

Article

Impact of Privacy Filters and Fleet Changes on Connected Vehicle Trajectory Datasets for Intersection and Freeway Use Cases

Enrique D. Saldivar-Carranza ^{*}, Rahul Suryakant Sakhare , Jairaj Desai , Jijo K. Mathew ,
Ashmitha Jaysi Sivakumar , Justin Mukai  and Darcy M. Bullock 

Joint Transportation Research Program, Lyles School of Civil and Construction Engineering, Purdue University, West Lafayette, IN 47907, USA; rsakhare@purdue.edu (R.S.S.); desaij@purdue.edu (J.D.); kjjo@purdue.edu (J.K.M.); sivaku28@purdue.edu (A.J.S.); jmukai@purdue.edu (J.M.); darcy@purdue.edu (D.M.B.)

* Correspondence: esaldiva@purdue.edu

Highlights:

What are the main findings?

- Data fuzzification to ensure privacy and fleet mix changes in a current commercial connected vehicle (CV) trajectory dataset resulted in a data penetration decrease of $\leq 0.5\%$, and almost 70% of analyzed road segments experienced a reduction in the number of CV samples available for analysis when contrasted with a comparable historic CV dataset without fuzzified records.
- Similarly, around 70% of 26,291 intersection movements evaluated showed a reduction in the number of CV trajectories available for analysis.

What are the implications of the main finding?

- Even though there is a general reduction in the number of samples available for analysis with the current dataset, the authors believe that it provides enough information to derive most relevant trajectory-based mobility studies.
- The implemented data fuzzification, which consists of the truncation of coordinates within $\frac{1}{2}$ mile of frequently visited destinations (i.e., location blurring), may induce unintended biases for movement level intersection studies. Therefore, estimated intersection performance and the interactions between movements need to be carefully evaluated before taking mitigation activities.

Abstract: Commercially available crowdsourced connected vehicle (CV) trajectory data have recently been used to provide stakeholders with actionable and scalable roadway mobility infrastructure performance measures. Transportation agencies and automotive original equipment manufacturers (OEMs) share a common vision of ensuring the privacy of motorists that anonymously provide their journey information. As this market has evolved, the fleet mix has changed, and some OEMs have introduced additional fuzzification of CV data around 0.5 miles of frequently visited locations. This study compared the estimated Indiana market penetration rates (MPRs) between historic non-fuzzified CV datasets from 2020 to 2023 and a 5–11 May 2024, CV dataset with fuzzified records and a reduced fleet. At selected permanent interstate and non-interstate count stations, overall CV MPRs decreased by 0.5% and 0.3% compared to 2023, respectively. However, the trend in previous years was upward. Additionally, this paper evaluated the impact on data characteristics at freeways and intersections between the 5–11 May 2024, fuzzified CV dataset and a non-fuzzified 7–13 May 2023, CV dataset. The analysis found that the total number of GPS samples decreased 10% statewide. Of the evaluated 54,284 0.1-mile Indiana freeway, US Route, and State Route segments, the number of CV samples increased for 33.8% and decreased for 65.9%. This study also evaluated 26,291 movements at 3289 intersections and found that the number of available trajectories increased for 28.3% and decreased for 70.4%. This paper concludes that data representativeness is enough to derive most relevant mobility performance measures. However, since the change in available trajectories is not uniformly distributed among intersection movements, an unintended sample bias may be introduced



Citation: Saldivar-Carranza, E.D.; Sakhare, R.S.; Desai, J.; Mathew, J.K.; Sivakumar, A.J.; Mukai, J.; Bullock, D.M. Impact of Privacy Filters and Fleet Changes on Connected Vehicle Trajectory Datasets for Intersection and Freeway Use Cases. *Smart Cities* **2024**, *7*, 2366–2391. <https://doi.org/10.3390/smartcities7050093>

Academic Editor: Pierluigi Siano

Received: 6 August 2024

Revised: 27 August 2024

Accepted: 28 August 2024

Published: 30 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

when computing performance measures. This may affect signal retiming or capital investment opportunity identification algorithms.

