JULY 2024

ite journal

A COMMUNITY OF TRANSPORTATION PROFESSIONALS

ITE Annual Meeting and Exhibition

Segment-Level Traffic Volume Estimation Incorporating Crowdsourced Data

By Syed Ahnaf Morshed, Ph.D., Kamar Amine, Ph.D., and Mohammed Hadi, Ph.D.

Public agencies have utilized crowdsourced traffic data based on probe vehicles for some time to estimate segment travel times. In recent years, these agencies have become increasingly interested in the prospect of using this data to estimate segment level volumes and origin-destination (O-D) matrices. Some third party vendors have used statistical or supervised machine-learning models to estimate segment-level traffic volumes by expanding the measured mobile data sampled on the segment to total segment volumes. Data from multiple sources could be used as inputs to develop the model including mobile vehicle data in combination with data from other sources such as traffic detectors, permanent traffic recorders (PTR), and even zonal demographics data, among others.[1] This paper examines the quality of the volume estimates based on data from one of the main third party vendors (TPV) data provider. The vendor uses advanced machine learning techniques to train their volume prediction models.

Shutterstock/metamorworks

Previous Efforts

Some agencies have already evaluated the volumes obtained from vendors that provide volume estimates of traffic volumes based on data from probe vehicles that constitute only a proportion of the total traffic on the segment. However, there are still a lot of questions about the validity of the traffic volumes estimates based on the data. Table 1 summarizes the approaches used in previous efforts in assessing the validity and usability of the data for different applications.

The Methodologies and findings of TPV based volume estimation and validation reported by the Virginia Department of Transportation (DOT), Minnesota DOT, and Oregon DOT are discussed below.

Virginia Department of Transportation Experience

A Virginia Department of Transportation (VDOT) study aimed at formulating a guideline on using a vendor data by measuring its performance in different application contexts.[2] The report assessed the quality of the provided metrics in six testing contexts covering the average annual daily (AADT), O-D trips, traffic flow on road links, turning movements at intersections, and truck traffic. The benchmark data sources were continuous count stations, toll transaction data, and VDOT's internal traffic estimations. The analysis showed that the AADT estimates had lower absolute percentage errors compared to the other assessed measures. The study also found that the estimates for lower volume levels had higher errors. Additionally, using multi-periods (i.e., multiple days, weeks, or months) rather than individual periods as the input for estimating traffic measures resulted in reduced errors especially in low-volume traffic segments. For low volume segments (with AADTs lower than 10,000 vehicle per day), the Mean of Absolute Percentage Error (MAPE) measure showed that the highest errors in AADT measurements for the years 2017 and 2018 were 18.2 percent and 10.2 percent, respectively. This possibly indicates that the quality of the prediction improved between 2017 and 2018. Significantly, higher errors were observed in the estimated hourly traffic volumes, particularly for segments with low volumes of less than 500 vehicles per hour (vph).

Minnesota Department of Transportation (MnDOT) Experience

MnDOT evaluated the accuracy of AADT estimates provided by a vendor which were compared against volume data from 442 permanent continuous counter locations in 2017 and 2019. Additionally, MnDOT conducted an analysis of several hundreds of low-volume sites for short-duration counts in 2019. Typically, short-duration count stations generate erroneous estimates of AADT. As a result, MnDOT conducted the evaluation study to address the uncertainty of AADT estimates based on two benchmark data sources (the permanent continuous counter locations and the short-duration count stations). The findings of the study also showed that the estimation of AADT based on the vendor data has improved significantly in 2019 compared to 2017 for segments with moderate to high volume ranges (AADT greater

Table 1. Previous Efforts in Assessing Volume Estimates based on Probe Data.

Public Agency: Department of Transportation (DOT)	Purpose	Benchmark Data	Performance Metrics
Virginia DOT[2]	Evaluation of AADT, O-D Trips, traffic counts, turn counts and truck volumes at intersections (VDOT 2020)	Traffic Count Database System (City of Virginia Beach 2019); O-D Trip based on Electronic Toll System (VDOT 2018)	Percentage Error (PE) & Absolute Percentage Error (APE)
Minnesota DOT[3]	Evaluation of AADT and Average Hourly Volume (2017)	MnDOT 69 permanent monitoring sites and 7837 short-duration count stations	MAPE[a], MAD[b], MSD[c]
Minnesota DOT[3]	Evaluation of AADT (2020)	PTR, permanent weight-in-motion, permanent traffic detectors	MAPE[a], MPE[d], Mean Error (ME)
Oregon DOT[4]	Evaluation of AADT (2019)	PTR	PE, APE, MAPE, RMSE[e], Normalized RMSE[e], Spearman's Rho, Paired Sample t-test
Georgia DOT[5]	Calculate O-D Matrix Indices, Freight patterns (2019)	N/A	N/A
Ohio DOT[6]	Estimation of daily truck volume	N/A	N/A

N/A indicates not applicable; [a] Mean absolute percentage error; [b] Mean Absolute Deviation; [c] Mean Standard Deviation; [d] Mean Percentage Error; [e] Root Mean Square Error.

than 10,000 vehicle per day). The MAPE decreased from approximately 42 percent to 10 percent for high volume locations and 68 percent to 34 percent for low volume locations in the year 2019 compared to 2017. This effort also explored the accuracy of hourly volume estimates and found high errors in the estimate.

