Spatio-Temporal Link Speed Correlations: An Empirical Study

Piotr RACHTAN (Corresponding Author)  
Department of Civil and Environmental Engineering  
University of Massachusetts Amherst  
Maryland State Highway Administration  
Office of Traffic and Safety  
7491 Connelley Drive  
Hanover, MD, 21076  
Phone: +1-410-787-4090  
Fax: +1-410-582-9469  
Email: prachtan@sha.state.md.us

He HUANG  
Singapore-MIT Alliance for Research and Technology  
Future Urban Mobility  
#09-02 CREATE Tower  
1 CREATE Way  
Singapore 138602  
Phone: +65-6601-1636  
Email: huanghe@smart.mit.edu

Song GAO  
Department of Civil and Environmental Engineering  
University of Massachusetts Amherst  
214 C Marston Hall  
130 Natural Resources Road  
Amherst, MA 01003  
Phone: +1-413-545-2688  
Fax: +1-413-545-9569  
Email: sgao@engin.umass.edu

Number of words: 5436  
Abstract: 249  
Number of figures: 3  
Number of tables: 2  
Total: 5436 + 249 + 5 x 250 = 6935  
Submitted on March 15, 2013  
To appear in Transportation Research Record
ABSTRACT

Traffic variables are known to be correlated over time and space due to traffic flow propagation. However, the correlation pattern is still largely unknown and most of the research in short-term travel time prediction, demand forecasting, and network modeling either ignore or assume correlation. In this paper investigated are the patterns of spatial and temporal correlations of average point speeds in a freeway setting. 5-minute speed aggregates are obtained for two directions of an urban freeway along a 12 mile segment. Other variables include traffic flow, ramp locations, number of lanes and the level of congestion at detector stations. Weighted least squares multivariate linear regression models are fitted to the data from 3 different times of day (morning, midday, and afternoon) along a shorter, 6.5 mile stretch of I-10 E freeway. Estimated regression models indicate that an increase in spatial and/or temporal distance reduces the expected value of Fisher Z (transformed correlation). The positive parameters of spatial and temporal distance interaction terms show that the reduction rate diminishes with spatial or temporal distance. Higher congestion tends to preserve higher correlation; variations in road geometry carry relatively small corrections to the models. Models are cross-validated on two locations: the remaining 5.5 mile stretch of I-10 E and the 6.3 mile segment of I-10 W. Cross-validation results show that models retain 75% or more of their original predictive capability when applied to independent samples. The developed regression models are thus transferable and are apt to predict correlation on other freeway locations.
INTRODUCTION

Congestion is a growing problem on urban highways over the world (1). It is becoming worse with the increasing number of commuters and random disruptions to the transportation system, such as incidents, road work, bad weather, etc. In order to model traveler’s route choice decisions and provide reliable prediction of future traffic conditions along the chosen path, a stochastic time-dependent network is required to capture the uncertainties.

There usually exist strong stochastic dependencies among link speeds (or travel times), largely due to traffic flow propagations over time and space, or an event that affects capacities in a wide area. Network stochastic dependencies are generally required to capture the benefits of real-time information for network routing, since only through the dependencies over time and space can the knowledge of an incident at the current time result in a better prediction of traffic conditions in the future at different location within the network. However, the shape of correlation patterns is still largely unknown and, as the literature review shows, most of the research in the related areas either ignore link correlation or base on simplifying assumptions.

Among the studies that ignore correlation are Zhang and Rice (2), who develop a linear prediction model with time-varying coefficients. Rice and van Zwet (3) continue to work on travel time prediction and use historical travel times to estimate regression parameters. They present a method to predict travel times that is computationally effective, but estimated level of congestion is based on an assumption that travel times for a given travel path aim to their historical mean. In other words, if congestion is severe at the beginning of a trip, by assumption it will relieve during the trip; and vice versa. By accounting for correlation, the prediction could perhaps be more computationally intensive, but also more adaptive to traffic conditions that change during the trip. In the study of characteristics of travel time distributions on urban roads (4), suspected is the existence of positive correlation among standardized travel times over different links.

