Assessing Connected Vehicle Data Coverage on New Jersey Roadways

Zijia Zhong, Liuhui Zhao, Branislav Dimitrijevic, Dejan Besenski, Joyoung Lee

John A. Reif, Jr. Department of Civil and Environmental Engineering

New Jersey Institute of Technology

Newark, NJ, USA

{zijia.zhong, liuhui.zhao, bxd1947, besenski, jo.y.lee}@njit.edu

Abstract—**The connected vehicle data (CVD) is one of the most promising emerging mobility data that greatly increases the ability to effectively monitor transportation system performance. A commercial vehicle trajectory dataset was evaluated for market penetration and coverage to establish whether it represents a sufficient sample of the vehicle volumes across the statewide roadway network of New Jersey. The dataset (officially named Wejo Vehicle Movement data) was compared to the vehicle volumes obtained from 46 weight-in-motion (WIM) traffic count stations during the corresponding two-month period. The observed market penetration rates of the Movement data for the interstate highways, non-interstate expressways, major arterials, and minor arterials are 2.55% (std. dev. 0.76%), 2.31% (std. dev. 1.07%), 3.25% (standard deviation 1.48 %), and 4.39% (standard deviation 2.65%), respectively. Additionally, the temporal resolution of the dataset (i.e., the time interval between consecutive Wejo vehicle trips captured at a given roadway section, time-of-day variation, day-of-month variation) was also found to be consistent among the evaluated WIM locations. Although relatively low (less than 5%), the consistent market penetration, combined with uniform spatial distribution of equipped vehicles within the traffic flow, could enable or enhance a wide range of traffic analytics applications.**

Index Terms—**Connected Vehicle Data, Probe Vehicle Data, Roadway Functional Classifications, Market Penetration, Average Daily Traffic**

I. INTRODUCTION

Probe vehicle data (PVD) is a mobile sensing method that has been widely used for ad-hoc traffic data collection. The scale of the PVD has evolved from the initial floating car runs using one or several probe vehicles, to fused data from embedded GPS trackers and cellphone mobile applications used by the participating drivers. The PVD has been used in numerous applications. Considering their spatial resolution, the PVD can be grouped into two groups: segment-based PVD and point-based PVD. The former provide an aggregate (prevailing) or disaggregate travel time and space-mean speed of vehicles along predefined roadway segments (e.g., traffic message channel (TMC) links, or segments divided by the position Bluetooth sensors), based on the positions and timestamps of tracked vehicles as they enter and exit the corresponding roadway segments. The segment-based PVD have been adopted by numerous transportation agencies, researchers, and practitioners. The point-based PVD report the waypoints of individual vehicles that constitute a vehicles trajectory as it travels the roadway network.

The PVD contain the data collected from a sample of vehicles in the traffic stream equipped with transponders that report the vehicle position at a certain preset frequency. Hence, the quality and reliability of analytics derived from the PVD depends on the size of vehicle sample relative to the total vehicle flow, which is referred to as "market penetration rate" (MPR). The GPS-transponder-based PVD can be augmented (fused) with the vehicle positioning data from smart devices, which can substantially boost the MP. However, such data is less consistent both in spatial and temporal respect due to the heterogeneity of the sensor pool [1]. For instance, the use of low-power consumption mode of mobile devices is an inherent disadvantage for sourcing PVD. The INRIX Trip Analytics is an example of a point-based PVD that uses fused sources of vehicle trajectory data, and reportedly covers 10% of all vehicle trips in the United States [2]. Other sources of point-based PVD typically have smaller MPR.

Even with its relatively sparse coverage at times (e.g., certain remote areas, night-time hours), the PVD have proven to provide substantial value for traffic analytics applications. Fan et al. compared 4-month INRIX trajectory data collected in 2015 against 84 continuous traffic count stations that were distributed throughout the State of Maryland. It was concluded that the PVD represented 1.5-2% of the total traffic flow on Maryland roadways and offered a promising approach to estimating vehicle-miles traveled (VMT) [3]. The INRIX trajectory data was used to develop GPS-based automated traffic signal performance measures (ATSPM) that was found to be spatially superior to segmented PVD (e.g., TMC travel time). The author also reported that an MP of the trajectory data in the range of 4% -10% was suitable for the proposed ATSPM, which requires a ping frequency of 5 s or less. Even with the MP below 0.04% (less than 1%), the proposed trajectory-based ATSPM allowed a quick diagnostics of immediate issues in a signalized corridor[4], [5].

