Correlation of link travel speeds

Michael Bernard, IVT, ETH Zürich
Jeremy Hackney, IVT, ETH Zürich
Kay W. Axhausen, IVT, ETH Zürich

Conference paper STRC 2006

STRC

6th Swiss Transport Research Conference
Monte Verità / Ascona, 15th-17th March 2006

Correlation of link travel speeds

Michael Bernard
IVT
ETH-Hönggerberg HIL F12.1
CH-8093 Zürich

Phone: +41 (0) 1 633 66 94
Fax: +41 (0) 1 633 10 57
email:
Bernard@ivt.baug.ethz.ch

Jeremy Hackney
IVT
ETH-Hönggerberg HIL F 51.3
CH-8093 Zürich

Phone: +41 (0) 1 633 33 25
Fax: +41 (0) 1 633 10 57
email:
Hackney@ivt.baug.ethz.ch

Kay W. Axhausen
IVT
ETH-Hönggerberg HIL F 32.3
CH-8093 Zürich

Phone: +41 (0) 1 633 39 43
Fax: +41 (0) 1 633 10 57
email:
Axhausen@ivt.baug.ethz.ch

March 2006

Abstract

Modern design concepts apply probability density functions to model the traffic flow and capacity. In the application travel times are calculated by drawing random realisations or by estimating an expected value. Generally it is assumed that the distributions of different infrastructure elements are independent and not correlated. In literature this property is so far not often tested.

During a project in November 2003 at IVT data was collected from floating cars using a device to record GPS-data on approximately 30,000 kilometres within the canton of Zurich. These floating car data (FCD) was aggregated to gain link travel times on the network of that area. In this paper the correlation of the link travel times or more generally the experienced travel speed is computed, grouped within different road types. This can be done by matching the GPS data to a network model.

Different network dependencies are investigated by calculating the correlation for travel speeds measured on paths for given numbers of intersections and distances. It can be seen that the correlation is high for pairs of speeds which are near to each other and decreasing with an increase in the distance.

The FCD of three vehicles over many days is combined by establishing a realistic space-time envelope based on theoretically possible routes to the position of other measurements. From each link with an aggregated travel time the FCD the shortest path (based on the travel time) to each later link information was calculated. If these two measurements on the links could have been produced by a car travelling that route at the time this link pair was also used for the calculation of the correlation. With these theoretical routes sufficient data pairs can be compared to get significant values for the correlation.

Keywords

Correlation – link travel speed – floating car data (FCD) –Swiss Transport Research Conference – STRC 2006 – Monte Verità

Citation

Bernard, M., J. Hackney and K. W. Axhausen (2006) Correlation of link travel speeds, 6th Swiss Transport Research Conference, Ascona.

1. Introduction

Transport engineering is dealing to large extends with travel times or more generally with travel speeds. Speaking of individual transport commonly static assignment models are applied using zone-based origin-destination matrices of the demand. Commonly used network assignment programs assume independency of the travel speeds between different links, as each network element is handled separately. The intention of this work is to test with speeds from floating car data (FCD) if the imposed independency is given. Generally speaking independence is given if the correlation between the speeds of neighboured links is (nearly) zero.

2. Data description of link speed data

The data describing the link speeds is generated from GPS measurements of floating cars in the Canton of Zurich. The data was collected from 3-week survey in November 2003 conducted by IVT, ETH Zurich. The chosen circuits were of approximately 500 km length with same starting and end points where the journeys were assembled of shortest paths of a random set of 50 traffic planning zones applying a Monte Carlo technique (see Hackney *et al*., 2004). From this survey data from 18 days were available containing 2.5 Mio seconds of GPS measurements with coverage of roughly 33,000 km.

The set of measured points of a directed link were aggregated to an average link speed on that link (Hackney and Axhausen, 2005).

