Spatiotemporal Traffic Volume Estimation Model Based on GPS Samples

Jack Snowdon
Massachusetts Institute of Technology
jsnowdon@mit.edu

Olga Gkountouna, Andreas Züfle, Dieter Pfoser
George Mason University
ogkounto,azufle,dpfoser@gmu.edu

ABSTRACT

Effective road traffic assessment and estimation is crucial not only for traffic management applications, but also for long-term transportation and, more generally, urban planning. Traditionally, this task has been achieved by using a network of stationary traffic count sensors. These costly and unreliable sensors have been replaced with so-called Probe Vehicle Data (PVD), which relies on sampling individual vehicles in traffic using for example smartphones to assess the overall traffic condition.

While PVD provides uniform road network coverage, it does not capture the actual traffic flow. On the other hand, stationary sensors capture the absolute traffic flow only at discrete locations. Furthermore, these sensors are often unreliable; temporary malfunctions create gaps in their time-series of measurements. This work bridges the gap between these two data sources by learning the time-dependent fraction of vehicles captured by GPS-based probe data at discrete stationary sensor locations. We can then account for the gaps of the traffic-loop measurements by using the PVD data to estimate the actual total flow.

In this work, we show that the PVD flow capture changes significantly over time in the Washington DC area. Exploiting this information, we are able to derive tight confidence intervals of the traffic volume for areas with no stationary sensor coverage.

CCS CONCEPTS

• **Information systems** → Data analytics; • **Applied computing** → **Transportation**; • **Mathematics of computing** → *Time series analysis*;

KEYWORDS

Modeling, Prediction, Traffic Volume, PVD data, Flow Capture Rate, Time Series

ACM Reference format:
Jack Snowdon and Olga Gkountouna, Andreas Züfle, Dieter Pfoser. 2018. Spatiotemporal Traffic Volume Estimation Model Based on GPS Samples. In *Proceedings of Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, Houston, TX, USA, June 10, 2018 (GeoRich'18),* 6 pages.
https://doi.org/10.1145/3210272.3210273

GeoRich'18, June 10, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-5832-3/18/06…$15.00
https://doi.org/10.1145/3210272.3210273

1 INTRODUCTION

The effective estimation and prediction of traffic conditions is crucial not only for short-term traffic management, but also for long-term transportation scheduling and, more generally, urban planning. Traditionally, traffic monitoring has been achieved via a network of stationary sensors such as induction traffic loops or microwave sensors. These sensors are not only costly to install and to maintain, but are also affected by erroneous measurements or equipment failures. With the advent of smartphones, a new and powerful traffic sensing technology became available. This so-called Probe Vehicle Data (PVD) refers to using data generated by individual vehicles as a sample to assess the overall traffic conditions ("cork swimming in the river"). Typically this data comprises basic vehicle telemetry such as speed direction and most importantly the position of the vehicle. Having large numbers of vehicles collecting such data for a given spatial area such as a city (e.g., taxis, public transport, utility vehicles, private vehicles, etc.) will create an accurate picture of the traffic speed condition in time and space [8]. Since PVD does not require a dedicated infrastructure, this data is easy to collect and provides ubiquitous coverage of the road network. However, as the name suggests, probe data only samples vehicular flow, thus an estimate of the actual traffic volume is not possible with this data.

In this work, we bridge this gap by joining volume information from stationary sensors and PVD data. We use discrete stationary sensor locations to learn the time-dependent fraction of vehicles captured by our GPS-based probe data. We can then use the PVD data to estimate the actual total flow, when the real measurements are missing due to temporary malfunctions of the sensor equipment. [4] We show that the PVD flow capture changes significantly over time in the Washington DC area. Exploiting this information, we are able to derive tight confidence intervals of the traffic volume for areas with no stationary sensor coverage.

Summary of Data

Traffic Loops: The Virginia Department of Transportation (VDOT) leverages their role in the maintenance of the road network and intermediary infrastructure by heading an initiative to harvest vehicular traffic data. VDOT is able to measure "traffic flow", or total vehicle volume, and average traffic speed through the installation of traffic loop detectors below the road surface. The relayed data is representative of all vehicles that travel over this area in the road network per five minute intervals. Issues lie in the reliability of this technology, as VDOT resorts to interpolation techniques to generate an entire day's, in extreme cases, worth of data to compensate for its malfunctions. In our model, the data extracted from the stations located at Lorton, on the I-95 highway heading north, and at Springfield on the I-395 highway heading south, is entirely authentic.