The emerging connected vehicles (CVs) with the built-in wireless connectivity are expected to greatly improve the collection of data about the transportation systems. Recently, a new form of commercialized probe vehicle data has emerged that is collected from vehicle fleets through exclusive agreements with automotive OEMs (e.g., Wejo, Otonomo, High-

Mobility). The scale of such commercialized CV data presents an immense potential for transportation planning, operation & maintenance, and incident management. Besides high-resolution waypoint data, these emerging data sets also provide vehicle telemetry data (e.g., fuel level, ABS engagement, hard braking, windshield wiper activation). While "connected car data" has been used at times to differentiate this data from traditional PVD, these two terms are used synonymously in this paper. Based on Wejo's internal estimation, it receives vehicle telemetry data from 1 in 20 vehicle in the U.S., and 1 in 50 vehicles in Europe. Otonomo reportedly has more than 4 billion data points ingested into its platform daily from over 40 million licensed vehicles [6].

Khadka et al. demonstrated an application of Wejo trajectory data for detection of queue propagation at freeway bottlenecks, as well as the assessment of traffic congestion on arterial roadways based on the percentage of slow moving vehicles [7]. For safety-related applications, Wejo harsh-braking event data was used as a safety surrogate measure in a before-and-after study of conversion of a left-turn phase from protected-permitted to protected. The same authors used Wejo event data in a different study and found a strong correlation between rear-end crashes and harsh-braking events [8]. In another study, the impact of a long-term work zone on trip diversions onto alternate routes (i.e., local arterial) was assessed using Wejo trajectory data. Owing to ubiquitous coverage of Wejo trajectory data, 100 intersections in the impact area of the work zone were evaluated during a 11-week period [9].

The market penetration of Wejo trajectory data in various areas has been reported. In the Dallas-Fort Worth area in Texas, the Wejo data presented 10%-15% of all moving vehicles [7]. In the State of Indiana, the Wejo data was compared to traffic counts at 24 locations on Indiana roadways. The MPR was found to be 4.3% with standard deviation (STDEV) of 1.0% on interstate highways, and 5.0% with STDEV of 1.36% on non-interstate highways [10]. A study that used Wejo trajectory data (collected in 2019) to estimate border crossing times, reported an insufficient MPR: around 30% of the test hours were found to have no Wejo samples at the Paso del Norte port of entry in El Paso, TX. However, the study also confirmed the strong correlation between the Wejo travel time and the Bluetooth travel time within the border crossing area.[11].

The goal of this study is to assess the spatial-temporal coverage and consistency of Wejo probe vehicle data, which are critically important when considering application of vehicle trajectory data both in established and novel frameworks for assessing the performance of a transportation system. The highlights of this study are the following: i) the study scope focused exclusively on the New Jersey roadways across four roadway functional classifications that span all NJ counties, except Cape May County, ii) the evaluation covers 46 permanent traffic count stations that provide hourly volume counts throughout a day, and iii) the temporal analysis of the data is performed at various levels of aggregation.

II. Evaluation Framework

A. Data

The CV data used in the study (acquired from Wejo Group Ltd.) is exclusively crowd-sourced from automotive OEMs. It provides vehicle waypoints (latitude and longitude) and associated timetimestamps, along with other vehicle telemetry metrics reported at 1-3 second intervals. The trajectory coordinates are reported at 6-digit decimal precision, which is at the lane-level (i.e., 3-meter radius). The Wejo data used in this study (officially named Vehicle Movement data) included complete vehicle trajectory dataset collected within the State of New Jersey from April 15 through June 13, 2021.