Oregon Department of Transportation (ODOT) Experience

Oregon DOT used AADT from the year 2017 based on Automatic Traffic Recorders (ATRs) data to assess the accuracy of the estimates of AADT from a vendor. The result showed that the median and mean absolute percentage errors are 18 percent and 26 percent, respectively. Like previous cases, low volume segments exhibited higher errors in the volume estimation. A similar study was conducted to evaluate estimates based on data collected between 2019 and 2021 from a different TPV. In 2019, higher traffic volume sites showed stronger correlation with ATR data, while lower volume sites exhibited higher error rates. In 2020, the accuracy decreased possibly due to the vendor's removal of March and April data, reflecting difficulties in capturing pandemic-related travel changes. In 2021, the accuracy further degraded across all volume bins, with no clear pattern observed. It remains uncertain whether the decrease in accuracy is due to the model becoming outdated or other structural issues.

Study Data

This study further investigated crowdsourced data based on probe vehicles from a TPV to provide estimates of daily and hourly traffic volumes. Such estimates, if accurate, can be very valuable as volume count data is expensive to collect. In addition, in many cases, counts are available only for short periods of time and may not cover all links in the network. This study evaluated the estimates based on crowdsourced data in combination with the continuous volume measurements collected at one or more locations in the network to estimate the link volumes on all network links. The utilized network in this study is located in downtown West Palm Beach, FL, USA. There are two permanent count stations (PCS) maintained by the Florida Department of Transportation (FDOT) in the case study network. The first is located near the Flagler Memorial Bridge (Site 0087), while the second is on I-95 (Site 0174), as indicated by the red spheres in Figure 1. The whole year, traffic volumes from these two permanent count stations are used in this study. The year-long data from 2020 is collected from Florida Traffic Online, which is an online tool established and maintained by the FDOT.

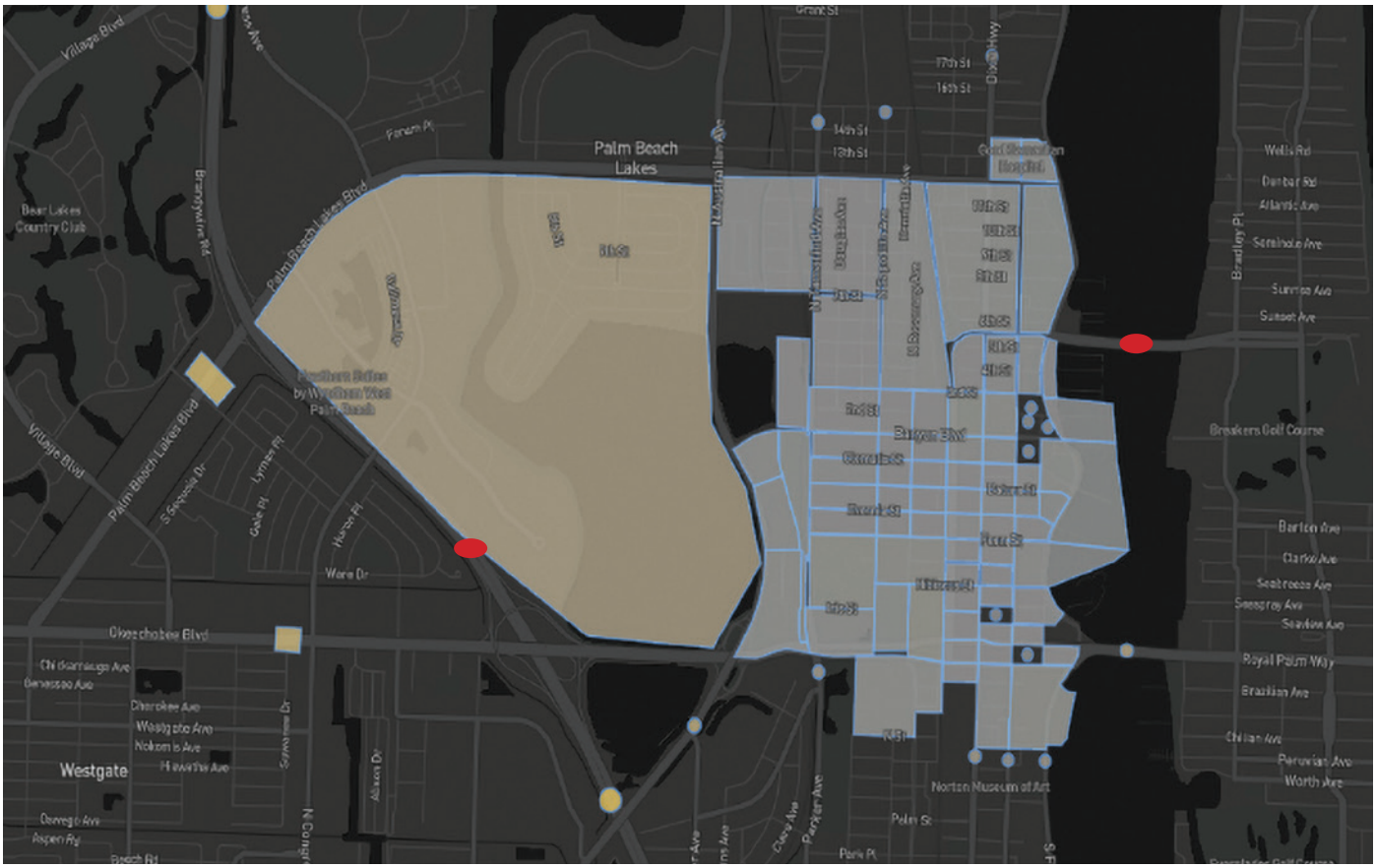


Figure 1. StreetLight analysis dashboard of the West Palm Beach network.

