power when uncorrelated from neighborhood effects.

7 Validation

The model results are validated by predicting speeds for the roads in the Winterthur sample using the parameter estimates for the network-based neighborhood models on the Zurich sample from Table 2 and comparing them to the withheld measurements. Because of the complicated multiple dimensions of spatial speed data, tests of "reasonableness" are more useful than finely tuned statistical hypotheses (U.S. Department of Transportation Federal Highway Administration 1997). At this stage, the models' agreement with observations is evaluated based on aggregate and qualitative tests.

The aggregate statistics for the predictions from the best WLS, SEM and SAC in Table 3 and the histogram in Fig. 5 show above all high consistency with one another. In transferring the Zurich model to the Winterthur area, all three model formulations tend to predict speeds for the non-fitted (holdout) sample that are too high $(\Delta\bar{\hat{v}} = \bar{\hat{v}} - \bar{v})$, though none of the differences are statistically significant. An average may not be a good measure of comparison because the speeds are distributed with positive skew and the estimated speed distribution is even bimodal. The SEM formulation has the lowest mean residual, while the SAC model seems to be slightly better at producing extreme values (see Fig. 5). To assess precision, the

Table 3 Characterization of model results based on the best network neighborhood neighborhood matrices (all units km/h)

Dataset	Model	$\bar{\hat{v}}$	$\Delta\bar{\hat{v}}$	$\sigma_{\hat{v}}$	SEP	SDR	$\hat{v}_{\min}$	$\hat{v}_{\max}$
Fitted dataset	WLS	53.9	0.0	22.3	30.4	20.5	23.2	126.0
N = 9,297	SEM	54.2	0.4	18.7	28.1	21.0	28.9	114.2
Zurich	SAC	54.6	0.7	22.8	30.4	20.1	8.4	127.6
	OBS	53.9	–	30.3	–	–	1.1	172.8
Holdout dataset	WLS	66.2	3.2	28.3	35.0	20.6	22.6	126.0
N = 1,209	SEM	64.8	1.9	24.0	31.8	20.8	25.1	122.2
Winterthur	SAC	67.5	4.6	29.4	36.1	20.9	22.4	132.5
	OBS	63.0	–	33.5	–	–	1.2	165.6

SEP Standard error of prediction, *SDR* standard deviation of the residuals, *OBS* is the set of observations corresponding to the link speeds estimated by each model, "–" means that the statistic is relevant to the model results, but not the set of observations

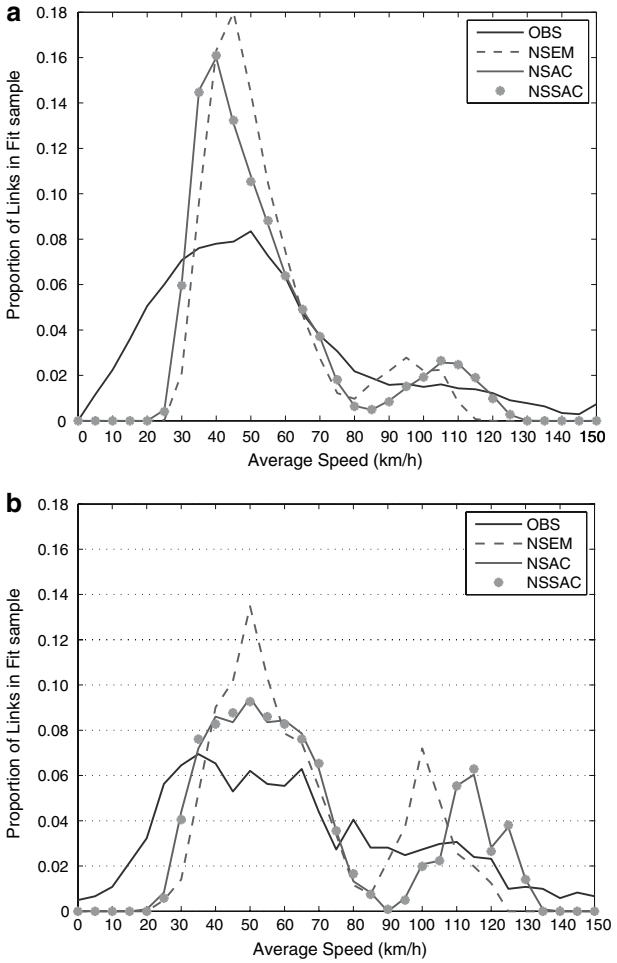

Fig. 5 Histograms of link speed estimates versus observations for all time periods (model results are for measured links only). *NSEM* network spatial error model, *NSAC* network spatial autocorrelation and error model; *NNSAC* hybrid model, *OBS* observations: **a** fitted sample (Zurich), N = 9,297, and **b** holdout sample (Winterthur), N = 1,209

standard error of prediction (SEP, Eq. 11) is the proper gauge of the model's ability to predict speed on a given link, by accounting for variance in the explanatory variables of the sample (National Institute of Standards and Technology NIST 2006).