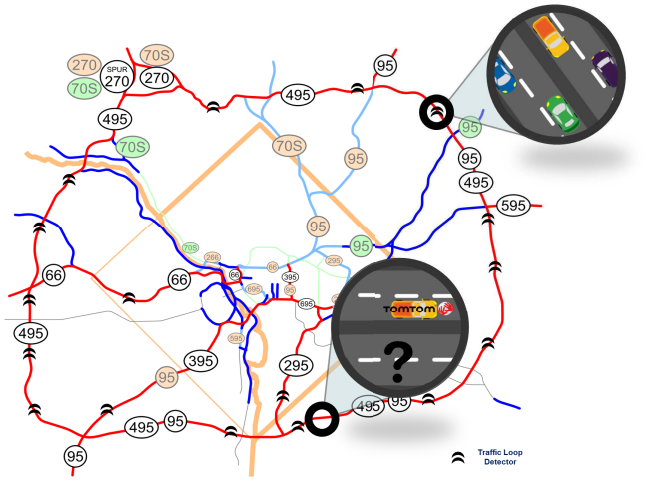


Figure 1: Road side traffic loops at various road segments.

PVD Data: We were also provided with proprietary PVD data, collected from navigation devices supported by a telematics company. We retrieved the data via a private database containing speeds mapped to individual navigation devices across the standard latitude and longitude grid. Each record contains the speed of an (anonymous) vehicle passing from a particular road segment and the time-stamp of the measurement. We aggregated this data to compute the average velocity and the number of vehicles (i.e., the flow) passing from each road segment at five minute intervals. While traffic flow and speed data from PVD samples is accessible at any location on the road network, it only represents a sample of the entire vehicle population.

Consider for example the road network of Figure 1, where traffic loop detectors are scattered in various locations of the network and depicted as double rounded black lines. Vehicles using navigation devices are depicted as yellow. In the top right circle, there is a loop detector, therefore both the time-series of the (partial) PVD volume and of the (total) VDOT volume are known. Thus, they can be used to learn the time-varying coverage of the PVD samples in that area. On the other hand, there is no traffic loop detector at the road segment shown at the bottom circle. While the PVD flow is still accessible, there are no measurements of the ground-truth total vehicular flow. The challenge is to learn a model that can infer the coverage and use it to predict the true total flow of cars at different parts of the road network.

2 RELATED WORK

Traffic volume, congestion, and other parameters of transportation networks are traditionally measured via static sensors such as traffic loop detectors [9, 13, 21] and surveillance cameras [3, 19, 22]. These devices provide collections of accurate measurements of the total traffic flow passing by the road segments that they monitor. These measurements can be used in order to train models for the prediction of the traffic volume at a segment [10, 15, 21]. However, this equipment is costly and impractical to install on every segment.

The fundamental diagram [5, 7, 13] describes the relationship among traffic density, speed and volume. It can be used to infer vehicular volume from traffic speed and density. However, learning these relationships requires a large amount of training data, which is not always easy to acquire. Furthermore, these quantities are not always available together everywhere on the road network; some devices may measure traffic flow but not speed.

To overcome the data sparsity issue, relevant research has recently shifted its focus to the traffic condition estimation using probe vehicle data, collected from vehicles equipped with GPS devices that transmit their geo-spatial coordinates in real time. Probe vehicle data has been used in literature to estimate travel times [3, 20, 23, 24] and traffic speed [16]. [11] proposes a map matching algorithm for low-sampling-rate GPS trajectories, considering the spatial geometric and topological structures of the road network and the vehicular speed constraints. Having matched the trajectories to specific road segments of the road network, the average speed [20, 24] at those segments can easily be derived. [1, 14] combine the speed, estimated by Probe Vehicle data, with the fundamental diagram to estimate traffic flow. However, directly inferring traffic volume from the average speed of sparse GPS samples leads to erroneous results, as the sample of vehicles is often a non-representative subset of the full set of vehicles on the streets. [2] study the case of learning a regression model from a roving sensor network of taxi probes. They demonstrate that the probe vehicle data are a biased sample, as using the taxi speeds leads to an underestimation of traffic flow during rush hours and overestimation otherwise. [18] adopt an unsupervised Bayesian model to learn the traffic volume from the PVD-estimated traffic speed on every road segment. [12] combines data collected by loop detectors and taxi trajectories in a machine learning model to infer the city-wide traffic volume. These approaches requires large amounts of data from various road segment locations of the city for the training phase. [17] uses anonymous phone call data to infer the number of vehicles moving from one cell to another. They model to the users' calling behavior and hourly intensity of calls and vehicles. The drawback of this approach is that it is only applicable to specific road segments that are crossing an intercell boundary.