The volume of the data, given its granularity, sample size, and coverage, posed a challenge in terms of data storage, processing, visualization, and analytics. For instance, the Movement data collected in New Jersey in June 2021 amounts to 8.17 billion data points from 22.18 million journeys. It should be noted that a unique journey ID is assigned each time the ignition is turned on in an equipped vehicle. As such, trip chaining could result in multiple unique journey IDs representing a single journey. It is assumed that within an hourly window, the vehicle traversing the bounding box is negligible.

The number of unique journey IDs derived from the trajectory data is compared to the hourly traffic counts collected at 46 weigh-in-motion (WIM) stations throughout the State of New Jersey as shown in Fig. 1. All of the WIM stations are classified as permanent count stations and are operated by the New Jersey Department of Transportation. The WIM data used in this study is publicly available via the NJTMS data portal [12]. As shown in Table I, the 46 WIM stations cover the roadways classified in four functional classes, from freeways to minor arterials.

Functional classification	Roadway category	Number of stations
1	Interstate	8
2	other freeway/expressway	12
3	other principle arterial	20
4	minor arterial	6
	Total	46

TABLE I: Roadway Classification & WIM Stations

To extract the vehicle waypoints from the trajectory dataset that correspond to each WIM station, the WIM station locations were geofenced by latitude/longitude boundaries covering approximately a 100 meter (300 ft.) buffer zone upstream and downstream. Manual verification was conducted to ensure the buffer zones do not include intersecting or adjacent roadways (e.g., underpass, overpass, frontage road), which could compromise the accuracy of the waypoint count corresponding to the WIM count. To process the massive amount of data in trajectory dataset, an Apache Spark clustering was employed.

B. Metrics

1) MPR: Market penetration is expressed as a percent of the entire (total) vehicle volume in a given area within a certain
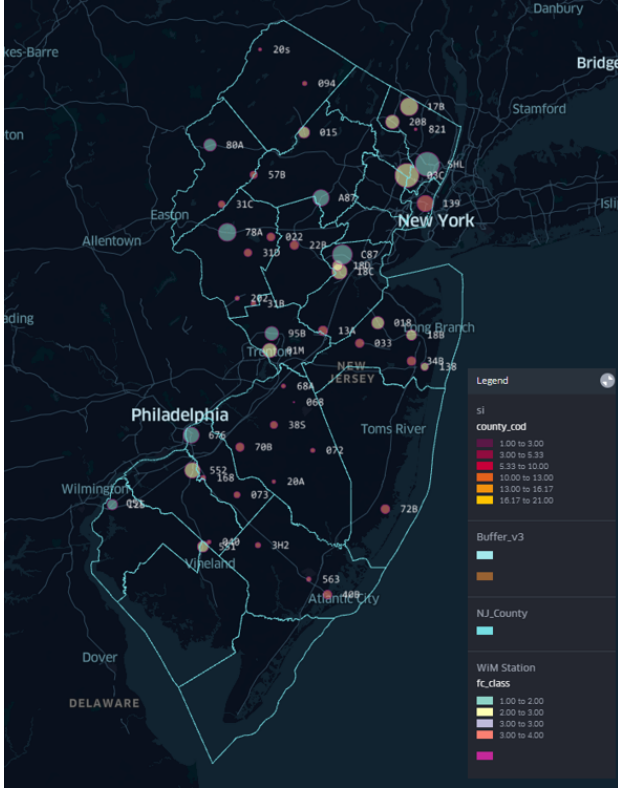
Fig. 1: 46 WIM permanent count stations

period contained in the trajectory dataset. In this study, the MPR for each hourly window was calculated using Eq. (1)

$$p^{ik} = \frac{V_m'^{ik}}{V^{ik}} \qquad (1)$$

where p^{ik} is the MPR observed during hour k at count station i, $k \in \{0, 1, ..., 23\}$ and $i \in \{1, 2, ..., 46\}$; $V_m'^{ik}$ is the observed Wejo volume at station i during hour k and V'^{ik} is the observed total volume at station i during hour k.