Other works assume correlation. Waller and Ziliaskopoulos (5) approach the adaptive routing problem with limited forms of spatial and temporal link cost dependencies. Given the cost of predecessor links, no further information is obtained through spatial dependence; limited temporal dependency assumes known link cost when the entrance node is reached. In contrast, Gao and Chabini (6) study optimal routing policy problems with an assumption of complete dependencies; it is recognized that capturing link correlations over time and space would potentially make the route choice models more realistic. Chandra and Al-Deek (7) investigate the effect of upstream and downstream location information by checking cross-correlation of speeds at these locations relative to the current location of hypothetical traveler. They find significant relationship of speed with stations both upstream and downstream of the traveler. Min and Wynter (8) develop volume and speed forecasting model using binomial spatio-temporal correlations based on simplified link state: traffic condition may be either congested or free flowing. Although based on simple assumptions, their model achieves reasonable accuracy in short-term prediction of speed and volume. Parent and LeSage (9) develop a space-time dynamic model that relates commuting times with highway infrastructure, gasoline taxes and congestion. It is found that neglecting of positive or negative correlations in analysis of variance of future travel times may lead to its underestimation as large as 75%, or its overestimation by 100%, respectively. In another recent work, a Bayesian network framework is used to learn its time-space dependencies using a structure learning algorithm (10). The algorithm is a simplified version of greedy-equivalence search algorithm based on assumptions that simplify traffic state
evolution; it accurately identifies changes of traffic state for short forecast horizons. Xing and Zhou (11) show how to incorporate spatial correlation into the reliable route search process.

Among the research based on empirical data, single vehicle data is used to establish time-headway distributions and speed-distance relations (12). It is found that free-flow and stop-and-go traffic states are characterized by a strong correlation between density and flow. Gajewski and Rilett (13) estimate link travel time correlation using Bayesian natural cubic splines – a nonparametric regression technique. Note that the correlations studied in this paper are at the link level with aggregate traffic and day-to-day randomness rather than at the vehicle level, as studied with probe-vehicle data (13). Historical travel time estimates are used together with their updated temporal variance-covariance relationships to predict the travel times in the next five-minute interval (14); it has been shown that usage of the updated temporal variance-covariance relationships of travel times can greatly improve the accuracy of the short-term travel time prediction. In the study of correlations between 24-hour segment volumes (15), it is shown that including correlation improves Average Annual Daily Traffic (AADT) prediction; however, large ratio of OD pairs to links may result in overestimation of correlation coefficients. A spatial regression model that considers spatial dependency effects (correlation) gains overall predictive capability and accuracy over the ordinary least squares regression when strong correlations exist (16). Frejinger and Bierlaire (17) capture correlation among alternatives in a route choice problem. A subnetwork is introduced to simplify the road network based on the original network’s road hierarchy. The proposed modified multinomial logit model captures correlation among paths by error components. Song et al. (18) model correlation between OD demands during given time period (e.g. morning peak). It is shown that mean traffic flows are very sensitive to correlation changes and thus correlations significantly influence travelers’ path choice behaviors. A method to evaluate uncertainty of future demand is shown by Duthie et al. (19).

In this paper, a statistical model of link speed correlation based on empirical data is presented and validated. This model contributes to the body of literature as a more realistic replacement to the ignored or assumed link correlation.

**METHODOLOGY**

The objective of this research is to develop a regression model for correlation of speeds on a freeway. By definition (20, 21), sample correlation is given by:

$$r_{xy} = \frac{cov(x, y)}{s_x s_y} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$  \hspace{1cm} (1)

where $s_x$ and $s_y$, and $\bar{x}$ and $\bar{y}$ are sample standard deviations and sample means of variables $x$ and $y$, respectively, and $n$ denotes sample size for estimation of $r_{xy}$. Here, $r_{xy}$ is estimated between the speed reading at a time-space location $x$ and the speed reading at a time-space location $y$. A time-space location is characterized by a unique combination of a post mile (PM) and a 5-minute time of day period (e.g. 7:00:00 to 7:04:59). 5-minute resolution of time is dictated by the data, as described in Section 3. As a dependent variable, correlation is bounded between -1 and 1 and is not normally distributed. Variations in data quality result in varied sample size for correlation coefficient estimation; thus, the assumption that variances of the error term $\varepsilon$ are constant in a linear model is violated (22). A transformation given by Fisher (23) is required to obtain a
The response variable from correlation that is unbounded and its distribution is approximately normal:

\[ z = \frac{1}{2} \ln \left[ \frac{1 + r(z)}{1 - r(z)} \right] \]  \( (2) \)

where \( z \) is the Fisher Z transform and \( r(z) = r_{yx}(z) \) is sample correlation coefficient estimated for a variable pair \( y \) and \( x \) for the given \( z \).

