TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Article

Transportation Research Record
2018, Vol. 2672(42) 148–158

DOI: 10.1177/0361198118772956
journals.sagepub.com/home/trr
SAGE

Reconstructing Vehicle Trajectories to Support Travel Time Estimation

Zheng Li[1], Robert Kluger[1], Xianbiao Hu[2], Yao-Jan Wu[1], and Xiaoyu Zhu[3]

Abstract

The primary objective of this study was to increase the sample size of public probe vehicle-based arterial travel time estimation. The complete methodology of increasing sample size using incomplete trajectory was built based on a *k*-Nearest Neighbors (*k*-NN) regression algorithm. The virtual travel time of an incomplete trajectory was represented by similar complete trajectories. As incomplete trajectories were not used to calculate travel time in previous studies, the sample size of travel time estimation can be increased without collecting extra data. A case study was conducted on a major arterial in the city of Tucson, Arizona, including 13 links. In the case study, probe vehicle data were collected from a smartphone application used for navigation and guidance. The case study showed that the method could significantly increase link travel time samples, but there were still limitations. In addition, sensitivity analysis was conducted using leave-one-out cross-validation to verify the performance of the *k*-NN model under different parameters and input data. The data analysis showed that the algorithm performed differently under different parameters and input data. Our study suggested optimal parameters should be selected using a historical dataset before real-world application.

Travel time plays a significant role in traffic planning, traffic management, and advanced traveler information systems (ATIS). In the past decades, there has been an increasing trend of using large public probe vehicle datasets for arterial travel time estimation (*1, 2*). A public probe vehicle dataset, or sometimes a "passive" dataset (*3*), comprises probe vehicle data that are collected from the public (navigation app, etc.) or public transport (taxis, transits, etc.). Because probe vehicle data are generally collected via crowdsourcing, the data can support travel time estimation on a large temporal and spatial scale but at a relatively low cost.

Although probe vehicle travel time estimation has plenty of advantages over traditional methods, it has several limitations. First, accurate travel time estimation requires a relatively high penetration rate and sampling rate. The penetration rate is defined as "the flow fraction of vehicles (unique devices) reporting to the probe dataset as compared with the total flow of vehicles along a road" and sampling rate is "the average rate at which any device reports its position and velocity" (*4*). Because probe vehicles are samples from all vehicles on the road, the travel time estimation result may not be statistically significant if the penetration rate is low, leading to a low confidence level in the estimates. Also, a low sampling rate is likely to result in poor accuracy of travel time estimation. Most current probe vehicle datasets have a relatively low sampling rate and penetration rate, which limit their applications. Public probe vehicle datasets also suffer from uneven temporal-spatial sample distribution. For example, more data are collected on major arterials but less on low-grade sections; and more data are collected during peak hours but less collected during off-peak times, or even none may be collected in the late night hours.

As a result of the limitation of current probe vehicle datasets, many current studies can be classified into two research areas. The first research area aims to estimate travel time when the sampling rate or the penetration rate is low (*1, 2, 5–8*). Another research area focuses on improving the accuracy of travel time estimation (*9–12*). Although there are already several research reports on travel time estimation using probe vehicle data, most of the previous studies were built on the scenario in which

[1]Department of Civil Engineering & Engineering Mechanics, The University of Arizona, Tucson, AZ
[2]Department of Civil, Architectural and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO
[3]Metropia Inc., Tucson, AZ

Corresponding Author:
Address correspondence to Yao-Jan Wu: yaojan@email.arizona.edu

the probe vehicle data have a relatively high penetration rate but low sampling rate. Travel time estimation when probe vehicle data have a good sampling rate but poor penetration rate is a more relevant problem, given the way probe vehicle data are collected today.

Accurate urban link travel time estimation can provide more useful information to agencies, researchers, and travelers. This study aims to increase the sample size of probe vehicle data by including partial vehicle trajectories. Partial trajectories already exist in the probe vehicle datasets but are currently treated as outliers or discarded in travel time estimation. If the partial trajectories can be utilized, the expected benefits are: (1) effectively increasing the sample size available for travel time estimation, (2) estimating more accurate travel time with low penetration rates of probe vehicles, and (3) reducing the costs of data collection.

Literature Review

Travel Time Data Sources

Traditionally, travel time estimation for an urban area relies on fixed sensors, including loop detectors (*13–15*), automated vehicle identification (*16–18*), Bluetooth/Wi-Fi devices (*19–21*), microwave sensors (*22*), and so on. All the above-mentioned data collection methods require corresponding sensors installed to retrieve data. Once the sensor is installed, it can continuously record data on the monitored road section. However, the cost of installing and maintaining fixed sensors is relatively high because many sensors are needed to achieve the appropriate accuracy level or cover a large research area.

An alternative approach is to measure travel time by mobile traffic sensors, for example, floating cars (*23*), probe vehicles (*24*), cellular data, and so on. Vehicles equipped with tracking devices (GPS or mobile phone) can be used for collecting travel time at any location without roadside equipment. However, mobile sensors are still costly because stabilized data collection needs operational vehicles running on the study area all the time. Hence, they can only cover a limited number of routes for a limited duration of time (*6*). As a result of cost considerations, there are only a few traffic studies using mobile sensors.

Many public vehicles (e.g., taxis, transit, etc.) are equipped with GPS devices. These public vehicles, to some extent, are probe vehicles, and they can collect travel time data on most of the network links during their service time with a low cost. In addition, with the popularity of mobile phones, trajectory data that are collected from mobile phones can also be used for travel time estimation. The appearance of these new data sources provides the possibility for large-scale and long-term travel time estimation. Along with the growth and availability of probe vehicle datasets, numerous studies have been conducted on travel time estimation using public datasets. Zhan et al. successfully estimated hourly travel time using New York City taxicab origin and destination trip data (*2*). Jenelius and Koutsopoulos discussed a statistical model for urban road network travel time estimation using vehicle trajectories obtained from low-frequency GPS probes. A case study was conducted on an arterial network in Stockholm, Sweden using taxi fleet data (*6*).

According to the location where data are collected, travel time can be classified into travel time on freeways or travel time on arterials. Whereas vehicular flow on freeways is often treated as uninterrupted flow, flow on arterials is much more complicated because it can be affected by signal delay, queue delay, pedestrians, and entry vehicles. On highway or urban environments, as travel time depends on the origin and destination, ATIS normally use methods that calculate travel time at a link or section level, rather than a trip level (*25*). Feng et al. proposed that the distribution of link travel time in an urban area can be approximated using mixtures of normal distributions. Although historical travel time data are available, probe vehicle data can be used to identify current traffic statement based on Bayes Theorem (*26*).

