Identification of spatiotemporal relationships in travel speeds along individual roadways using probe vehicle data

Reihaneh Kouhi Esfahani
Department of Physics
Isfahan University of Technology
Isfahan 84156-83111, Iran
r.koohiesfahani@ph.iut.ac.ir

Vikash V. Gayah
Department of Civil and Environmental Engineering
The Pennsylvania State Unviersity
231L Sackett Building
University Park, PA 16802
gayah@engr.psu.edu
phone: 814-865-4014

Paper submitted to TRB Annual Meeting 2019
November 15, 2018
$5,729$ words + 11 figures + 27 references

ABSTRACT

The existence of spatiotemporal correlations in traffic behavior across links in a transportation network can be incredibly useful. However, travel speeds are often strongly correlated simply due to natural variations in travel demand patterns. Such temporal trends might obstruct more meaningful relationships in link performance caused by the physics of traffic. To alleviate this problem, the present paper proposes a non-parametric, moving-average detrending method that can be used to remove these background trends, even during non-stationary periods in which traffic states are changing with time. Cross-correlations performed on the detrended data can then be used to identify more meaningful trends. The proposed method can also account for temporal lags in traffic performance, which account for the time it takes for information to travel. Links that exhibit strong correlations after detrending can then grouped into communities that behave together using graph theory methods. The proposed methodology is applied to a case study network using real-time link travel speeds obtained from probe vehicles. The results reveal that the 40 links in the network can be grouped into 8 to 12 communities, depending on the day of the week. This suggests that only a handful of links may need to be monitored to estimate travel speeds across the entire network. Furthermore, the significant overlap in the community structure across these days reveals that the network structure plays a large role in spatiotemporal correlations in link travel speeds in a network. This community structure can be leveraged to improve speed prediction algorithms.

INTRODUCTION

The rapid development of communications and vehicular technologies has provided new data sources that can be used to improve our ability to monitor transportation networks and forecast traffic conditions (*1*, *2*, *3*, *4*). The availability of large amounts of accurate traffic data is essential to obtain travel times and traffic states to inform Advanced Traveler Information Systems (ATIS) and provide routing guidance to drivers (and soon automated vehicles). However, even with these newly available sources, traffic data often suffers from numerous deficiencies that must be addressed. For example, most data-sets are not entirely accurate or complete; they may contain missing or incorrect values, duplicate data, and incorrect data linkages. Cleaning or filtering these data to identify only the most useful pieces of information is costly and time consuming. One study estimates that bad data costs the United States $3 Trillion Per Year (*5*). Furthermore, the cost associated with obtaining and storing these data in real-time (often using cloud computing and storage methods) can be large.

Instead of trying to filter through as much data as possible, another strategy is to identify network-specific relationships that can be leveraged to reduce the amount of data that needs to be collected and maintained. For example, if traffic performance on several links share the same general spatiotemporal patterns, it might be possible to gain an accurate depiction of current traffic states throughout the network by monitoring only a subset of these connected links. In this way, the number of observations that are needed can be reduced significantly, which can decrease the cost associated with data collection and analysis. Correlations in traffic performance across individual links is one method that can be used to identify the most important links and reduce data requirements. Previous studies have leveraged correlations between traffic metrics over time and space in transportation networks to improve forecasting, route choice and other transportation problems. For example, Gao and Chabini showed that route choice models become more realistic when link correlations over time and space are included (*6*). Leveraging correlations between volumes have also been shown to improve predictions of average annual daily traffic (AADT) on roadway links (*7*).

Many recent studies have examined spatiotemporal correlations in traffic metrics across transportation networks. The earliest works focused on correlations across adjacent links. These studies found that—perhaps unsurprisingly—traffic conditions on adjacent links are often highly correlated (*8*, *9*, *10*, *11*). More recent studies have focused on correlations on links that are not adjacently connected. One class of studies focused on "l^{th}-order neighbors", which are those that are located l links away in the transportation network (*12*, *13*). Others have found that all links on a network exhibit positive correlations in traffic performance, even those located far from each other. For example, Sun et al. analyzed correlation coefficients of traffic metrics on a grid network containing 31 traffic links and found positive and distance-independent correlations (*14*).

However, traffic patterns generally follow the same temporal trends due to peaks in travel demands that are independent of the spatial configuration of a traffic network. On a given day, traffic volumes on all links are generally low in the morning, peak during the AM rush, flatten during the middle of the day, peak during the PM rush and fall during the evening. Similar patterns occur during individual days of the week and throughout the year. Failure to account for these naturally occurring trends might yield correlations that inflate the relationship between traffic performance on arbitrary links in a network or lead to identification of relationships that are not physically meaningful.

To alleviate this concern, Ermagun et al. examined correlations after detrending traffic data

(*15*). Detrending is a method used to remove the naturally occurring fluctuations in traffic metrics (e.g., flows) by first identifying and removing the temporal trend and then assessing correlations of the fluctuations around these trends (i.e., the residuals), which might better represent how performance of individual links may be correlated. While this presents a step in the right direction, (*15*) employed an autoregressive (AR) detrending model applied to data over a single hour of the day for a 1-year period. Unfortunately, AR models rely on the existence of stationary conditions in which traffic is relatively stable throughout the analysis period. In general, traffic patterns are highly variable and can change significantly (e.g., from freely flowing to congested conditions) even over the course of an hour. Furthermore, the AR model cannot accommodate non-linear trends in traffic data. Lastly, the authors only considered correlations between data observed at the same time and did not consider the impacts of temporal lags. Such lags are important as it takes information time to travel across links in a network; failure to account for these temporal shifts can underestimate correlations between link performance.

In light of this, this paper proposes the use of more sophisticated detrending methods to study spatiotemporal travel patterns. A non-parametric, non-linear moving-average method is proposed that can more accurately describe traffic data, especially during the transition states that might occur throughout a day, compared to linear or parametric models. Seasonality effects can also be accounted for to adjust for the use of a fixed-time window in the moving-average method. Once the data are detrended, cross-correlation values can then be computed using different potential time lags, which accounts for the time it takes information to travel across links. The proposed method is applied to a case study example using probe vehicle speeds from several links in downtown Philadelphia, Pennsylvania. The results are used to identify communities of links that share similar spatiotemporal patterns using graph theory methods (*16, 17*). These communities are shown to be remarkably consistent across the days in a week—and even throughout a day—suggesting that they represent relationships due to the underlying traffic network structure. These communities can be used to improve traffic state estimation and forecasting with lower data requirements since only data from one link within each community is necessary, instead of data from each link in the network. To demonstrate the usefulness of this approach, the communities are leveraged to improve a simple prediction model of link travel speeds. Specifically, we demonstrate that prediction accuracy increases when input data for the prediction comes only from a link's community as opposed to from all links in the network.

The remainder of this paper is organized as follows. First, the proposed detrending and cross-correlation methodology is explained. Then, the methods are applied to a case study example to demonstrate how they can be used to identify communities of links that behave similarly. Next, an example of how the community structure can improve speed predictions is presented. Finally, some concluding remarks are provided.

METHODOLOGY

This paper proposes a generic detrending method to remove natural correlation between traffic variables (specifically, link travel speeds) that occur due to temporal fluctuations in travel demand. For example, travel speeds on across any network will be highly correlated throughout a day as speeds generally decrease in a similar pattern due to peak rush hour periods. Detrending removes these temporal variations to more accurately capture meaningful spatiotemporal correlations between link-level data. Many detrending methods exist in the data mining literature. With respect to detrending applied to traffic data, previous studies (*15*) have applied detrending to traffic data using simple linear methods like autoregressive (AR) models, which can identify trends when traffic conditions are stationary throughout the analysis period. This assumption is reasonable for short time periods that do not occur when traffic states are transitioning (e.g., changing from peak to non-peak periods). However, traffic conditions are non-stationary over longer time periods (e.g., throughout the course of a day) when transitions between unique traffic states occur and thus the linear, AR models are not generally applicable to describe traffic data.

Instead, we implement here a non-linear detrending method that is better suited to non-stationary data. The remainder of this section describes the proposed detrending method in more detail, as well as the methods propose to use the detrending data to identify cross-correlations between individual link speeds and identify communities of links that behave similarly during any time period.

Detrending method

The detrending method proposed here consists of three steps: 1) identification of a non-linear temporal trend; 2) identification of repeating seasonality effects; and, 3) estimation of the remaining residuals. The residuals resulting from the detrending can then used to assess correlations between traffic data (i.e., link speeds) between different elements to identify those that exhibit spatial correlation.