According to Bartlett (24), for moderately large samples we may assume that \( z \) is approximately normally distributed and that the variance of \( z \) is:

\[ V \var {ar} (z) \cong \frac{1}{n - 3} \]  \( (3) \)

where \( n \) denotes sample size for correlation coefficient estimation as in Eq. (1). Bartlett points out, however, that Fisher Z transformation alone does not eliminate unequal variances of the error terms even though \( z \) is normally distributed (24). Since the variance depends on the sample size behind each sample correlation estimate, weighted least squares are applied with weights \( W_i = \frac{1}{\text{var}(z_i)} = (n_i - 3) \), where \( i = 1, 2 \ldots N \) and \( N \) is the total number of correlated pairs. The general weighted least squares regression model is then defined as:

\[ z_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i \]  \( (4) \)

where \( z_i \) is the value of the Fisher Z transform for the \( i \)-th correlated pair of speeds; \( \beta_0, \beta_1, \ldots, \beta_{p-1} \) are estimable parameters; \( x_{i1}, \ldots, x_{ip-1} \) are known predictors; and error term \( \epsilon_i = N(0, \sigma_i^2) \). To compare observed correlation to the predicted values produced by Eq. (4), transformation back to the correlation domain is necessary with the following formula:

\[ \hat{r}_i = \frac{\exp(2z_i) - 1}{\exp(2z_i) + 1} \]  \( (5) \)

where \( \hat{r}_i \) is the correlation predicted from \( z_i \), the expected value of \( z \) for the \( i \)-th correlated pair of speeds. The methodology laid out in this section is then applied to the empirical data as described in the section that follows.

**EMPIRICAL SETTING**

The data used for model estimation are obtained from the Performance Measurement System (PeMS) operated by the California Department of Transportation (Caltrans). The location chosen for the study is a 12.04 mile (19.38 km) segment of I-10, two directions; between post mile (PM) 0.17 in Santa Monica and PM 12.21 at the interchange with I-110 in Los Angeles. The period of analysis consists of all non-holiday weekdays between January 1, 2010 and June 30, 2010, a total of 127 days. There are 41 and 34 detector stations along the eastbound and westbound segment, respectively. Loop detector readings are provided in 5-minute aggregated format; reported data of interest are speed and total mainline flow. Sample size \( n \) for estimation of correlation coefficients reflects the number of readings (days) for a given time-space location; it depends on the length of the analysis period and data quality.
All detectors have experienced temporary maintenance shutdowns during the 6-month period, but there are also detector stations that have not been operating for weeks or longer. PeMS provides imputed readings from inoperative stations using modified interpolated values based on neighboring stations’ readings (25). Detection and exclusion of (or filtering out) imputed data is necessary to avoid artificial stabilization of the variance of speed. The filtering process begins with an assessment of data quality at each station. Percentages of reporting detectors within each station are provided by PeMS at 5-minute resolution (e.g., if a 3-lane freeway indicates 67% quality, it means that 2 out of 3 detectors are working during the given 5-minute period at a given station). Stations that remained inoperative (i.e. data quality=0%) during the entire analysis period are removed from the data set. To ensure that reliable data are used, all readings with data quality below 50% are removed from the data set and replaced by missing values. Setting the filtering threshold above 50% would significantly decrease the amount of available data; at the same time, we have decided that 50% quality gives a reasonable estimate of the traffic conditions on all lanes. In other words, situations when one encounters congested and free flowing traffic streams on different lanes on the same direction at the same location are rare.