As part of the data processing, some hourly intervals were excluded from the analysis. Specifically, an hourly interval was excluded if the corresponding WIM station had no count report for that hour, or had reported hourly volume of less than 50 vehicles. For WIM stations configured to collect and report two-way vehicle counts, the hours in which they reported only one-way counts were also excluded from the comparative analysis (assuming the sensor(s) were offline). Eventually, 52,496 hourly counts were included in the comparative analysis (with 3,206 hourly records removed).

2) Spatial Distribution: The interval between the consecutive vehicles within the WIM station geofence was also analyzed for each WIM station. This concept is akin to time headway and can indirectly provide insights into the spacing between the tracked vehicles within the traffic stream. Assuming that their speed remains constant at a given roadway segment (e.g., equal to speed limit), the distribution of ping intervals is in essence equivalent to the distribution of time headway between the consecutive vehicles captured in the

trajectory dataset. To remove the possible long ping intervals, the sample points during nighttime hours (from 9 PM to 5 AM) were excluded from the analysis.

$$h_n^{ik} = \frac{\sum_2^n (t_n^{ik} - t_{n-1}^{ik})}{n} \qquad (2)$$

where h_n^{ik} is the nth time interval for station i during hour k, t_n^{ik} is the timestamp when the nth vehicle entered the WIM station bounding box, and t_{n-1}^{ik} is the timestamp when the $(n-1)th$ vehicle entered the WIM station bounding box, $n \geq 2$

III. RESULTS

A. Statewide MPR

The MPR statistics aggregated by roadway functional classification (FC) is summarized in Table II and the linearity for each FC between the Wejo count and the WIM count is shown in Fig. 2. For FC=1 (Interstate highways), the MPR was 2.55% with a STDEV of 0.75%. The MPR for FC=2 (other freeways and expressways) was 2.31% with a STDEV of 1.07%. The MPR for FC=3 (other principal arterials) and FC=4 (minor arterials) are noticeably higher than those for FC=1 and FC=2: 3.25% (Std. Dev 1.48%) for FC=3, and the 4.39% (STDEV 2.65%) for FC=4. The low STDEV for FC=1 indicates that the representation of traffic stream by the vehicle trajectory dataset on interstate highways is relatively consistent. The data also shows that the MPR for FC=4 stations experienced higher variation as compared to the MPR at stations in other functional classes, which can be attributed to relatively low vehicle volumes recorded at these WIM stations (averaging 280 vehicles per hour). It is also of interest to note that the average MPR values are within the 2.5-4.5% range, but their variations and probability distributions differ between different roadway functional classes. This can be seen in the cumulative distribution curves for each functional classification, shown in Fig. 3. For example, the MPR for FC=1 caps at round 5.74%, while the MPR for FC=3 was as high as 36.7%.

TABLE II: MPR Statistics

FC	Mean	Median	r^2	Std. Dev.	Avg. Hourly Vol.	Avg. Daily Vol.
1	2.55%	2.62%	0.949	0.76%	2,913	69,906
2	2.31%	2.28%	0.829	1.07%	2,329	55,890
3	3.25%	3.15%	0.836	1.48%	853	20,465
4	4.39%	3.69%	0.571	2.65%	282	6,763

