RC25230 (W1111-031) November 2, 2011
Mathematics

IBM Research Report

Real-Time Traffic Prediction Using GPS Data with Low Samping Rates: A Hybrid Approach

Wei Shen, Laura Wynter
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
wshen@us.ibm.com
lwynter@us.ibm.com

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Abstract

This paper presents an approach for real-time traffic speed prediction via GPS speed readings. The approach combines techniques from data mining with traffic speed estimates available from other sources. In particular, we consider GPS data that is provided in the form of point speeds, rather than trajectories. This is the case when GPS data from consumers is sampled at discrete points by a service provider, e.g. to protect privacy of the consumers by not permitting a reconstruction of their trajectories. In the context studied in this paper as well as others observed in practice, such GPS sampling rates are quite low and hence the GPS-based speed readings can be quite unreliable. Our method recognizes this fact and uses the GPS speed readings in a novel way in conjunction with another source of speed data for the network. The example studied is drawn from the 2010 IEEE International Competition on Data Mining (ICDM) traffic prediction competition, in which the authors were part of a team that finished second worldwide.

Keywords: traffic prediction, GPS, hybrid, nearest neighbor, cross-validation

1 Introduction

Real-time road traffic prediction is an important component of modern traffic management and information systems. Reliable prediction of near-term traffic conditions in road networks allows traffic management agencies to generate proactive traffic operation strategies to alleviate congestion and allows both public agencies and private companies to provide accurate travel time estimates to road users.

In the past two decades, much research effort has been invested in developing accurate and robust traffic prediction models. The modeling approaches can be classified into parametric methods and non-parametric methods. The former category relies primarily on statistical techniques, including historical average and smoothing techniques (e.g., Smith and Demetsky, 1997; Williams et al., 1998), autoregressive moving average models (e.g., Ahmed and Cook, 1979; Levin and Tsao, 1980; Al-Deek et al., 2001; Smith et al., 2002; Kamarianakis and Prastacos, 2003; Min and Wynter, 2011), and Kalman filter algorithms (e.g., Okutani and Stephanedes, 1984; Guo and Williams, 2010). The main non-parametric approaches published to date include non-parametric regression (e.g., Smith and Demetsky, 1996; Clark, 2003; Huang and Sadek, 2009) and artificial neural networks (ANN) (e.g., Clark et al., 1993; Vythoulkas, 1993; Yun et al., 1998; van Lint et al., 2005; Vlahogianni et al., 2005; Khosravi et al., 2011). See Vlahogianni et al. (2004) for a somewhat dated review of different traffic prediction models.

The above approaches have been designed in general for traffic prediction based on fixed-location data sources such as inductive loops, roadside radar sensors, and traffic cameras. When using fixed-location sensor data only, taking into account spatial and temporal correlations of traffic flow can be done in a straightforward manner. For example, traffic measurements on neighboring links can be used to formulate a univariate or mutlivariate linear/nonlinear prediction model.

While fixed-location traffic sensors are commonly available on expressways and other major roadways, real-time traffic data using fixed-location sensors is far less ubiquitous on urban networks. The proliferation of GPS devices in fleets of vehicles, passenger cars in the form of in-vehicle navigation systems, and more recently from applications on smart phones, has led to an increasing emphasis on using GPS data for the identification of traffic speeds on the roads. Some products exist in the marketplace to determine traffic "color maps" for GPS-enabled mobile phones equipped with dedicated applications that transmit periodically locations to a server. Such traffic estimation is related to the traffic prediction problem that we are interested in but is different in two important ways. On the one hand, we are interested in quantitative traffic prediction that produces a set of future speeds on the various road links, rather than the types of ranges produced for use on "color maps". Secondly, we are interested in going beyond real-time traffic estimation to future traffic prediction, which in general requires more data than the real-time estimation problem.

There are two major issues in using GPS data for traffic prediction. First, disclosing detailed vehicular trajectory data is always associated with privacy and security concerns. Even if trajectory data is broadcasted in an anonymous manner, it is still possible to identify individuals from the trajectory data (Hoh et al., 2006). Hence, in some cases, GPS measurements are sampled at random times/locations/devices so that vehicular trajectories cannot be inferred. One example of such an approach is the Virtual Trip Line proposed by Hoh et al. (2008), which is essentially a spatial trigger for GPS devices to collect

and report measurements when pre-defined virtual lines in the network are crossed. The data collection procedure for our study is similar in nature to the idea of Virtual Trip Line. The only difference is that sampling is performed on devices rather than on locations. More specifically, for each given interval, only a sampled collection of GPS devices report their location and speed information. The second challenge is that traditional traffic prediction methods which are based on reliable traffic observations from fixed locations are usually inapplicable in this context. Indeed, the proportion of vehicles who disclose their location or speed information to any given application provider is typically very low and hence the amount of real-time and historical data on each road link is insufficient for standard traffic prediction approaches to apply.

The method presented in this paper is based on the approach developed for such a problem as part of the 2010 IEEE International Conference on Data Mining (ICDM) TomTom Traffic Prediction Contest for Intelligent GPS Navigation. The authors were part of the team which placed second worldwide in that contest. The remainder of this paper is structured as follows. Section 2 introduces the problem setting and the data used for developing our approach. Section 3 describes the hybrid method developed. Section 4 presents the prediction performance of the proposed method on the test data set, and Section 5 concludes the paper by summarizing the major findings.

2 Problem Setting and Data

The traffic prediction problem investigated in this paper is as follows: Suppose that GPS data is available from a sampling of drivers at random points on a road network. Suppose further that trajectory data cannot be gathered, as is the case with numerous smartphone applications that sample user locations and speeds at separate points in time far enough apart to be unable to redraw the users' paths. This is done primarily to protect the users' privacy as part of an opt-in approach to the application. Since the data is available only to the particular application provider, the subset of the population available to the provider is limited. In addition, given the sampling that the provider agrees to do to protect users' privacy, the available population is reduced further still.

Because of the sparsity of available signals on all of the urban road links at all points in time, a secondary source of data is useful. Specifically, we make use of a set of additional information on average travel speeds for the links of interest, which may have been obtained from other sources but not available as a fixed-sensor-based real-time data feed. We term these travel times "historical" because in our setting they are available for the training data but not as part of the testing data set.

The goal, as stated previously, is to perform near-term prediction of traffic speeds on selected road links.