Step 1: Trend identification

A non-parametric, non-linear moving average method is proposed to identify and remove the natural temporal trend in traffic data from each link (18, 19, 20, 21). This moving average method is applicable to any type of data and can readily accommodate situations in which the trend is non-linear or traffic states are not stationary. It is also computationally efficient and simple to apply.

In this moving average method, a fixed window length is chosen and average speeds are obtained by sliding this window forward in time to obtain consecutive moving average values. For example, the moving average for any given link k at time t is obtained by averaging the speeds values from that link within a window centered at time t. The next element obtained by shifting the window forward in time by one unit, which excludes the first speed value of the series and includes the next available speed value from that link.

Mathematically, the moving average method is expressed as follows. Denote the time-series data from each link k on day d as $Y(k,d) = [v_{k,0,d}, v_{k,1,d}, v_{k,2,d}, ..., v_{k,T,d}]$, where $v_{k,t,d}$ represents the speed on link k at discrete time period t on day d and T is the total number of time periods available. Applying the moving average converts the vector $Y(k,d)$ into a new vector of speed trends, $D(k,d)$:

$$D(k,d) = [D_{k,l,d}, D_{k,l+1,d}, D_{k,l+2,d}, ..., D_{k,T-l,d}]. \tag{1}$$

and

$$D_{k,t,d} = \frac{\Sigma_{i=-l}^{l} v_{k,t+i,d}}{(2l+1)}, \tag{2}$$

where $2l+1$ is equal to the window length of moving average.

The length of the window presents a trade-off between reducing statistical fluctuations and meaningful observable trends. Longer windows have less variation but might minimize the presence of these trends. Thus, several windows must be tested to determine the length that provides the best balance.

Step 2: Seasonality adjustments

The use of a sliding window for the moving average means that a single speed value impacts $2l + 1$ data elements in the time series of speed trend, $T(k, d)$. This can lead to regular, periodic impacts caused by higher or lower values observed in the dataset. To remove these period impacts, additional steps are necessary to identify and remove 'seasonal' impacts. To do so, a convolution filter is applied to the dataset it identify regularly occurring trends with the period $2l + 1$. The average of the smoothed series identified for each period is the returned as the seasonal component. For more details, see (22, 23).

In this work, we consider a functional form in which the seasonlity adjustments influence the observed speed values in a multiplicative manner. In this way, the relationship between the observed speeds, speed trends, seasonality adjustment and speed residuals takes the following form:

$$Y(k, d) = D(k, d) * S(k, d) * r(k, d) \tag{3}$$

where $S(k, d)$ provides the seasonality impacts and $r(k, d)$ provide the remaining impacts. The length of the time series here is $T - 2l$, which removes the first l elements at the beginning and end of the original data due to the moving-average being applied.

Note that other functional relationships can be assumed between the trend, seasonality and error terms. For example, an additive model was also considered in this work in which $Y(k, d) = D(k, d) + S(k, d) + r(k, d)$. However, the trends in the error terms were remarkably similar across the two model structures and thus both functional forms provided very similar results in terms of spatiotemporal correlations of individual link speeds. For brevity, only the results of the multiplicative model are provided in this paper since it provided slightly more precise results in a predictive test.

Once the seasonality adjustment is accounted for, a new time series of residuals is obtained which takes the following form for any given link K and day d:

$$r(k, d) = [r_{k,l,d}, r_{k,l+1,d}, r_{k,l+2,d}, ..., r_{k,T-l,d}] \tag{4}$$

Step 3: Obtaining average residuals across many days

At the end of Step 2, a unique time series of speed residuals is available for each day. These residuals are then averaged across several days to reduce the impacts of missing data for any one time period on any given day. The resulting residuals array, $\hat{r}(k) = [\hat{r}_{k,l}, \hat{r}_{k,l+1}, ..., \hat{r}_{k,T-l}]$, has elements that take the following form:

$$\hat{r}(k) = \frac{\Sigma_{i=1}^{N} r(k, i)}{N} \tag{5}$$

where N is the number of days for which data that are available.

Cross-Correlation and link communities

Cross-correlation is a standard method in signal processing to estimate the degree to which two series are correlated as a function of a temporal displacement of one relative to the other. In this work, we propose to calculate cross-correlation values for the detrended speed residuals obtained at the end of Step 3 described above. The cross correlation γ between links k and k' with a temporal displacement of τ is defined as:

$$\gamma_{\hat{r}(k)\hat{r}(k')}(\tau) = \frac{\sum_{t=0}^{T-\tau}(\hat{r}_{k,t} - \bar{\hat{r}}(k))(\hat{r}_{k',t+\tau} - \bar{\hat{r}}(k'))}{\sqrt{\sum_{t=0}^{T-\tau}(\hat{r}_{k,t} - \bar{\hat{r}}(k))^2}\sqrt{\sum_{t=0}^{T-\tau}(\hat{r}_{k',t+\tau} - \bar{\hat{r}}(k'))^2}} \tag{6}$$

where $\bar{\hat{r}}(k)$ and $\bar{\hat{r}}(k')$ are sample means of variables $\hat{r}(k)$ and $\hat{r}(k')$ respectively. Cross-correlation is bounded between -1 and 1 and values with absolute values closer to 1 representing a stronger relationship. The resulting cross-correlation matrices can be used describe the relationship between traffic conditions between any link pair.

Various time-lags are tested and the value that maximizes the cross-correlation between any link pair on a given day is defined as the optimal time lag, $\tau^c_{k,k'}$. These optimal time lags and cross correlation values are used to identify links that have the strongest spatiotemporal connection with other links in the network, which can provide an indication of the most important links in the network.

Identifying communities in networks, especially in large-scale networks, is an important task in many scientific areas, such as graph theories (*24*) or in transportation networks (*25*, *26*, *27*, *28*). The cross-correlation values can also be used to identify communities of links over which traffic performs similarly. To do so, a threshold cross-correlation value, $\gamma\prime$, is identified that indicates high correlation in traffic performance across two links. The specific value of the threshold is obtained by examining the cumulative probability function of the cross-correlation matrix elements to identify the cross-correlation value for which some fraction of the elements are greater than a certain value. A new cross-correlation matrix is then created in which link pairs that have a correlation that exceed this critical value are retained. The remaining link pairs are removed, which assumes that no information is shared between these links. The result is a directed and weighted graph in which the weights represent the strength of the highly correlated link pairs and information flow within the transportation network.

This graph is then used to identify communities of links that behave similarly. A general complex network is said to have community structure if the nodes of the network can be clustered into sets of nodes such that each set of nodes is densely connected internally but has only sparse connections with the rest of the network. Communities give a large-scale overview of the network and its function since each community acts like meta-node in the network which can simplify its representation (*29*, *30*). In this paper, the Infomap technique is applied to identify communities in the directed and weighted graph. In the Infomap method, communities are detected with respect to how the information or resources flow through that network. The details are omitted here for brevity, particularly since the method used to identify the communities is not the primary purpose of this paper. More information on this approach can be found at (*31*).

APPLICATION OF METHODOLOGY TO A CASE STUDY EXAMPLE

As a case study example, the proposed detrending and community identification methods were applied to real-time speed data obtained from a private-sector company. This section describes the data and the results of the proposed methods.

Dataset

Speed data were obtained from probe vehicles traveling within the transportation network that communicated their location using GPS-enabled devices and either fleet or cellular communication technology. The probe vehicle locations were used to estimate speeds on segments or links defined by the service provider at regular intervals. Speed data were obtained from the urban area of Philadelphia, Pennsylvania for a one-month period (March 2018). The specific network considered is illustrated in the top of FIGURE 1 and consists of a 0.893 miles by 1.360 mile area in downtown Philadelphia. The numbered circles in the figure represent the start and end points of each link (defined by the service provided) included in the study. The bottom of FIGURE 1 provides a graph representation of this network using the numbered circles as nodes in the network. The solid arrows in the graph represent links with available probe travel speeds, while the dotted arrows represent significant travel links for which probe speed data were not available. Each link with data available are assigned a link number, which is illustrated in blue. Speed data were available for 40 links in the network and these links varied in lengths from 0.0072 miles to 0.6842 miles, with average length of 0.161. Links are generally defined based on the existence of traffic control devices and merge/diverge points along roadway segments, which are anticipated to interrupt traffic flow and change travel speeds.