Seasonal Variations

First, the seasonal variation in the crowdsourced was checked. This is important because an analyst may need to conduct multi-scenario analysis or need to estimate the traffic volumes in the peak season(s) of the year. This section compares seasonal variation in average daily traffic (ADT) from the PCS (used in this case as benchmark data) with the variations based on the crowdsourced-based volume, in vehicles per day (vpd). The monthly average daily traffic (MADT) of June, July and August are aggregated to represent the summer season ADT. The MADT for December, January, and February are aggregated to represent the winter season ADT. Typically, the traffic demand during the winter in West Palm Beach, FL, is expected to be higher compared to the summer, due to the high number of visitors and temporary residents in the winter months.

Table 2 compares the seasonal variation in the ADT using seasonal factors (SF) based on the PCS and crowdsourced data at the two PCS locations i.e., the Flagler Memorial Bridge for eastbound and westbound traffic, and the I-95 Congress location for northbound and southbound traffic. The SF were calculated by dividing the seasonal (summer or winter) ADT based on the

Table 2. Seasonal ADT Comparison Crowdsourced-based vs PCS-based Volumes.

	Flagler Memorial Bridge				I-95 Congress			
	Eastbound		Westbound		Northbound		Southbound	
	Winter	Summer	Winter	Summer	Winter	Summer	Winter	Summer
PCS (vpd)	28,004	17,823	32,073	21,072	322,701	258,266	326,074	259,769
Crowdsourced (vpd)	25,969	18,922	27,316	19,873	223,363	179,689	241,207	197,272
PCS SF	0.61	0.39	0.60	0.40	0.56	0.44	0.56	0.44
Crowdsourced SF	0.50	0.50	0.50	0.50	0.49	0.51	0.48	0.52
Percentage Difference	18%	28%	17%	27%	12%	15%	13%	16%

3 months over the ADT based on all 6 months (summer and winter). It is evident that there is a significant difference between the estimates based on the two sources in some cases.

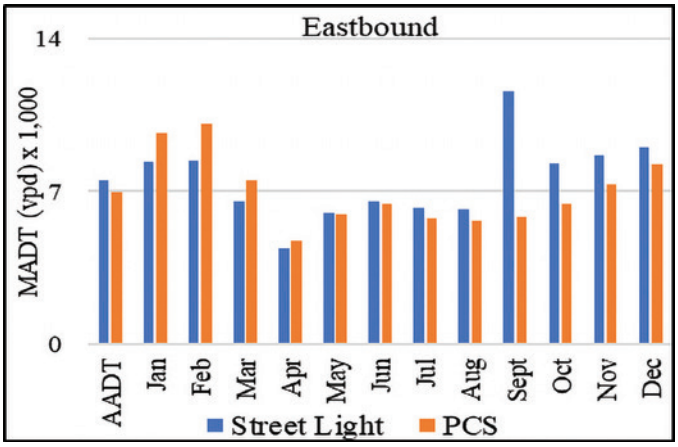
Comparison of Monthly Average Daily Volumes

Figure 2 shows a comparison of the MADT for each month of the year based on the data from PCS with crowdsourced-based estimates. For the eastbound (EB) and westbound (WB) directions at the Flagler Memorial Bridge location, the MADT estimates based on TPV crowdsourced data are similar to the PCS data, except for the months of January, February, and September, as displayed in Figure 2(a) and Figure 2(b). The AADT for the eastbound direction is 6,961 vpd and 7,507 vpd based on PCS and crowdsourced data, respectively. The AADT for the westbound direction is 7,958 vpd and 8,157 vpd based on PCS and crowdsourced-based estimates, respectively. However, for the northbound (NB) and southbound (SB) directions of the I-95 location, the crowdsourced-based values underestimate the MADT compared to the PCS data in most cases, as seen in Figure 2(c) and Figure 2(d). In terms of AADT at the I-95 location, high discrepancies are observed for both directions of I-95. The underestimation of the volume is possibly due to the expansion of the partial data collected from mobile sources using data collected from other locations that have different characteristics or are less congested than freeway facilities in South Florida.

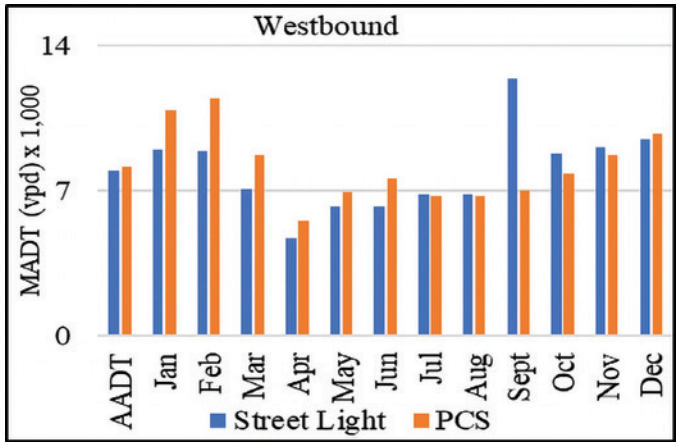
Estimating Daily Volume at PCS locations Using Regression Analysis

The comparison in the previous section indicates that there are large discrepancies between the AADT and MADT estimated based on TPV crowdsourced data and PCS data. This section presents an investigation of an enhanced methodology to estimate the traffic volumes based on a regression model to expand the data using local PCS data rather than using data already expanded by the vendors.