The specific data used in this paper was obtained as part of the 2010 IEEE ICDM Traffic Prediction competition. In this case, the GPS data was generated by a traffic simulation program rather than from live users. The simulator, called the Traffic Simulation Framework (TSF), was developed at the University of Warsaw, and is based on the well-known Nagel-Schreckenberg's cellular automata traffic flow models (Nagel and Schreckenberg, 1992). The functionality of TSF was described in detail in Gora (2009).

Traffic simulation was performed for the network of Warsaw, Poland (Figure 1), and contains in all 18716 nodes and 35170 links. Among the links, 8631 are classified as major roads, and 100 are further termed "critical" links. Our goal was to perform traffic prediction on the 100 critical links.

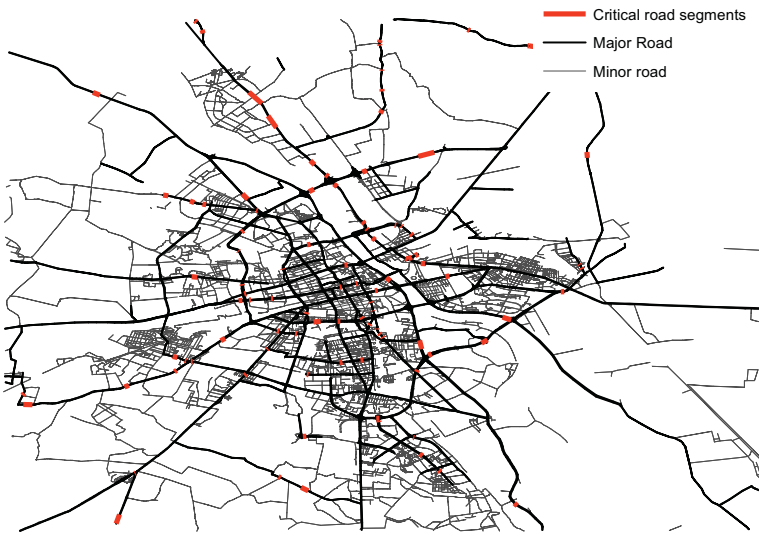


Figure 1: Network of Warsaw, Poland

The simulation was run for essentially 500 simulation-hours, thereby representing 50 10-hour daytime periods on similar days, such as 10 weeks' worth of five-day weeks, during the daytime hours. That data was considered to be historical data to be used for training. An additional set of data of the same quantity was generated and was used as testing data. However, for each hour of the testing data set, only GPS information of the first 30 minutes of each of the 500 hours was revealed. In terms of simulated sampled GPS data, 1% of the vehicles simulated were sampled during each 10-second interval and the instantaneous speed and location recorded.

Each GPS record contains a timestamp, latitude and longitude coordinates, and the instantaneous speed of the sampled vehicle. Before any prediction model can be applied, a procedure is needed to map GPS location to the road segments on the network. Since GPS data are generally noisy, the reported coordinates may not necessarily fall precisely on any link. Map-matching algorithms (e.g., Greenfeld, 2002; Alt et al., 2003) are therefore needed to accurately approximate the location of the GPS points on the links. Since the simulated traffic data was less noisy, we were able to employ the built-in spatial join function in ArcGIS for the map matching function; specifically, each GPS data point was associated with the closest link within a 20-meter radius. If no such link is found, the GPS data point was discarded. To reduce computational time, only major links were included as candidate links for matching. The map-matching procedure was performed for the GPS data points of both the training and test data sets. Roughly 80% of the GPS points were able to be map-matched of the approximately 280,000 GPS points provided during an average training hour.

As the secondary source of average-case data, 6-minute (harmonic) average speeds were provided for all of the 100 critical links for each hour of the training data. No average-case speed information

was available for the testing data to reflect the fact that such data is not typically available on the links in question as a real-time feed.

As part of the IEEE ICDM competition, the objective was to perform two traffic speed predictions on the full set of 100 critical links: a 6-minute-ahead prediction and a 30-minute-ahead prediction, in both cases with the speeds being harmonic averages of the preceding 6 minutes.

Figure 2 illustrates all the GPS points collected during a sample 10-hour simulation cycle. As shown, GPS data points cover most of the Warsaw network, with a preponderance on the central region.

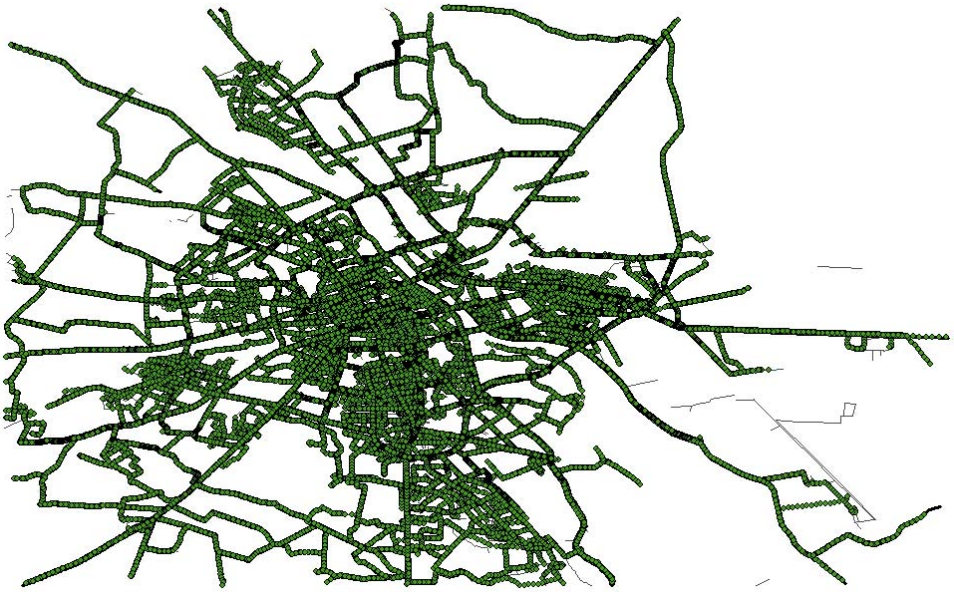
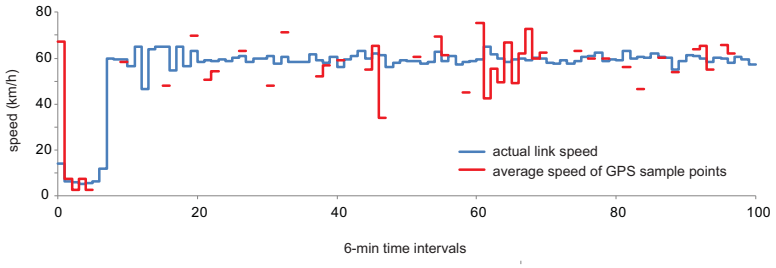