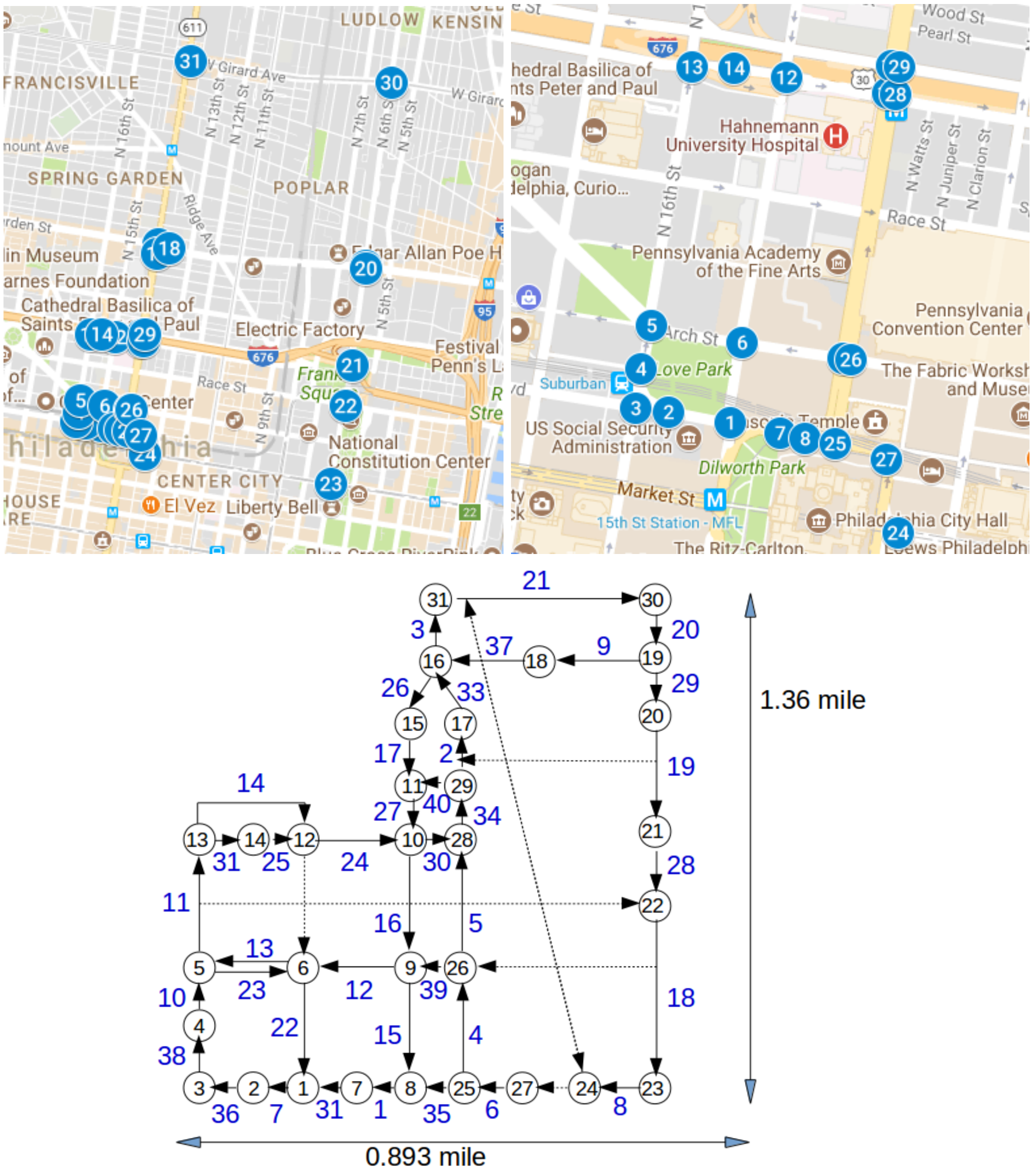

FIGURE 1 Top: Map of case study network used for this paper. Each symbol represents start and end points of the links under study. Bottom: Network representation of the case study network. The solid arrows represent links for which speed data are available. The dashed line arrows represent important links that can affect the network, but for which speed data are not available.

Speed data are available for each link at regular 5-minute intervals. Each speed measurement also includes an indication of whether the measurement represents a real-time speed measurement or is an estimated speed using historical travel time information. Time periods for which real-time speed data were not regularly available were excluded from this analysis. Thus, only data from 6AM to 8PM were considered as probe data were generally not available for late evening and early morning periods.

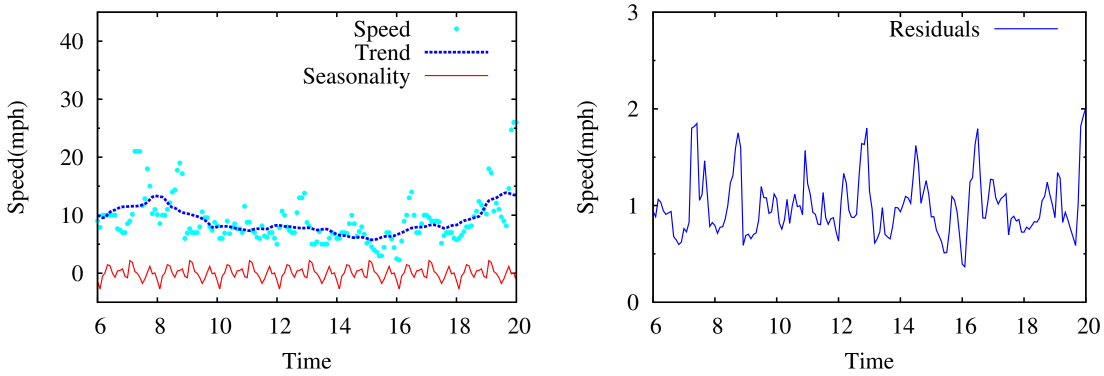

FIGURE 2 Left: Illustration of detrending results for Link $26 \rightarrow 28$ for all Thursdays within study period. Right: Final residuals obtained after detrending.

Detrending results

Speed data for each link was detrended as described in the Methoodology section. FIGURE 2 provides an illustration of the detrending process and results. The cyan dots on the left hand side represent the observed speed data on link $26 \rightarrow 28$ over time of day for all Thursdays in the dataset. The moving average representing the temporal trend is illustrated by the solid blue line in FIGURE 2 for a window length of 2 hours. Note that several windows lengths were considered here and that the results were consistent for window lengths between 1.5 and 2.5 hours. Window lengths less than 1.5 provided too much randomness in the trend, while windows larger than 2.5 hours did not show significantly fluctuations in speed throughout the day. Notice that the blue line suggests that speeds are generally high in the morning, decrease during the day, and then increase after the PM peak hour.

The seasonality adjustment was obtained using the convolution filter. These values had a mean value of 1 and fluctuated in a cyclical pattern with period equal to $2l + 1 = 2$ hour. Note that the fluctuations illustrated here are actually multiplied by 30 to better visualize the seasonality trend that exists in the data.

The final residual values were then obtained by removing the trend and seasonality adjustments from the average travel speed on each link as per equation 3. The right hand side of FIGURE 2 illustrates this for link $26 \rightarrow 28$. Notice that these values also fluctuate around 1 since the residuals are considered in a multiplicative manner as per Equation 3. This process was repeated for all links and these detrended residuals are then used to identify spatiotemporal correlations in link behaviors in the case study network.

Cross-correlation matrices

The left hand side of FIGURE 3 provides the correlation values obtained between all links considering travel speeds across the entire time period (6AM to 8PM). (Please note that color is used to illustrate these correlation values and other pertinent features in the figures of this paper; the reader is strongly encouraged to view the electronic file or a colored printout to better observed these features.) Notice that correlation values are both positive and negative. Positive values represent links in which speeds move in the same direction (i.e., increase or decrease at the same time), while negative values indicate links for which speeds move in opposite directions. This is reasonable as

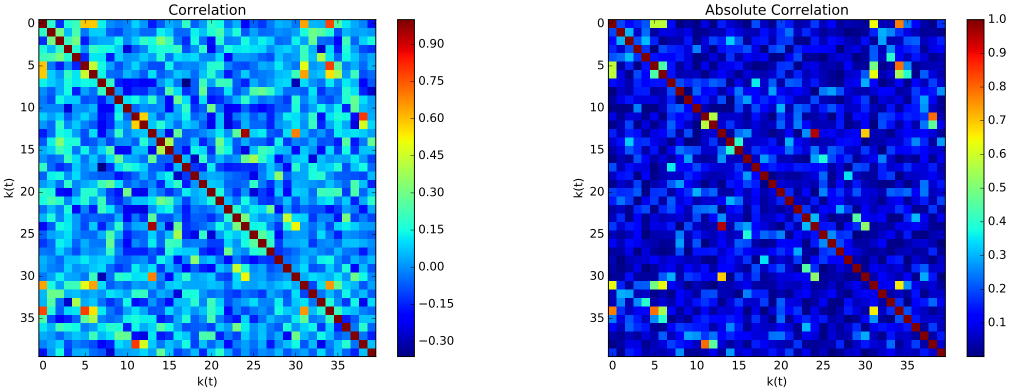

FIGURE 3 Left: Correlation value between each link pair. Right: Absolute correlation value.

in general traffic networks some links would be expected to behave similarly while others would not. However, the majority of the link correlations (58%) have positive correlation values.