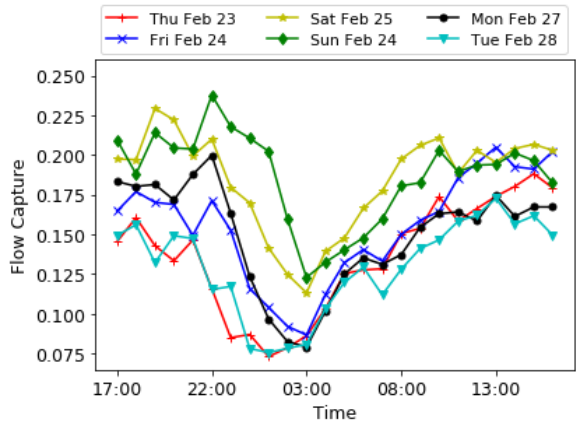
To the best of our knowledge, this is the first work that uses the time-varying flow capture rate of probe vehicle data on road segments, to predict the total traffic flow.

3 PROBLEM DEFINITION

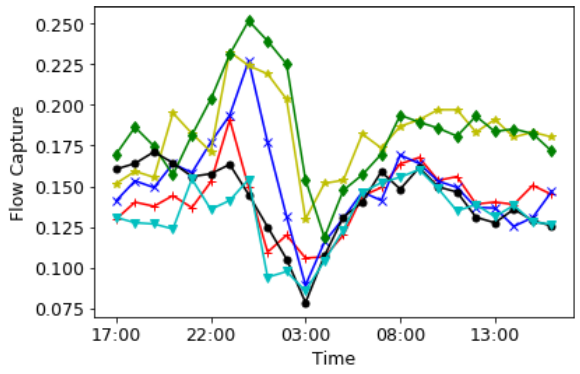
In this section, we formalize the problem of estimating traffic volume given only a sample of PVD observations. For this purpose, we define the challenge of estimating the *flow capture rate*, i.e., the fraction of vehicles that are captured by the PVD sample.

We assume a finite set of locations $\mathcal{L} = \{L_1, L_2, ...L_m\}$ where traffic loop detectors are located. At these locations, the exact traffic volume, over time, is given. In our traffic loop data obtained from VDOT, this volume of vehicles passing a road segment is aggregated at $\Delta t = 5$ minute intervals. Thus, each tuple of the traffic loop dataset is of the form $<N(L,t), t>$, where t is the time-stamp of the measurement and $N(L,t)$ is the total number of vehicles observed by a traffic loop detector located at location L during a five minute interval $[t, t + \Delta t]$.

The data that were collected from sparse GPS signal samples, were aggregated at the same time intervals. Each tuple from this



(a) Lorton (I-95)



(b) Springfield (I-395)

Figure 2: Hourly flow capture rate of the proprietary PVD data at Lorton (I-95 North direction) and at Springfield (I-395 South direction).

aggregated dataset is of the form $<N_S(L,t), t>$, where $N_S(L,t)$ is the number of vehicles that passed from a road segment at a location L, during a five minute interval starting at time t *and* were using a navigation device or service.

Since the vehicles captured by the PVD data sample are always a subset of the total volume of cars in the streets, it holds for every road segment L and any timestamp t that: $N_S(L,t) \leq N(L,t)$.

The challenge of this work is to predict the total traffic flow of vehicles using the probe vehicle data. Formally, we use the GPS sample traffic flow $N_S(L,t)$ and the past values of total volume $N_D(L,t)$ measured from traffic-loop detectors, in order to predict the future values of traffic flow $N(L,t')$ for a future time t'.

The main idea of our approach is to calculate to what degree the GPS samples are representative of the total vehicle flow at each road segment of $\mathcal{L}$ and use it to estimate the total vehicular flow in the remaining segments of the road network. To this end, we introduce the notion of *flow capture rate* of the GPS sample, which we define below.