The measurements were separated by following time periods:

- Peak hours: 6:30 - 8:30 and 16:30 - 18:30 (week days)
- Off peak: 8.30 - 16:30 and 18:30 - 20:30 (week days)
- Shoulder: 6:00 - 6:30 and 20:30 - 21:00 (week days)
- Saturday: 6:00 - 21:00

3. Correlation of link speeds on network

For this type of analysis it is obvious, that network information must be used to determine the contiguity instead of Euclidean distances. Of course it may be true for some cases that the travel speeds of two parallel roads are highly correlated due to a high demand for a origin-destination pair (OD pair), but in this study no information of OD-matrices should be used to assess the correlation. That is, if these two roads in this example are not connected to each other e. g. by a bypass near the points of the actual measurements one must assume that the traffic volume as a proxy for the travel speed on the first road cannot directly influence the volume on the second one and therefore the measured travel speed. As the drivers using one of these two roads cannot interact, these two roads are not regarded as near neighbours, even if they are located close to each other.

3.1 Correlation

To calculate the correlation of links speeds the common expression for correlation is used:

$$\rho = \frac{1}{N} \frac{\sum_{i=1}^{N} (v_{1i} - \overline{v_1})(v_{2i} - \overline{v_2})}{\sigma_{v1}\sigma_{v2}}$$

where ρ is the correlation coefficient, v_{1i} is the measured speed on link of group 1, v_{2i} the speed on link of group 2, $\bar{v}$ (for group 1 and 2) is the mean speed of that group and σ_v (group 1 and 2) is the standard error within the corresponding group, and N indicates the total number of pairs. The identifiers (vehicle number, day, and timestamp) of v_1 and v_2 must be different.

The groups of pairs (1 and 2) were built using constraints for the time and the location of the measurement. The main constraints for all groups are:

- $\text{timestamp}_1 \leq \text{timestamp}_2$ (e. g. 8:01.30)
- $\text{day type}_1 = \text{day type}_2$ (weekday or weekend)
- $\text{time window}_1 = \text{time window}_2$ (peak hour, shoulder, off peak)

The setups tested introduce additional constraints to the basic constraints shown above. These constraints group the travel speed measurements into distance classes – or more general resistance classes – that describe the resistance on the time-shortest path from a measurement on link 1 to another on e on link 2, which should be compared to the first one.

By defining these classes the correlation can be tested over network distance to prove the intuitive assumption of decreasing correlation when looking at larger distances, as the influence is supposed to be higher on nearby links.

3.2 Path setup

Each class is defined by 500 m steps resulting in following classes:

$class_i$: $(500 \bullet i)+1 \leq dist_{1->2} \leq 500 \bullet (i+1)$, $i=0, 1, \ldots I$

e. g.:

- $class_1$: $1\text{ m} \leq dist_{1->2} \leq 500\text{ m}$
- $class_2$: $501\text{ m} \leq dist_{1->2} \leq 1000\text{ m}$

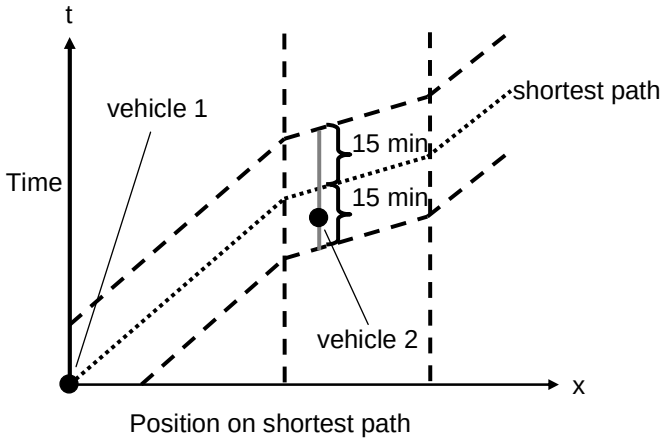
…

The first setup, the "path analysis" is looking at the driven path of the vehicles where the data is enriched by virtual paths that a vehicle might have taken. As measurements are available for several days when looking at weekdays, which are handled as one day type, and mostly three cars were driving at the time a theoretical path could be found from the position of the first measurement to the second one. For these paths it has to be verified, if the timings could have happened in reality. To test this an additional constraint is introduced:

- $timestamp_1 + tt_{1->2} - 15\text{ min} \leq timestamp_2 \leq timestamp_1 + tt_{1->2} + 15\text{ min}$

where $tt_{1->2}$ denotes the estimated travel time on the time-shortest path from $link_1$ to $link_2$ based on the average travel speed of the links given by the Canton's road network model KVM98 (Jenni and Gottardi AG, 1998). A graphical interpretation of this constraint can be seen in Figure 1. The points labelled with vehicle 1 and vehicle 2 represent two link speed measurement at a given location (x) and certain time (t). The dotted line marks the shortest path from link 1 (position of vehicle 1) to link 2 based on average link speeds. The measurement of vehicle 2 is set to be valid, if the position in space and time is within the ±15 min window delimited by the dashed lines. Measurements outside this corridor are not regarded for the calculation of correlation, as these vehicles are not likely to be on a path that could have been driven in reality.