The magnitude of the correlation values (i.e., absolute values) represent the strength of the relationship between speed residuals between any two links. These absolute values are plotted in FIGURE 3. Notice that this figure reveals that most of the absolute values are in the range of about 0.1, which does not indicate very strong relationships. If actual speed data were used without detrending, correlations would be much higher as previously discussed.

A cumulative distribution of the absolute value of correlation values is shown by the red line in FIGURE 4. These low values can be partially explained by the fact that correlations are being calculated on the residuals across an entire day. Larger values would be obtained when examining correlations over shorter time periods (e.g., over the course of several hours instead of the entire day). Furthermore, no time lags are considered here. In reality, information takes time to travel between links on a network. For example, fluctuations in speed will move downstream in free flow traffic as vehicles travel from one link to the next. In congested traffic, information generally travels upstream and takes time to propagate from one link to the next.

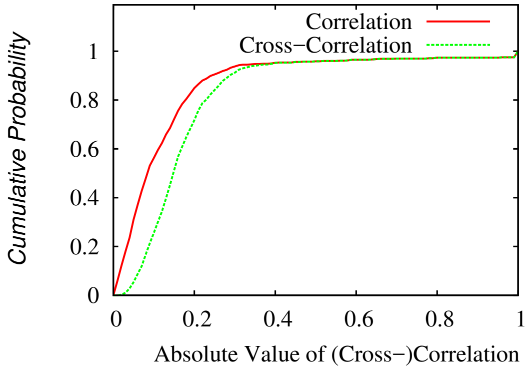

FIGURE 4 Cumulative probability of observing values in correlation (no time lag) and cross-correlation (with optimal time lag) matrices.

To better capture these temporal relationships in speed residuals on individual links, the correlation values were re-calculated considering various time lags that might exist. For the purposes of this paper, we considered only time lags from 0 to 15 minutes in 5-minute increments; however, in practice, finer time resolutions could (and likely should) be considered, especially for a small network of this size. The top left of FIGURE 5 shows the maximum (for positive correlations) or minimum (for negative correlations) cross-correlation values (MCC) obtained across all time lag values considered. The absolute values of the strongest cross-correlations with optimal time lags are provided in the top right of FIGURE 5. For each matrix, the x-axis represents the link $r(k)$ at time t while the y-axis represents link $r(k')$ with optimal time lag, $\tau^c_{k,k'}$, applied. So, if one is interested in considering the information flow sent from a particular link to others, the column of that link should be studied. On the other hand, if one is interested in the information flow received from a link from others, the row of that link should be studied.

Visual comparison of FIGURE 5 with FIGURE 3 reveals that adding the optimal time lag improves the correlation results; however, the differences might be difficult to observe. To better illustrate this improvement, FIGURE 4 also provides the cumulative distribution of the absolute values of the correlation results when the optimal time lag for each link is applied. As shown, the magnitude of the correlations are significantly larger when the optimal time lag is applied, as evidence by a higher proportion of larger values. Across the all links and considering correlations of link speeds across the entire time periods, adding the optimal time lag improve correlation values by 38% compared to the case in which no time lag is considered. The optimal time lag values are illustrated in the bottom of FIGURE 5. Notice that most optimal time lags are 0 or 5 minutes, likely due to the small network size. Note that this optimal time lag could differ based on the time period of analysis (e.g., across the different days of the week); however, several tests (not shown here for brevity) suggest that these optimal values are fairly stable across analysis days, especially for links that are in close spatial proximity.

We now use the cross-correlation metrics to examine how link speeds are correlated spatially and temporally in the network. As an illustrative example, three links ($8 \rightarrow 7$, $14 \rightarrow 12$ and $20 \rightarrow 21$) are randomly selected and the optimal time lag and associated cross-correlation values between these links and all other links in the network are examined to reveal spatiotemporal patterns; see FIGURE 6. In this figure, the color of each link represents the cross-correlation value between that link and the subject link (illustrated with a star), while the number near each link represents the optimal time lag, τ^c, between that link and the starred link.

FIGURE 6 reveals the magnitudes of the cross-correlation values generally (but does not always) decrease with the distance between the subject link and all other links. Similarly, the optimal time lags generally increase with the distance between the subject link and all other links. For example, links on the same (right-to-left) arterial as the subject link have a very strong correlation with the subject link with zero optimal time lag. The spatial relationship is reasonable as links in closer proximity should behave more similarly than links that are further away. Links nearby also tend to have a positive correlation with the subject link, while negative correlations are generally only observed with links that are located further away in the network. The temporal relationship is also reasonable as it takes a shorter time for information to travel to links in nearby proximity as it does for information to travel to links further away in the network. Note the optimal time lag for some links on the other side of the network are zero, indicating that information takes no time to travel between these links and the subject links. In general, these links have similar correlation values for the range of time lags considered and zero simply happens to have the lowest value so is

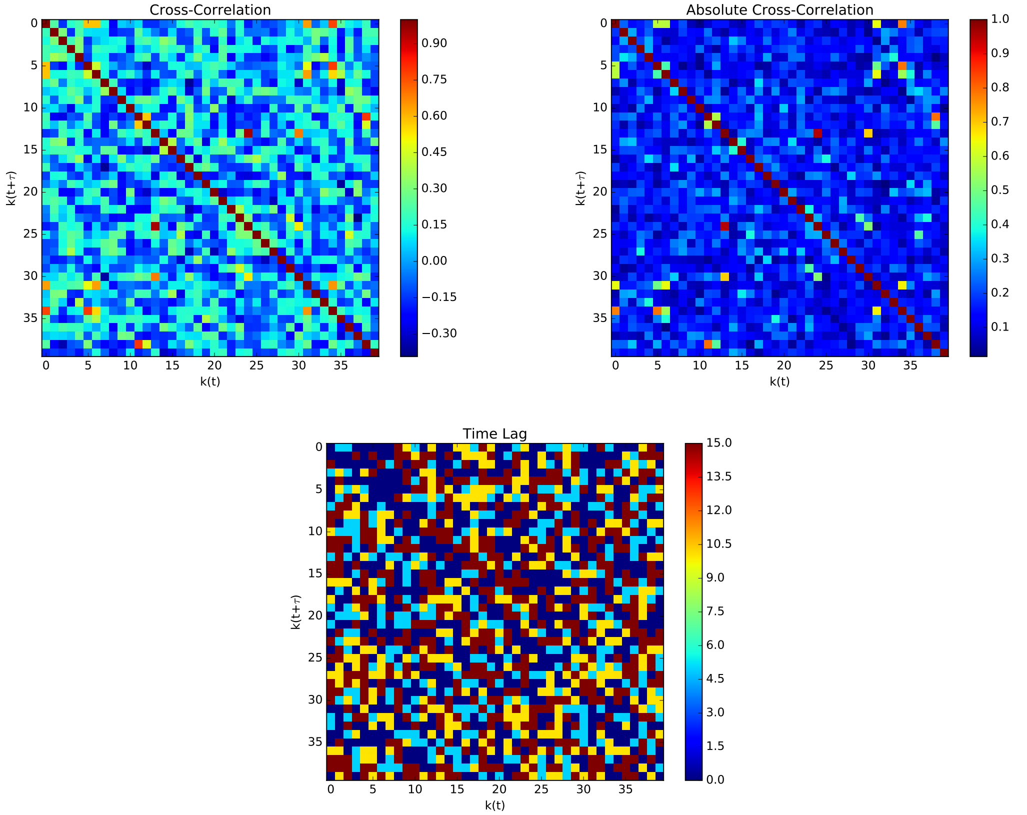

FIGURE 5 Top-left: MCC value between each link pair when optimal time lag is applied. Top-right: Absolute MCC value between each link pair when optimal time lag is applied. Bottom: Optimal time lag for each link pair.

reported here for complete accuracy. However, this does not negate the general trend that optimal time lag is zero for nearby links and higher for links further away.

Link community identification

We now use the cross-correlation values to identify communities of links in the network that behave similarly. Knowledge of these communities can be leveraged to improve prediction of individual link speeds, dynamically partition the network based on how information travels (*32*), or target traffic control in a more intelligent manner.