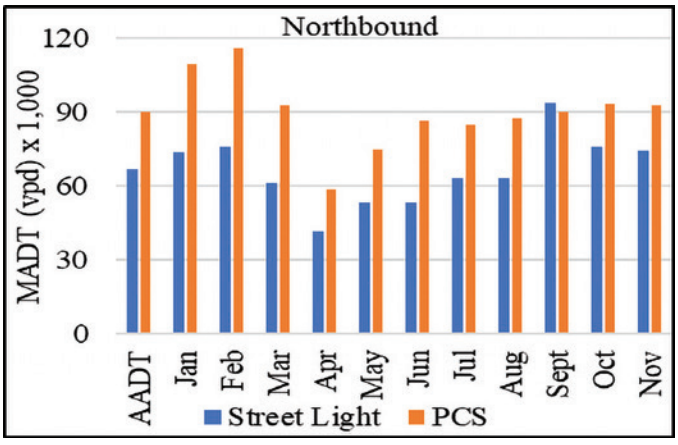
The method for expansion based on local network data involves deriving the relationship between the PCS measures and what is referred to in this study as a relative volume index (RVI). RVI is a relative measure of the volume of the trips on a link calculated based on the number of detected mobile devices normalized based on



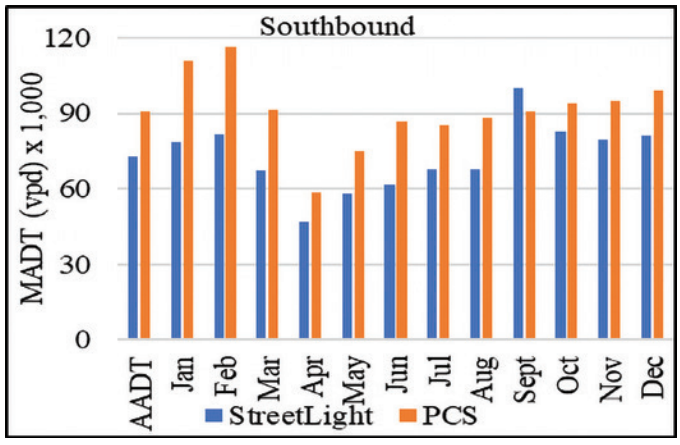
(a) — PCS on Flagler Memorial Bridge



(b) — PCS on Flagler Memorial Bridge



(c) — PCS on 1-95



(d) — PCS on 1-95

Figure 2. Comparison of MADT based on TPV Crowdsourced data vs. PCS Volumes.

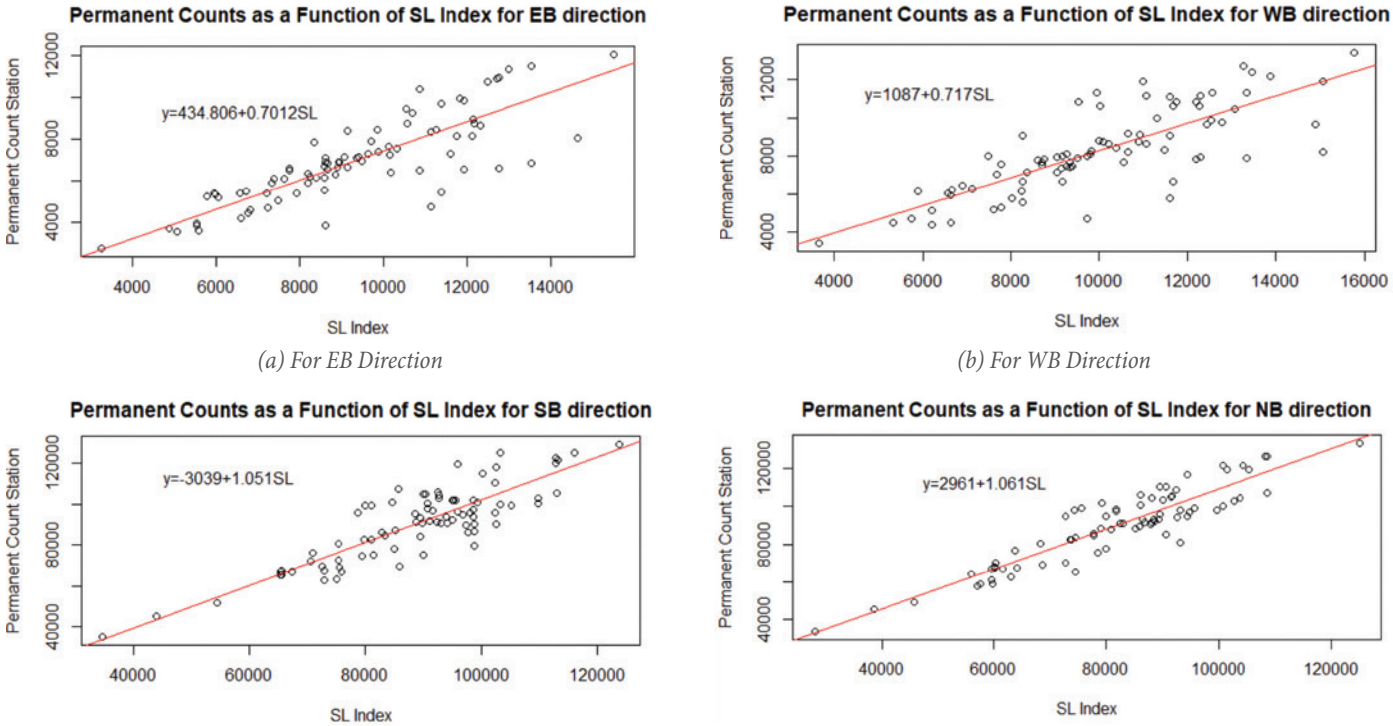
several parameters. In the regression models, the volumes based on the PCS values are used as as the response or dependent variables (indicated as PCS in the regression equation in the remainder of this paper) and RVI as the explanatory or independent variable (indicated as RVI in the regression equations in the remainder of this paper). Data aggregated daily from 11 months chosen randomly out of the 12 months in the year 2020 are used in developing the model and data from the remaining month is used for testing the results of the regression model. For each month, data collected from the PCS and RVI data for the same locations discussed in the previous section are averaged for each of the seven days in the week over the whole month, resulting in seven data points per month. For instance, all of the Mondays in a certain month are averaged.