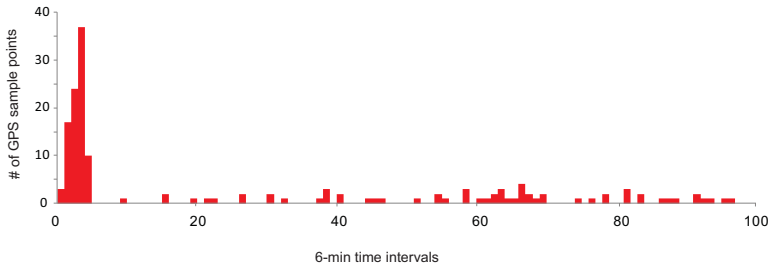
Figure 2: GPS points in a 10-hr simulation cycle

Suppose the speed estimate for a link during any 6-min interval is regarded as missing only if no GPS sample points fall on the link during the corresponding time interval. In the training data, the overall missing data percentage is 76.94% on the major links and 68.24% on the 100 critical links. In fact, 22 of the 100 critical links have no GPS records at all in any of the training data history.

For time intervals with at least one GPS sample point on a critical link, we investigate whether the GPS instantaneous speed records provide reliable estimates of the link-level space-mean, or harmonic-averaged, speed readings available as part of the "historical" data on speeds from a "different" data source. A comparison of the average of the GPS-based speeds and the historical average speeds on two sample links during a randomly chosen 10-hour training period is shown in Figure 3(a) and Figure 4(a). The average of the GPS-based speed readings is depicted by a red line, and gaps on the red lines mark the time intervals without any GPS speed samples. In addition, the actual number of GPS samples collected for each corresponding 6-min interval on these two sample links are shown in Figure 3(b) and Figure 4(b).

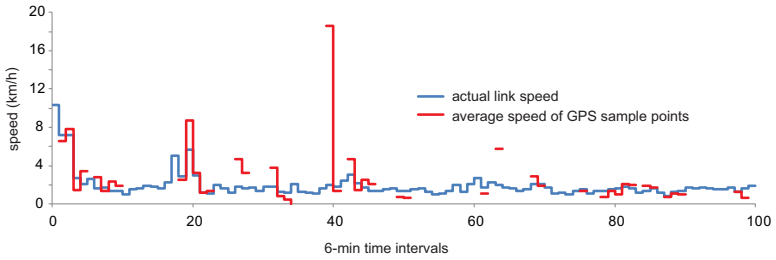


(a) Actual link speed v.s. average speed of GPS sample points

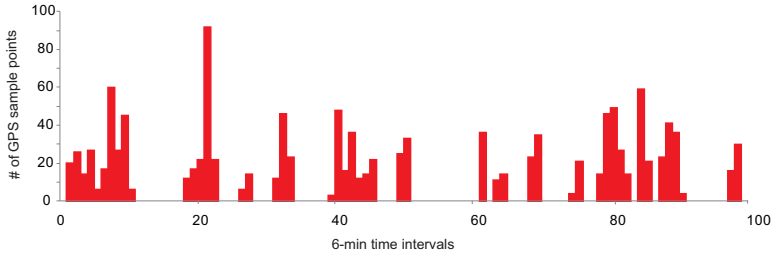


(b) # of GPS sample points during each 6-min interval

Figure 3: Comparison of the actual speed profile and speed estimates based on GPS samples on link 1 during one simulation cycle



(a) Actual link speed v.s. average speed of GPS sample points



(b) # of GPS sample points during each 6-min interval

Figure 4: Comparison of the actual speed profile and speed estimates based on GPS samples on link 2 during one simulation cycle

Figures 3 and 4 highlight some interesting features of the GPS-based speed data set. First, the red lines of both links have many gaps, indicating that many time intervals have no GPS records. Secondly, for time intervals which do have GPS sample points, the speed estimates based on the limited number of

available individual speed samples can be very far from the link speed provided by the other data source, which we consider to be accurate. Finally, observe that time intervals with lower average speed tend to have more GPS points. This is likely to occur as congested links usually accommodate more vehicles at any point in time, and hence are likely to contain more GPS samples.

3 Methodology

A straightforward approach for traffic prediction based on GPS data is to construct good estimates of link speed for each time interval on each link, based on its corresponding GPS speed samples, and treat them as fixed location observations. Then any number of traffic prediction approaches could be used on the aggregated data.

The question is whether this method is feasible with such a low sampling rate (i.e., 1%). Figures 3 and 4 imply that such a method will not fare well due to the unreliable nature of the sampled GPS speed data. As such, averaging the values does not tend to replicate observed speeds from other data sources.

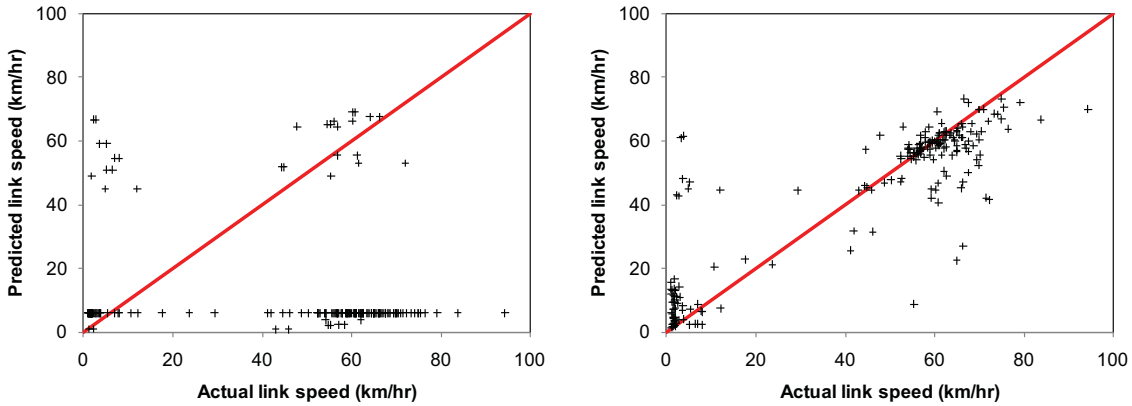
We construct two baseline values for use in the prediction step. The first baseline corresponds also with that provided by the contest and is constructed based on the GPS data only. Namely, the average speed of all the vehicles passing through each critical link in the first 30min interval of an hour is used as the 6min-ahead and 30min-ahead predicted link speed. If no GPS point is recorded on a critical link during that initial 30min interval, the average speed over all the GPS sample points in the network is used as the baseline for that link. The second baseline is based exclusively on the historical observed link speed. Namely, the average speed of the corresponding time intervals in an hour (i.e., 30-36 min, and 54 - 60min) for a given link and during all the training hours are used as the 6min-ahead and 30min-ahead prediction.