Keywords: connected vehicle; trajectory; OEM; penetration; highway; intersection

1. Introduction

Roadway infrastructure is critical for moving an ever-increasing number of people and goods. There are over 4 million miles of public roadways in the United States [1]. In 2023, there were over 3.2 trillion vehicle miles traveled in the country [2], a 13% increase from 2003 [3]. It is estimated that each year, motorists pay over \$1000 in wasted time and fuel while traveling the transportation networks [1]. Therefore, it is important for transportation agencies to actively monitor the performance of their managed mobility infrastructure to identify improvement opportunities and determine the best allocation of limited funds.

Interstates, US Routes, State Routes, and arterials serve most of the traffic demand [2]. Several state agencies liaise with Traffic Management Centers (TMCs) to monitor these types of roadways and coordinate incident response [4]. Historically, a network of roadside sensors and intelligent transportation system (ITS) cameras installed statewide, as well as segment-based crowdsourced speed and travel time data, have been used to assess prevailing traffic conditions and identify challenges [5,6]. However, sensors and ITS cameras only provide location-specific information, require regular maintenance, and are difficult to observe on a large scale. Furthermore, crowdsourced probe vehicle speeds and travel time data are usually aggregated, which complicates the detailed analysis of traffic conditions often requiring granular information.

Roadway intersections are another critical component of transportation networks. There are over 400,000 signalized intersections in the United States, which are estimated to contribute up to 10% of all traffic delay on the National Highway System [7]. A properly managed traffic signal can reduce congestion, enhance mobility, and decrease delays and the number of vehicles stops [8,9]. Over the last two decades, agencies have implemented Automated Traffic Signal Performance Measures (ATSPMs) to proactively assess intersection operations. ATSPMs are visualizations and tools derived from traffic signal controller high-resolution (tenth-of-a-second) output and detector data [10,11]. Actionable insights are derived thanks to the data's high reporting intervals [12]; however, ATSPMs are difficult to scale, and the estimation of performance measures is sensitive to traffic conditions [13] and detector configuration [14].

In recent years, commercial connected vehicle (CV) trajectory data has emerged as an alternative dataset to actively monitor roadway and intersection performance [13,15–17]. The main benefit of this dataset is that it allows for the accurate estimation of traffic conditions at a variety of levels, ranging from the analysis of localized intersection movements [13] to nationwide mobility [18].

1.1. Connected Vehicle Trajectory Data

It is anticipated that in 2025, 470 million CVs will be in operation in the US, Europe, and China [19]. A study of perhaps the largest provider of CV data in 2022 reported that, on average, one in every twenty vehicles in the United States provided telematics-based CV data through a data broker that could be used to estimate interstate and arterial performance measures [20].

1.1.1. Description

Crowdsourced CV trajectory data consist of sets of waypoints that describe the journey that equipped vehicles undertake as they traverse the roadways. The waypoint reporting interval for the same vehicle is usually in the order of a few seconds, and the spatial accuracy is usually in the order of 2–3 m. Every waypoint has the following descriptive

information attached: latitude, longitude, timestamp, speed, heading, and an anonymous trip identifier. By chronologically linking individual waypoints with the same trajectory identifier, the journey of a vehicle can be obtained.

1.1.2. Applications in Transportation

Commercial CV trajectory data provide accurate information of vehicles' journeys at virtually any scale. This characteristic makes them a good candidate for a variety of transportation studies.

Sakhare et al. have leveraged the CV dataset to measure and visualize freeway conditions, evaluate incident response, and assess work zone performance [15]. State transportation agencies, such as the Indiana Department of Transportation (INDOT), use CV data, in conjunction with other ITS assets, to monitor roadway performance and safety [21].