We first select a threshold MCC value to identify links with a significant connection between them. To ensure that only the most strongly connected links are considered in the community identification, we select the 90th percentile for the observed MCC values, which ensures that only the strongest 10% of the link pairs are considered. Based on the results of FIGURE 4, an MCC threshold of 0.3 was considered here. Note that while this threshold is not very high, previous papers that applied linear detrending methods consider much lower threshold values to identify strongly correlated links. For example, (*8*) used a critical threshold of 0.1 to identify highly correlated locations. Additionally, this threshold was applied to correlation values when considering speeds throughout the entire day. Higher thresholds would be needed to identify the strongest 10%

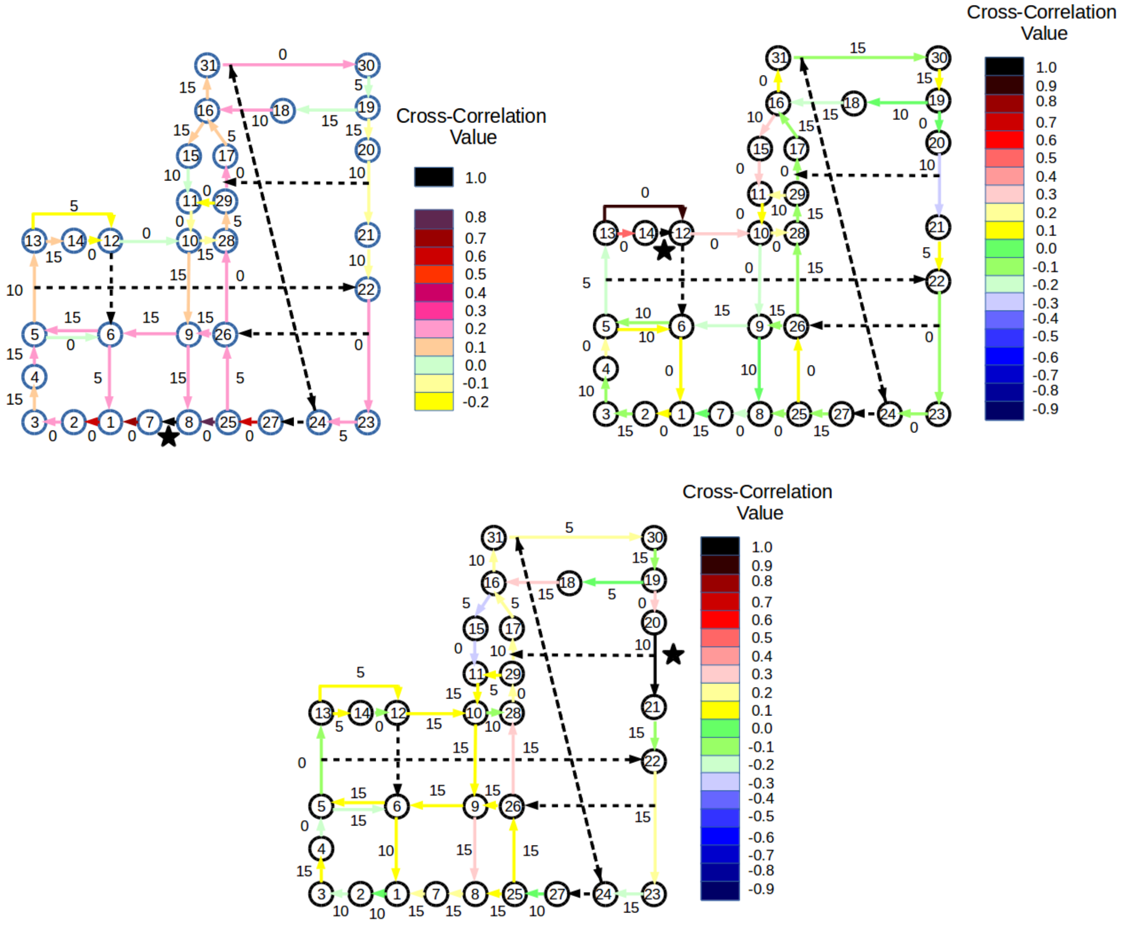

FIGURE 6 Spatiotemporal cross-correlation between links $8 \rightarrow 7$, $14 \rightarrow 12$ and $20 \rightarrow 21$ and other links. The color of each link shows the cross-correlation value between that link the starred link. The number near each link shows the associated τ^c value between the link pair.

of link relations for correlations over shorter time periods; e.g., analysis of link travel speeds during 2-hour periods required an MCC threshold of 0.6 to identify the most correlated link pairs.

The corresponding graph based on the MCC threshold of 0.3 is shown on the left of FIGURE 7. Using this graph, the Infomap technique is used to identify communities within this graph network that contain the most information overlap within the community and with minimal information sharing outside of the community. The resulting 11 communities are illustrated on the right-hand side of FIGURE 7. For each community, one can find the links which have the higher cross-correlation with nodes in the same community. Based on this information, it might be possible to observe traffic speeds on only these 11 most influential links to gain an accurate picture of traffic conditions across the entire network.

To explore the relationships between these communities and the individual links, the 11 communities are also illustrated on map of the network in FIGURE 8. As can be seen, these communities generally consist of links that are in close spatial proximity. Furthermore, communities are generally made up of links on the same arterial (e.g., the black or purple links at the bottom of the network). Again, this is reasonable as one would expect links on certain segments of the same street to behave in the same way. In some cases, a community can be made up of links on nearly parallel streets; e.g., the dark blue links in the middle of the network. This also seems reasonable, especially when vehicles can distribute themselves to respond to lower speeds on one link by taking and alternative route. This re-routing is usually done using both real-time (i.e., on the ground) information and historical information from drivers using the same network over the course of many days.

It should be noted that the lengths of individual links vary considerably. This is due to the fact that the links are pre-defined by the provided of the probe speed data and generally start/end at traffic control devices and merge/diverge points along the roadway. Many of the communities identified exist of shorter links that are on the same roadway segment, and it is not terribly surprising that their speed profiles share similar fluctuations throughout the day. Still, the fact that the algorithm groups these links together helps to verify that it can accurately identify these spatiotemporal patterns. Additionally, some short links on the same roadway segment are not always grouped into the same community; examples include links $2 \rightarrow 3$ and $1 \rightarrow 2$ and links $3 \rightarrow 4$ and $4 \rightarrow 5$. Thus, even though one might expect that these shorter links be always grouped together, on several days this is not the case due to prevailing spatiotemporal speed patterns.

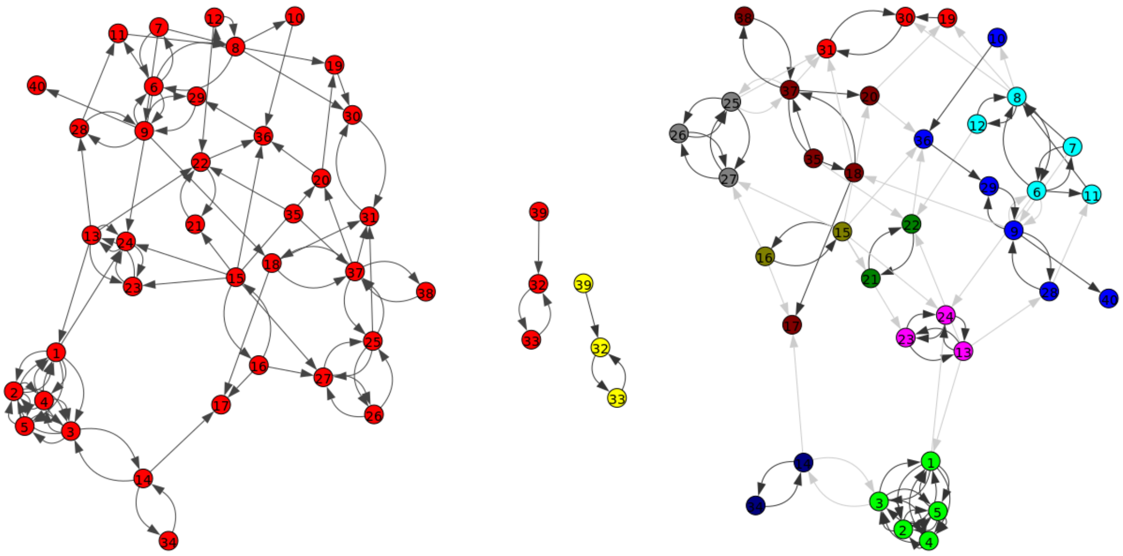

FIGURE 7 Left: Network of the graph. Right: Community of the network

Dependence of the community structure on days of the week

The previous results also considered cross-correlations and communities based on traffic conditions that occurred on Thursdays during the study period. In this section, we repeat the previous analyses to examine how the community structure changes based on the day of the week. The resulting community structures are presented in FIGURE 9 for each of these other days. In this figure, each unique color represents a link community. Interestingly, although there are some slight differences across the seven days in the week, the community structure is remarkably similar across these days. In fact, subsets of some communities persist exactly the same across all days in the week