Fig. 4 shows the MPR pattern based on the daily vehicle volume recorded at the WIM stations across the roadway functional classes. With the circle size representing the values of standard deviations, one can observe that the higher the daily volume at a station, the lower the MPR and the standard deviation. For instance, the WIM stations at the FC=4 roadways tend to have lower daily volume and their MPR could be higher at times and typically have higher variance. For FC=1 and FC=2 stations, the MPR generally falls in the lower spectrum, but with smaller standard deviations.
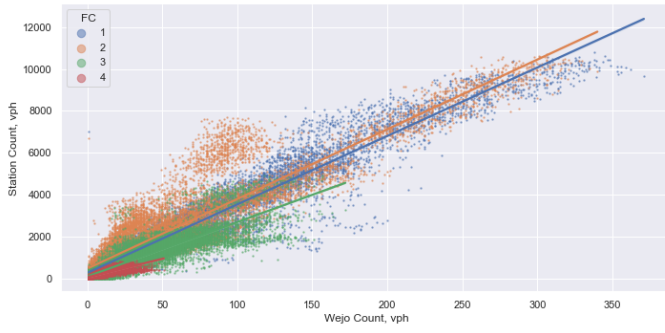
Fig. 2: Comparison of Wejo and hourly volume
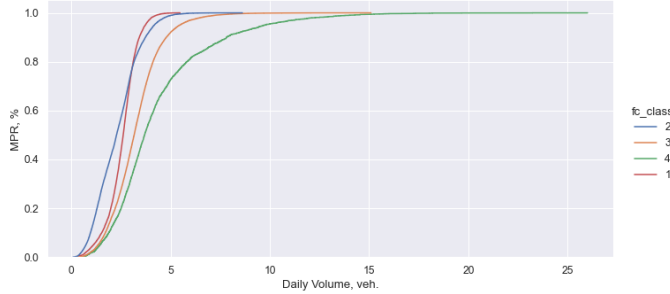

Fig. 4: MPR vs. station daily traffic count

Fig. 3: Cumulative distribution curve for MP values

B. Spatial Distribution

The cumulative distribution curves of headways between the probe vehicles (i.e., the ping intervals) are plotted in Figs. 5(a), 5(b), and 5(c) for WIM stations located at roadways of functional classes FC=1, FC=2, and FC=3, respectively. The WIM stations at FC=1 roadways, with the exception of station "80A", all have more than 80% of the ping intervals that are less than 60 seconds (some stations even have more than 80% of the headways that are less than 40 seconds). Based on the analysis it can be concluded that the spatial distribution of the trajectory data is sufficiently representative of the overall traffic stream during daytime. Due to relatively long ping intervals and lower overall volume during nighttime, the trajectory data may not provide sufficient data points for conclusive analytics during nighttime hours.

C. Temporal Variation

To investigate the time-of-day (ToD) variation, the MPR statics are grouped into four periods: "AM-peak" (6AM-10AM), "mid-day" (10AM-3PM), "PM-peak" (3PM-7PM) , and "off-peak" (7PM-6AM). The MPR comparison among the four ToD windows is shown in Fig. 6. The roadways in FC=1 have the narrowest interquartile range (IQR) among the four classes, followed by FC=3, then FC=2. The WIM locations at roadways in FC=4 have the largest IQR in each ToD window and the MRP tends to skew towards the higher values. The median MPR tends to edge higher during daytime (6AM to 7PM) at all locations.
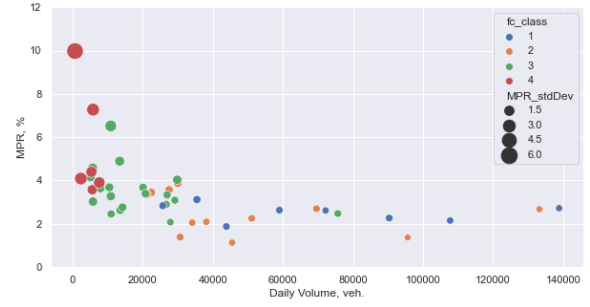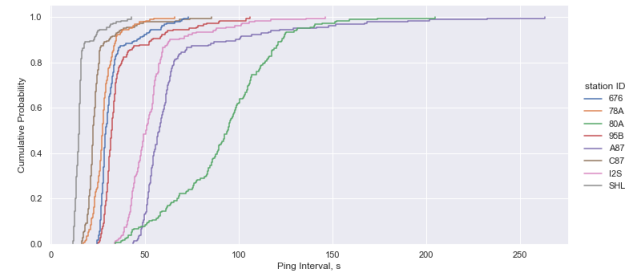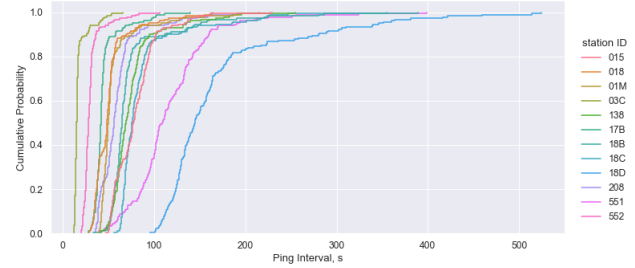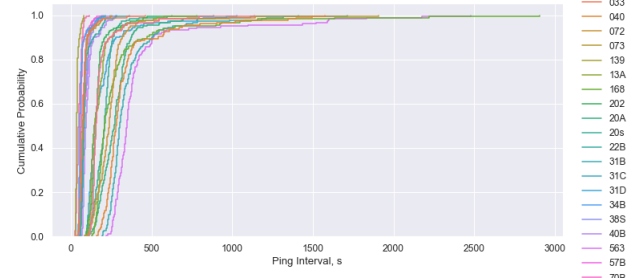