CV trajectory data have also been used to evaluate intersections. Various techniques have been developed to derive traffic signal performance measures [13,16,17] with the objective of identifying challenges and signal retiming opportunities [13]. Additionally, CV-derived traffic signal, roundabout, and stop-controlled intersection performance measures have been used to locate statewide capital investment opportunities [22], helping agencies perform data-driven investment decisions.

Other studies have used CV trajectory data for a wide range of purposes. Desai et al. used CV data to assess electric vehicle (EV) usage and charging infrastructure [23]. Alsaifi et al. created an algorithm that can create and update road maps and identify their characteristics from vehicle trajectories [24]. Further research has focused on the estimation of infrastructure characteristics from CV data, such as traffic volumes [25], vehicle miles traveled [26], roadway speeds [27], and traffic signal timing [28,29].

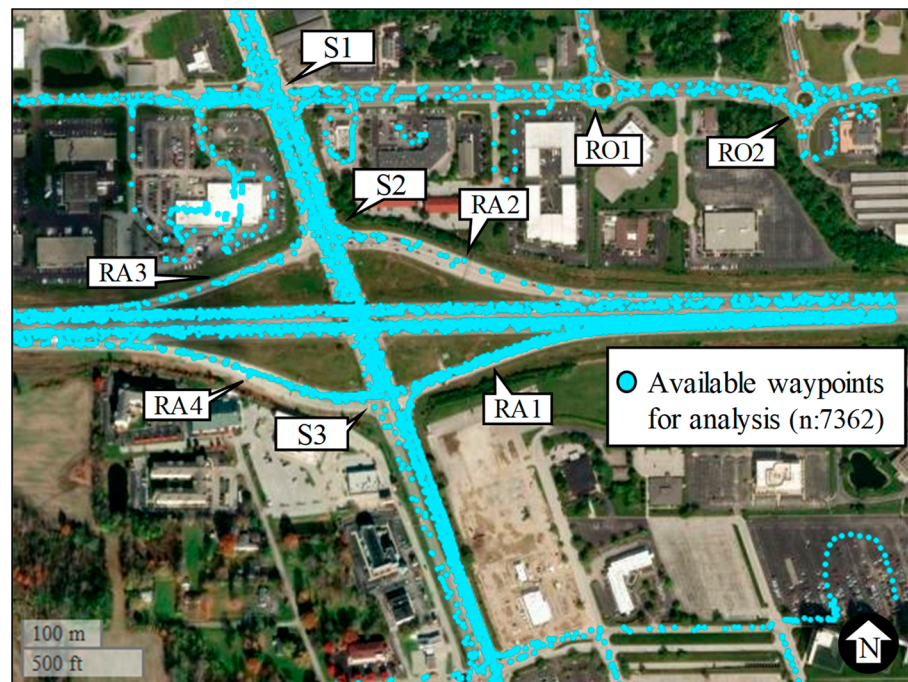
1.2. Motivation and Objective

As the commercial CV data industry matures, the dataset characteristics also evolve with the objective of ensuring the privacy of motorists that anonymously provide their journey information while attempting to maintain the scale and granularity needed for transportation studies. One such change in the CV dataset that occurred in 2024 is the fuzzification of trajectory waypoints. Waypoint fuzzification entails the distortion of selected records to protect sensitive information in a manner that attempts to minimize information loss.