Definition 3.1. (PVD flow capture rate) We define the PVD flow capture rate C as the fraction of vehicles that participate in the GPS sample and are passing from a road segment location L at time t, over the total flow of vehicles passing from that segment at t:

$$C(L,t) = \frac{N_S(L,t)}{N(L,t)}$$

Consider Figure 2 where the flow capture rate of sparse GPS samples is depicted for two different road segments of the Washington Metropolitan Area road network, over the course of 24 hours. Six different days were used, corresponding to the week of February 23 – March 1, 2017. Days 3 and 4 are a weekend, while the remaining are weekdays. The first graph corresponds to a segment of interstate I-95 northbound, located at Lorton. The second is a segment of interstate I-395 southbound, located at Springfield.

It can be easily observed that PVD flow capture rate is time-dependent. For this category of roads, it appears to vary from about 7.5% (on weekday nights) to more than 20% (on weekend evenings). Furthermore, it appears to follow the patterns of total flow, gaining its highest value around rush hours when the flow is at its peak, while dropping at times when the road occupancy is lower. This gives us the intuition that drivers tend to turn their navigation systems on more frequently during congestion. A reason for this may be to get alternative routes to avoid heavy traffic. The flow capture rate also appears to depend on the type of the day, i.e. weekday vs. weekend. On weekends, the fraction of drivers who use their navigation devices is larger than weekdays. A possible reason for this is that on weekdays, most of the drivers follow well-known routes, eg. to commute to and from work, while on weekends they may be exploring new destinations.

The problem we are approaching in this work is to estimate the traffic volume for a road segment for which we have no current traffic loop data, using PVD samples only. However, we assume, that traffic loop data was available at this location at a previous time to learn the time-dependent flow capture. This problem is formally defined as follows.

Definition 3.2. Let L be a location for which PVD sample data $N_S(L,t)$ is available for a time interval $t \in T$. At the same location, we assume that the exact traffic volume $N(L,t)$ is available for a smaller time interval $t \in T' \subset T$. Our task is to estimate $N(L,t)$ during all other times $t \in T \setminus T'$, given the PVD sample $N_S(L,t)$.

4 METHODOLOGY

This section describes our approach to obtain a point estimate of the traffic volume $\hat{N}(L,t)$, and how to build a confidence interval around this point estimate by modeling the traffic volume, given the PVD sample, by a negative binomial distribution.

4.1 Traffic Volume Point Estimates

For any time $t \in T \setminus T'$ where we have exact traffic volume data, we can compute the PVD flow capture rate $C(L,t)$ using Definition 3.1. To estimate the traffic volume at a future time T' of day d at a road segment located at L, we use a set of n previous days $\{d-1, d-2, \ldots, d-n\}$ to calculate the mean flow capture rate of that time of the day for that location as the predicted flow capture rate of this location at time t of day d:

$$\hat{C}(L,t) = \frac{1}{n}\sum_{i=1}^{n} C(L,(t - i \cdot 24 \cdot h)) \quad (1)$$

where h is duration of the one hour. For example, if the measurements are taken in 5-minute intervals, then $h = 12$.

The total flow at time t of day d at a road segment located at L can be predicted as:

$$\hat{N}(L, t) = \frac{N_S(L, t)}{\hat{C}(L, t)} \tag{2}$$

The above equation gives us a point estimate. However, during some times of the day we may have an extremely low number of PVD samples, either due to a low capture rate, or simply due to low traffic volume. During those times, our confidence about this point estimate will be lower than at other times where our PVD samples may be much larger and more representative of the traffic volume. To capture this uncertainty, the next section shows how to obtain confidence intervals for the estimated traffic volume $\hat{N}(L, t)$.

4.2 Confidence intervals

Consider the problem of flipping a (biased) coin, which has a known probability of p to yields "heads" until you observe the event "heads" for the k'th time. The distribution of the random number of coin-flips n required to obtain k heads follows a *negative binomial distribution* [6]. In the case of $k = 1$, we obtain the special case of a geometric distribution. The difference to the traditional binomial distribution is that is describes the distribution of the number of trials given a number of successes, rather than the number of successes given a number of trials.

Our problem is of similar nature. At a location L and time t We assume to know the likelihood $\hat{C}(L, t)$ of any individual being captured in the PVD sample (as learned from previous days in Equation 1). Further, we observe the number of individuals $N_S(L, t)$ that are captured in the PVD sample: The distribution of the total volume $N(L, t)$, given this information follows the same negative binomial distribution, having $k = N_S(L, t)$ and having $p = \hat{C}(L, t)$.