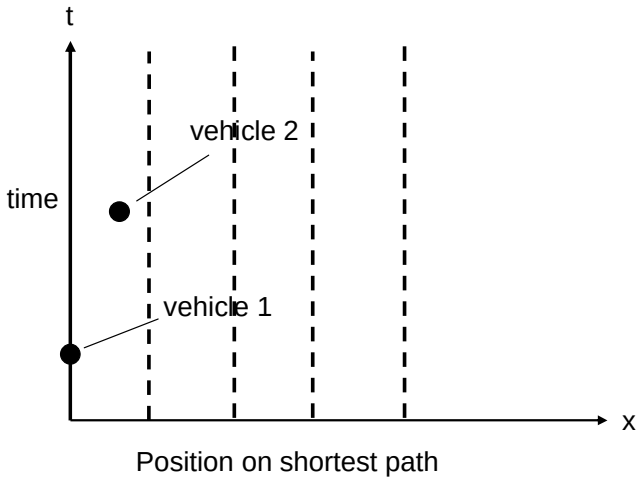
Figure 1 Path constraint



3.3 Snapshot setup

In contrast to the path setup above the snapshot setup is simpler and more general. There is no additional time-dependent constraint apart from the general constraints. That is, all vehicles driving during a given day type (weekday) and time window (peak hour, shoulder or off peak) are regarded for the calculation of the coefficient of correlation. Figure 2 shows two vehicles that are in one group. The only restrictions are given by the general constraints. As in the path setup vehicle 1, which is by definition the earlier vehicle, defines the origin for computing the distance from point 1 to point 2. Contrasting from Figure 1 no additional time constraint is given and therefore the time-gap between vehicle 1 and vehicle 2 could become any value, as long as the general constraints – especially the time window – are not violated.

Figure 2 Snapshot constraint



3.4 Alternative path setup

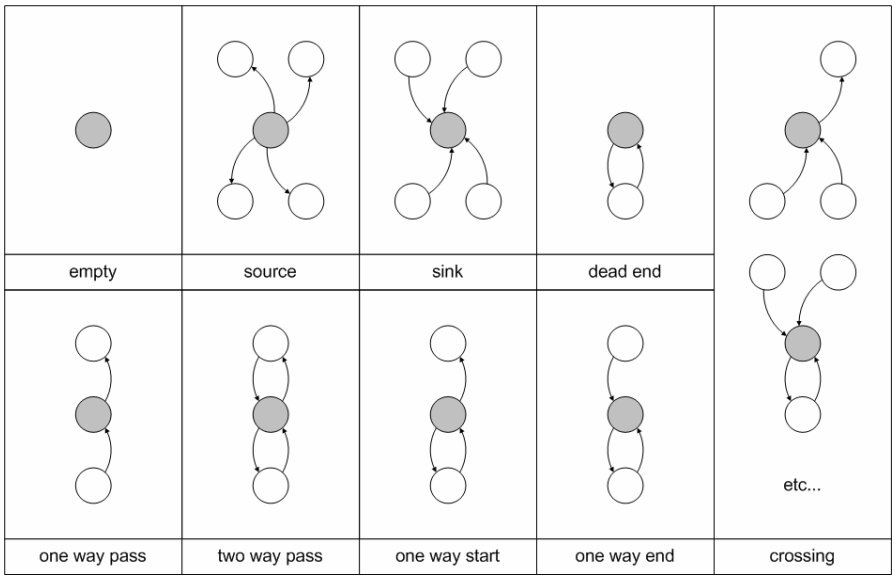
As a third setup, an alternative path setup was defined. This setup eliminates a shortcoming of the path setup of section 3.2 based on distance classes. As the density of intersections is higher for road-types of lower priority in comparison to those of higher priority, the disturbances resulting from changes of traffic volumes at the intersections is likely to influence the coefficient of correlation. With this setup the problems due to differences in the density of intersections are avoided by using the number of passed intersections to specify the distance classes. The constraints were kept equal to those of the path setup (see Figure 1).