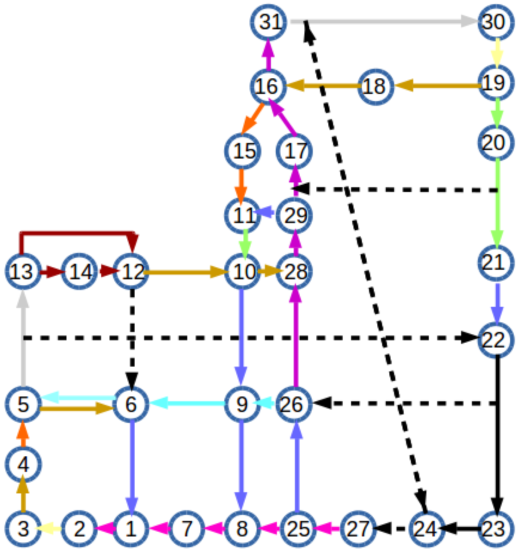

FIGURE 8 Link communities identified based on the data of all Thursdays of March 2018. Links of the same color represent a single community.

(i.e., always share the same color every day). Specifically, the following groups of links (identified by the nodes that they connect) are always included in communities together:

$$\text{Grout } 1 = 27 \rightarrow 25 \rightarrow 8 \rightarrow 7 \rightarrow 1 \rightarrow 2$$
$$\text{Group } 2 = 26 \rightarrow 9 \rightarrow 6 \rightarrow 5$$
$$\text{Group } 3 = 13 \rightarrow 14 \rightarrow 12, 13 \rightarrow 12$$
$$\text{Group } 4 = 3 \rightarrow 4 \rightarrow 5$$

This is very interesting finding given that there are obvious differences in demand patterns that occur across these different days. The significant overlap between the communities created from traffic data on different days of the weeks suggest that the network structure and traffic control properties play a large role in the strong correlations between travel speeds on individual links within the network.

FIGURE 10 shows the number of clusters on each day of week. As illustrated, there are generally more communities on weekdays than on weekends. This might suggest that there is more diversity in speed distribution and traffic phases on weekdays when compared to weekend days, perhaps due to more complicated traffic and demand patterns. The lone exception appears to be Wednesday, which has very few communities. This could be due to prevailing traffic patterns on Wednesday or the use of a variable, day-specific MCC threshold when defining communities. Note that the number of communities on any day is highly influenced by the MCC threshold selected. FIGURE 11 illustrates this relationship showing the average of communities based on threshold for all days of the week. Errors bars represent average deviation of days of the week from the mean value. As expected, the number of communities increases with the threshold selected. This makes sense as larger thresholds results in a sparser network used to identify these communities. The tradeoff here is that larger thresholds helps to identify communities in which speeds are more

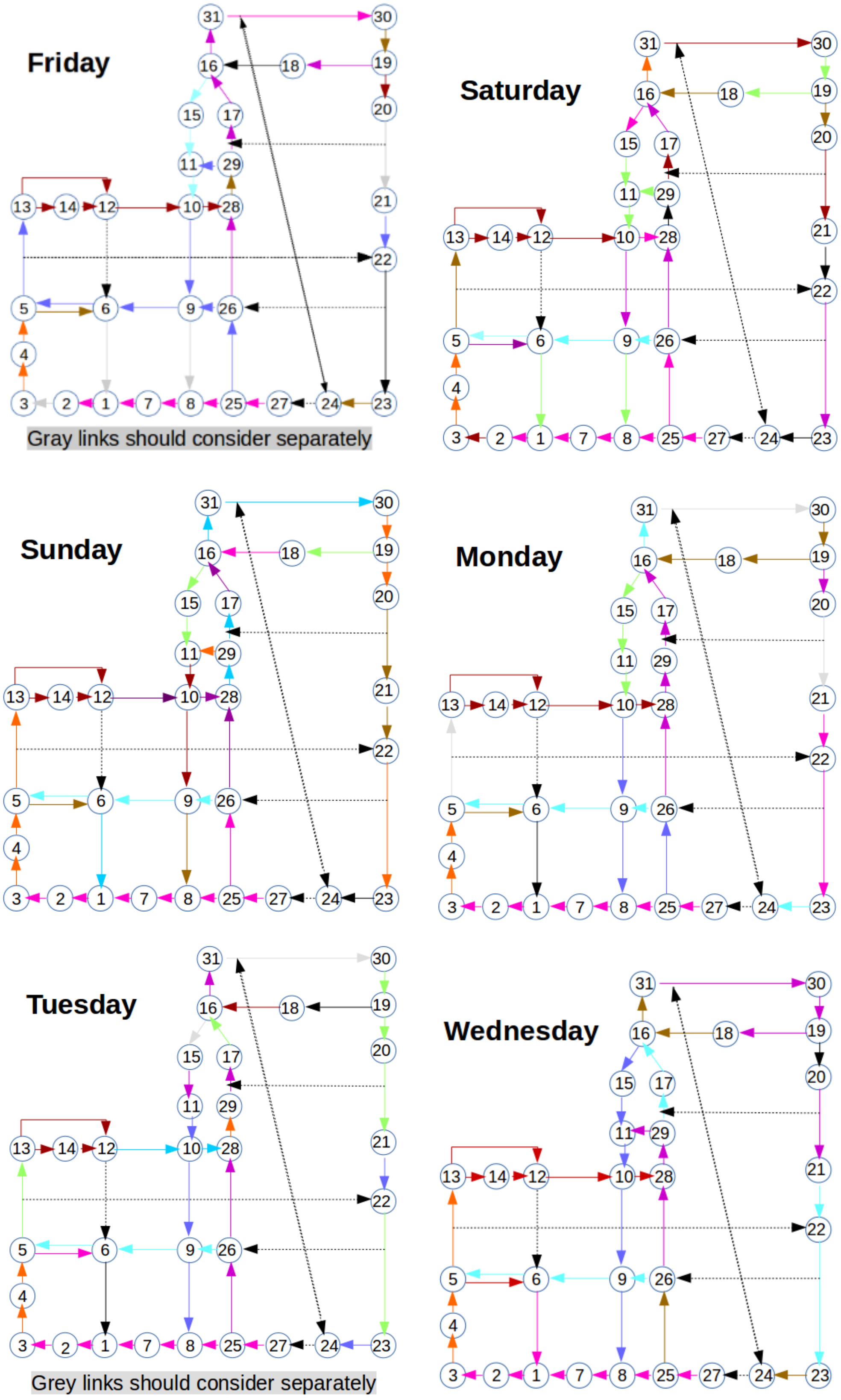
Friday
Gray links should consider separately
Saturday
Sunday
Monday
Tuesday
Grey links should consider separately
Wednesday

FIGURE 9 Community structures in days of a week.

closely related. As the threshold increases, the overlap between the individual links that make up the communities across different days of the week decreases. This makes sense: the more strict the criteria becomes for identifying the communities, the less likely these communities will be similar across the days of the week. Nevertheless, there is still significant overlap even when the critical threshold is increased above the value used for FIGURE 10, which suggests strong spatiotemporal correlations in link travel speeds exists in urban traffic networks.

Similar findings were also obtained when examining community structures for shorter time periods during a single day. For example, the Thursday time period was broken up into 7 two-hour periods and cross-correlation values obtained for each of these shorter time intervals were used to identify community structures during each interval. In doing so, the critical threshold used to identify strong correlations had to be increased (from 0.3 to 0.6) since correlations were generally stronger during the shorter time intervals. The results (omitted here for brevity) revealed that the community structure for these shorter intervals differed from those obtained for the entire day. However, significant overlap still existed among these two-hour communities and groups of links were still common among the daily and two-hour communities. The results confirm the existence of strong spatiotemporal correlations that result in similar traffic behaviors across links in a network.

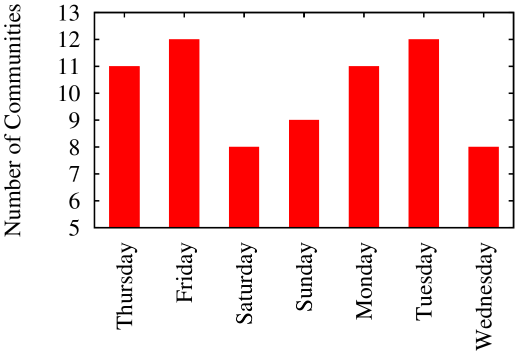

FIGURE 10 Number of communities in each day of a week using the 0.3 MCC threshold.