Definition 4.1. (Negative Binomial Distribution Let L be a location at a time t. Assume that we have observed $N_S(L, t)$ individual vehicles in our PVD sample. Futher, let $C(L, t)$ denote the likelihood that any individual vehicle is captured in our data. The total traffic volume follows a negative binomial distribution, where $N_S(L, t)$ is the number of successes, and $C(L, t)$ is the hit probability.

Then, the probability mass function of the number $N(L, t)$ of trials, is:

$$P(N(L, t) = x) = \binom{x - 1}{N_S(L, t) - 1}(1 - C(L, t))^{x - N_S(L, t)} C(L, t)^{N_S(L, t)}$$

We note that the negative binomial distribution can be approximated by a Gaussian distribution for sufficiently large $N_S(L, t)$. However, in our data set we have various observations having $N_S(L, t) < 10$ PVD samples per five minute interval, particularly during night-times. Thus, a simple normal approximation would a highly biased distribution.

Using the probability distribution of the estimated traffic volume $N(L, t)$ at location L at time t, as obtained in Definition 4.1 we can now proceed to construct confidence intervals, such that, at a given level of significance α, we expect $N(L, t)$ to fall into this interval.

Definition 4.2 (Confidence Intervals). Let $P(N(L, t) = x)$ be a probability mass function, and let α be a specified level of significance (for example, $\alpha \in \{0.01, 0.05\}$). We derive a confidence interval $[min, max]$ as follows:

$$min = \arg\max_x P(N(L, t) \leq x) \leq \frac{\alpha}{2}.$$

$$max = \arg\min_x P(N(L, t) \geq x) \leq \frac{\alpha}{2}.$$

Intuitively, min is the largest number of vehicles, such that less than $\frac{\alpha}{2}$ probability mass is to the left of min, and max is defined symmetrically. The resulting interval (min, max) is guaranteed to contain a probability mass of at least $1 - \alpha$.

5 EXPERIMENTS

In this section, we describe the data sets that we use, the baseline solution that we employ, and the results of our experimental evaluation comparing our approach against these baseline solutions.

5.1 Data

To evaluate our approach, we used real traffic volume data from the Virginia Department of Transportation (VDOT) and proprietary PVD data from a telematics company, as described in Section 1. We were provided with anonymous raw observations of vehicles passing from two different interstate road segments: (i) a segment of I-95 located at Lorton, northbound and (ii) a segment of I-395 located at Springfield, southbound. These observations correspond to the time period of the week of February 23, 5pm – March 1, 5pm of 2017.

We then aggregate this dataset at five minute intervals to calculate the sample volume $N_S(L, t)$. We also collected a series of traffic-loop detector measurements of the total vehicular volume $N(L, t)$ at the same road segments, over the course of the same time period (week February 23, 5pm – March 1, 5pm of 2017), also collected at five minute time intervals.

5.2 Approaches

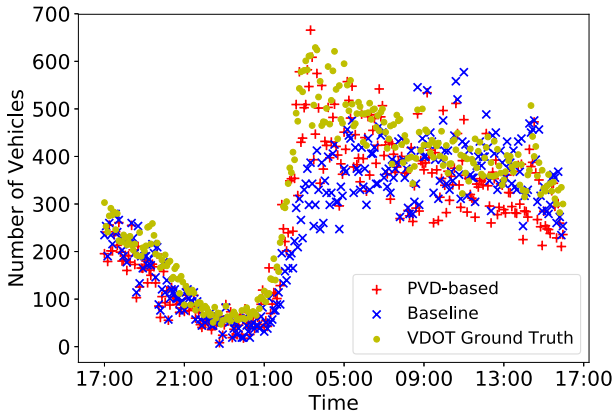
PVD-based Prediction. To evaluate our approach, we compute the PVD flow capture values over the time series of the first six days of the week and we use them to predict the total flow on the last day. To achieve this, we calculate the mean flow capture of every timestamp, for the same time of the day over the six days. We then combine this flow capture with the partial PVD flow of the last day to estimate the expected values of the time series of total flow for that day, using Equations 1 and 2.