Definition of Intersection

As street networks often contain nodes (vertices) to reproduce the run of the roads, the number of intersections on a path cannot be computed by counting the number of network-nodes along that path. A classification of the network nodes can be found in Balmer *et al.* (2005), see Figure 3. To identify an intersection (crossing) all other possible roles of a network node have to be ruled out. These types are listed in Figure 3: empty, source, sink, dead end, one way pass, two way pass, one way start, and one way end. A crossing is given, if the traffic volume can change at this node. For the calculation of the number of intersections

on a time-shortest path a Dijkstra algorithm was implemented and modified to return the number of passed intersections.

Figure 3 Classification of a network-node



Source: Balmer *et al.* (2005)

4. Correlation results

Generally speaking the assumption that the coefficient of correlation based on the travelling speeds on links is higher the closer the links are located using network paths can be validated. All three setup types show a decline in the coefficient of correlation with higher resistances (distances between two measurements on links). This decline appears to be non-linear.

It must be remarked that the number of pairs for resistance classes within short distances is lower than those of classes containing pairs with larger distances. The significance (95%) for all coefficients of correlation was calculated, where those of the major road types (motorways, trunk roads, and collector roads) were significant even for the short-distance classes. For the comparison of the time windows (peak hours, off peak and shoulder) one should notice that most measurements were collected during the off peak and least during the

shoulder. As a result of this for some resistance classes the coefficient of correlation was not significant to the desired level. In these cases the coefficient of correlation was not shown in the result figures (Figure 4 - Figure 6).

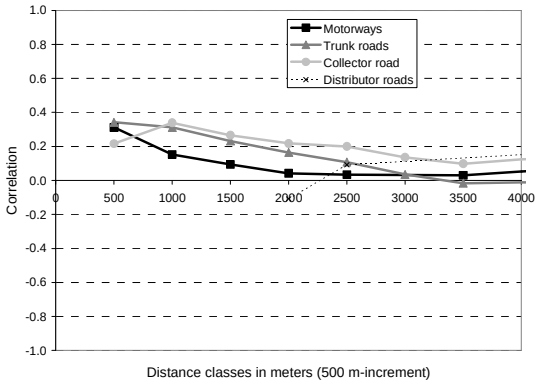
4.1 Path setup: distance

Using the path setup with distance classes of 500 m-increments an exponential decline in the coefficient of correlation could be seen for all road types and all time windows (Figure 4 a, b, and c).

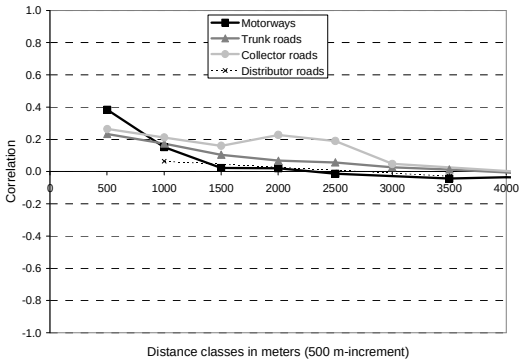
The focus will be on the roads with higher priority, the motorways and trunk roads. The measured speeds of the roads of less priority (collector roads and especially distributor roads) are not very reliable. From the graphs in Figure 4 one could say that the coefficient of correlation approaches zero near 2500 meters (distance: 2001-2500 m). This finding is important for simulations that assume independency of travel speeds. As long as the resolution of the network is low enough these effects may be neglected. But especially when high-resolution models are applied the correlation effects should be considered.

The slope of the coefficient of correlation of the collector roads during the peak hours (Figure 4 a) may be due the relative low number of observations, though being highly significant. But this slope could not be explained without further analysis of the network and the chosen circuit.