Prediction accuracies are evaluated by calculating the root mean squared error (RMSE) of the inverse of a prediction. That is, predicted speeds are transformed - through inverting and multiplying by 60 - into predicted travel time over 1km of the road segment, expressed in minutes. These travel times are then compared with speed as provided through an alternative data source, using the RMSE measure.

Using the actual link speed provided by the contest website for post-competition analysis, we can easily calculate the prediction performance of the inverse of the predictions for both baselines. The RMSE values for the first and second baselines are 18.065min/km and 14.785min/km, respectively.

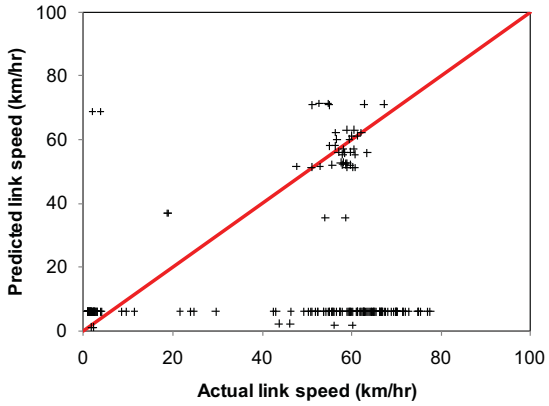
Figure 5 further illustrates the predicted v.s. the actual link speed for the two baselines on two sample hours. The X-Y coordinates of each point in Figure 5 represent the predicted and actual speed on one critical link. Both the 6min-ahead and 30min-ahead predictions are included.

The four illustrations in Figure 5 indicate that the second baseline is superior to the first. As shown in Figures 5(a) and 5(c), it is quite common that no GPS samples appear on some critical links during the entire 30min interval. For those links, predicted values based on network wide average GPS sample speed can be very far from the actual. The better prediction performance of the second baseline may be due to the recurrent feature of traffic evolution and congestion development in road networks. As is reported frequently in the literature, historical data can carry valuable information for predicting future traffic condition during the analogous time intervals.

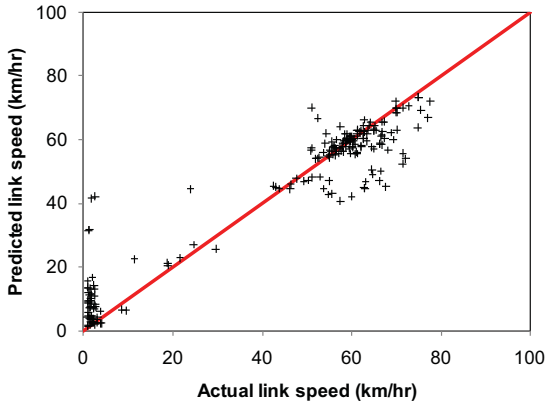


(a) Baseline I, sample test hour 1

(b) Baseline II, sample test hour 1



(c) Baseline I, sample test hour 2



(d) Baseline II, sample test hour 2

Figure 5: Prediction performance of two baselines

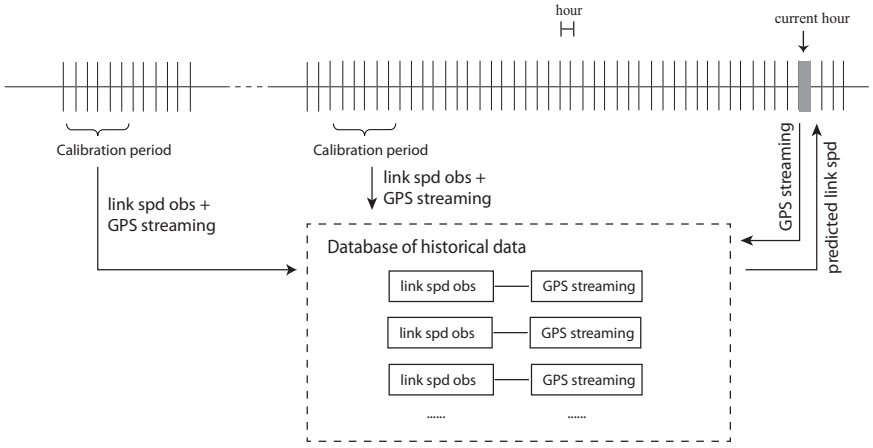


Figure 6: Data framework of the prediction model

Our hybrid approach is thus motivated by this observation: a long-running data source with broad coverage but low sampling rate (i.e., GPS records) along with another periodic short-term data source

which collects traffic observations on critical links may be combined to generate reliable traffic predictions. As shown in Figure 6, during calibration periods, actual link speed observations on critical links are collected. Together with the GPS data received during the same period, these data are stored as prediction candidates. The GPS records received in real time can then be used to determine which prediction candidate is most appropriate.

The data analysis indicated that although the GPS speed samples may not be sufficient to construct reliable speed estimates, the *number* of GPS samples received in each time interval may be a good indicator of how congested the network is, both globally at the network level and locally at the link level.

Figures 7 and 8 plot the relationship between the total number of GPS samples received globally during the entire 30min interval and the actual 6min-ahead (30min-ahead) speed on six sample critical links, as obtained from the alternative data source. Each point in a plot corresponds to 1 out of 500 simulation hours in the training data set.