PVD-based Prediction(hourly). The previous approach estimates the flow capture $C(L, t)$ at each five minute interval. The resulting model may overfit to random error due to small sampling size in five minute intervals, especially at night times. Thus, we propose to compute the flow capture hourly.

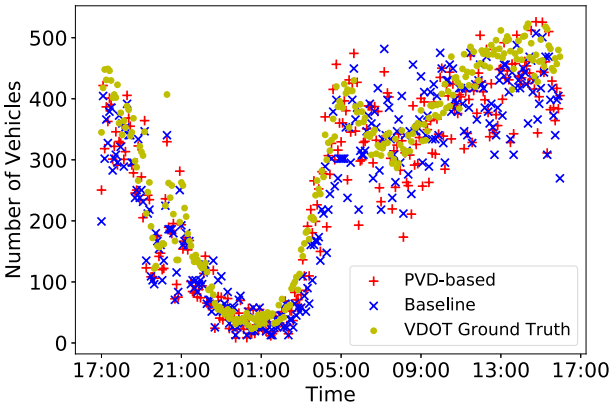
$$C_{\text{hourly}}(L, t) = \frac{\sum_{t' \in H} N_S(L, t')}{\sum_{t' \in H} N(L, t')},$$

where H is the set of all five minute intervals in the same hour as t. Thus, for each five minute interval, this approach uses the average flow capture of the corresponding hour.

Baseline. As a simple baseline, we use an approach that uses a constant flow capture C. To obtain C, we use the average flow capture rate over the first six days.



(a) Lorton (I-95)



(b) Springfield (I-395)

Figure 3: Traffic volume prediction.

VDOT Ground Truth. We compare the predicted values to the ground-truth total volume that was measured by traffic-loop detectors of VDOT during the seventh day of our dataset. We evaluated our results in terms of the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the r^2 coefficient of determination.

5.3 Traffic Volume Prediction

Figures 3(a) and (b) show the point estimates of the traffic volume, i.e., the number of cars that passed from each of the two road segments, for each five minute interval on the seventh day of our dataset. For the traffic at Lorton, the baseline approach underestimates the flow when it is at its highest peak, around 5:00am, while it overestimates flow later in the afternoon. This underestimation is due to a consequence of the relatively low fraction of navigation devices used during this time. Since the baseline uses the global average flow capture, it overestimates this flow capture, and thus, underestimates the traffic volume.

Our prediction follows the behavior of the ground-truth more accurately. As shown in Table 1, the mean absolute percentage error (MAPE) of our proposed approach is 21.94% vehicles, while the baseline yields a MAPE of 23.42%. Furthermore, the coefficient

Apporach	Location	RMSE	MAE	MAPE	R^2
Baseline	Lorton	99.53	71.37	23.42%	0.51
	Springf.	65.76	51.87	23.30%	0.80
PVD-based (5-min)	Lorton	74.30	59.65	21.94%	0.76
	Springf.	58.55	46.52	21.00%	0.84
PVD-based (hourly)	Lorton	72.27	57.57	21.11%	0.77
	Springf.	58.68	45.63	20.62%	0.85

Table 1: Traffic Volume Prediction Error.

of determination (R^2) is 0.76, which shows a higher prediction accuracy than the 0.51 of the baseline approach. At the Springfield traffic loop, both predictions introduce smaller errors. We still manage to outperform the baseline with a MAPE of 21% and R^2 score of 0.84, compared to 23.3% MAPE and 0.8 R^2 score of the baseline.

In addition, we employed the *PVD-based Prediction(hourly)*, to avoid overfitting to large sampling variance. Thus, we compute the average flow capture rate, at hourly intervals. Again, we compute the mean hourly flow capture for every hour of the previous six days, to use as the predicted hourly flow capture of the last day. The results are also shown in Table 1. At Lorton, using the hourly flow capture rate reduces the mean absolute percentage error by 4%, thus outperforming the baseline by 9.9%. It also raises the R^2 score to 0.77. At Springfield the hourly approach achieves a coefficient of determination of 0.85 and a a MAPE of 20.62%, which is 2% lower than the 5 minute approach and 11.5% better than the baseline.

Next, we evaluate the quality of the employed confidence intervals as described in Section 4.2. First, Figure 4 depicts a visualization of the 90% confidence intervals and the ground truth VDOT measurements. Table 2 shows the fraction of total flow values $N(L,t)$ that were captured by the confidence intervals. Therefore, we scale the level of confidence of the confidence interval, and count the fraction of (location, time)-pairs where this confidence contains the true traffic volume $N(L,t)$.