$$SEP = \sqrt{\sigma_{\hat{y}}^2 + SDR^2}. \tag{11}$$

The residuals of the WLS, SEM, and SAR predictions of speeds in Zurich and in Winterthur were analyzed for systematic bias with respect to categories of travel period, road type, combined travel period and road type, road densities (by type of

road), and spatially by categorized values of the regional structure variables and spatial plots of standardized residuals. The categories of time and road type are defined above; the other categories were chosen to have equal widths. Space does not permit reporting on detailed summaries of the analysis, but while the absolute value of the mean of the residuals in certain cases exceeds 10 km/h, the mean residuals are insignificantly different from zero in nearly all categories. Thus, there is no systematic indication of circumstances in which the models perform better or worse than in other circumstances. Skill score comparisons of speed estimates against observations by road type and time of day, compared with static assignment, favor the regression for highways and show no difference between the methods for other road types.

The majority of the estimated link speeds in the three models are within ±10 km/h of the observations. When larger differences occur, they generally have the same sign for all models on the same links, indicating a problem of missing variables rather than in the treatment of correlation structure. Detailed quantitative spatial or network topological analysis of the results has not been carried out, however two observations of spatial residuals can be made: first, speed over- (under-) estimates occur more frequently in areas of less (more) dense development, and second, the treatment of spatial correlation does not change the qualitative spatial distribution of speed over- or underestimates.

8 Conclusions and outlook

This article reports on an approach to estimate link speeds employing both structural variables and the network context, with correction for the spatial error and autocorrelation terms. The method is related to direct demand models and is intended to provide easier, more scalable estimates of speed on all links than more sophisticated efforts. The link-hectare dataset was assembled and the first regressions were estimated by a student during a summer internship. The calculation of network neighborhoods required a program (written in Java) which generates the eight matrices by searching down branches of the network for each measured link. The estimation of three SAC models (see Fig. 4) was an investigative undertaking performed by brute force with no effort to optimize calculation. Estimating models with 16×16, 8×8, and 8×16 combinations of neighborhood matrices in Matlab required 18 days (Pentium 4, 2.4 GHz). The simple features of the resulting log-likelihood surfaces show that a simple implementation of a gradient search would probably have given a global optimum solution in all three cases in much less time. Finally, applying an estimated model to systemwide link speeds using regression parameters, including scenarios based on changed structure variables, takes only a few seconds.

The validation with a large hold-out sample shows that the carefully implemented approach produces an acceptable fit of the mean, with weakness in predicting very low or very high speeds. Low transfer error is a result of fairly consistent relationships between spatial structure and road speeds, which means that

application of the model across the whole Cantonal network is plausible, including non-measured links, with the appropriate neighborhood matrices.

Estimating the range of spatial models reveals that there are substantial spatial correlations which need to be accounted for. A simple linear regression is not appropriate and is likely to bias the conclusions. Spatial autocorrelation and spatial error correlation models are to some extent substitutes in terms of improving model fit, but they assume different understanding of the underlying processes, which is reflected in the parameters and the speed estimates. Indeed the autocorrelation model itself exhibits residual spatial error correlations which must be treated. A network neighborhood explains speeds better than a Euclidean spatial neighborhood, and an intuitive hybrid model using both types of matrix, yields similar results to the network neighborhood model.

While the time period and road type interactions did not reveal any surprises, the different values estimated for the different model formulations highlight the need to be careful in the interpretation of spatial regressions for policy making. Likewise, the new results on the impacts of the structural variables show that they must be taken into account in order to understand variation in local speeds. While the macroscopic speed/space relationships yielded by this regression model are statistically significant and spatially detailed in a way useful for system monitoring, in many applications the model results cannot substitute for the explicit representation of dynamic OD matrices, agent behavior, or land use-transportation system coupling. In these cases, the spatial regression results may be useful for initial values or baselines of comparison in conjunction with the traditional approaches.

Acknowledgments The authors thank Thomas Niederöst of Zurich's planning office and Jean Wolf of GeoStats for support in gathering GPS data, and James LeSage for help in applying the Econometrics Toolbox for Matlab to this task. The remarks of anonymous reviewers greatly improved the clarity and helped place the work in its most relevant context.