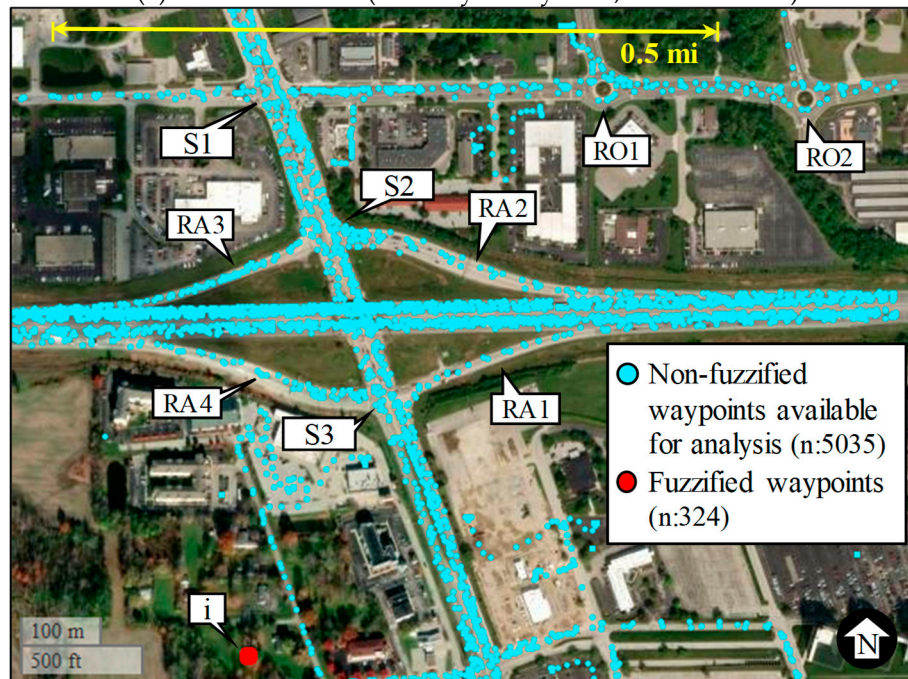
The fuzzification approach implemented in a current CV dataset truncates latitude and longitude coordinates to two decimal points (location blurring) when vehicles are located within 0.5 mi of frequently visited locations. Furthermore, when a waypoint is fuzzified, speed and heading values are not available.

Qualitatively, the impact that this fuzzification has on data availability and distribution is shown in Figure 1. Figure 1 compares 10 min of data collected on the same day-of-week (DOW) and time-of-day (TOD) during the second week of May between a historic 2023 dataset with non-fuzzified waypoints (Figure 1a) and a current 2024 dataset with fuzzified waypoints (Figure 1b). It is important to note that, in addition to the fuzzification difference, the historic dataset is comprised of more OEM fleets than the current dataset, which is by itself expected to affect the level of representativeness.

In Figure 1, the overall waypoint sample size decreased 27% from the historic to the current dataset. Of all waypoints available for the region shown in Figure 1b, 6% are fuzzified (callout i shows the location of the truncated GPS coordinates), and the rest are available for analysis. Areas on and between ramps (RA), signalized intersections (S), and roundabouts (RO) saw fewer sampled waypoints. In particular, RA1 and S1 show a noticeable decrease of traversing vehicles.



(a) Historic dataset (Tuesday 9 May 2023, 17:00–17:10 h.)



(b) Current dataset (Tuesday 7 May 2024, 17:00–17:10 h.)

Figure 1. Non-fuzzified waypoints available for analysis and fuzzified waypoints sampled during a 10-min period.

As transportation agencies, the private sector, and academia continue to use and invest in CV trajectory data, it is important to assess the impact that privacy filters (i.e., fuzzification) and fleet changes may have on derived studies. Since no previous study has provided such an analysis, the objective of this study is threefold:

1. Evaluate the current CV market penetration rate (MPR) and compare it to previous years' estimations.
2. Assess the impact of privacy filters and fleet changes on interstate, US Route, and State Route coverage.

- Evaluate the change on available vehicle trajectories for analysis by movement at traffic signals, roundabouts, and all-way stops.