(a) FC-1(interstate highway), n=1,440

(b) FC-2 (freeway/expressway), n=2,160

(c) FC-3 (principle arterial), n=3,958

Fig. 5: Equipped vehicle in traffic stream

The comparison of MPR during weekdays and weekends is shown in Fig. 7. We observe that the MPR values are slight higher during weekends for FC=1. This could be attributed to the optional trips (non-commute trips) on interstate highways. It was observed that the weekends typically had sample points with higher MPR values, as high as 9%.

The following analysis also shows that the trajectory data, though being only a subset of the overall traffic stream, reflects
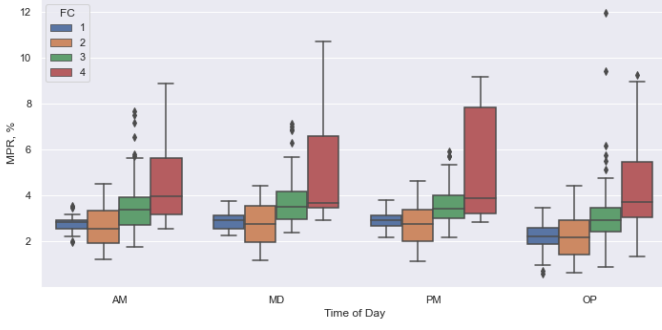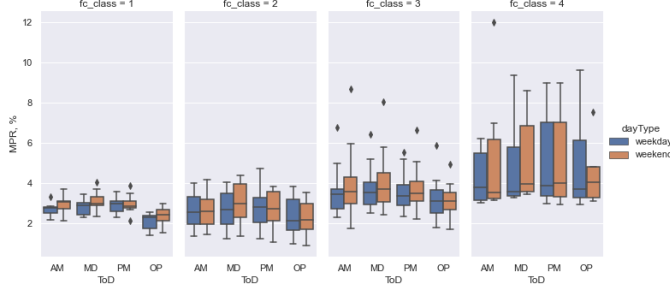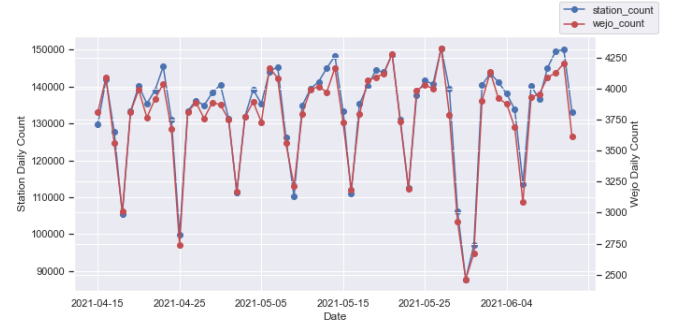
Fig. 6: MPR comparison for time of day

(a) Daily variation

Fig. 7: Weekday-weekend comparison

(b) Time of day variation

Fig. 8: Variation (station "03C" on NJ-3, $r^2 = 0.977$)

well the time-of-day and day-of-month traffic patterns. The observation from WIM station "03C" (NJ Route 3 in Clifton Township) is presented as an example of high correlation between the WIM counts and the corresponding trajectory counts ($r^2 = 0.977$). Fig 8 (a) shows the daily volume observed at the WIM station (left axis) and the Wejo count (right axis). Even though the two axes are of different scale, one can clearly see the Wejo volume closely tracking the daily variation of the ground-truth (WIM counts).