We note that, if our confidence interval model would perfectly capture the uncertainty of the traffic, we would expect that a level of confidence of x, would capture exactly a fraction of x measurements. We observe in Table 2, that the these fractions are significantly lower. For example, at Lorton, for a level of confidence of 90%, only 60.07% of the observed volume values were captured within the confidence intervals, while for Springfield 70.83% of the true measurements were inside the confidence range, thus, leaving a large amount of unexplained uncertainty. In contrast, the 99% confidence are more accurate, capture 98.6% and 97.9% of the measurements, which is close to the expected result of 99%.

As can be observed from Figure 2, the flow capture rate differs between weekdays and weekends. Thus, we repeated the same experiments using only the weekdays of the training set (the set from which we estimated the flow capture rate in Equations 1 and 2). In this case, the percentage of true values of Lorton and Springfield flow captured within the 90% confidence intervals rises to 70.83% and 74.31%, respectively.

6 CONCLUSIONS

Real time volume estimation allows for more comprehensive targeted marketing strategies, traffic prediction models, and infrastructure planning. Yet, obtaining traffic volume by counting individual

Level of confidence	10.0%	20.0%	30.0%	40.0%	50.0%	60.0%	70.0%	80.0%	90.0%	95.0%	99.0%	99.9%
Lorton hits	6.94%	10.76%	14.24%	17.71%	24.31%	29.51%	35.07%	44.10%	60.07%	77.08%	98.61%	99.65%
Springfield hits	5.21%	10.07%	14.93%	21.53%	28.47%	34.72%	45.49%	59.72%	70.83%	86.46%	97.92%	98.26%

Table 2: Percentage of the ground truth values captured within the confidence intervals.

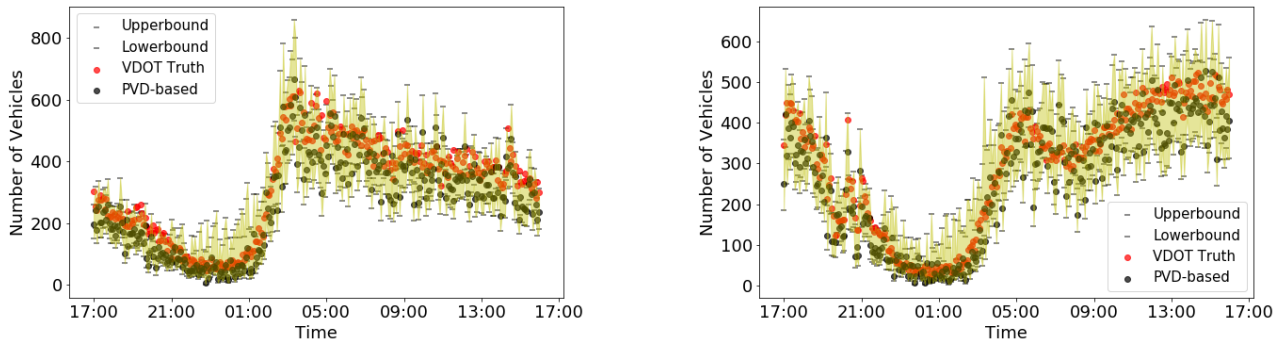


Figure 4: Traffic flow prediction with 95% confidence intervals at Lorton (I-95) and Springfield (I-395) over all days.

vehicles crossing a segment is expensive. We propose a first approach towards estimating traffic volume using PVD data, which is easily and cheaply collected from mobile devices. We found that the main challenge of estimating volume from PVD samples is the constantly changing volume capture rate. During high traffic, more people employ their navigation devices, thus giving us a larger PVD sample, whereas when the traffic is flowing free, more people turn off their navigation device as they already know their way. Our research presented in this paper shows, however, that recurring patterns of PVD volume capture allows to obtain a fairly good traffic volume estimation, without incurring the high cost of installing road-side traffic loops. As future work, we want to be able to use volume information learned at one location to be applied to another location. This step is challenging, as spatial auto-correlation (in the Euclidean space) does not hold: In our data, two locations only meters from each other, but measuring traffic volume in different directions of the same road, had entirely different daily patterns of volume capture. Furthermore, in our future work we would like to incorporate other variables, such as how the "weather" affects the flow-capture.