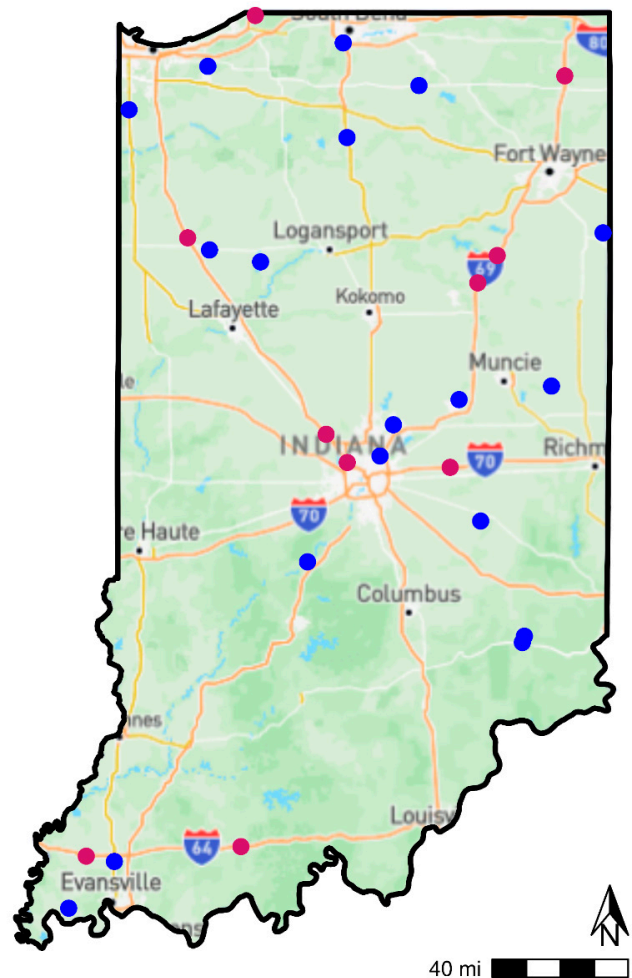
These analyses provide stakeholders with insights on the data representativeness changes and possible effects on related studies. All assessments are conducted using statewide Indiana CV trajectory data.

2. Market Penetration Rates

The MPR provides agencies with a key metric to answer how representative the data are of the actual traffic and is essential for building confidence in the data. The MPR is the estimated percentage of the vehicles on the roadways that provide their trajectory information.

The MPR of CV data with fuzzified records was evaluated over a week from 5–11 May 2024. The actual traffic volume information was collected from INDOT’s count stations. A majority of the count stations in Indiana use loop detectors [30–32] to count and classify vehicles. A total of 28 count stations that were operational during the entirety of the same week, as shown in Figure 2, were chosen for analysis. Of the 28 stations, 10 were along interstates and the remaining 18 were along non-interstate roadways that cover various geographies in Indiana.

Count station location: ● Interstate (10) ● Non-interstate (18)



Basemap (c) Mapbox, (c) OpenStreetMap

Figure 2. Locations of twenty-eight count stations in Indiana.

A virtual box a quarter-mile long and as wide as the road width was created around every count station. Unique journey identifiers were counted within this box and assumed as the trajectory counts for the CV data. Heading information from individual waypoints from the CV data was used as a filter to exclude journeys along a different route and direction. Since the GPS coordinates of fuzzified records are truncated, they were excluded from the MPR analysis. The MPR for a count station was calculated using the following equation:

$$P_n^{CV} = \left(\frac{T_n^{CV}}{V_n} \right) * 100 \quad (1)$$

where P_n^{CV} is the MPR of count station n , T_n^{CV} is the number of unique trajectories from the CV data within the quarter-mile-long bounding box near count station n , and V_n is the volume of vehicles from the same count station over the same time period.

Figures 3 and 4 show vehicle volumes, trajectories from CV data, and MPR by count station along interstate and non-interstate roadways. On 10 interstate locations, the vehicle volume ranged from 0.11 to 0.61 million over a week. During the same week, the identified CV trajectories ranged from 5.1 to 25.4 thousand. The MPR ranged from 3.5% to 6%. On 18 non-interstate locations, the vehicle volume ranged from 0.007 to 0.53 million and the identified CV trajectories from 0.28 to 19 thousand. The MPR ranged from 2.4% to 10.2%.