The ToD analysis for station "03C" is shown in Fig. 8(b). As expected, each period has the volume corresponding to the typical ToD variation with high traffic volume observed mostly in the AM and the PM peaks. Interestingly, both counts exhibit a high linear correlation. In another words, regardless the ToD, the MPR of the Wejo counts are highly consistent. It shoudl be noted that the WIM station "03C" is somewhat of an ideal case. For comparison, another WIM station is showcased, located on NJ-168, with less daily traffic. The daily WIM counts and trajectory counts at this station are less correlated (r^2=0.735), and the daily and ToD patterns are less aligned, as shown in Fig. 9.

IV. DISCUSSION

Based on the comparative analysis of vehicle counts and Wejo trajectory counts at 46 WIM stations across New Jersey, the spatial-temporal coverage and resolution of the trajectory data, make this data a great candidate for numerous traffic analytics applications. It was also observed that the trajectory data reflects traffic patterns even when the underlying MPR is less than 5%.

Previous studies reported that the critical value of MPR varies by the types of applications. In another words, the trajecotry data with low MPR could still contribute to specific operational analysis. For instance, in the case of signal performance measures, 1% penetration rate could provide insight into cyclic flow profiles of an intersection and similar performance could be achieved with the MPR as low as 0.1% when PVD is stacked over multiple days [13]. The signal coordination issues can be identified with less than 0.04% MPR [5]. For estimating vehicle miles traveled (VMT), 1.5-2% MPR was found to be sufficient [3]. Using the trajectory data is a promising and cost-effective approach for the volume and VMT estimation at roadways where traffic count data may be less reliable or infrequently collected, due to the ubiquitously coverage of Wejo data.

A study using Wejo data concluded that, based on the analysis of 8 intersections, one month of harsh-braking data is adequate for a reliable correlation with over 4-5 years' worth of crash data [8]. In certain cases, using the Wejo data for traffic analytics may present a challenge. For instance, a study of using Wejo data (2019) to estimate border crossing times found that the underlying Wejo MPR was insufficient, as 30% of test hours lacked Wejo samples [11]. It is our assessment, although this should be verified for specific locations and time periods, that the Wejo data generally reflects the actual traffic patterns, and could even be feasible source for real-time applications in active traffic monitoring and management practices, owing to the high resolution and consistent spatial and temporal representation of the traffic stream.
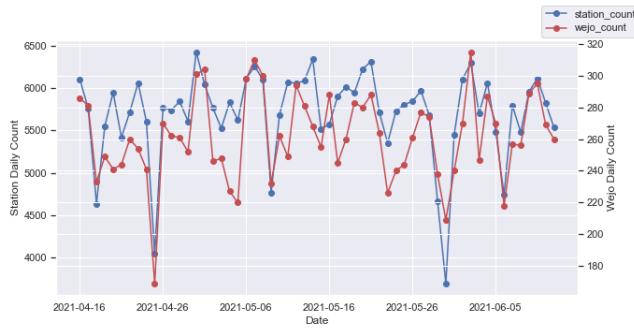
(a) Daily variation
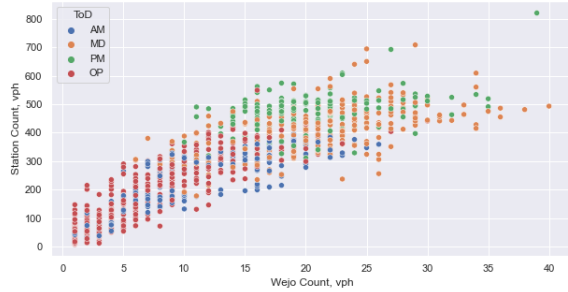
(b) Time of day variation