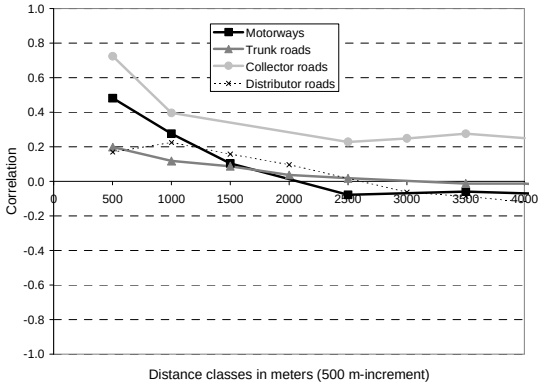
Figure 4 Path setup: correlation of speeds over distance classes (500 m-increment)



(a) Peak hours



(b) Shoulder



(c) Off peak

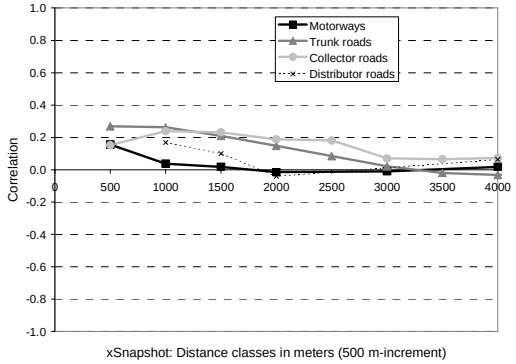
The form of the coefficient of correlation of the collector roads during the peak hours (Figure 4 a) may be due the relative low number of observations, though being highly significant. But this incline could not be explained without further analysis of the network and the chosen circuit.

4.2 Local snapshot: distance

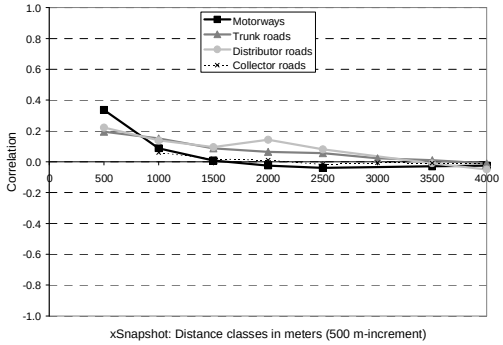
The local snapshot setup is performed to verify the results of the path setup using distance classes. The intention is to test the restriction imposed on the previous setup, that the timestamps of two vehicles must be within a 30-minute time-windows on the time-shortest path. Note that average network speeds were used to calculate the paths. These values may be wrong in some cases and could produce errors and therefore wrong coefficients of correlation. Comparing the results of Figure 5 with those of Figure 4, it is obvious that the shape and the level of the curves are qualitative equal. As it had to be expected the coefficients of correlation are lower. This is due to the lack of one restriction. The similarity of the curves is still given, for the measurements are still separated by peak hours, shoulder, and off peak. As

the reduction of the values is small, one can conclude that the restriction is not too restrictive and can be chosen.

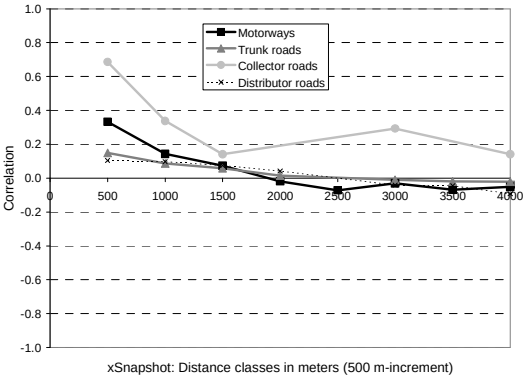
Figure 5 Snapshot setup: correlation of speeds over distance classes (500 m-increment)



(a) Peak hours



(b) Shoulder



(c) Off peak

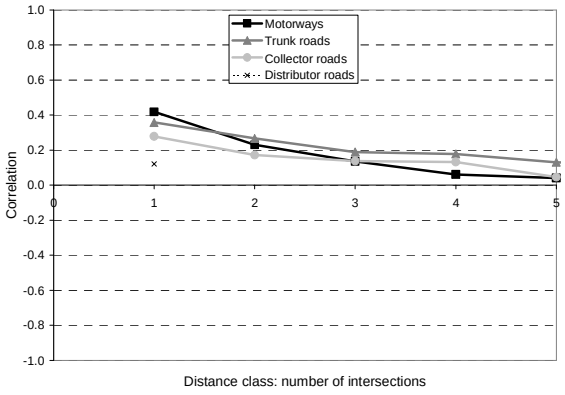
4.3 Path setup: intersections

In contrast to the path setup using distance classes the results of the path setup using the number of intersections as resistance classes seems to be more stable when looking at the shape of the curves. But the interpretation of the results is different to the one of the previous setups. With every intersection a change in traffic volumes and conditions may occur. This must not necessarily be the case for the distance setups.