Probe Vehicle Travel Time Estimation Methods

Probe vehicles equipped with GPS systems can collect position, speed, and time stamp data every few seconds (*27*). Theoretically, probe vehicles can provide all the information needed to calculate travel time on any area at any time. However, because of the shortcomings of current probe vehicle datasets, this approach still has many limitations with respect to applications of probe vehicle travel time estimation. The limitations primarily come from two aspects: low sampling rate and/or low penetration rate.

Low sampling rate has made it difficult to measure travel time directly because little information is known between every two continuous data points. As most current datasets have a low sampling rate, there are many papers that seek to calculate accurate travel time using a sparse probe vehicle dataset. Wan et al. proposed a method to reconstruct maximum likelihood trajectory of probe vehicles between sparse updates based on the expectation maximization algorithm (*1*). Another method is to use models to estimate travel time (neural networks, etc.). Zheng and Van Zuylen built a three-layer neural network model to estimate complete link

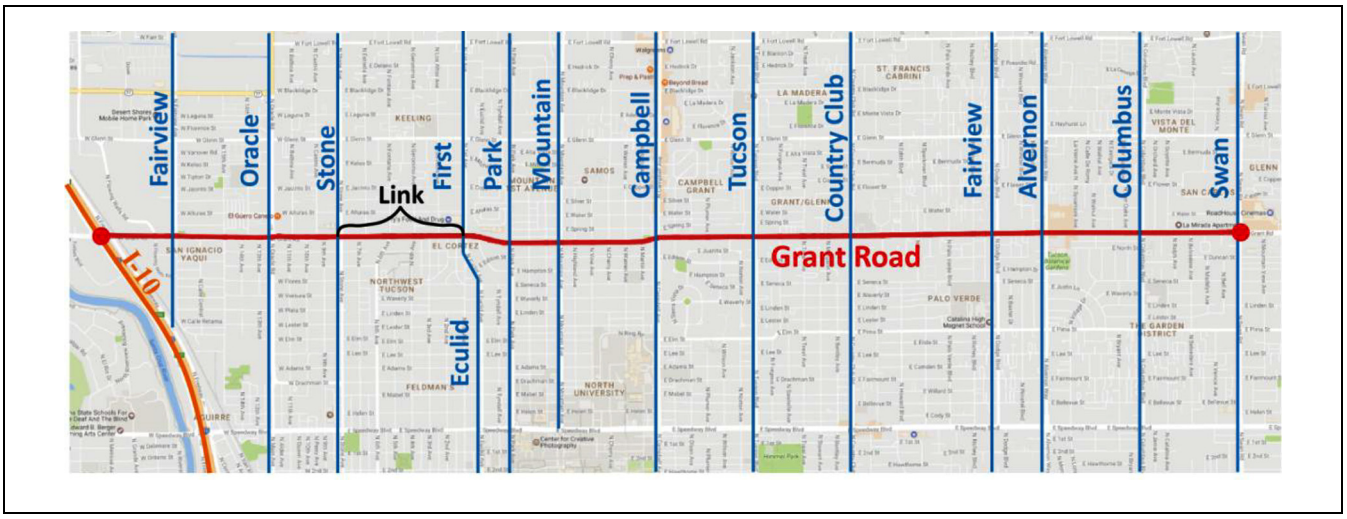

Figure 1. Study corridor.

travel time for individual probe vehicles traversing the link, and both simulation data and real-world data were used to verify the result of the model (*5*).

When penetration rate is low, probe vehicle samples cannot represent the entire population and the estimation may not be accurate. There is much research on the relationship between sample size and estimation error. Patire et al. analyzed the estimation error when sampling rate and penetration rate are different by a data fusion approach (*4*). Bucknell and Herrera analyzed estimation error of different combinations of penetration rate and sampling rate on highways using a NGSIM dataset (*7*). However, few research studies focus on how to increase the sample size of probe vehicle datasets. The appearance of public probe vehicle datasets increases the penetration rate of probe vehicles, which is important for the application of probe vehicle data. However, the problem of low penetration rate is still very common, and this means a way to increase probe vehicle sample size based on existing datasets is beneficial.

In summary, there are several studies on the penetration rate requirement for probe vehicle travel time estimation (*4, 7, 8, 28*), but only one study was identified that focused on increasing penetration rate (*29*). In addition, there are few valid methods to increase probe vehicle samples without adding new data sources. The primary objective of this paper is to outline and evaluate a process for increasing the number of usable samples when calculating travel time using probe vehicle trajectories. Currently vehicles must traverse an entire link for inclusion in travel time calculations; however, many vehicles do not traverse the entire link and thus are left out of the calculation. Including those vehicles' trajectories increases the sample size without requiring an increased penetration rate.

Data

Study Corridor

The study corridor is focused on Grant Road between I-10 and Swan Road in Tucson, Arizona. Grant Road is a major east–west direction arterial with annual average daily traffic of 36,000 vehicles per day (*30*). The study corridor is shown in Figure 1 with primary cross-streets labeled. Most of the roads are five lanes in total, with two lanes in each direction and a two-way left-turn lane. At the time of data collection, the only six-lane sections extended from Fairview Ave. to Stone Ave. and starting at Swan Ave. heading eastward. All study links have a speed limit of 40 mph (64 km/h). The links in the study refer to the one-direction segment between each of the contiguous primary cross-streets on Grant Road. For example, the eastbound Oracle–Stone link refers to the road on Grant Road between Oracle Road to Stone Road in the east direction as shown in Figure 1. Data collection was conducted in both directions.

Probe Vehicle Trajectories

Vehicle trajectory data are used to extract travel time information and to further build a historical database. The data are collected by a smartphone app, "Metropia"; when a user starts a trip using the app, the internal GPS module built into the smartphone is activated and starts to record the second-by-second data. These data, including detailed position such as latitude, longitude, heading, timestamp, velocity, and corresponding link in the roadway network, are collected at a fine time interval and sent back to the cloud server, where they are stored and will be used for further analysis.