**MODEL ESTIMATION RESULTS**

Various combinations of variables are used to estimate a general model using the entire segment of I-10 eastbound and 24-hour data. The primary objective is to achieve best possible fit using variable combinations that make physical sense; the secondary objective is to obtain the simplest model possible. The general model is found not to be able to capture most of the variance in the Fisher Z dependent variable. Thus, it is decided to reduce estimation sample in spatial dimension to approximately 6.5 miles (10.5 km) between Venice Boulevard and I-110 interchange; temporal dimension has initially been reduced to 14 hours between 6:00 and 20:00, but ultimately has been split into three times of day: morning (7:00-10:59), midday (11:00-14:59) and afternoon (15:00-19:59). Division of data into times of day promotes the estimation of models that are more responsive to sudden changes in the dependent variable. Division of the physical setting leaves part of the data available for independent validation of the models. Table 1 shows the results of weighted least squares estimations for each of the three times of day at the partial, 6.5 mile (10.5 km) stretch of I-10 eastbound. Please note that adjusted R-square values reported in the table pertain to the fit of \( \hat{Z}_t \) transformed back to correlation with Eq. (5) and compared to the observed correlation coefficients.

The primary variables in the models are **miles** (spatial distance, absolute value of difference in PM between correlated locations \( x \) and \( y \); unit is 1 mile) and **timediff** (temporal distance in minutes between \( x \) and \( y \); due to resolution of the data, **timediff** increment is equal to 5-minutes). All subsequent variables are combined (i.e. interacted) with **miles**, **timediff**, or both. When spatial and temporal differences are equal to zero, only the intercept is left in a model. However, neither of the intercepts correctly transforms to 1 to reflect correlation with self. This is due to the fact that all self-correlations have been removed from the estimation samples to avoid division by zero during Fisher transformation. Regardless of time of day, increase in either **miles** or **timediff** has a negative effect on correlation. The positive sign of **miles*****timediff** interaction term suggests that the rate of correlation drop decreases with an increase of spatial-temporal distance. Inclusion of quadratic terms, **miles**^2 and **timediff**^2, was suggested in literature (24); these terms also have positive signs and provide minor corrections to relieve the correlation drop. Combined effects of variables are assessed by spreadsheet projections: the variable of interest is varied over its range (**miles** and **timediff**) or its extreme values (**state**, **lanes**, **temporal** interaction term **miles**^2, or **hours**^2 respectively).
While all other variables are held constant. The value of the response is estimated for all three models. In general, the effect associated with spatial distance tends to be stronger than the effect of temporal distance. For example, the overall effect of 1 mile has a correlation-reducing effect equivalent to approximately 29 minutes for all three models.

Predictor variable state is only used in interaction with one or both of the primary variables. State is a continuous variable with values between 0 and 1; derivation process of state is as follows. A speed-flow diagram is plotted for each station using the historical data of best-possible quality; then, speed at the maximum flow is set as the threshold for a given station. Using that, any speed below the threshold is an indicator of unstable (or congested) conditions, while speeds above the threshold indicate stable (or free-flow) traffic state. Within our 6-month data set, a proportion of stable/unstable conditions is determined for each 5-minute period at each detector station. Since this analysis deals with correlated pairs of time-space locations, state is an average proportion of the historical conditions for a given pair. Extreme values occur either when both time-space locations are always under free-flow (state=0), or, when both time-space locations are always congested (state=1); here, “always” means “all historical observations within the two 5-minute periods of interest”. State carries the congestion information for both time-space locations and its main benefit is the simplification of the model by avoiding two additional indicator variables (e.g. one covering Peak–Off-Peak and the other covering Off-Peak–Off-Peak pairs) along with their family of interaction variables. Interaction variable miles*state between spatial distance and the level of congestion has a positive sign for all three models – this suggests that increasing average congestion level between two locations reduces the effect of spatial distance. This is opposite to the timediff*state interaction, while the second level interaction term miles*timediff*state has varying sign depending on the model. The combined effect of all three variables associated with state is the same for all times of day (projected): higher congestion level increases the expected value of the response. This finding does not agree with Gajewski and Rilett (12), who find that heavier congestion reduces the correlation of travel times between links. However, as mentioned in the Introduction, the research setup of that paper was quite different. Authors of this paper are convinced that findings presented here agree with intuition as higher congestion implies higher traffic density and thus stronger vehicle to vehicle interactions.