Fig. 9: Variation (station "168" on NJ-168, $r^2 = 0.735$)

V. Conclusion

This study conducted a large-scale evaluation of temporal-spatial converge and consistency of high-resolution commercial connected vehicle trajectory data (acquired from Wejo Group Ltd) on the New Jersey roadways. 17 billion records of trajectory points (approximately 9 TB, uncompressed) were analyzed to extract the volume data for 46 permanent (WIM) count stations across New Jersey. The results show that the sample of the trajectory probe vehicles relative to the total vehicle volume (referred to as market penetration or representation) resides within the range of 2.4-4.5%, on average, during daytime hours. In addition, the reviewed previous studies demonstrated general uniformity of the distribution of equipped vehicles in the traffic stream over various roadways, which could greatly increase the practicality of the trajectory data. Analysis of the temporal data variation (e.g., time-of-day, weekday-weekend comparisons) also shows the good representation of the traffic stream in general. The traffic analytics that can immediately benefit from the vehicle trajectory data include roadway risk profiling, crash timeline reconstruction, traffic incident impact monitoring, and driving behavior analysis.

References

[1] M. L. Pack, N. Ivanov, J. Bauer, E. Birriel *et al.*, "Considerations of current and emerging transportation management center data," United States. Federal Highway Administration, Tech. Rep., 2019.

[2] National Renewable Energy Laboratory, "NREL Examines U.S. Transportation Patterns During COVID-19 Pandemic." [Online]. Available: https://www.nrel.gov/news/program/2020/nrel-examines-us-transportation-patterns-during-covid-19-pandemic.html

[3] J. Fan, C. Fu, K. Stewart, and L. Zhang, "Using big gps trajectory data analytics for vehicle miles traveled estimation," *Transportation research part C: emerging technologies*, vol. 103, pp. 298–307, 2019.

[4] J. M. Waddell, S. M. Remias, J. N. Kirsch, and T. Trepanier, "Utilizing low-ping frequency vehicle trajectory data to characterize delay at traffic signals," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 8, p. 04020069, 2020.

[5] J. M. Waddell, S. M. Remias, and J. N. Kirsch, "Characterizing traffic-signal performance and corridor reliability using crowd-sourced probe vehicle trajectories," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 7, p. 04020053, 2020.

[6] Otonomo, "The Promise of Connected Vehicle Data." [Online]. Available: https://info.otonomo.io/hubfs/PDF/OOOO-Smart-Cities-Survey-Promise-of-data.pdf

[7] S. Khadka, P. i, and Q. Wang, "Developing novel performance measures for traffic congestion management and operational planning based on connected vehicle data," *Journal of Urban Planning and Development*, vol. 148, no. 2, p. 04022016, 2022.

[8] M. Hunter, E. Saldivar-Carranza, J. Desai, J. K. Mathew, H. Li, and D. M. Bullock, "A proactive approach to evaluating intersection safety using hard-braking data," *Journal of Big Data Analytics in Transportation*, vol. 3, no. 2, pp. 81–94, 2021.

[9] E. D. Saldivar-Carranza, M. Hunter, H. Li, J. Mathew, and D. M. Bullock, "Longitudinal performance assessment of traffic signal system impacted by long-term interstate construction diversion using connected vehicle data," *Journal of transportation technologies*, vol. 11, no. 4, pp. 644–659, 2021.

[10] M. Hunter, J. K. Mathew, E. Cox, M. Blackwell, and D. M. Bullock, "Estimation of connected vehicle penetration rate on indiana roadways," 2021.

[11] X. Li, M. Martin, B. Dadashova, D. Salgado, and S. Samant, "Exploring crowdsourced big data to estimate border crossing times," 2021.

[12] New Jersey Department of Transportation, "New Jersey Weight in Motion." [Online]. Available: https://www.njtms.org/njwimsite/

[13] C. M. Day and D. M. Bullock, "Detector-free signal offset optimization with limited connected vehicle market penetration: Proof-of-concept study," *Transportation Research Record*, vol. 2558, no. 1, pp. 54–65, 2016.