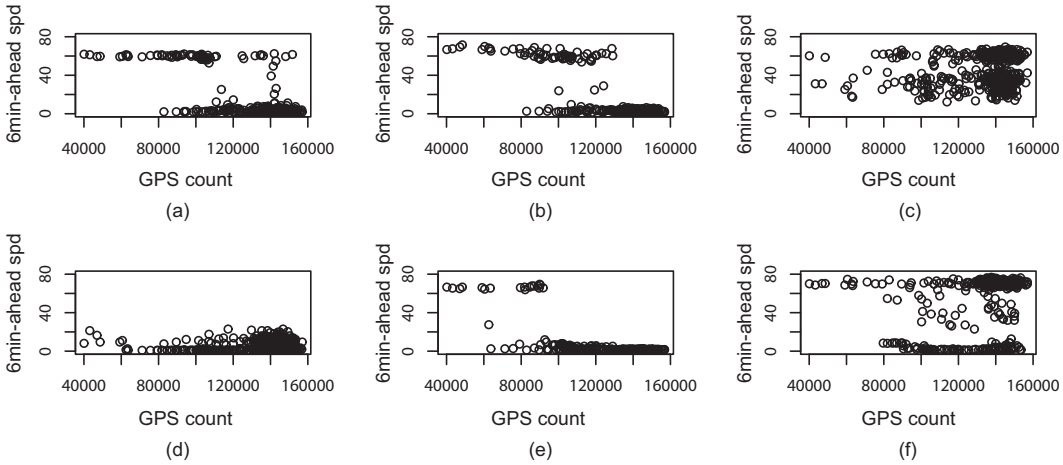


Figure 7: 30min GPS count v.s. 6min-ahead speed

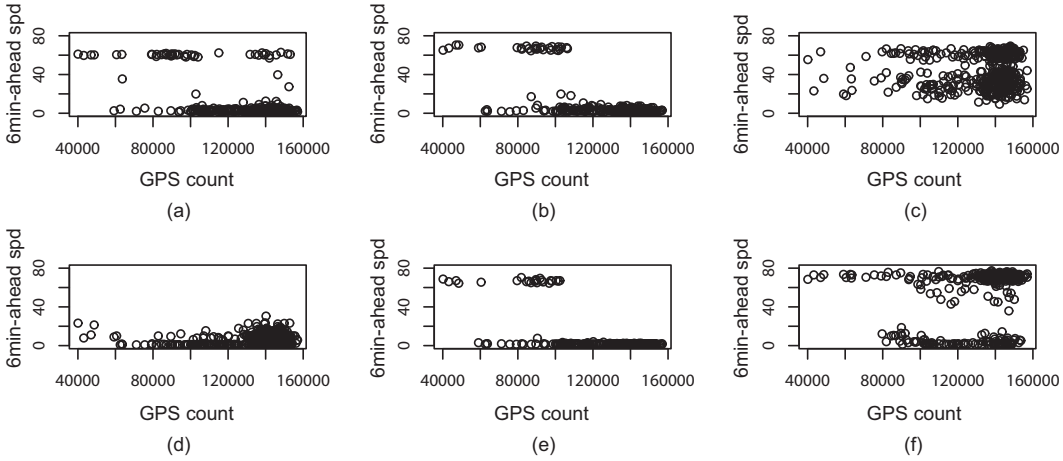
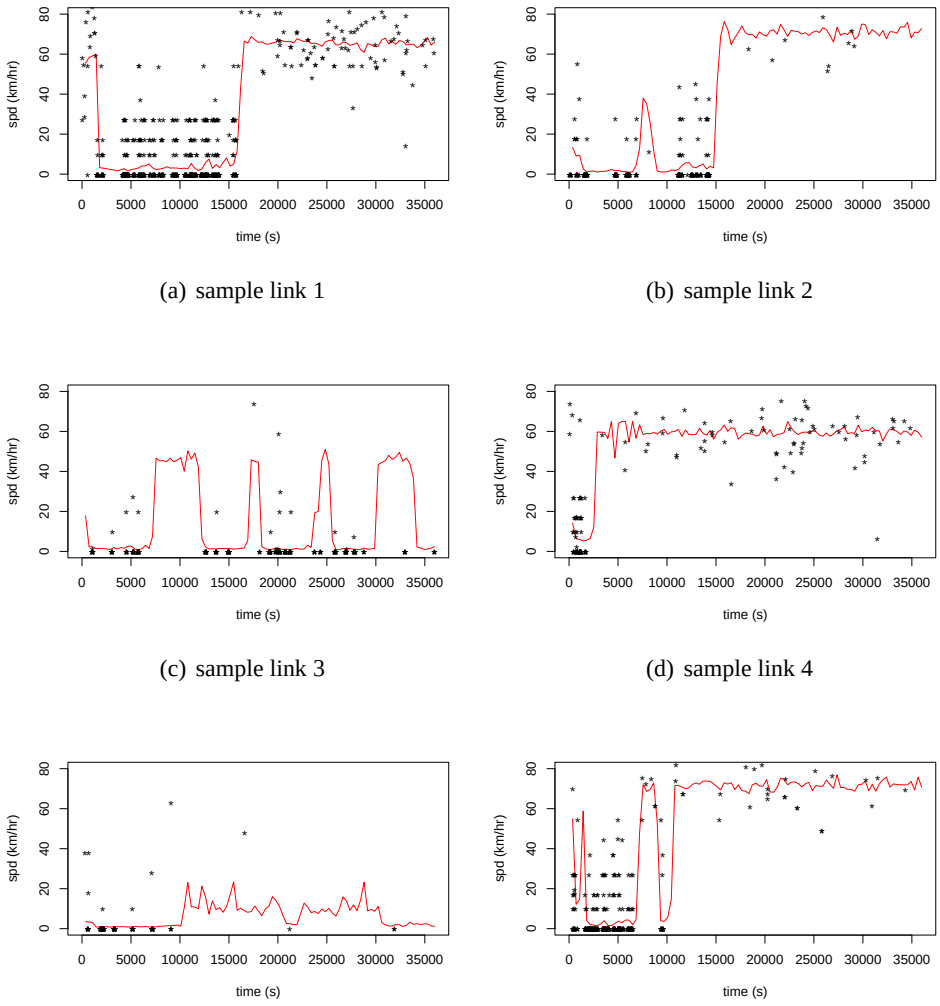


Figure 8: 30min GPS count v.s. 30min-ahead speed

As shown, the global GPS count seem to carry seem valuable information for determining the link level traffic condition. For most links in Figures 7 and 8, the congested state appears when the number of global GPS counts reach a certain level. On the other hand, the figures also indicate that multiple states (congested, uncongested) may still exist for many links for cases with large numbers of global GPS counts.