Another predictor, lanes, is derived and applied similarly to state. Lanes is a regressor associated with the freeway geometry; it is the average number of mainline lanes for a given pair of time-space locations. On this segment of I-10, lanes can vary between 3 and 5 (eastbound) or 3 and 6 (westbound). The regression shows that a higher number of lanes reduces the spatial dimension’s effect, but also brings more negative drop with time-space interaction term. The effect of lanes*timediff varies from model to model and is insignificant during midday period. Simulations showed that – in the overall effect – adding lanes increases the expected value of the response in both morning and afternoon models. Diagnostics indicate that having lanes in the models carries multi-collinearity issues; most likely due to limited variation in numerical values in lanes and the fact the predictor is interacted with miles and timediff. Excluding lanes from the models reduces variance inflation factors to low levels (ca. 10, depending on the model), but one of the considerations in having lanes in the model was the belief that it may be an important factor when a model is transferred to a different physical setting. Thus, until models are successfully cross-validated without lanes, it is decided that model transferability is more important than increased standard errors and bias.
RN_dummy and NN_dummy variables are both ramp presence indicators. For a pair of locations, if a ramp is present (regardless if access or egress) at both locations, then RN_dummy and NN_dummy are equal to 0. RN_dummy is equal to 1 if one, and only one location in the pair has a ramp; NN_dummy is equal to 1 if none of the locations in the pair have ramps. In regression, the overall effect of the ramp presence has been shown to slightly reduce or slightly increase the estimated response in morning and midday model, respectively, when compared to situation when both locations have ramps. In both cases, the effect is smaller when only one reference location has ramp, i.e. NN_dummy induces higher overall change than RN_dummy. The exception is the afternoon model, which seems to be unaffected by the ramp presence in the simulated scenarios.

Of the three models listed in Table 1, the morning model achieves the best fit. This is not surprising, as the morning rush period is much shorter than in the afternoon (at least on this segment of I-10) and, typically, trips made by morning commuters are more predictable and involve less variability than afternoon trip patterns. The ‘more ordered’ morning trip pattern translates into easier-to-predict correlation pattern. The two plots in the top row of Figure 1 provide insight into morning model’s performance in fitting the observed correlation in both spatial and temporal dimension. Each plot has a reference point described by two numbers: first number being the horizontal axis reading when observed correlation is equal to 1; second number (indicated in plot heading) is either time of day or PM for spatial and temporal correlation plot, respectively. Spatial correlation plot is derived by holding timediff equal to 0 while varying miles and other variables associated with geometry changes. The plot shows that corrections to the function from geometry variations can sometimes predict the location, but not necessarily the exact shape of the disturbance in the general trend. Temporal correlation plot is derived by holding miles equal to 0 while varying timediff and state. Similarly to spatial dimension, this plot shows that the morning model captures the trend rather well but does not perform as well in delineating minor variations in the correlation pattern. P-values are for the hypothesis test of no correlation against the alternative that there is a nonzero correlation; they are associated with the observed correlation estimates. A small p-value indicates that the corresponding correlation coefficient is significantly different from zero.

The midday model has an adjusted R-squared slightly smaller than that of the morning model. Similarly to the morning model, it is also estimated for a 4 hour period; it encompasses most of the time between the morning and afternoon peak and ends well into the afternoon peak. Estimated vs. observed plots are in the middle row of Figure 1. Observed values on the spatial correlation plot tend to be more scattered than those in the morning period. In temporal domain, the model performs with similar accuracy as in the spatial domain except for the slightly underestimated correlation when the temporal distance exceeds 2 hours.

The afternoon model has the smallest adjusted R-squared, but also deals with the largest sample from 5-hour period. The more chaotic nature of afternoon peaks in general is believed to be the underlying cause of the loss in this model’s fit. Despite worse numerical performance indication – which may be related to the model’s more general applicability (5 hour time of day period) – correlations tend to be less variable from the general trend and are quite accurately captured by the model in the bottom row of Figure 1.