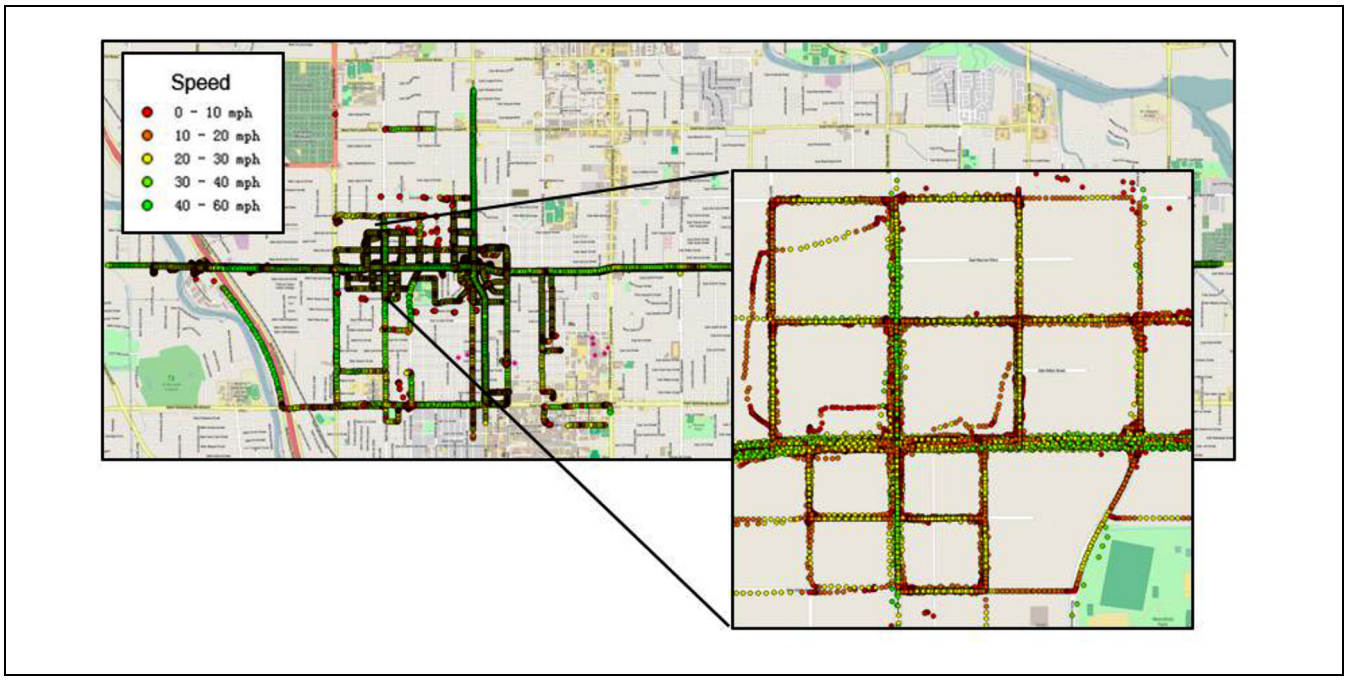

Figure 2. Visualization of example probe vehicle data (7:30–9:30 AM, Nov 17th, 2015).

The original data were collected from January 1 to December 31, 2015. There were 57,645,478 GPS points collected from 1,837 users and 43,315 trips in Tucson. The sampling rate of probe vehicle data is 1 second. In addition, roadway geometry data were acquired for the entire city of Tucson. An example of probe vehicle raw data on Grant Road is shown in Figure 2.

The data underwent an extensive selection and cleaning process including the following steps:

- Selection of trajectories at the study corridor
- Data cleaning through heuristic checks
- Map matching, to pair probe vehicle to roadway links
- Movement determination to identify travel movements
- Dimension reduction of trajectories to distance-time attributes

Methodology

Definition of Trajectories

There are two types of trajectories: complete trajectories and incomplete trajectories. Complete trajectories are defined as those from probe vehicles that passed through both an upstream and downstream intersection surrounding a link, whereas incomplete trajectories are from the probe vehicles that only traverse part of the link.

The concept of this research is to utilize incomplete trajectories in travel time estimation. For this paper, three types of incomplete trajectories are defined based on their relationship to the two intersections surrounding a link, as shown in Figure 3. The first type of trajectory is the one that passes the upstream intersection but exits the link before reaching the downstream intersection. The second type passes through neither intersection surrounding the link. The third type enters the link after the upstream intersection, and passes through the downstream intersection.

k-*Nearest Neighbors Regression*

The *k*-Nearest Neighbors regression algorithm (*k*-NN) is a non-parametric technique and it has been widely used in travel time estimation. Handley et al. used flow, occupancy, and other variables as inputs of the *k*-NN algorithm to estimate travel time on freeways (*31*). Robinson and Polak successfully used single loop detector data as inputs of *k*-NN to estimate travel time within an urban area. They compared different parameters of the *k*-NN algorithm and the results of the *k*-NN algorithm with other algorithms such as neural network. They also inferred that there is a high potential to use the probe vehicle GPS data as the input of the *k*-NN algorithm (*14*). Zhou et al. applied sparse probe vehicle data as the input of the *k*-NN algorithm to estimate link travel time in an urban area. The study suggested that the *k*-NN