Figure 9 depicts the sampled speed received from the GPS records and the actual speed on six sample links during one 10-hr simulation period in the training data. Similar to the number of global GPS records, the congested state is correlated with a higher number of GPS samples than the uncongested state. Second, the GPS samples with zero instanteneous speed seem to only appear during the congested state, e.g. where stop-and-go traffic patterns are likely to occur.



(a) sample link 1

(b) sample link 2

(c) sample link 3

(d) sample link 4

(e) sample link 5

(f) sample link 6

Figure 9: GPS sampled speed v.s. actual speed on six sample links during one simulation day

Our method therefore works by selecting from the historical link-level speed observations the K most similar hours and uses a linear combination of the corresponding speed observations as prediction values. Parameters of the selection criterion and the coefficients of the values in the weighted average are optimized through a 5-fold cross-validation framework. The final predictions are then generated by grouping several solutions generated by different neighboring criterions.

4 Optimized Model and Results

As shown, both the overall sample counts of GPS records and the link-level GPS counts carry valuable information for determining the link-level traffic state. Therefore, the nearest-neighbor distance criterion we employ in our prediction model is constructed by taking into account both a global and a local similarity index, defined as follows.

1) Global similarity: S_{ij}^{g} measures how close the total number of GPS counts in one test hour is to that of a training hour. To construct S_{ij}^{g}, let c_i^t and C_j^t be the total number of GPS points received during every 1-min interval $t = 1,\ldots,30$ of test hour $i = 1,\ldots,500$ and training hour $j = 1,\ldots,500$, respectively. The global similarity between a test hour i and a training hour j, denoted as S_j^g, is measured by the RMSE of c_j^t and C_j^t. Namely,

$$S_{ij}^{g} = \sqrt{\frac{\sum_{t=1}^{30}(c_i^t - C_j^t)^2}{30}}. \tag{1}$$

2) Local similarity: We construct two versions of local similarity, zero and non-zero: S_{ijk}^{l1} and S_{ijk}^{l2}. A local similarity measure based on the total number of GPS records with zero and nonzero values on any critical link k, S_{ijk}^{l1} and S_{ijk}^{l2} measuring the similarity of a test hour $i, i = 1,\ldots,500$ and a training hour $j, j = 1,\ldots,500$ on link $k, k = 1,\ldots,100$, is computed as follows:

$$S_{ijk}^{l1} = |p_{ik} - P_{jk}|, \quad S_{ijk}^{l2} = |q_{ik} - Q_{jk}|, \tag{2}$$

where

p_i and P_j are the total number of GPS records with zero values during the first half of test hour i and training hour j, respectively;

q_i and Q_j are the total number of GPS records with nonzero values during the first half of test hour i and training hour j, respectivley.

Given link $k = 1,\ldots,100$ and test hour $i = 1,\ldots,500$, the overall similarity measure S_{ijk} for each training hour $j = 1,\ldots,500$ is then computed as the weighted sum of the ranks of the global similarity and the local similarities. Namely,

$$S_{ijk} = \alpha_k \text{rank}(S_{ij}^{g}) + \beta_k \text{rank}(S_{ijk}^{l1}) + \gamma_k \text{rank}(S_{ijk}^{l2}), \tag{3}$$

where the rank of a training hour is measured by its position when the corresponding similarity measure for all training hours is sorted in ascending order.

Finally, the harmonic average speeds of the first and last 6-min intervals of the second half of each test hour are estimated as the weighted harmonic average speeds of the corresponding intervals of the K

most similar training hours. The inverse of the similarity metric of each candidate training hour is used as the weight.

One potential problem in using the harmonic mean of the K nearest neighbors is that if all the candidate hours in the neighbor list have high speeds except for a few small outliers, the harmonic mean can be very small. The existence of such cases contributes to quite a significant portion of the error. To avoid the outlier effect, a conditional trimmed harmonic mean is used by filtering out the rare small outliers when most of the neighbors have high velocity values.

Note that there is some flexibility in constructing the estimator. For example, in the global similarity measure, we may use time aggregation granularity other than 1 minute; When combining solutions from the K nearest neighbors, besides harmonic mean, other choices may include arithmetic mean, median, etc. Our final solution is an ensemble of six different estimators constructed from combinations of two time granularity levels (1min and 6min) and three different aggregation methods (arithmetic mean, median, and harmonic mean).

For each link k of interest, our K nearest neighbor method with the outlier filter has seven parameters in total:

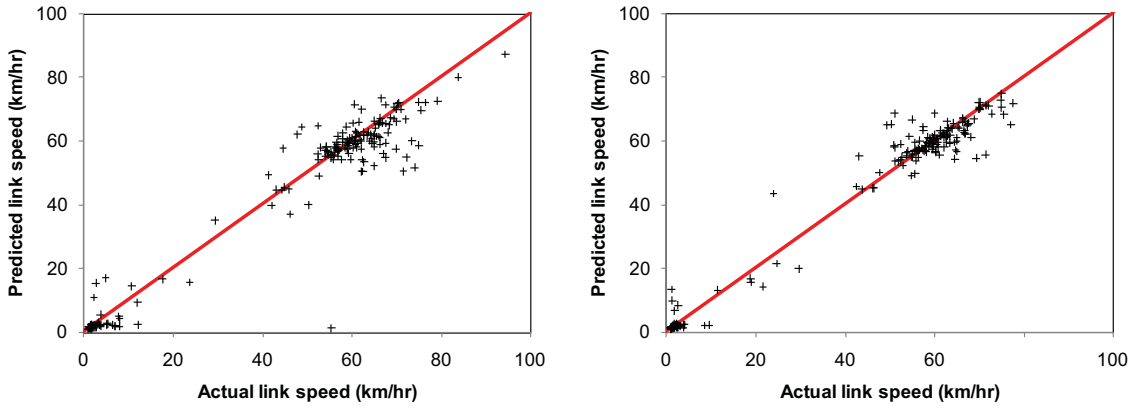
1) K - the total number of neighbors used in constructing the velocity estimate;
2) α_k - weight of the global similarity measure;
3) β_k - weight of the local congested (zero-speed) similarity measure;
4) γ_k - weight of the local uncongested (higher-speed) similarity measure;
5) n_k - the total number of high speed candidates for the outlier filter to be initiated;
6) h_k - the high cut-off value of the outlier filter;
7) l_k - the low cut-off value of the outlier filter.