MODEL CROSS-VALIDATION

Background and Motivation
The regression models described in the previous section provide a satisfactory fit to the data taken from a single location. To answer the question whether correlation patterns are location-specific, the models need to be transferred to other freeway settings with different traffic patterns and geometrical arrangements. Cross-validation aims to test the robustness of the models and to provide insight into how correlation patterns are dependent on location. Two distinctive new settings are discussed below.

**Location 1: I-10 Westbound**

The models are first cross-validated on the opposite direction of the freeway between the similar range of post miles. Thus, a segment of I-10 W between PM 11.96 and PM 5.66 is chosen. It is likely that traffic patterns are different from the corresponding stretch of I-10 E, since this is the outbound link from downtown Los Angeles. Road geometries between the two directions are comparable when one notices the frequent ramps, but by no means are the geometries mirror reflections of each other. When one compares the freeway segments on two directions between the same post miles, one notices distinctive differences in the number and configuration of lanes, as well as the ramp arrangement.

Figure 2 shows the predicted vs. observed spatial and temporal correlation for all three models applied to the I-10 W sample. Morning, midday and afternoon models’ plots are in top, middle, and bottom row of Figure 2, respectively. Despite the different shapes of the observed spatial correlation (plots on the left), all three models retain most of their capability. In temporal dimension (plots on the right) morning and midday models tend to be slightly underestimating and afternoon model slightly overestimating correlation. Again, while models are hardly sensitive to minor variations in correlation patterns – especially in the temporal dimension – the ability to predict the trend is retained.

**Location 2: I-10 Eastbound (Beginning)**

The models are next cross-validated on the eastbound stretch that has not been used for estimation. This segment spans over 14 detector stations starting in Santa Monica (PM 0.17) and ends before the major interchange with Venice Boulevard (PM 5.04); it is later referred to as “I-10 E (beginning)”.

Figure 3 presents observed vs. predicted correlation for the three time-of-day models applied to the I-10 E (beginning) sample. During all three times of day, spatial correlation patterns (plots on the left) tend to be much more variable than on locations discussed before. The fit of the regression lines is visibly worse, especially for the midday model. Temporal correlation pattern in the morning (Fig. 2. top right) is characterized by a very steep slope; correlation reaches near-zero values within 30 minutes from the reference period. In contrast, midday and afternoon temporal correlation plots (Fig. 2, middle and bottom right) tend to drop less dramatically; the fit of the regression lines is noticeably better.

**Cross-validation Summary**

For proper assessment of the general predictive performance of each model, common and objective measures are needed. Note that R-squared can be interpreted as adjusted R-squared when the sample size used for model estimation is very large. Table 2 lists statistics associated with each model at each location. Per Table 2, the capability of morning model to correctly predict correlation on the westbound direction is about at the same level as it has achieved on its own sample, although it is perhaps slightly more biased; standard error remains roughly
unchanged. When the same model is applied to the beginning stretch of the eastbound direction, all performance measures indicate a reduction of performance. Nevertheless, the model is still able to explain approximately half of the variance in the observed correlation.

Cross-validation of the midday model at the westbound sample indicates the same level of very low bias and standard error as at its estimation sample, with a slight increase in R-squared. While delivering an impressive result at the opposite direction, the midday regression experiences the largest loss of predictive performance at the second cross-validation location.

The afternoon model achieves the smallest R-squared when fit to the observed data in its estimation sample is assessed, but its bias and standard error are at the same time the smallest of all models. Although cross-validation to the westbound direction significantly reduced the fit, the afternoon model experiences the least performance loss when applied to the beginning stretch of the eastbound direction. At both cross-validation locations, it retains the very low level of bias and standard error.

SUMMARY AND CONCLUSIONS
Correlation can be a useful metric when applied in traffic operations; it can also be a measure of the drivers’ constraint in traveling at their desired speed (13). The analysis presented in this paper results in the development of three regression models to describe correlation variation with the resolution of 5-minutes in time and one detector station in space. The primary factor for correlation is the spatial distance. The slope of correlation tends to be more affected by the traffic state in temporal domain than in the spatial domain. Models agree that an increasing congestion level creates a ‘retaining effect’ on correlation – the rate of decrease tends to be lower. Relatively small corrections to time-space effects are added when freeway geometrical arrangement changes occur; however, it is believed that these parameters improve the overall fit and play an important role – together with traffic state – when a model is transferred to a new physical setting. Each model has a slightly different behavior when applied to an independent cross-validation sample; however, all models retain 75% or more of their original predictive ability. The developed regression models are capable of explaining significant percentage of variance in the observed correlation and the results show that these models can be transferred to new locations with some success.