References

Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht

Balmer M, Bernard M, Axhausen KW (2005) Matching geo-coded graphs. Swiss Transport Research Conference, Swiss Federal Institute of Technology, Ascona

Bernard M, Hackney J, Axhausen KW (2006) Correlation of link travel speeds. Swiss Transport Research Conference, Swiss Federal Institute of Technology, Ascona

Bhat C, Guo J (2004) A mixed spatially correlated logit model formulation and application to residential choice modeling. Transportation Res B 38:147–168

Bivand R (1998) A review of statistical techniques for location studies. CEPR Symposium on New Issues in Trade and Location (2277), Norwegian School of Economics and Business Administration, Lund, Sweden

Black WR, Thomas I (1998) Accidents on Belgium's motorways: a network autocorrelation analysis. J Trans Geogr 6:23–31

Bolduc D, Dagenais MG, Gaudry MJI (1989) Spatially autocorrelated errors in origin-destination models: a new specification applied to aggregate mode choice. Trans Res B 23:345–359

Bolduc D, Laferrière R, Santarossa G (1992) Spatial autoregressive error components in travel flow models. Reg Sci Urban Econ 22:371–385

Bradley M, Bowman J (2006) A summary of design features of activity-based micro simulation models for U.S. MPOs. TRB Conference on Innovations in Travel Demand Modelling, Austin

Griffith DA (1996) Practical handbook of spatial statistics. CRC, Boca Raton

Guo JY, Bhat C (2007) Operationalizing the concept of neighborhood: application to residential location choice analysis. J Trans Geogr 15:31–45
Hunt JD, Kriger DS, Miller EJ (2005) Current operational urban land-use-transport modelling frameworks: a review. Trans Rev 25:329–276
Kim S-E, Niemeier D (2001) A weighted autoregressive model to improve mobile emissions estimates for locations with spatial dependence. Trans Sci 35:413–424
LeSage JP (2000) Spatial econometrics. Working Paper, University of Toledo, Toledo
LeSage JP (2005) Applied econometrics using Matlab. Working Paper, University of Toledo, Toledo
Maddala GS (2001) Introduction to econometrics, 3rd edn. Wiley, West Sussex
Marchal F, Hackney J, Axhausen KW (2006) Efficient map-matching of large GPS data sets—tests on a speed monitoring experiment in Zurich. Trans Res Rec 1935:93–100
Miller HJ (1999) Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T). Geogr Anal 31:373–399
Nagel K, Esser J, Rickert M (2000) Large scale traffic simulations for transportation planning. Annual review of computational physics. World Scientific, Singapore
Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam modelling. Oper Res 51:681–710
National Institute of Standards and Technology NIST (2006) e-Handbook of statistical methods. Retrieved 30 January 2006, from http://www.itl.nist.gov/div898/handbook/
Navteq Corporation (2004) High resolution digital road network of Canton Zurich. Frankfurt
Okabe A, Okunuki K, Shiode S (2006) SANET: a toolbox for spatial analysis on a network. Geogr Anal 38:59–66
Páez A, Scott D, Potoglou D, Kanaroglou P, Newbold KB (2007) Elderly mobility: demographic and spatial analysis of trip making in the Hamilton CMA, Canada. Urban Stud 55:123–146
Páez A, Scott DM (2004) Spatial statistics for urban analysis: a review of techniques with examples. Geojournal 61:53–67
Páez A, Suzuki J (2001) Transportation impacts on land use change: an assessment considering neighborhood effects. J East Asia Soc Trans Stud 4:47–59
Peeta S, Ziliaskopoulous AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. Netw Spat Econ 1:233–265
Public Works Office of the Canton Zurich, P. u. S. (2002) Der Einsatz des Kantonalen Verkehsmodells im Rahmen der Zweckmässigkeitsbeurteilungen, Synthesebericht (The evaluation of the Cantonal transportation model, Summary Report). Zurich, Baudirektion Kanton Zurich
Salvini PA, Miller EJ (2005) ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. Netw Spat Econ 5:217–234
Steenberghen T, Dufays T, Thomas I, Flahaut B (2004) Intra-urban location and clustering of road accidents using GIS: a Belgian example. Geogr Inf Sci 18:169–181
Stetzer F (1982) Specifying weights in spatial forecasting models: the results of some experiments. Environ Plan A 14:571–584
Swiss Federal Statistical Office (2001) Travel behaviour results from the 2000 Microcensus travel. Mobility in Switzerland. Swiss Federal Office for Spatial Development and Swiss Federal Statistical Office. Berne, SFSO and ARE
Tobler WR (1970) A computer model simulating urban growth in the Detroit region. Econ Geogr 46:234–240
U.S. Department of Transportation Federal Highway Administration (1997) Model validation and reasonableness checking manual. Travel Improvement Program, U.S. Department of Transportation, Federal Highway Administration
Waddell P, Ševcíková H, Socha D, Miller EJ, Nagel K (2005) Opus: an open platform for urban simulation. Computers in Urban Planning and Urban Management Conference (CUPUM), London
Wegener M (ed) (2004) Overview of land use transport models. Handbook of transport geography and spatial systems. Elsevier, Oxford
Zhao H, Bhat C (2002) The spatial analysis of activity stop generation. Trans Res B 36:557–575
Zhou B, Kockelman KM (2005) Neighborhood impacts on land use change: a multinomial logit model of spatial relationships. Paper presented at the 52nd North American Regional Science Association International (RSAI) Conference, Las Vegas, Nevada, (2005)