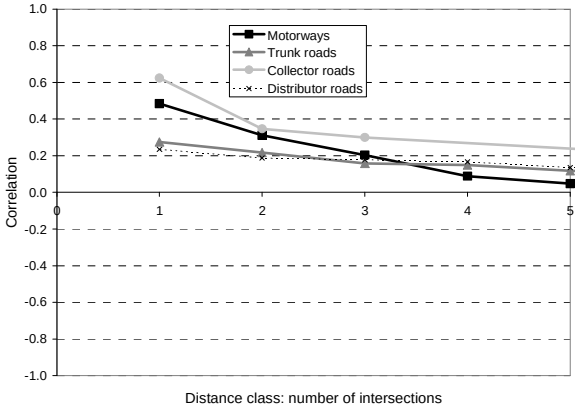
When looking at the coefficient of correlation of the travel speeds on motorways in Figure 6 (a, b, and c) one can see that the values are lowest during the peak hours, higher during the shoulder, and highest during the off peak. The previous setups (Figure 4 and Figure 5) do also show this behaviour. The interpretation of this finding is the strong diversification of traffic states during hours of higher traffic volumes, as the speeds were measured of roads of the whole canton of Zurich with urban and rural areas. The travel speeds in rural areas do not vary to the extent measured in urban areas but are roughly at the same level.

Figure 6 Path setup: correlation of speeds over number of intersections

(a) Peak hours



(b) Shoulder



(c) Off peak

5. Application: estimation of link travel speeds

The findings of the correlation of link travel speeds should be used in modelling approaches. One area of application are regression models to estimate link travel speeds.

Starting with an ordinary least squares model (OLS), information about spatial correlations can be integrated to correct the model. Exemplarily the spatial error model (SEM) should be used here, that includes a spatial weighting matrix W, which represents the contiguity of measurements. (For more information see Anselin, 1988 or LeSage, 1998).

The spatial error model (SEM) includes the spatial interdependence among observed variables by the spatially lagged error term, which is analogous to a correlated time series:

$$y = \beta X + u$$

$$u = \lambda W_e u + \varepsilon \quad , \qquad \varepsilon \sim N(0, \sigma)$$

with y: observations
β: parameter estimates
X: independent explanatory variables
W_e: weighting matrix

5.1 Explanatory variables

The dependent variables, the link travel times were used as described in Section 2. As explanatory variables are road network density, regional structure, and dummies for road type and time period are used. The variables are chosen for their qualitative value in explaining speeds, correlation with speed, correlation with each other, the improvement they bring to model fit, and parameter significance. (For more information see Hackney *et al.*, 2006).

The road density is a proxy for the local number of alternatives. In addition to that the number of highway access points was chose as an explanatory variable, which plays a similar qualitative role, though it is smoothed to de-emphasize effects over a distance. The population and employment opportunities were calculated over different smoothed kernel density radii.

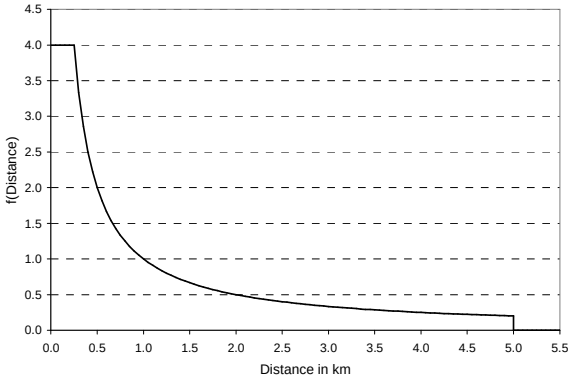
5.2 Constructing the weighting matrix W

The findings from Section 4 suggest using an exponentially decreasing function to weight the influence of neighbouring links. The option chosen here is a reciprocal of the distances of the following form:

$$d_{ij} = \begin{cases} 0 & if\ dist_{ij} = 0\ (same\ link) \\ 4 & if\ 0 \le dist_{ij} < 250\,m \\ 1/dist_{ij} & if\ 250\,m \le dist_{ij} \le 5\,km \\ 0 & if\ dist_{ij} > 5\,km \end{cases}$$