Certain disruptions to the traffic stream like inclement weather, roadwork, and incidents have not been included in the analysis. Although certainly affecting correlation, these events can change the users’ decisions before entering the traffic stream (with a priori knowledge about an event they can postpone the trip, change mode, etc.). The main interest of this research was to explore correlation-affecting factors within the traffic stream during typical network usage; all while retaining reasonable simplicity and maximizing the transferability of the models.

Potential venue to make the models more accurate and – possibly – to allow the usage of a general model for all times of day could be the constriction of the analysis to a shorter, 1 or 2 hour prediction horizon. Among the tasks that should be considered in the future is an extension of the cross-validation to different time period. Also, the analysis detected some multicollinearity issues associated with the presence of variable lanes. Thus, cross-validation of the models that do not include variable lanes should be attempted to determine whether the transferability would be affected.

This paper fills the gap in the literature with an empirical analysis of speed correlations in a freeway setting. On freeways, the traffic stream between the consecutive exits can be reasonably assumed to be a continuous flow. Extending the research onto more chaotic arterials
where there is no control for accessing or exiting the stream is expected to be a very daunting task itself. Hopefully, this work provides a useful foundation for extension onto arterial roadways, and then onto entire transportation networks.

REFERENCES


LIST OF TABLES

TABLE 1  Weighted least squares estimation results for Fisher Z during three time periods of the day ................................................................. 15

TABLE 2  Predictive ability of the regression models on three locations .......... 16

LIST OF FIGURES

FIGURE 1  Observed vs. estimated spatial (left) and temporal correlation (right) on the model estimation sample; morning model (top), midday model (middle), afternoon model (bottom) ................................................................. 17

FIGURE 2  Observed vs. predicted spatial (left) and temporal correlation (right); cross-validation on I-10 W sample; morning model (top), midday model (middle), afternoon model (bottom) ................................................................. 18

FIGURE 3  Observed vs. predicted spatial (left) and temporal correlation (right); cross-validation on I-10 E (beginning) sample; morning model (top), midday model (middle), afternoon model (bottom) ................................................................. 19
TABLE 1  Weighted least squares estimation results for Fisher Z during three time periods of the day