LEVERAGING COMMUNITY STRUCTURE TO IMPROVE SPEED PREDICTION

This section illustrates how the knowledge of a link's community can be used to improve overall travel speed prediction by removing extraneous information. The specific prediction applied here is a Gaussian Process (GP) model, which is a supervised learning technique used to solve regression and probabilistic classification problems based on Bayesian inference. This model type can solve hard machine learning problems due to its flexible, non-parametric nature and relative computational simplicity (*33*). This method has been previously applied in many fields, such as marketing (*34*), biology (*35*), and industrial engineering (*36*). The specific details on this modeling technique are omitted here for brevity but can be found in the preceding references.

Travel speed predictions for links $1 \rightarrow 2$ and $13 \rightarrow 14$ are performed using the following three sets of input data used to train and apply the GP model:

1) Speed data from all other links of the network.
2) Speed data only from the links in the subject link's community.

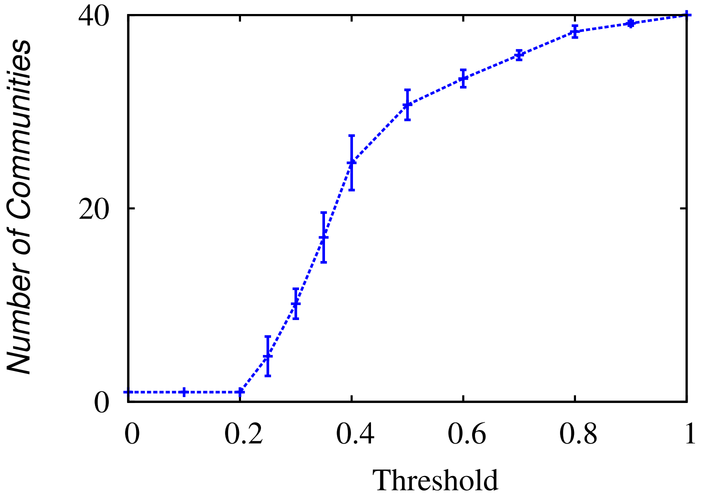

FIGURE 11 **Number of communities and average deviation for different threshold values over different days of the week.**

Link	RMSD-Network	RMSD-Community	RMSD-Most important link of the community
1 → 2	7.07	7.03	7.94
13 → 14	5.83	3.07	4.13

TABLE 1 **RMSD based on different methods**

3) Speed data from the most important link in the subject link's community.

In all methods, speed data from 6 AM to 8 PM on four days $(1, 8, 15, 22$ of March $2018)$ are used to train the model. Data from the last day (29 March 2018) are then used to validate the predictions.

FIGURE 12 illustrates the prediction results on the two subject links along with the observed speeds for comparison. As shown, all methods seem to predict the data fairly well, although some abnormal outliers are not well-captured for all models. The root-mean-square-deviation (RMSD) is also used to quantify the overall prediction performance across the entire one-day prediction horizon. TABLE 1 provides these values for the three methods applied to both links. Notice that prediction using speed data from the subject link's community outperforms speed predictions using speed data from all links in the network. This occurs because data from all links in the network will include some non-useful information into the prediction model, and this decreases the efficiency of the prediction algorithm. The non-useful information belongs to links that are in other communities, which are more or less independent of the subject links. In other words, adding the information of the links that have low correlation with the study link will only serve to damage the prediction result. Thus, the community structure can be leveraged to make sure only the most useful information is included in the prediction model.

CONCLUDING REMARKS

This paper proposes the use of novel detrending methods to study spatiotemporal correlation of link travel speeds on an urban traffic network using probe vehicle data. The use of non-parametric moving-average detrending methods to identify temporal trends in individual link speed data while accounting for seasonality effects and classification of detrended data into link communities represents a significant contribution over previous studies in this area. The proposed approach can readily account for non-linear trends and does not require stationary traffic conditions. Thus, this method can be applied to both short time periods during which traffic conditions are relatively stable and long time periods during which traffic conditions might vary widely; e.g., an entire day including morning and evening peak periods as demonstrated here. After detrending the data, correlations between individual link speeds were found to improve when a time lag was introduced (i.e., when using cross-correlations). The time lag better accounts for the physics of traffic and the fact that information takes time to travel between links in a transportation network.

The cross-correlation values were then used to identify groups of links (communities) that behaved similarly with respect to traffic speeds throughout the day using graph theory methods. For links in a community, traffic states throughout the community can be estimated fairly well using only data from a single link. For example, since there are 11 communities identified for Thursdays, travel speeds on all 40 links in the network can be estimated fairly well using only data from 11 links and travel speeds on links in a given community can be used to predict speeds on other links within that community. This latter fact is verified using a simple Gaussian Process speed prediction model.

The number of communities generally changed from day to day and weekdays were shown to have higher number than weekends, perhaps due to more complex travel patterns. However, there was remarkable similarity among the community structures across the days of the week. In fact, three groups of links were identical across all days. This similarity suggests that the spatiotemporal relationships between travel speeds on individual links might representing underlying relationships about the structure of the network, in addition to travel demand patterns. It is worth noting that existing network partitioning methods can also be used to identify these groups/communities if applied to data from different days; however, these methods generally do not apply (non-linear) detrending methods to account for the impact of daily traffic patterns.

While this study revealed interesting findings about spatiotemporal correlations that exist between travel speeds on individual links in a network across the days of the week, additional work is needed to verify that these relationships hold during other time periods. For example, this study only considered speed data obtained for a one-month period. A more extensive dataset is required to ensure that these patterns hold for longer time periods that might include seasonal variations in travel demands. Every effort was also made to use traffic data taken from links with the same land use patterns. However, due to the nature of the probe data, links were often of different lengths. Further research in this area should consider links with more uniform lengths. Additionally, changes in land use patterns should be considered when applying these methods to larger spatial regions.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Transportation Systems Management and Operations (TSMO) group at the Pennsylvania Department of Transportation for their guidance and access to probe speed data that were used in this work. The research was supported by NSF Grant CMMI-1749200.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: R. Kouhi and V. Gayah; data collection and analysis: R. Kouhi; interpretation of results: R. Kouhi and V. Gayah; draft manuscript preparation: R. Kouhi and V. Gayah. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Ma, X., Z. Tao, Y. Wang, H. Yu, and Y. Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, Vol. 54, Elsevier, 2015, pp. 187–197.

[2] Ma, X., H. Yu, Y. Wang, and Y. Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, Vol. 10, No. 3, Public Library of Science, 2015, p. e0119044.

[3] Peng, C., X. Jin, K.-C. Wong, M. Shi, and P. Liò. Correction: Collective Human Mobility Pattern from Taxi Trips in Urban Area. *PloS one*, Vol. 7, No. 8, Public Library of Science, 2012.

[4] Noulas, A., S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. Correction: a tale of many cities: universal patterns in human urban mobility. *PloS one*, Vol. 7, No. 9, Public Library of Science, 2012.

[5] *Bad Data Costs the U.S.* $3 *Trillion Per Year*, 2016. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

[6] Gao, S. and I. Chabini. Optimal routing policy problems in stochastic time-dependent networks. *Transportation Research Part B: Methodological*, Vol. 40, No. 2, Elsevier, 2006, pp. 93–122.

[7] Goel, P. K., M. R. McCord, and C. Park. Exploiting correlations between link flows to improve estimation of average annual daily traffic on coverage count segments: Methodology and numerical study. *Transportation research record*, Vol. 1917, No. 1, SAGE Publications Sage CA: Los Angeles, CA, 2005, pp. 100–107.

[8] Ermagun, A. and D. Levinson. *Spatiotemporal traffic forecasting: Review and proposed directions. Retrieved from the University of Minnesota Digital Conservancy*, 2016.

[9] Okutani, I. and Y. J. Stephanedes. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B: Methodological*, Vol. 18, No. 1, Elsevier, 1984, pp. 1–11.

[10] Stathopoulos, A., L. Dimitriou, and T. Tsekeris. Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 23, No. 7, Wiley Online Library, 2008, pp. 521–535.

[11] Chandra, S. and H. Al-Deek. Cross-correlation analysis and multivariate prediction of spatial time series of freeway traffic speeds. *Transportation Research Record: Journal of the*

Transportation Research Board, , No. 2061, Transportation Research Board of the National Academies, 2008, pp. 64–76.

[12] Kamarianakis, Y. and P. Prastacos. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 1857, Transportation Research Board of the National Academies, 2003, pp. 74–84.

[13] Cai, P., Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*, Vol. 62, Elsevier, 2016, pp. 21–34.