Figure 7 shows a graph of the shape of the function. The idea of this function is to weight close links (up to 250 m distance) equally, since the influence might not change very much in the very near region. Additionally, the chosen function (1/distance) requires a cap for low distance values to avoid extremely high weightings, which are approaching infinity for distances near zero. Higher values are less weighted until a distance of 5 kilometres. Having the graphs of Figure 4 in mind, where one can see that the coefficient of correlation approaches zero near 2.5 kilometres, an upper limit of 5 kilometres for the weighting radius seem sufficient.

Figure 7 Weighting function: $d_{ij} = f(distance_{ij})$



The values (d_{ij}) of the temporary NxN-weighting matrix D are standardized to get the final weighting matrix W with a row sum of 1:

$$w_{ij} = \frac{d_{ij}}{\sum_{j=1}^{N} d_{ij}} \quad \text{to gain:} \quad \sum_{j=1}^{N} w_{ij} = 1.$$

This shape of the weighting matrix W ensures that nearby links on the basis of network distances have a much higher impact on the local speeds than far off links, as these weights are decreasing with the reciprocal of the distance.

For the assembly of the matrix the model specification has to be regarded. For the estimation presented here different time periods (Saturday, peak, shoulder and off peak) are modelled within the same model, which are represented via dummy variables. As the measurements of different time periods cannot be compared to each other, there must not be an influence from neighbouring links where the speed is measured during different time periods. This omission of influences is modelled by marking the corresponding values w_{ij} in the weighting matrix with zeros, resulting in a W-matrix, which contains non-zeros for links being in the given distance range (0 to 5 km) and being driven during the same time period.

5.3 SEM-Model results

As the residuals of the travel speed calculated with an OLS were heteroscedastic (not of the same variance) the dependent and independent variables were weighted (divided) by the standard deviation of the residuals grouped by road types. The results of the weighted least squares estimation (WLS) can be seen in Table 1. The dummy variables for road types and

time periods are modelled as interactions (a * b), where the values are multiplied with each other. Since the dummy variables could only take the values zero or one, all interaction terms with the dummy variables can be interpreted as: "under condition of dummy variable being true". This is also the case for combinations of the dummy variables with scalar values like the number of highway accesses, logarithm of employment opportunities, or the logarithm of the population within a given radius.

Comparing the results of the WLS with the SEM model the coefficients do not vary to a large extent. But when looking at the adjusted R^2 value it is obvious that the SEM model performs better without additional independent explanatory variables apart from the contiguity information in the weighting matrix. The significance for most variables persists even in the SEM model (see probability in Table 1). But for the variables describing the density of road types the significance drops for the density of urban collector roads (17.8% error probability) and even the sign changes. The coefficients of the remaining density variables are lower in the SEM model in comparison to the WLS model, indicating a reduction in the sensitivity for these variables. Whereas the road types dummy variables interacting with the employment opportunities and the population are getting an higher impact in the SEM model, as the absolute values of the coefficients are higher. The reason for this must be the inclusion of the spatial interdependence given by the W-matrix. As the λ-parameter is an indicator for the influence of neighbouring errors on the local error, a strong influence must be present, since the theoretical maximum is one.