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MORNING (7:00-10:59)</th>
<th>MIDDAY (11:00-14:59)</th>
<th>AFTERNOON (15:00-19:59)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 411778</td>
<td>N = 411778</td>
<td>N = 619941</td>
</tr>
<tr>
<td></td>
<td>adj. R² = 0.6501</td>
<td>adj. R² = 0.6189</td>
<td>adj. R² = 0.5405</td>
</tr>
<tr>
<td>Parameter</td>
<td>Estimate t-Statistic</td>
<td>Estimate t-Statistic</td>
<td>Estimate t-Statistic</td>
</tr>
<tr>
<td>intercept</td>
<td>1.146  893.62</td>
<td>1.201  961.52</td>
<td>1.023  1199.53</td>
</tr>
<tr>
<td>miles</td>
<td>-0.3859 -121.8</td>
<td>-0.4229 -137.09</td>
<td>-0.2075 -110.57</td>
</tr>
<tr>
<td>timediff</td>
<td>-0.006456 -70.5</td>
<td>-0.008089 -325.99</td>
<td>-0.006885 -166.88</td>
</tr>
<tr>
<td>miles^2</td>
<td>0.007377 57.42</td>
<td>0.007032 56.4</td>
<td>0.008900 103.53</td>
</tr>
<tr>
<td>timediff^2</td>
<td>0.00001314 130.16</td>
<td>0.00001449 147.41</td>
<td>0.00001409 326.85</td>
</tr>
<tr>
<td>miles*timediff</td>
<td>0.001616 40.03</td>
<td>0.002297 73.95</td>
<td>0.001147 59.32</td>
</tr>
<tr>
<td>miles*state</td>
<td>0.05589 59.43</td>
<td>0.02229 24.35</td>
<td>0.02027 30.33</td>
</tr>
<tr>
<td>timediff*state</td>
<td>-0.0002632 -9.49</td>
<td>-0.0003689 -15.1</td>
<td>-0.0004872 -34.01</td>
</tr>
<tr>
<td>miles<em>timediff</em>state</td>
<td>-0.0002045 -16.97</td>
<td>0.00007099 6.24</td>
<td>-7.917E-05 -11.24</td>
</tr>
<tr>
<td>miles*lanes</td>
<td>0.03593 54.31</td>
<td>0.04913 76.29</td>
<td>0.009453 22.61</td>
</tr>
<tr>
<td>timediff*lanes</td>
<td>-0.0004985 -26.67</td>
<td>-0.00002986 -45.52</td>
<td>-0.0001322 -30.23</td>
</tr>
<tr>
<td>miles<em>timediff</em>lanes</td>
<td>-9.987E-05 -11.67</td>
<td>-0.0002986 -45.52</td>
<td>-0.0001322 -30.23</td>
</tr>
<tr>
<td>miles*RN_dummy</td>
<td>0.01206 26.48</td>
<td>-0.01510 -34.06</td>
<td>-0.001714 18.7</td>
</tr>
<tr>
<td>timediff*RN_dummy</td>
<td>0.0001859 14.13</td>
<td>-0.0002644 -20.88</td>
<td>-0.0004015 -57.04</td>
</tr>
<tr>
<td>miles<em>timediff</em>RN_dummy</td>
<td>-8.995E-05 -14.99</td>
<td>0.0001600 27.57</td>
<td>0.0001397 54.41</td>
</tr>
<tr>
<td>miles*NN_dummy</td>
<td>0.02336 22.91</td>
<td>-0.03120 -31.42</td>
<td>0.006763 9.95</td>
</tr>
<tr>
<td>timediff*NN_dummy</td>
<td>0.0004351 16.01</td>
<td>-0.0002547 -9.83</td>
<td>-0.0005855 -39.93</td>
</tr>
<tr>
<td>miles<em>timediff</em>NN_dummy</td>
<td>-0.0002060 -16.5</td>
<td>0.0002331 19.38</td>
<td>0.0001711 24.68</td>
</tr>
<tr>
<td>MODEL</td>
<td>Statistic</td>
<td>Est. Sample</td>
<td>Cross-validation at:</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-10 E (PM 5.64 - 11.96)</td>
<td>1-10 W (PM 11.96-5.66)</td>
</tr>
<tr>
<td>MORNING</td>
<td>bias</td>
<td>-0.01466</td>
<td>0.07839</td>
</tr>
<tr>
<td></td>
<td>standard error</td>
<td>0.1600</td>
<td>0.1744</td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.6501</td>
<td>0.6198</td>
</tr>
<tr>
<td>MIDDAY</td>
<td>bias</td>
<td>-0.01408</td>
<td>0.0186</td>
</tr>
<tr>
<td></td>
<td>standard error</td>
<td>0.1437</td>
<td>0.1418</td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.6189</td>
<td>0.6345</td>
</tr>
<tr>
<td>AFTERNOON</td>
<td>bias</td>
<td>-0.01200</td>
<td>-0.04899</td>
</tr>
<tr>
<td></td>
<td>standard error</td>
<td>0.1323</td>
<td>0.1720</td>
</tr>
<tr>
<td></td>
<td>R squared</td>
<td>0.5405</td>
<td>0.4192</td>
</tr>
</tbody>
</table>
FIGURE 1  Observed vs. estimated spatial (left) and temporal correlation (right) on the model estimation sample; morning model (top), midday model (middle), afternoon model (bottom)
FIGURE 2  Observed vs. predicted spatial (left) and temporal correlation (right); cross-validation on I-10 W sample; morning model (top), midday model (middle), afternoon model (bottom)
FIGURE 3  Observed vs. predicted spatial (left) and temporal correlation (right); cross-validation on I-10 E (beginning) sample; morning model (top), midday model (middle), afternoon model (bottom)