The reported overall MPR for interstate and non-interstate is given by Equation (2):

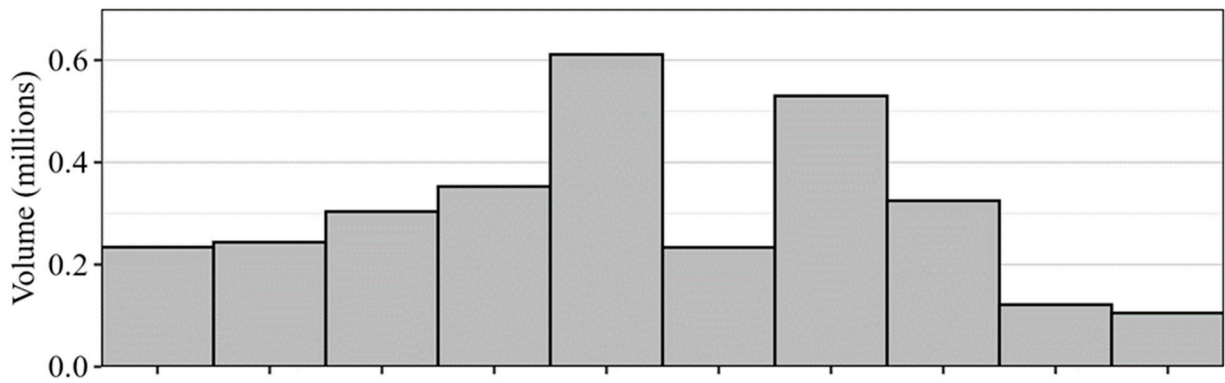
$$P_r^{CV} = \left(\frac{\sum_{n \in R} T_n^{CV}}{\sum_{n \in R} V_n} \right) * 100 \quad (2)$$

where $r \in \{Interstate, Non-Interstate\}$, R is the list of count stations along r , and P_r^{CV} is the overall MPR of the CV data. Due to the variation in volume at select count stations, total CV trajectories and traffic volumes across the analysis period were considered for the overall MPR calculations. This is also a comparable metric with changing stations and analysis period days across the years. In 2024, the overall MPR on interstates was reported as 4.6% and on non-interstates it was 5%.

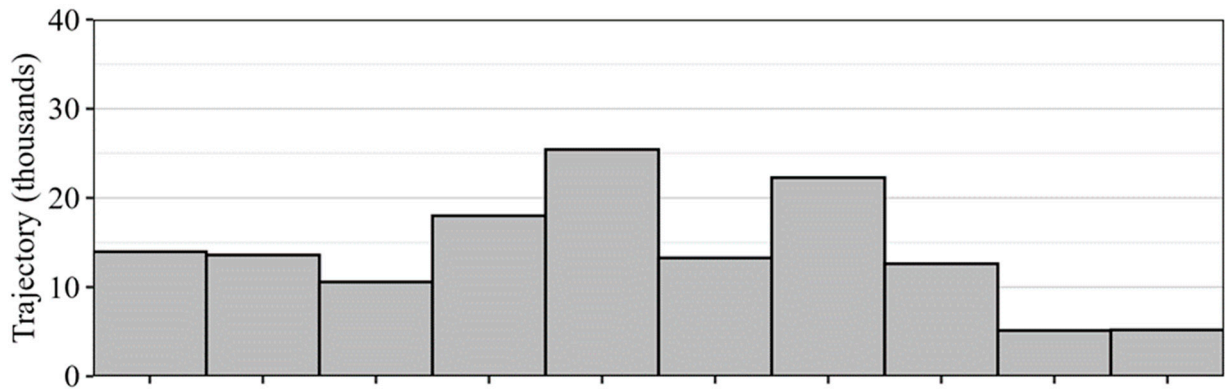
MPR Comparison with Previous Years

The 2024 MPR was compared to reported values in previous studies from 2020 to 2023 [20,33]. Table 1 presents a summary of the MPR analysis across the five years. The number of days analyzed during each year and the number of count stations evaluated varied depending upon the availability of the data. However, a minimum of a contiguous one-week period and at least seven count stations by road type were assessed during any of the years. Analysis was conducted during the month of August in 2020 and 2021 and in May for 2022, 2023, and 2024.