In addition, for the ensemble, we also determine the weight w_i for each estimator $i = 1, \ldots, 6$. This requires another 5 parameters as the sixth one can be determined as $w_{6k} = 1 - \sum_{i=1}^{5} w_{ik}$.

A 5-fold cross validation framework was employed to determine the parameters of our prediction model. Within a 5-fold cross validation framework, the entire training data set is evenly divided into 5 subsets. For each of the five test-training data sets, one subset is used as "test data" and the remainders as "training data". A set of parameters are regarded as optimal if it generated the best average performance over the five test-training data sets. Finally, the optimal parameter settings are applied to the real test data to obtain the final predicted values.

Our prediction model results in the overall prediction performance measure of inverted RMSE = 7.46 min/km, which is significantly better than both of the two baselines discussed previously, namely 18.065min/km and 14.785min/km. Figure 10 through Figure 12 provides a closer look at the prediction performance of our model.

Figure 10 illustrates the relationship between predicted v.s. actual values on two sample test hours. Both the 6min-ahead and 30min-ahead predictions on all the 100 critical links are included. Comparing Figure 10 to Figure 5, we can see that the instances which performed poorly using the second baseline are corrected using our approach.

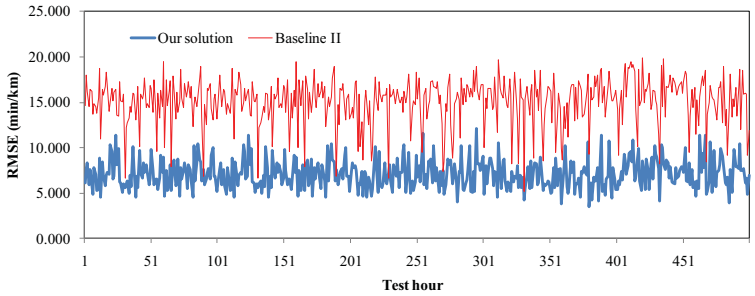


(a) Our solution, sample test hour 1

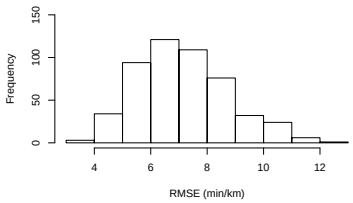
(b) Our, sample test hour 2

Figure 10: Prediction performance of the solution

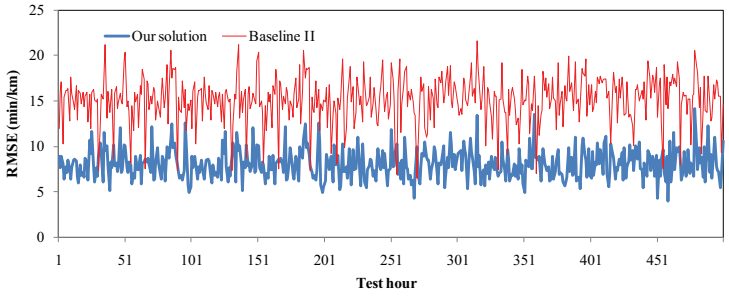
Illustrations in Figure 11 depict the RMSE of inverted predictions by test hours, for both the 6min-ahead and 30min-ahead predictions. Figures 11(a) and 11(c) on the left show the exact RMSE values for all the 500 test hours while Figures 11(b) and 11(d) on the right present the RMSE's in histograms. The same predicted measure of baseline II is included in the value plots for comparison.



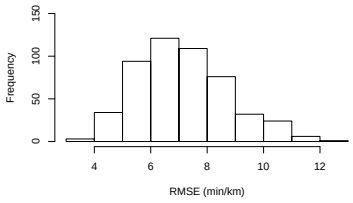
(a) RMSE values of 6min-ahead prediction



(b) RMSE histogram of 6min-ahead prediction by our solution



(c) RMSE values of 30min-ahead prediction



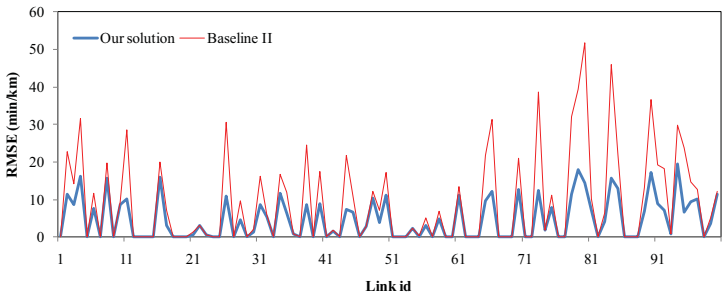
(d) RMSE histogram of 30min-ahead prediction by our solution

Figure 11: Prediction performance by test hours

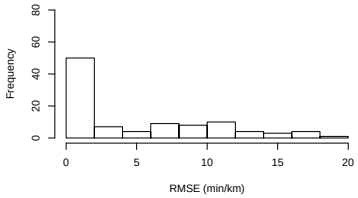
As shown, our prediction model provides much better performance than baseline II in almost all the

test hours. The RMSE values are in general lower by a large margin (approximately 7 min/km) and are less volatile. Similar patterns can be seen for both the 6min-ahead and 30min-ahead predictions. The overall performance of 6min-ahead prediction (RMSE = 7.30 min/km) is slighly better than that of 30min-ahead prediction (RMSE = 8.23 min/km), which can be expected as prediction far into the future is usually more difficult to make.