Regression models are fitted for the two location sites in five variations, including a model for each of the four directions and a model with all four directions combined. The statistical inferences of the models are summarized in Table 3. The R^2 value is a goodness-of-fit measure that indicates the deviations in the dependent variable are explained by the independent variable in the fitted model. The higher the R^2, the better and more accurate the model is in predicting the real-world values RVI as an input. The fitted regression models, for every direction, have an R^2 greater than 60 percent, indicating that the models can predict more than 60 percent of the variation in the PCS values. The model developed with all directions has an R^2 of 97.6 percent, which is the best fitted model among all of the models. The p-value determines if the explanatory variable is significant in the model. A p-value less than a certain significance level, usually 5 percent, indicates that the explanatory variable is significant. In this case, all p-values are close to zero, indicating that, for all directions, the RVI aggregated daily for every month is a significant predictor of PCS values. The R^2 and p-value results in Table 3 indicate that there is a significant relationship between the PCS count and RVI.

Table 3. Regression statistical inferences for PCS counts as a function of RVI using daily data.

Direction	Regression Equation	R^2	p-value
EB	PCS = 434.8060 + 0.7012RVI	69.0%	0.000
WB	PCS = 1087.0000 + 0.717RVI	62.0%	0.000
SB	PCS = -3039.0000 + 1.051RVI	77.8%	0.000
NB	PCS = 2961.0000 + 1.061RVI	83.8%	0.000
All Directions	PCS = -2531.0000 + 1.082RVI	97.6%	0.000

The regression plots for the two different location sites in five variations, including a model for each of the four directions and a model with all four directions combined are shown in Figure 3 and Figure 4, respectively.



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 3. PCS counts vs. RVI fitted regression model using daily volumes.