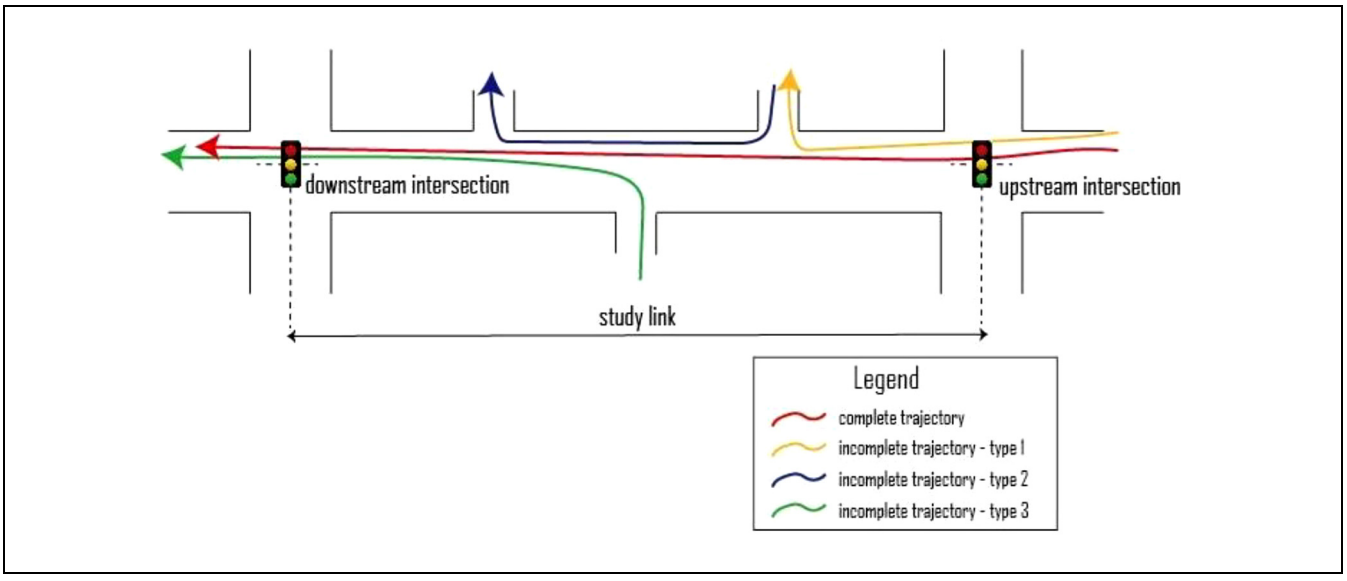

Figure 3. Definitions of complete and incomplete trajectories.

algorithm performed better than the neural network model (*32*).

k-NN has many advantages over other regression algorithms in terms of probe vehicle data analysis. The assumption under *k*-NN regression is that target value is represented by *k* closest samples. Compared with parametric techniques such as linear regression, *k*-NN has no target function. The lack of a target function makes it a more suitable approach to modeling to probe vehicle data in urban areas, because probe vehicles are greatly affected by surrounding environment (road geometry, signal timing, time of day, and other vehicles, etc.). Because of that variation in road environment, a fixed target function may not be able to fit the data well.

The problem in regression is to predict labels $y' \in \mathbb{R}^d$ for new patterns $x' \in \mathbb{R}^q$ based on a set of N observations, that is, labeled patterns $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. For an unknown pattern x', *k*-NN regression computes the mean of the function values of its *k*-nearest neighbors, shown in Equation (1) (*33*).

$$f_{\mathrm{k-NN}}(x') = \frac{1}{K} \sum_{i \in \mathcal{N}_k(x')} y_i \tag{1}$$

where $\mathcal{N}$ is neighborhood set, set $\mathcal{N}_k(x')$ containing the indices of the k-nearest neighbors of x'.

The aim is to use *k*-NN to estimate link travel time using one or more incomplete trajectories of a probe vehicle traveling on the link. Because the incomplete trajectory did not cross the whole link, the link travel time cannot be directly calculated. However, there are likely historical complete trajectories under similar traffic conditions that can be used to represent the incomplete trajectory. *k*-NN regression uses the *k* most similar complete trajectories to represent the link travel time, $f_{\mathrm{k-NN}}(x')$. For example, if the link travel time of the *k* most similar complete trajectories are $y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}$ separately, the predicted link travel time of the incomplete trajectory is, using Equation 1, $\frac{(y_{i1} + y_{i2} + y_{i3} + y_{i4} + y_{i5})}{5}$.

The distance between an incomplete trajectory and each complete trajectory is used to find the *k*-nearest neighbors. The procedure to calculate the distance between an incomplete trajectory and a complete trajectory is shown using Figure 4. The length of the study link is L. After the dimension reduction, vehicles trajectories have a position on the link at each timestamp. Dimension reduction makes the *k*-NN algorithm simpler and allows for direct comparison of travel time-related attributes rather comparing the exact path taken. There are two trajectories; the longer one is a complete trajectory and the shorter one is an incomplete trajectory. The study link is divided into n segments (n is 8 in Figure 4) on average and each segment has a length of $\frac{L}{n}$. In Figure 4, the complete trajectory passed all the segments and the incomplete trajectory passed five segments, three of which were fully passed. T_i is the segment travel time for segment i that the complete trajectory has fully traversed, and t_i is the segment travel time for segment i that the incomplete trajectory has fully traversed. The i segments used are those the incomplete trajectory fully passed. Segment travel time is only calculated when a probe vehicle's trajectory has fully traversed the segment. For example, the red, dashed line in Figure 4 is the trajectory that did not fully pass any segments and those red portions are not used to calculate segment travel time. The red part of the trajectory represents the vehicle entering

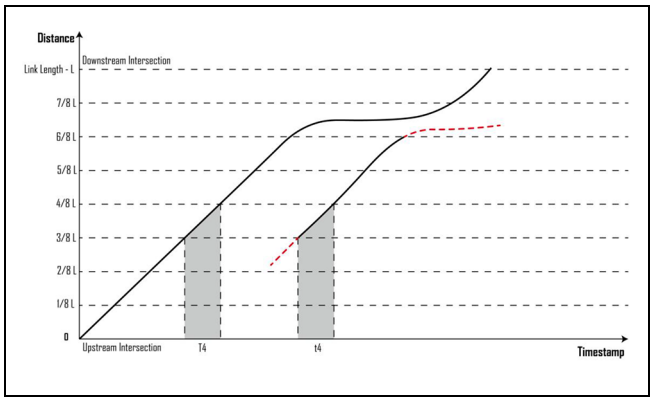

Figure 4. Trajectory similarity calculation.

and/or exiting the link, so it is unlikely to reflect actual traffic conditions.

The distance between two trajectories S is defined by Equation (2):

$$S = \sum_{i \in M} (T_i - t_i)^2 \tag{2}$$

where S is the distance between two trajectories, T_i is the time interval for the complete trajectory to pass the i_{th} segment; $t_i$is the time interval for the incomplete trajectory to pass the i_{th} segment; i is the sequence of segments; and M is the set of segments that the incomplete trajectory has fully passed.

Validation of Method

Theoretically, there is no ground truth for an incomplete trajectory because the link travel time for the incomplete trajectory cannot be calculated. However, a method was developed to evaluate the algorithm by using an incomplete trajectory generated from a complete trajectory as the input. A complete trajectory was cut-off to simulate an incomplete trajectory with a known ground truth. Using this method, the performance of the algorithm can be evaluated based on the ability to accurately recreate trajectories.