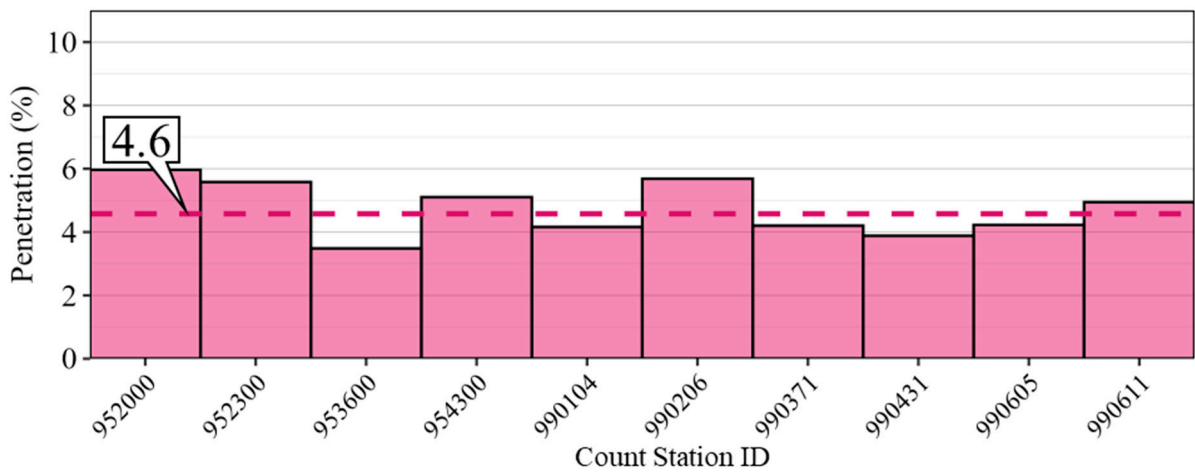
The MPR increased every year from 2020 to 2023 along both interstate and non-interstate roadways. Interstate MPR increased from 4.4% to 5.1%, whereas non-interstate MPR increased from 4.6% to 5.3% during this period. The increase may be due to the adoption of newer CVs and/or changes in commercial arrangements in the data supply chain. In 2024, the MPR dropped 0.5% on interstates and 0.3% on non-interstates. The reduction in MPR is possibly due to the fuzzified records, a reduced fleet, and/or a change in the data supply landscape.



(a) Vehicle volume.



(b) Unique CV trajectories.



(c) Penetration of CVs.

Figure 3. CV penetration in 2024 across ten interstate count stations in Indiana.