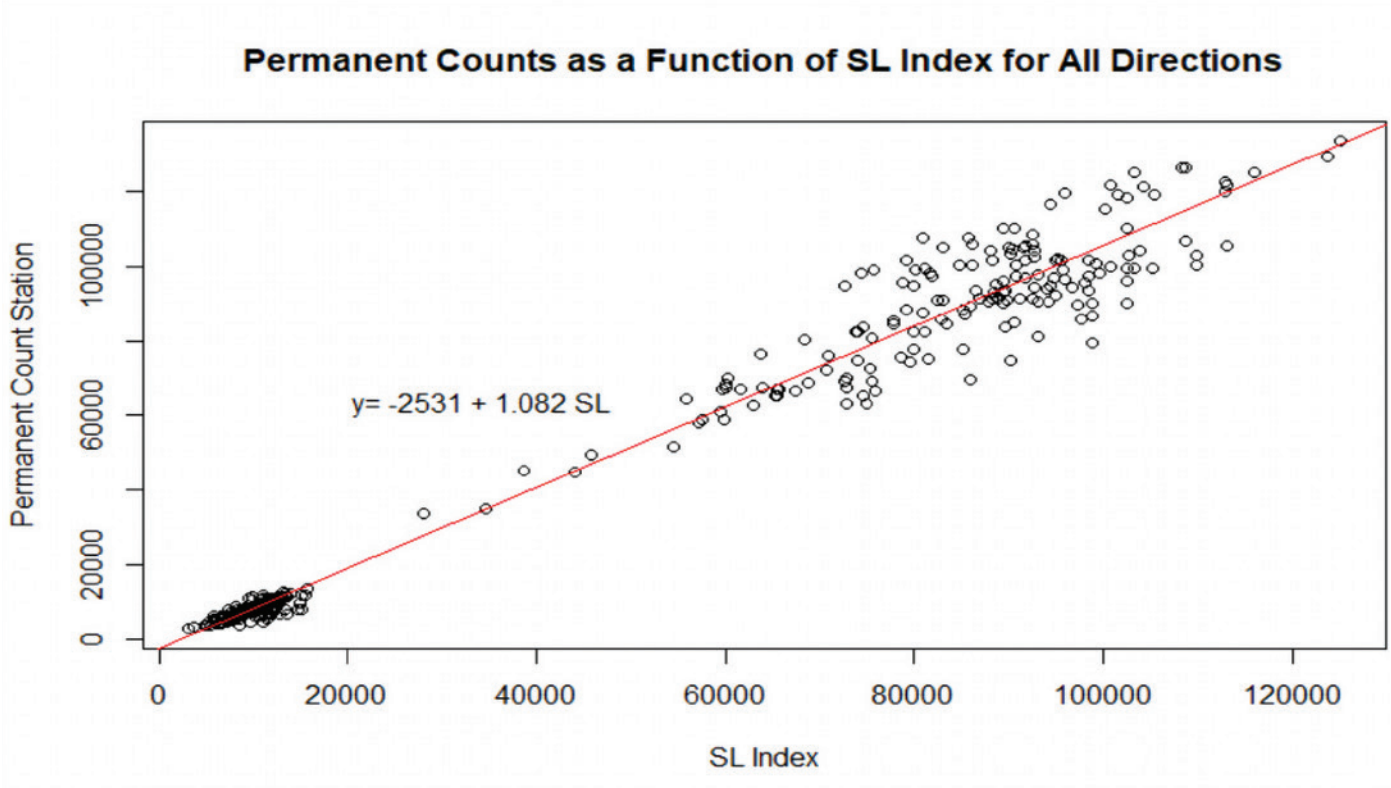


Figure 4. PCS counts vs. RVI fitted regression model for all directions using daily volumes.

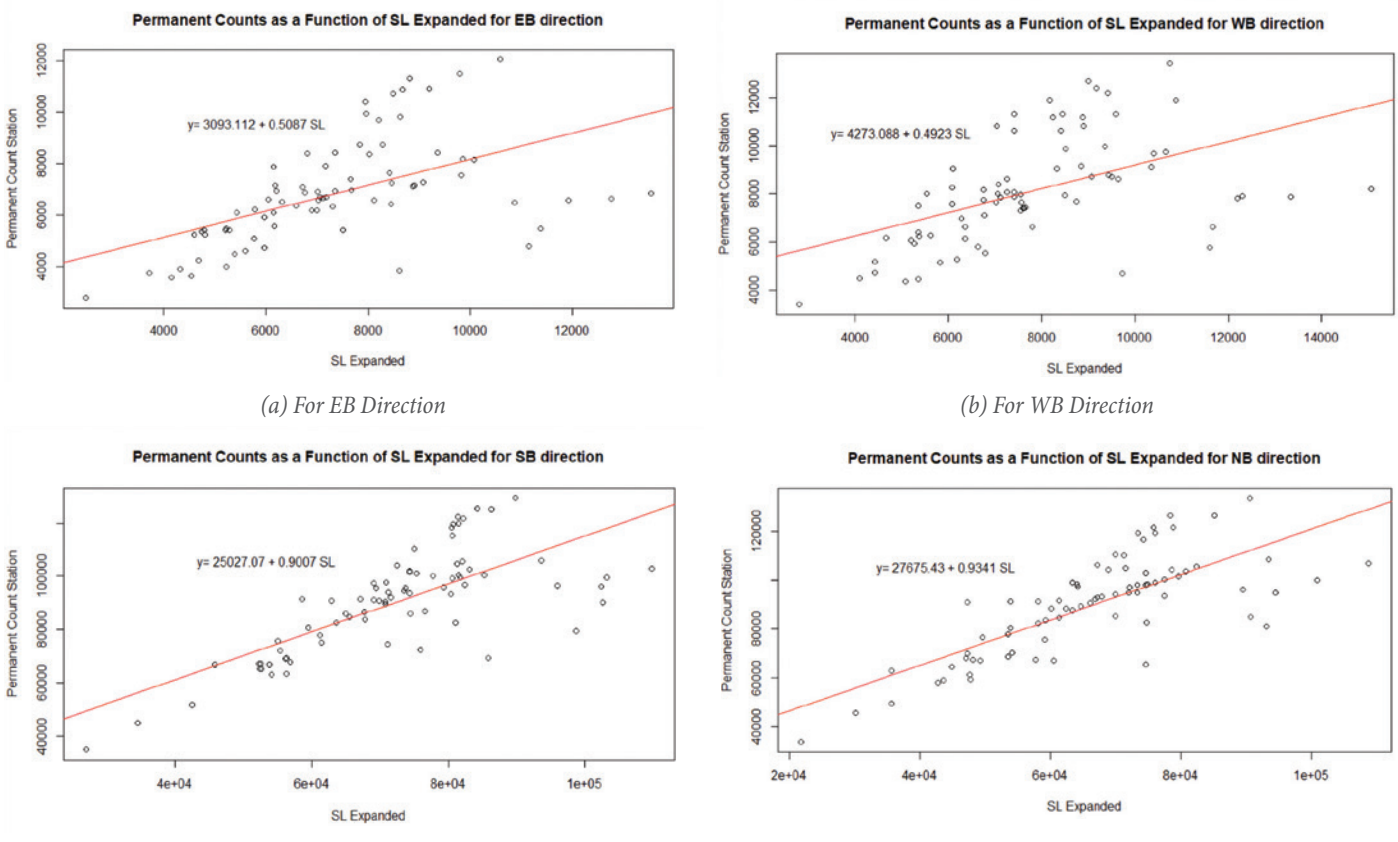
A similar approach was developed to correct the TPV volume estimates based on PCS measures instead of using the RVI as described above, again at the daily aggregation level, with the volumes based on the PCS values as the response or dependent variable (referred to as PCS in the regression equations) and TPV volume estimates as the explanatory or independent variable The statistical inferences of the fitted models are summarized in Table 4. The R^2 values are relatively lower for all five models compared to the models developed earlier using the RVI. The values are particularly low for the models developed based on the eastbound and westbound data, which are 24 percent and 29 percent, respectively. The lower R^2 values indicate that the models developed per direction for TPV volume estimates are not as accurate to predict the variations in the counts as the models developed based on the RVI. However, the All Directions model still produced a very good fit with an R2 of 94.1 percent. The reason that the All Directions model produces better results than the individual direction model is expected to be due the coverage of a wider range of volumes from low to high volumes in the data used for the All Directions model. Overall, the regression models between the PCS counts and the TPV volume estimates were less significant than those between the PCS counts and the RVI.