Leave-one-out cross-validation (LOOCV) (*34*) was utilized to evaluate the algorithm with the historical complete trajectory dataset. In each round of LOOCV, one complete trajectory was converted into an incomplete trajectory and this incomplete trajectory was input into the k-NN algorithm. Two measures of accuracy were used to verify the algorithm's performance: mean absolute error (MAE) and mean absolute percentage error (MAPE). MAE shows the average error of each round in LOOCV. As link travel time is related to link length, MAE shows an average time difference between algorithm output and ground truth but it cannot reflect the performance difference between links. MAPE shows the error as a percentage and the performance can be compared between links. The definitions of the two measures are shown in Equations (3) and (4), respectively.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |g_i - e_i| \tag{3}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|g_i - e_i|}{g_i} \tag{4}$$

where N is the number of LOOCV times, and g_i is the ground truth of link travel time in the i_{th} LOOCV, and e_i is the estimated link travel time in i_{th} LOOCV.

Sensitivity analysis was conducted to determine the preferred inputs for the k-NN algorithm, using the LOOCV approach. Numerous inputs can be tested, but four were selected for preliminary testing. The selected parameters for this research were the length of the incomplete trajectories, incomplete trajectory type generated in LOOCV, and the values of k and n in the k-NN regression algorithm.

Results

Effectiveness of Sample Size Increasing

Incomplete trajectories in November 2015 were used as the input and complete trajectories from January to October 2015 were used as the historical dataset. Although a specific month was selected, the sensitivity analysis can indicate how well this approach may hold during other months. The sample size comparison is shown in Figure 5. After the implementation of the k-NN algorithm, around half of the links have a sample size that has increased more than 30% and some links even have a sample size increased more than 100%. There are two links that have no improvements, westbound through movement from First Ave. to Stone Ave. and eastbound through movement from Dodge Blvd. to Alvernon Way. A visual inspection of these two links shows that the performance of the algorithm is related to link geometry characteristics and land use around the link. Land use on the north side from First Ave. to Stone Ave. is primarily residential, which may explain the shortage of incomplete trajectories. The link of Dodge Blvd. to Alvernon Way was very short and there were very few access points, and this could be the reason why incomplete trajectories were not captured. These results show that sample size can be substantially increased, depending on roadway geometry features such as access points.

Sensitivity Analysis of the k-NN Model

Sensitivity analysis was conducted to determine the preferred inputs for the k-NN algorithm, using the LOOCV

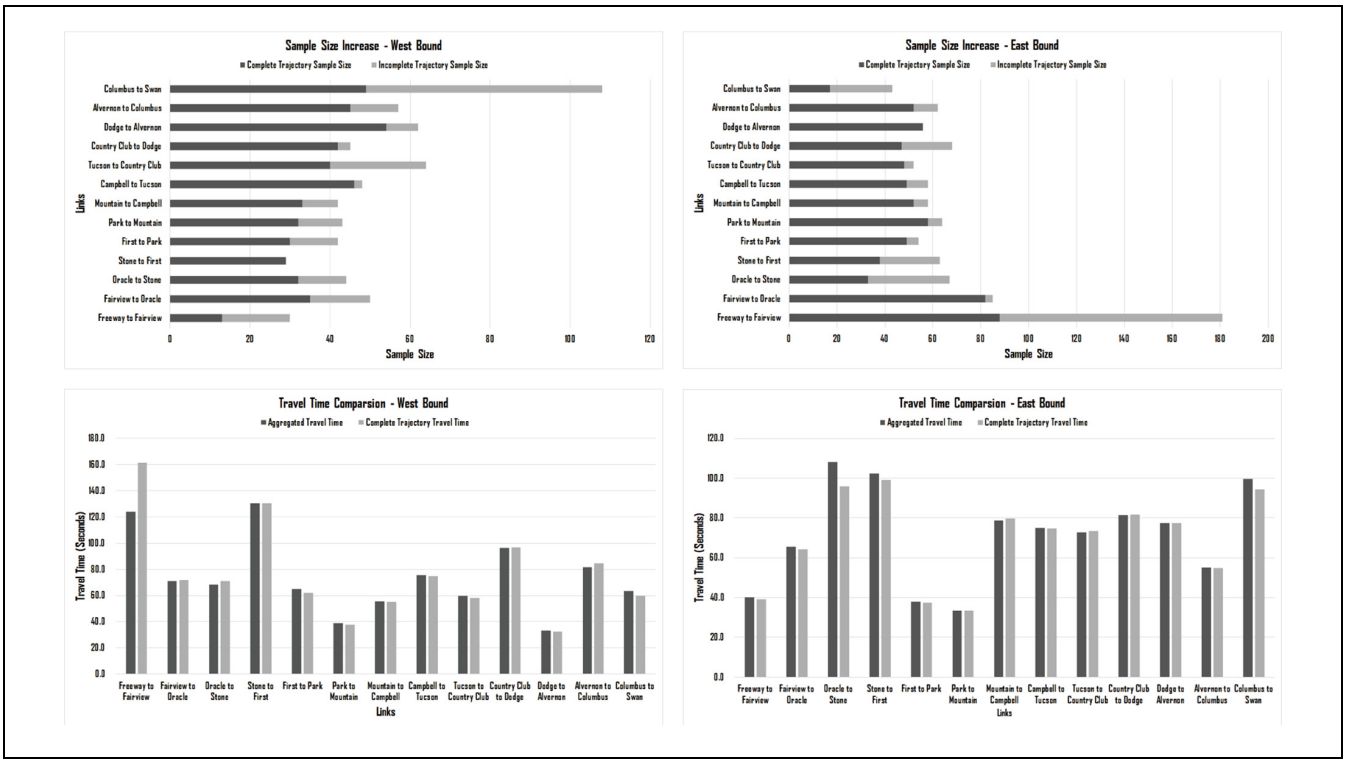

Figure 5. Sample size and travel times by links.

approach. Four parameters were selected for preliminary testing. The selected parameters for this research were the length of the incomplete trajectories, incomplete trajectory type generated in LOOCV, and the values of k and n in the k-NN regression algorithm. These parameters were selected because of their importance in laying the groundwork for future application of this method.