ACKNOWLEDGEMENTS

This research has been supported by National Science Foundation "AitF: Collaborative Research: Modeling movement on transportation networks using uncertain data" grant NSF-CCF 1637541.

We would like to thank the Virginia Department of Transportation for providing us with their data.

REFERENCES

[1] K. A. Anuar, F. G. Habtemichael, and M. Cetin. Estimating traffic flow rate on freeways from probe vehicle data and fundamental diagram. In *IEEE 18th International Conference on Intelligent Transportation Systems, ITSC*, pages 2921–2926, 2015.

[2] J. A. Aslam, S. Lim, X. Pan, and D. Rus. City-scale traffic estimation from a roving sensor network. In *10th ACM Conference on Embedded Network Sensor Systems*, pages 141–154, 2012.

[3] A. Bhaskar, E. Chung, and A. Dumont. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. *Comp.-Aided Civil and Infrastruct. Engineering*, 26(6):433–450, 2011.

[4] Federal Highway Administration. Traffic Detector Handbook: Third Edition—Volume II. *McLean, VA*, 2006.

[5] B. Greenshields, W. Channing, H. Miller, et al. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. Nat. Research Council, 1935.

[6] J. M. Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

[7] B. S. Kerner. Three-phase traffic theory and highway capacity. *Physica A: Statistical Mechanics and its Applications*, 333:379–440, 2004.

[8] R. Kuehne, R.-P. Schaefer, J. Mikat, K.-U. Thiessenhusen, U. Boettger, and L. S. New approaches for traffic management in metropolitan areas. In *Proc. IFAC CTS Symposium*, 2003.

[9] J. Kwon, P. Varaiya, and A. Skabardonis. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. *Transportation Research Record: Journal of the Transportation Research Board*, (1856):106–117, 2003.

[10] S. Lee and D. Fambro. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Trans. Research Rec.: Journal of the Trans. Research Board*, (1678):179–188, 1999.

[11] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, pages 352–361, 2009.

[12] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng. City-wide traffic volume inference with loop detector data and taxi trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1:1–1:10, 2017.

[13] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez. Traffic density estimation with the cell transmission model. In *American Control Conference*, volume 5, pages 3750–3755. IEEE, 2003.

[14] T. Neumann, P. L. Bohnke, and L. C. T. Tcheumadjeu. Dynamic representation of the fundamental diagram via bayesian networks for estimating traffic flows from probe vehicle data. In *16th International IEEE Conference on Intelligent Transportation Systems, ITSC*, pages 1870–1875, 2013.

[15] I. Okutani and Y. J. Stephanedes. Dynamic prediction of traffic volume through kalman filtering theory. *Transp. Research Part B: Methodological*, 18(1):1–11, 1984.

[16] Q. Ou, R. L. Bertini, H. van Lint, and S. P. Hoogendoorn. A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data. *IEEE Trans. Intelligent Transportation Systems*, 12(3):747–756, 2011.

[17] N. C. Sánchez, L. M. R. Pérez, F. Benítez, and J. M. D. Castillo. Traffic flow estimation models using cellular phone data. *IEEE Trans. Intelligent Transportation Systems*, 13(3):1430–1441, 2012.

[18] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu. Inferring gas consumption and pollution emission of vehicles throughout a city. In *The 20th ACM SIGKDD*, pages 1027–1036, 2014.

[19] G. S. Thakur, P. Hui, and A. Helmy. Modeling and characterization of urban vehicular mobility using web cameras. In *Proceedings IEEE INFOCOM Workshops*, pages 262–267, 2012.

[20] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *The 20th ACM SIGKDD*, pages 25–34, 2014.

[21] D. Wilkie, J. Sewall, and M. C. Lin. Flow reconstruction for data-driven traffic animation. *ACM Trans. Graph.*, 32(4):89:1–89:10, 2013.

[22] X. Zhan, R. Li, and S. V. Ukkusuri. Lane-based real-time queue length estimation using license plate recognition data. *Transportation Research Part C: Emerging Technologies*, 57:85–102, 2015.

[23] X. Zhan, S. V. Ukkusuri, and C. Yang. A bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data. *Automation in Construction*, 72:237–246, 2016.

[24] F. Zheng and H. Van Zuylen. Urban link travel time estimation based on sparse probe vehicle data. *Trans. Res. Part C: Emerging Technologies*, 31:145–157, 2013.