Table 4. Regression statistical inferences for PCS counts as a function of TPV Volume Estimates using daily data.

Direction	Regression Equation	R^2	p-value
EB	PCS = 3093.1120 + 0.5087TPV	29.0%	0.000
WB	PCS = 4273.0880 + 0.4923TPV	23.7%	0.000
SB	PCS = 25,027.0000 + 0.9007TPV	53.9%	0.000
NB	PCS = 27,675.4300 + 0.9341TPV	58.4%	0.000
All Directions	PCS = -547.0 + 1.281TPV	94.1%	0.000

The regression plots along with the regression models are shown in Figure 5 for EB, WB, SB, and NB and in Figure 6 for all directions, respectively. Based on the statistical inferences from the above results, regression models developed based on the RVI for daily volumes were selected for further analysis.

The performance of the regression models developed based on RVI to estimate the daily volumes are compared to the ground truth data collected from the PCS. Table 56 shows the MAPE computed for the training data and testing data for the two methods (expansion of RVI based on local volume data, and the original TPV volume estimate data). As shown in Table 5, the



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5. PCS counts vs. TPV Volume Estimates fitted regression model using daily volumes.

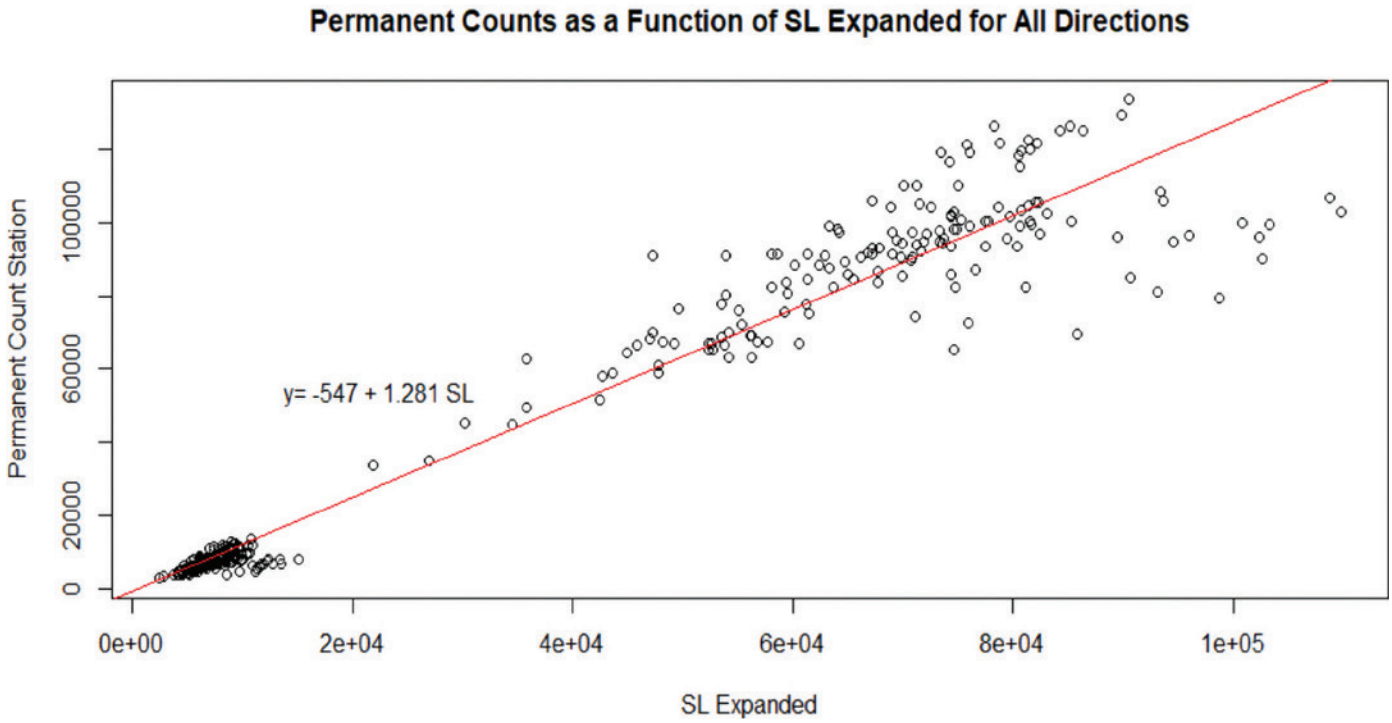


Figure 6. PCS counts vs. SL Expanded fitted regression model for all directions using daily volumes.

Table 5. MAPE of RVI regression models for daily volume estimation.