Table 1 Estimation of link travel speeds – comparison of WLS and SEM results

Variable			WLS Coeff	WLS Prob	SEM Coeff	SEM Prob
Highways	*	Saturday	127.14	0.000	137.13	0.000
Highways	*	Peak Period	116.37	0.000	126.03	0.000
Highways	*	Peak shoulder	125.37	0.000	134.99	0.000
Highways	*	Off Peak Period	128.37	0.000	135.12	0.000
Trunk roads	*	Saturday	119.71	0.000	121.62	0.000
Trunk roads	*	Peak Period	113.89	0.000	115.98	0.000
Trunk roads	*	Peak shoulder	116.12	0.000	118.23	0.000
Trunk roads	*	Off Peak Period	119.65	0.000	122.19	0.000
Collector roads	*	Saturday	106.83	0.000	115.14	0.000
Collector roads	*	Peak Period	100.50	0.000	108.44	0.000
Collector roads	*	Peak shoulder	103.52	0.000	111.40	0.000
Collector roads	*	Off Peak Period	102.30	0.000	113.47	0.000
Distributor roads	*	Saturday	111.01	0.000	110.57	0.000
Distributor roads	*	Peak Period	106.05	0.000	106.39	0.000
Distributor roads	*	Peak shoulder	107.43	0.000	108.25	0.000
Distributor roads	*	Off Peak Period	111.33	0.000	111.51	0.000
Other roads	*	Saturday	62.96	0.000	64.29	0.000
Other roads	*	Peak Period	60.26	0.000	60.23	0.000
Other roads	*	Peak shoulder	62.53	0.000	62.87	0.000
Other roads	*	Off Peak Period	63.03	0.000	66.99	0.000
Highways	*	Highway access points, r=1km	-2.39	0.000	-1.48	0.000
Highways	*	LN(Employment Opport., r=5km)	-4.46	0.004	-7.68	0.000
Trunk roads	*	LN(Employment Opport., r=1km)	-7.34	0.000	-6.96	0.000
Trunk roads	*	LN(Population, r=5km)	-4.66	0.000	-5.19	0.000
Collector roads	*	LN(Employment Opport., r=1km)	-2.86	0.006	-3.19	0.003
Collector roads	*	LN(Population, r=3km)	-5.53	0.000	-6.51	0.000
Distributor roads	*	LN(Employment Opport., r=1km)	-5.32	0.000	-5.12	0.000
Distributor roads	*	LN(Population, r=5km)	-5.65	0.001	-5.69	0.001
Other roads	*	LN(Employment Opport., r=1km)	-4.63	*0.075*	-4.47	*0.082*
Density highways [m/ha]			3.40	0.000	0.03	0.000
Density trunk roads [m/ha]			0.64	0.004	0.01	0.001
Density distributor roads [m/ha]			1.31	0.001	0.01	0.003
Density urban collector roads [m/ha]			-0.57	0.007	0.00	*0.178*
Density urban distributor roads [m/ha]			-0.81	0.000	-0.01	0.001
λ			-		0.55	0.000
$\overline{R}^2$			0.5347		0.5544	

6. Conclusion and outlook

The existence of correlation of link travel speeds should be regarded in most applications dealing with travel speeds or times. Estimation models that neglect this correlation are not only poorly designed but the results of those models are likely to be even wrong. With the regressions demonstrated here, a basic understanding of the correlation of travel speeds and a way to incorporate these effects into link speed estimation models was shown. A closer description of regression models including spatially correlated data can be found in Hackney *et al.* (2006).

Estimating link travel times on the basis of structural variables and network information can help engineers to get speed estimates on the basis of commonly available data without applying e. g. assignment models. With increasing availability of floating car data parameters can be calculated that could be applied to different type of regions, which can be used as an alternative estimate for travel speeds that do not depend on a well fitted function to map traffic volumes to travel speeds or times.

7. Literature

Anselin, L. (1988) Spatial Econometrics: Methods and Models, Kluwer Academic Publishers, Dorddrecht.

Balmer, M., M. Bernard, and K. W. Axhausen (2005) Matching geo-coded graphs, Conference Paper, STRC, Monte Verità.

Hackney, J. K., F. Marchal, and K. W. Axhausen (2004) Monitoring a road system's level of service: The Canton Zurich floating car study 2003, paper presented at the 84th Annual Meeting of the TRB, Washington D. C., January 2005.

Hackney, J. K., M. Bernard, S. Bindra, and K. W. Axhausen (2006) Prediction of average road speeds with regional structure variables, *Arbeitsberichte Verkehr und Raumplanung*, **351**, IVT, ETH Zürich, Zürich.

Hackney, J. K. and K. W. Axhausen (2005) Speed of transit in Zurich, Conference Paper, STRC, Monte Verità.

Hackney, J. K. and M. Bernard (2005) A spatial regression model of traffic speed in Zurich, IVT Seminar, ETH, December 2005, Zürich.

LeSage, J. P. (1998) Spatial Econometrics, Working paper, http://www.spatial-econometrics.com/html/wbook.pdf, accessed November 2005.

LeSage, J. P. (2005) Econometrics Toolbox for Matlab Version 7, http://www.spatial-econometrics.com/html/jplv7.zip, accessed November 2005.