[14] Sun, S., C. Zhang, and Y. Zhang, Traffic flow forecasting using a spatio-temporal bayesian network predictor. In *International conference on artificial neural networks*. Springer, 2005, pp. 273–278.

[15] Ermagun, A., S. Chatterjee, and D. Levinson. Using temporal detrending to observe the spatial correlation of traffic. *PloS one*, Vol. 12, No. 5, Public Library of Science, 2017, p. e0176853.

[16] Imani, F., A. Gaikwad, M. Montazeri, H. Yang, and P. Rao. Layerwise in-process quality monitoring in laser powder bed fusion. *ASME Paper No. MSEC*, Vol. 6477, 2018.

[17] Imani, F., A. Gaikwad, M. Montazeri, P. Rao, H. Yang, and E. Reutzel. From Process Condition to Build Quality through Modeling and Monitoring of In-process Layerwise Images in Laser Powder Bed Fusion Additive Manufacturing Process. *Journal of Manufacturing Science and Engineering*, 2018.

[18] Marple, S. L. and S. L. Marple. *Digital spectral analysis: with applications*, Vol. 5. Prentice-Hall Englewood Cliffs, NJ, 1987.

[19] Kingma, D. P. and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Shih, Y.-T., H.-M. Cheng, S.-H. Sung, W.-C. Hu, and C.-H. Chen. Application of the N-Point Moving Average Method for Brachial Pressure Waveform–Derived Estimation of Central Aortic Systolic Pressure. *Hypertension*, Am Heart Assoc, 2014, pp. HYPERTENSIONAHA–113.

[21] Chen, F., K. Tian, X. Ding, Y. Miao, and C. Lu. Finite-size effect and the components of multifractality in transport economics volatility based on multifractal detrending moving average method. *Physica A: Statistical Mechanics and its Applications*, Vol. 462, Elsevier, 2016, pp. 1058–1066.

[22] Ozaktas, H. M., B. Barshan, D. Mendlovic, and L. Onural. Convolution, filtering, and multiplexing in fractional Fourier domains and their relation to chirp and wavelet transforms. *JOSA A*, Vol. 11, No. 2, Optical Society of America, 1994, pp. 547–559.

[23] Burrus, C. S. and T. Parks. *and Convolution Algorithms*. Citeseer, 1985.

[24] Ronhovde, P. and Z. Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, Vol. 80, No. 1, APS, 2009, p. 016109.

[25] De Leo, V., G. Santoboni, F. Cerina, M. Mureddu, L. Secchi, and A. Chessa. Community core detection in transportation networks. *Physical Review E*, Vol. 88, No. 4, APS, 2013, p. 042810.

[26] Harenberg, S., G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, K. Padmanabhan, and N. Samatova. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 6, No. 6, Wiley Online Library, 2014, pp. 426–439.

[27] Qian Ge, P. W. and D. Fukuda. A community detection method for identifying neighborhoods of MFD. *Hong Kong Society of Transport Studies annual conference*, Hong Kong(HKSTS),oral, 2016.

[28] Lopez, C., P. Krishnakumari, L. Leclercq, N. Chiabaut, and H. Van Lint. Spatiotemporal Partitioning of Transportation Network Using Travel Time Data. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 2623, Transportation Research Board of the National Academies, 2017, pp. 98–107.

[29] Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, Vol. 74, No. 3, APS, 2006, p. 036104.

[30] Zare, H., P. Shooshtari, A. Gupta, and R. R. Brinkman. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC bioinformatics*, Vol. 11, No. 1, BioMed Central, 2010, p. 403.

[31] *Source code for multilevel community detection with Infomap*, 2017. http://www.mapequation.org/code.html.

[32] Ji, Y. and N. Geroliminis. On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological*, Vol. 46, No. 10, Elsevier, 2012, pp. 1639–1656.

[33] Seeger, M. Gaussian processes for machine learning. *International journal of neural systems*, Vol. 14, No. 02, World Scientific, 2004, pp. 69–106.

[34] Patel, J., S. Shah, P. Thakkar, and K. Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, Vol. 42, No. 1, Elsevier, 2015, pp. 259–268.

[35] Tonner, P. D., C. L. Darnell, B. E. Engelhardt, and A. K. Schmid. Detecting differential growth of microbial populations with Gaussian process regression. *Genome research*, Cold Spring Harbor Lab, 2016.

[36] Imani, F., C. Cheng, R. Chen, and H. Yang, Nested Gaussian Process Modeling for High-Dimensional Data Imputation in Healthcare Systems, 2018.

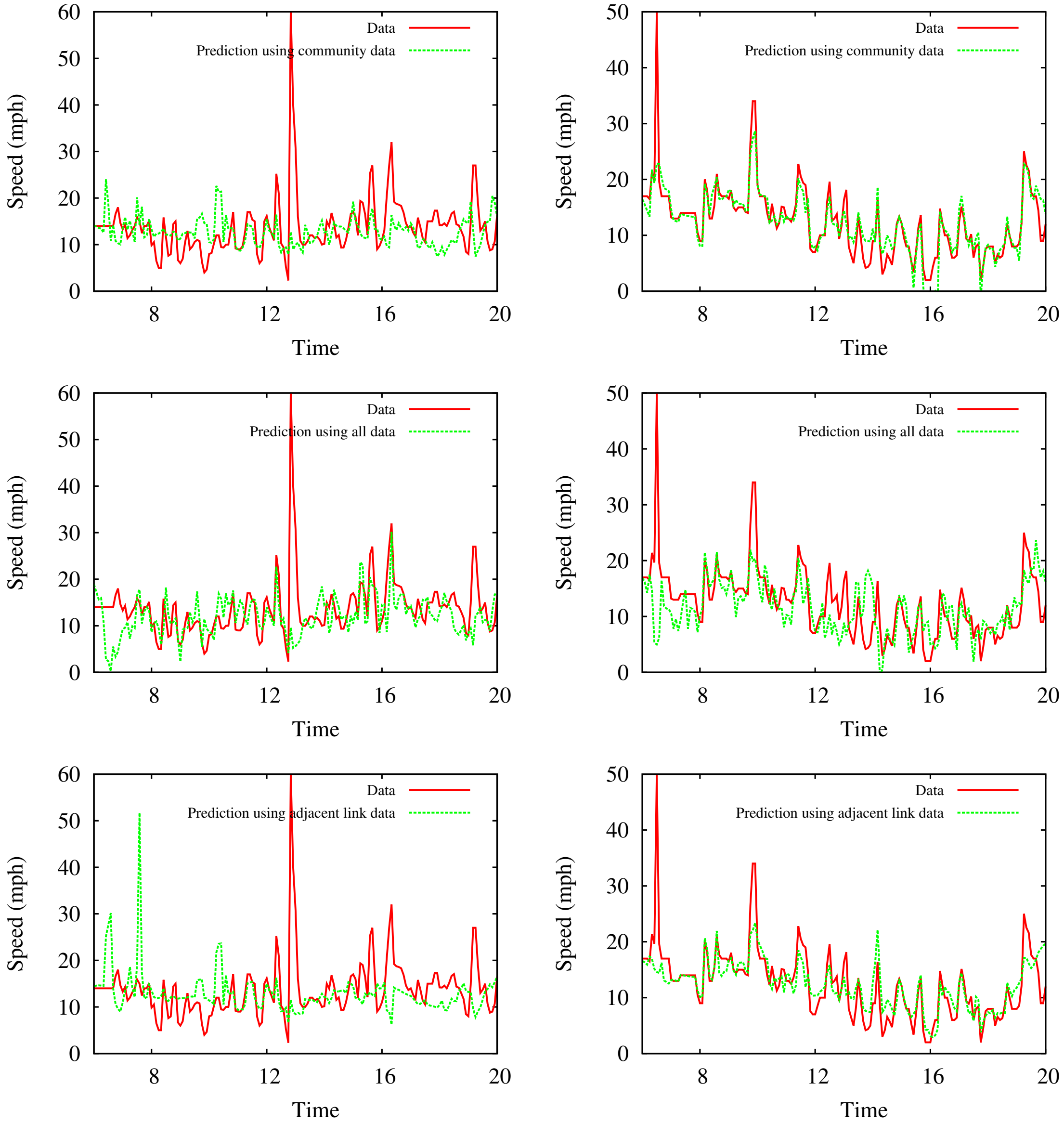

FIGURE 12 Prediction results for links $1 \rightarrow 2$ (Left column) and $13 \rightarrow 14$ (right column) Top row: Using only speed data from other links in the community for prediction. Middle row: Using speed data from all link in the network for prediction. Bottom row: Using speed data only from the most important link of the community for prediction.