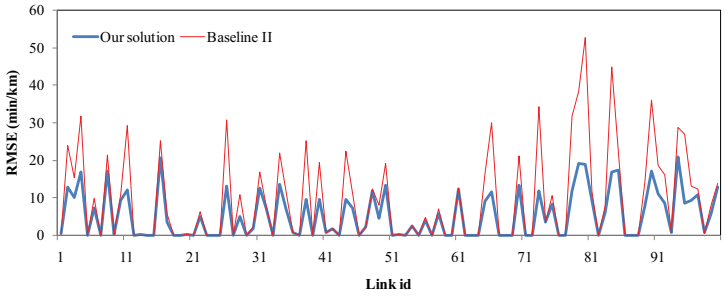
A similar comparison of the prediction performance of our solution and baseline II to Figure 11 is included Figure 12. This time, RMSE values are compuated by links instead of by test hours. As shown, the variance of RMSE by links is much larger than the variance of RMSE by test hours. Our solution is able to significantly improve the prediction performance especially for links with huge RMSE values in baseline II. We also notice that a quite large portion of links have very small RMSE values in both our solution and baseline II. Basically, these are links which are either congested or uncongested all the time.



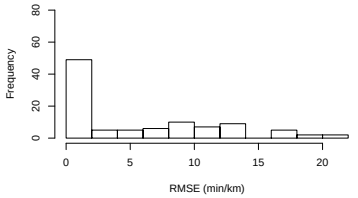
(a) RMSE values of 6min-ahead prediction



(b) RMSE histogram of 6min-ahead prediction by our solution



(c) RMSE values of 30min-ahead prediction



(d) RMSE histogram of 30min-ahead prediction by our solution

Figure 12: Prediction performance by links

5 Conclusions

This paper proposes an approach for the traffic prediction problem using GPS data with low sampling rates, where typical prediction models based on fixed-location data sources break down, in conjunction with limited speed data as obtained from fixed data sources, but unavailable in real-time. We propose a hybrid approach that combines these two data sources in a novel manner.

Main observations from our results are that the speed data as obtained from sampled, isolated (i.e.

non-trajectory-based) GPS readings are in some cases oscillatory and hence unreliable as surrogates for observed speeds either in their raw form or averaged over multiple such readings. On the other hand, data obtained from more traditional, e.g. fixed, sensors can provide a stabilizing element to the real-time GPS-based speed information.

In our case, we used the real-time GPS readings primarily to identify traffic state, and used the speed data from alternate sources to determine the likely speeds given the real-time estimated state.

As far as future work is concerned, one may wish to explore whether further improvements can be obtained from a finer categorization of the GPS speed readings and/or from the incorporation of information from other links in geographic proximity to the prediction links. Alternatively, an in-depth exploration of the optimal frequency and time span for calibration periods would be of use. Finally, in future studies, it would likely be valuable to make use of other information such as day of week, time of dayweather, etc.

References

M. S. Ahmed and A. R. Cook. Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Board*, 722:1–9, 1979.

H. Al-Deek, S. Ishak, and M. Wang. A new short-term traffic prediction and incident detection system on i-4, vol. i. final research report. Technical report, Transportation Systems Institute(TSI), Department of Civil and Environmental Engineering, University of Central Florida, 2001.

H. Alt, A. Efrat, G. Rote, and C. Wenk. Matching planar maps. *Journal of Algorithms*, 49:262 – 283, 2003.

S. Clark. Traffic prediction using multivariate nonparametric regression. *Journal of Transportation Engineering*, 129:161–168, 2003.

S. D. Clark, M. S. Dougherty, and H. R. Kirby. The use of neural networks and time series models for short-term traffic forecasting: a comparative study. *Proceedings of the PTRC 21st Summer Annual Meeting*, 1993.

P. Gora. Traffic simulation framework - a cellular automaton-based tool for simulating and investigating real city traffic. *Recent Advances in Intelligent Information Systems, Warsaw*, pages 642–653, 2009.

J. Greenfeld. Matching gps observations to locations on a digital map. In *Proceedings of the 81st Annual Meeting of the Transportation Research Board*, Washington D.C., 2002.

J. Guo and B. M. Williams. Real-time short-term traffic speed level forecasting and uncertainty quantification using layered kalman fiters. *Transportation Research Record*, 2175:28–37, 2010.

B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5:38–46, 2006.

B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J. C. Herrera, and A. Bayen. Virtual trip lines for distributed privacy-preserviing traffic monitoring. In *The Six Annual International Conference on Mobile Systems, Applications and Services (MobiSys 2008)*, Breckenridge, U.S.A., June 2008.

S. Huang and A. W. Sadek. A novel forecasting approach inspired by human memory: The example of short-term traffic volume forecasting. *Transportation Research Part C*, 17:510–525, 2009.

Y. Kamarianakis and P. Prastacos. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record*, 1858:74–84, 2003.

A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. Van Lint. A genetic algorithm-based method for improving quality of travel time prediction intervals (in press). *Transportation Research Part C*, 2011.

M. Levin and Y.-D. Tsao. On forecasting freeway occupancies and volumes. *Transportation Research Record*, 773:47–49, 1980.

W. Min and L. Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C*, 19:606–616, 2011.

K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *Journal de Physique I*, 2:2221–2229, 1992.

I. Okutani and Y. J. Stephanedes. Dynamic prediction of traffic volume through kalman filtering theory. *Transportation Research Part B*, 18:1–11, 1984.

B. L. Smith and M. J. Demetsky. Multiple-interval freeway traffic flow forecasting. *Transportation Research Record*, 1554:136–141, 1996.

B. L. Smith and M. J. Demetsky. Traffic flow forecasting: comparison of modelling approaches. *Journal of Transportation Engineering*, 123(4):261–266, 1997.

B. L. Smith, B. M. Williams, and R. K. Oswalsd. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C*, 10:303–321, 2002.

J. van Lint, S. Hoogendoorn, and H. van Zuylen. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C*, 13:347–369, 2005.

E. Vlahogianni, J. C. Golias, and M. G. Karlaftis. Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews*, 24:533–557, 2004.

E. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C*, 13:211–234, 2005.

P. C. Vythoulkas. Alternative approaches to short-term traffic forecasting for use in driver information systems. In *Transportation and Traffic Theory, Proceedings of the 12th International Symposium on Traffic Flow Theory and Transportation*, Berkeley, CA, July 1993.

B. M. Williams, P.K. Durvasula, and D. E. Brown. Urban traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record*, 1644:132–144, 1998.

S.-Y. Yun, S. Namkoong, J.-H. Rho, S.-W. Shin, and J.-U. Choi. A performance evaluation of neural network models in traffic volume forecasting. *Mathematical Computer Modelling*, 27:293–310, 1998.