To identify the algorithm's sensitivity to incomplete trajectory length, the size of the incomplete trajectories was varied between 5% and 90%. Complete trajectories were converted as type 2, which is the type of incomplete trajectories that enter after the upstream intersection and leave before the downstream intersection. Only complete trajectories that were in peak hours were used. The performance of the algorithm continuously drops with the shortening of incomplete trajectory length. Both MAPE and MAE reach to their minimum values when incomplete trajectory length is 90% of the link length. The minimum and maximum value of MAPE is 6.8% and 33.3%, respectively, and the minimum and maximum value of MAE is 5.7s and 22.1s, respectively. MAPE and MAE were negatively associated with incomplete trajectory length. As longer incomplete trajectories contain more information about the complete trajectories the error is smaller, on average, for those longer trajectories. These results are shown in Figure 6.

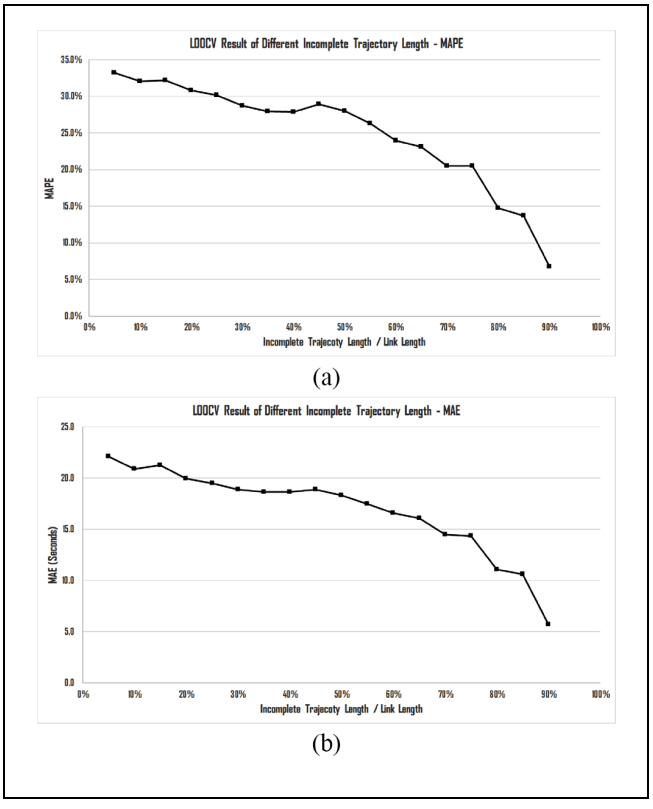

Figure 6. Sensitivity analysis on varying incomplete trajectory lengths; (*a*) MAPE (*b*) MAE.

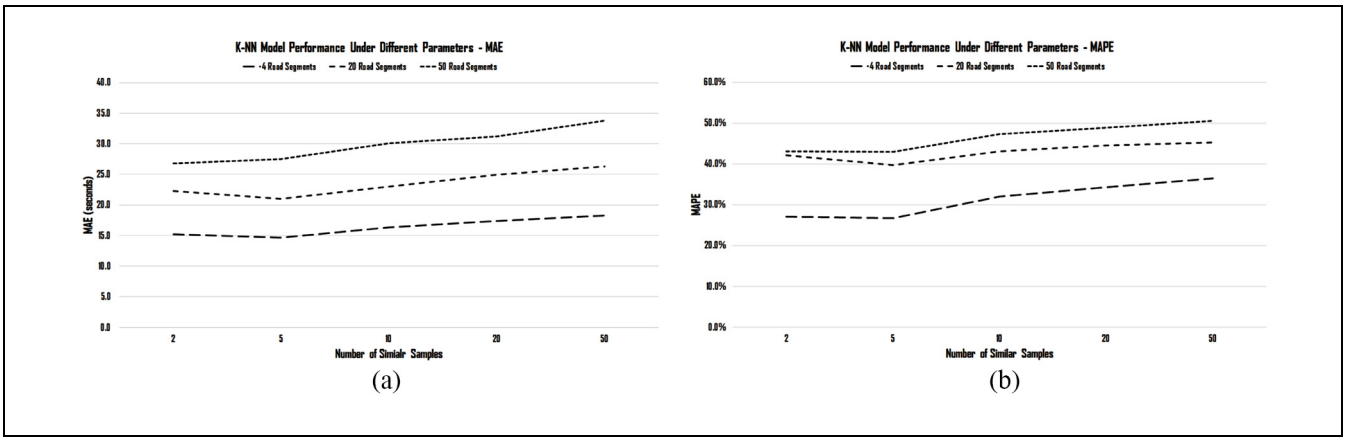

Figure 7. *k* and *n* sensitivity analysis: (*a*) MAE sensitivity and (*b*) MAPE sensitivity.

To understand how the algorithm performs based on the trajectory type, the MAE and MAPE were calculated for each of the three types independently, shown in Table 1. Complete trajectories were converted into 50% of the link length. Only complete trajectories that were in peak hours were used. Incomplete trajectory type 1 performed worst with an MAE of 18.3s and an MAPE of 39.1%. Incomplete trajectory type 2 and type 3 performed almost the same with an MAE around 7s and an MAPE around 20.7%. Incomplete trajectory type 2 and type 3 performed better than type 1 on average. It can be inferred from the result that incomplete trajectory that contains queue information performs better.

Finally, two input parameters to *k*-NN regression were simultaneously varied in the number of similar segments (*k*) and the segment length (*n*). The value of *k* was varied between 2 and 50. This was done for three different scenarios, breaking the links into 4, 20, and 50 segments. The MAE and MAPE are shown in Figure 7. With the increase in the number of similar samples, the performance of the algorithm improved at first. When the number of similar samples reached a threshold, the algorithm performed worse with the increase of the number of similar samples. The result was similar to results found in many other *k*-NN applications. We suggest that an appropriate number of similar samples needs to be selected before real-world application of the algorithm.

The results of the sensitivity analysis show that one can successfully recreate the complete trajectories using incomplete trajectories in certain cases using *k*-NN regression. Although there is no ground truth to evaluate the impact the algorithm has on the accuracy of the travel time calculation, the fact that the complete trajectories were able to be recreated likely means the travel time calculation would not change significantly for segments that already have a sufficient number of samples. However, for segments where the sample size is sparse, including incomplete trajectories should be able to help provide more accurate estimates by virtue of the increased sample size. In addition, it can be used in more specific scenarios where sample size may be an issue, such as movement-based travel time calculation.