Used Model	Measure	Flagler Bridge EB	Flagler Bridge WB	I-95 NB	I-95 SB
TPV Volume Estimates	MAPE Training Data	22.28%	19.96%	17.62%	20.20%
	MAPE Testing Data	8.00%	17.00%	36.00%	29.00%
Direction-Specific Linear Regression based on RVI	MAPE: Training Data	12.63%	12.88%	7.53%	8.14%
	MAPE: Testing Data	7.40%	8.50%	6.10%	5.35%
All Direction Linear Regression based on RVI	MAPE: Training Data	18.19%	16.05%	5.81%	9.01%
	MAPE: Testing Data	9.43%	7.58%	5.51%	8.87%

use of the direction specific regression models based on the RVI produced a MAPE of 5.35 percent to 8.14 percent for I-95 and 7.40 percent to 12.88 percent for the Flagler Bridge. Depending on the specific applications, these errors may or may not be acceptable. The above results indicate that the development of regression models between PCS data and TPV data has the potential of providing acceptable results for daily volume estimation, particularly for higher volume segments. However, it should be remembered that the testing here is done using PCS Count and RVI data from the same location from which the data is collected for the regression. For the models to be useful, they need to be tested to estimate the volumes for locations other than the locations of the PCS station used in the regression. Such testing should be done in a future study.

Estimating Hourly Volume at the PCS Locations Using Regression Analysis based on RVI

The study aimed to enhance AADT estimates by developing regression models based on the hourly RVI and PCS hourly volumes. The data was divided into model development and testing datasets, with PCS values as the dependent variable and the RVI as the independent variable. Despite efforts to improve the accuracy by considering different variations in the regression models, such as averaging volumes over typical weekdays and peak hours, the results revealed that the models, when applied to the same PCS location, exhibited large errors in volume estimation.

Summary

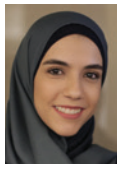
The article highlights significant discrepancies in the estimation of AADT, seasonal ADT, and MADT based on data from one PTV compared to ground truth data for two locations. To address this, the study refines volume estimation using regression models linking TPV data to PCS data. Notably, the regression models with RVI prove more statistically significant than those with the TPV volume estimates, producing a MAPE of 5.35 percent to 8.14 percent for I-95 and 7.40 percent to 12.88 percent for Flagler Bridge daily counts. While promising for daily volume estimation, challenges persist for hourly volumes, as all attempted regression models failed to produce good results.

The results suggest that developing regression models between PCS and TPV data holds promise for achieving acceptable results in daily volume estimation, especially for higher volume segments. However, it is crucial to note that the testing focused on PCS Count and RVI from the same location used in regression. To maximize usefulness, future studies should extend testing to estimate volumes in locations beyond the PCS station, ensuring broader applicability. Testing beyond the PCS location is crucial for the broader applicability of these models, emphasizing the need for future studies in this direction. While hourly volume regression models in this study showed large errors, the positive takeaway lies in the potential improvement and refinement that future research can bring to enhance accuracy. It is also possible that the TPVs will further improve their data and models in future years, allowing more accurate estimates of the volumes. **itej**

Reference

1. Morshed, S. A. (2022). "Enhanced Methods for Utilization of Data to Support Multi-Scenario Analysis and Multi-Resolution Modeling."
2. Yang, H., Cetin, M., & Ma, Q. (2020). Guidelines for using StreetLight data for planning tasks (No. FHWA/VTRC 20-R23). Virginia Transportation Research Council (VTRC).
3. Turner, S., Tsapakis, I., & Koeneman, P. (2020). "Evaluation of StreetLight Data's traffic count estimates from mobile device data" (No. MN 2020-30). Minnesota. Dept. of Transportation. Office of Policy Analysis, Research & Innovation.
4. Roll, J. (2019). "Evaluating Streetlight Estimates of Annual Average Daily Traffic in Oregon" (No. OR-RD-19-11).
5. Granato, S. "Various uses for INRIX/Streetlight data: Ohio plus border area. Ohio Department of Transportation" (ODOT), 2017.
6. Georgia Department of Transportation. "Existing Volume Development and Origin-Destination Data." Downtown Connector Study, 2016.

Syed Ahnaf Morshed, Ph.D. (M) *is a specialist in transportation systems, management, and operations, having recently completed his Ph.D. in Civil Engineering (Transportation) from Florida International University. Originally from Bangladesh, his expertise includes data analytics, transportation simulation, resilience, sustainability, planning, and operations. Before joining ITE, Dr. Ahnaf focused on traffic data analysis and statistical modeling. In his free time, he enjoys playing soccer, cooking, and cheering for Manchester United.*



Kamar Amine, Ph.D., EIT (S) *is a Traffic and ITS Engineer at Mead & Hunt. She has graduated from Florida International University with a Ph.D. in Civil Engineering, with a focus on Transportation Engineering. Her work involves federal and state-funded projects in the fields of Transportation Systems Management and Operation, simulation modeling, connected and automated vehicles, ITS technologies, and traffic data analytics in support of agencies' decision making. She is involved in traffic signal optimization as well as Advanced Traffic Management Systems. Kamar is an active member of ITE's TSMO council and serves as its social media liaison. In her free time, she is takes up gardening and enjoys a good old history book.*

Mohammed Hadi, Ph.D. *is the Director of the Leman Center for Transportation Research at Florida International University. He has more than 35 years of extensive experience in Transportation Systems Management and Operations, ITS, connected and automated vehicles, cooperative driving automation, simulation and dynamic assignment modeling, signal control, performance measurement, data analytics, and decision support systems. He has developed frameworks, methods, algorithms, and tools for data-based and model-based decision support of transportation agencies. He has worked on ITS planning, design, operations, and evaluation projects from around the United States and Puerto Rico; performed evaluation and testing of ITS technologies; and used the generated data in developing decision support tools for off-line and real-time operations. He currently serves as the chair of the Traffic Simulation Committee (ACP80) in the Transportation Research Board.*

Answer to "Where in the World" on page 11: Shibuya Scramble Crossing, Tokyo, Japan. Photo submitted by Stephen Byrd.