Table 1. Algorithm Performance by Incomplete Trajectory Type

	Type 1	Type 2	Type 3
MAE/Second	18.3	6.5	7.5
MAPE	39.10%	20.50%	20.90%

As the model could be affected by different traffic conditions, such as free-flow or congested conditions, the model result was analyzed under different times of day. As the time of day cannot be simulated, input data were classified into several categories by time of day. Morning peak was defined as 7:30 AM to 9:30 AM on weekdays and evening peak was defined as 4:00 PM to 6:00 PM on weekdays. Note that, even though the input data are classified as being at peak, the training set still covers all time periods because the assumption is that a trajectory can reflect all traffic conditions.

Table 2 shows that the algorithm performed better during peak hours. The average of MAPE during peak hour is around 14% and the average of MAE during peak hour is about 10s. The algorithm performed worse during non-peak hour with a MAPE of 26.4% and a MAE of 12.4s. The reason the algorithm performed better during peak hour may be because vehicles have similar trajectories during peak hour, as traffic is more congested. Vehicle trajectories during non-peak hour may depend more on drivers' behavior. Similar trajectories during non-peak hour may be caused by similar driving behavior rather than traffic condition. These

Table 2. LOOCV Result of Different Time of Day

Time of day	Morning peak	Evening peak	Peak	Non-peak	All
Sample size	372	113	485	470	955
MAPE	13.6%	15.6%	14.0%	26.4%	20.1%
MAE/seconds	10.3	9.49	10.1	12.4	11.2

results may also indicate how well one could expect the method to perform at different times of year. As the method performs better in congested conditions, one could expect the months with more frequent congestion to see better performance. For Tucson, those months are when the University of Arizona is in session and during the winter when the population grows because of the pleasant climate.

Conclusion

This study proposed a *k*-NN based travel time calculation method using incomplete trajectories to generate additional travel time samples. Incomplete trajectories were compared with historical complete trajectories and link travel times of incomplete trajectories were represented by these similar complete trajectories. Both the feasibility of the model and the sensitivity of the model were considered in the study. The case study showed that sample size increased 42% on average after implementing *k*-NN method. The results showed that the effect of sample size increasing is related with road geometry and driving behavior of the study area. As the corridors all had pre-timed signals, the conclusion only apply to pre-timed control, and corridors with other control methods require further analysis and validation.

The sensitivity analysis of the *k*-NN algorithm showed that the algorithm performed differently under different parameters and input data. The research into the key parameters and input data concluded the following:

(1) Both the number of similar samples and the number of road segments influence the accuracy of the algorithm. The study suggests that although a small number of road segments can improve the performance of the algorithm in the study, a small number of road segments would also reduce the accuracy of similarity calculation between trajectories.
(2) The length of incomplete trajectory has a positive correlation with performance of the algorithm. The study suggests that longer trajectory contains more traffic information so complete trajectories found can better estimate the travel time of incomplete trajectories.
(3) Incomplete trajectories that contain queue information performed better. Link travel time is the summary of free-flow travel time and delay, in which delay is the key element that decides link travel time. Queue information is highly correlated with delay so the algorithm performed better when the input data reflected queue information.

The *k*-NN algorithm developed in this study can help increase public probe vehicle-based travel time collection without collecting extra data. The study also gives some suggestions on the performance of the algorithm under different parameters and input data. This study suggests that, before real-world application, optimal parameters need to be selected using a historical dataset for an accurate result. In future studies, quantitative analysis on the sample size increasing potential under different road geometry and driving behavior can be conducted. The result can help decision makers choose a better strategy to collect more data. In addition, the performance of the model can be further verified in a full connected vehicle environment. The authors recommend that future studies could focus on those issues.

Acknowledgments

This research was supported by The University of Arizona, Metropia, Inc., Pima Association of Governments, and the City of Tucson. The authors would like to thank to the reviewers for their comments and valuable suggestions for improving the quality of the article.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Yao-Jan Wu, Zheng Li; data collection: Xianbiao Hu, Xiaoyu Zhu; analysis and interpretation of results: Zheng Li, Robert Kluger, Yao-Jan Wu, Xianbiao Hu; draft manuscript preparation: Zheng Li, Robert Kluger. All authors reviewed the results and approved the final version of the manuscript.

References

1. Wan, N., A. Vahidi, and A. Luckow. Reconstructing Maximum Likelihood Trajectory of Probe Vehicles between

Sparse Updates. *Transportation Research Part C: Emerging Technologies*, Vol. 65, 2016, pp. 16–30. https://doi.org/10.1016/j.trc.2016.01.010.
2. Zhan, X., S. Hasan, S. V. Ukkusuri, and C. Kamga. Urban Link Travel Time Estimation Using Large-Scale Taxi Data with Partial Information. *Transportation Research Part C: Emerging Technologies*, Vol. 33, 2013, pp. 37–49. https://doi.org/10.1016/j.trc.2013.04.001.
3. Turner, S. M., W. L. Eisele, R. J. Benz, and D. J. Holdener. *Travel Time Data Collection Handbook*. Report FHWA-PL-98-035. Texas Transportation Institute, 1998.
4. Patire, A. D., M. Wright, B. Prodhomme, and A. M. Bayen. How Much GPS Data Do We Need? *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 325–342.
5. Zheng, F., and H. Van Zuylen. Urban Link Travel Time Estimation Based on Sparse Probe Vehicle Data. *Transportation Research Part C: Emerging Technologies*, Vol. 31, No. 111, 2013, pp. 145–157. https://doi.org/10.1016/j.trc.2012.04.007.
6. Jenelius, E., and H. N. Koutsopoulos. Travel Time Estimation for Urban Road Networks Using Low Frequency Probe Vehicle Data. *Transportation Research Part B: Methodological*, Vol. 53, 2013, pp. 64–81. https://doi.org/10.1016/j.trb.2013.03.008.
7. Bucknell, C., and J. C. Herrera. A Trade-Off Analysis between Penetration Rate and sampling Frequency of Mobile Sensors in Traffic State Estimation. *Transportation Research Part C: Emerging Technologies*, Vol. 46, 2014, pp. 132–150. https://doi.org/10.1016/j.trc.2014.05.007.
8. Argote-Cabañero, J., E. Christofa, and A. Skabardonis. Connected Vehicle Penetration Rate for Estimation of Arterial Measures of Effectiveness. *Transportation Research Part C: Emerging Technologies*, Vol. 60, 2015, pp. 298–312. https://doi.org/10.1016/j.trc.2015.08.013.
9. Zhang, Y., and A. Haghani. A Gradient Boosting Method to Improve Travel Time Prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 308–324. https://doi.org/10.1016/j.trc.2015.02.019.
10. Cao, P., T. Miwa, & T. Morikawa. Modeling Distribution of Travel Time in Signalized Road Section Using Truncated Distribution. *Procedia-Social and Behavioral Sciences*, Vol. 138, 2014, pp. 137–147. https://doi.org/10.1016/j.sbspro.2014.07.189.
11. Seo, T., and T. Kusakabe. Probe Vehicle-Based Traffic Flow Estimation Method without Fundamental Diagram. *Transportation Research Procedia*, Vol. 9, 2015, pp. 149–163. https://doi.org/10.1016/j.trpro.2015.07.009.
12. Hellinga, B. R., & L. Fu. Reducing bias in probe-based arterial link travel time estimates. *Transportation Research Part C: Emerging Technologies*, Vol. 10, No. 4, 2002, pp. 257–273. https://doi.org/10.1016/S0968-090X(02)00003-7.
13. Coifman, B., and M. Cassidy. Vehicle Reidentification and Travel Time Measurement on Congested Freeways. *Transportation Research Part A: Policy and Practice*, Vol. 36, No. 10, 2002, 899–917. https://doi.org/10.1016/S0965-8564(01)00046-5.
14. Robinson, S., and J. Polak. Modeling Urban Link Travel Time with Inductive Loop Detector Data by Using the k-NN Method. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1935: 47–56.
15. Wu, C., C. Wei, D. Su, M. Chang, & J. Ho. Travel Time Prediction with Support Vector Regression. *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, Vol. 2, 2004, pp. 1438–1442. https://doi.org/10.1109/ITSC.2003.1252721.
16. Park, D., and L. R. Rilett. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1998. 98: 63–170.
17. Li, R., & G. Rose. Incorporating Uncertainty into Short-Term Travel Time Predictions. *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 6, 2011, pp. 1006–1018. https://doi.org/10.1016/j.trc.2011.05.014.
18. Sherali, H. D., J. Desai, & H. Rakha. A Discrete Optimization Approach for Locating Automatic Vehicle Identification Readers for the Provision of Roadway Travel Times. *Transportation Research Part B: Methodological*, Vol. 40, No. 10, 2006, pp. 857–871. https://doi.org/10.1016/j.trb.2005.11.003.
19. Wang, Y., Y. Malinovskiy, Y. Wu, and U. K. Lee. *Error Modeling and Analysis for Travel Time Data Obtained from Bluetooth MAC Address Matching*. Research Project Agreement No. 61-8390, Final Research Report. Department of Civil and Environmental Engineering University of Washington, 2011.
20. Haghani, A., M. Hamedi, K. Sadabadi, S. Young, and P. Tarnoff. Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2160: 60–68.
21. Park, S., A. Saeedi, D. S. Kim, & J. D. Porter. Measuring Intersection Performance from Bluetooth-Based Data Utilized for Travel Time Data Collection. *Journal of Transportation Engineering*, Vol. 142, No. 5, 2016, 1–9. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000836.
22. Yeon, J., L. Elefteriadou, and S. Lawphongpanich. Travel Time Estimation on a Freeway Using Discrete Time Markov Chains. *Transportation Research Part B: Methodological*, Vol. 42, No. 4, 2008, pp. 325–338. https://doi.org/10.1016/j.trb.2007.08.005.
23. Byon, Y. J., A. Shalaby, and B. Abdulhai. GISTT: GPSGIS Integrated System for Travel Time Surveys. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.
24. Boyce, D. E., A. M. Kirson, and J. L. Schofer. Advance: The Illinois Dynamic Navigation and Route Guidance Demonstration Program, Chapter 11. In *Advanced Technology for Road Transport: IVHS and ATT* (I. Catling, ed.), Artech House, London, 1994, pp. 247–270.
25. Cheu, R. L., C. Xie, and D. H. Lee. Probe Vehicle Population and Sample Size for Arterial Speed Estimation. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 17, No. 1, 2002, pp. 53–60. https://doi.org/10.1111/1467-8667.00252.
26. Feng, Y., J. Hourdos, & G. A. Davis. Probe Vehicle Based Real-Time Traffic Monitoring on Urban Roadways. *Transportation Research Part C: Emerging Technologies*,

Vol. 40, 2014, pp. 160–178. https://doi.org/10.1016/j.trc.2014.01.010.
27. Li, Y., and M. McDonald. Link Travel Time Estimation Using Single GPS Equipped Probe Vehicle. *Proc., IEEE 5th International Conference on Intelligent Transportation Systems*, IEEE, Singapore, 2002, pp. 932–937.
28. Srinivasan, K., and P. Jovanis. Determination of Number of Probe Vehicles Required for Reliable Travel Time Measurement in Urban Network. *Transportation Research Record: Journal of the Transportation Research Board*, 1996. 1537: 15–22.
29. Liu, H. X., and W. A. Ma. Virtual Vehicle Probe Model for Time-Dependent Travel Time Estimation on Signalized Arterials. *Transportation Research Part C: Emerging Technologies*, Vol. 17, 2009, pp. 11–26.
30. Pima Association of Governments. *PAG Traffic Counts*. 2014. http://www.pagregion.com/Default.aspx?tabid=909. Accessed May 11, 2016.
31. Handley, S., P. Langley, and F. A. Rauscher. Learning to Predict the Duration of an Automobile Trip. *Proc., 4th International Conference on Knowledge Discovery and Data Mining*, Conference Proceedings. New York, 1998, pp. 219–223.
32. Zhou, X., Z. Yang, W. Zhang, X. Tian, and Q. Bing. Urban Link Travel Time Estimation Based on Low Frequency Probe Vehicle Data. *Discrete Dynamics in Nature and Society*, Vol. 2016, 2016, pp. 1–13.
33. Kramer, O. *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Intelligent Systems Reference Library 51. Springer, New York; Berlin, 2013.
34. Rizzo, M. L. *Statistical Computing with R*. CRC Press, Boca Raton, Fla., 2007.

The Standing Committee on Urban Transportation Data and Information Systems (ABJ30) peer-reviewed this paper (18-03631).