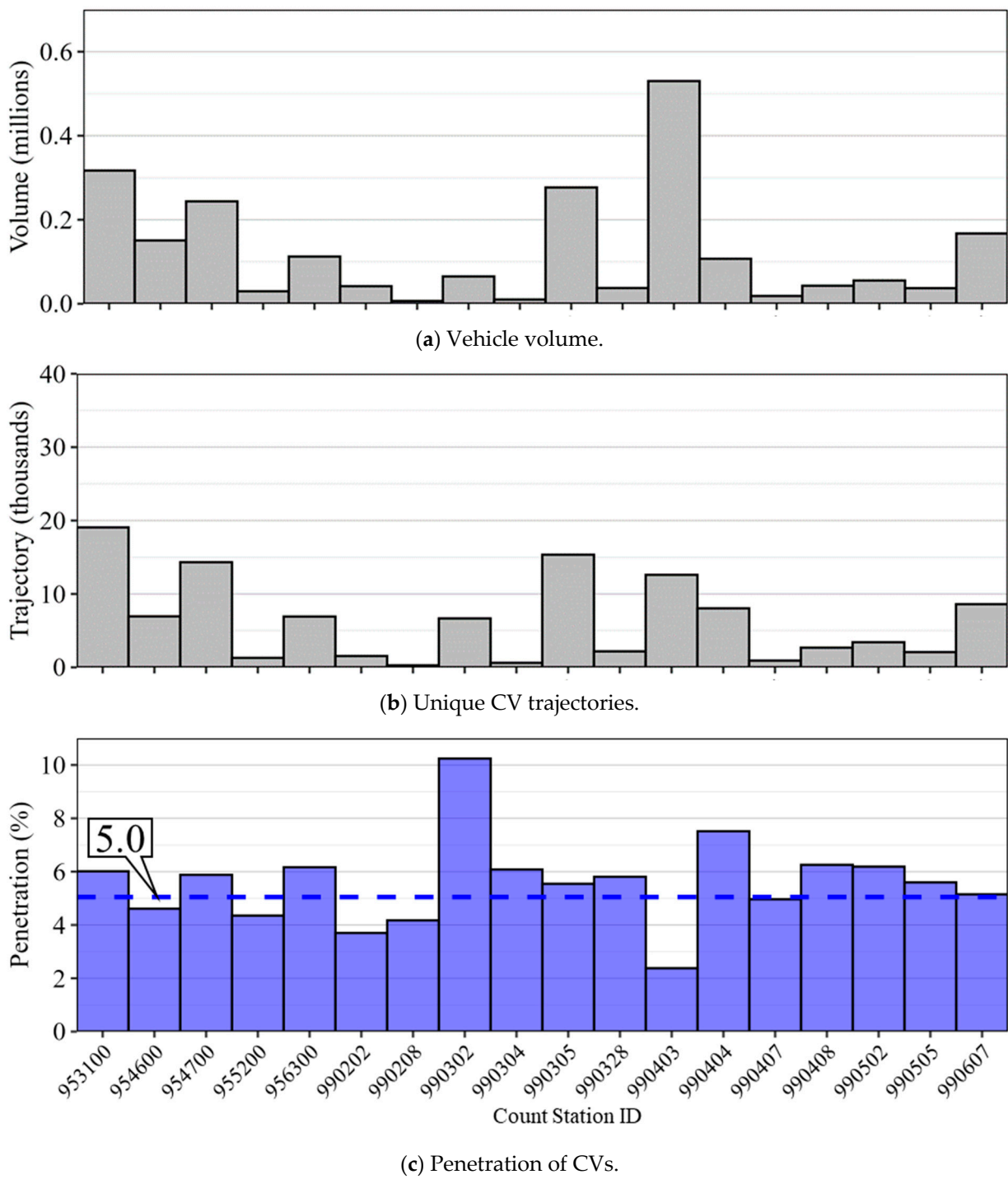


Figure 4. CV penetration in 2024 across eighteen non-interstate count stations in Indiana.

Table 1. Summary of longitudinal traffic penetration analysis of connected vehicles.

	Interstate					Non-Interstate				
	2020	2021	2022	2023	2024	2020	2021	2022	2023	2024
Count stations analyzed	21	29	18	7	10	32	29	25	10	18
Analysis period (days)	31	31	7	7	7	31	31	7	7	7
Total traffic volume (millions)	28.5	60.6	8.1	2.1	3.1	18.8	24.6	2.7	1.3	2.2
Average daily volume per station	43,706	67,331	65,057	42,794	43,723	18,948	27,323	15,503	18,096	17,857
Total CV trajectories (millions)	1.2	2.7	0.4	0.1	0.1	0.7	1.2	0.1	0.07	0.1
Average CV trajectories per station	1258	2958	3197	2175	1112	876	1319	798	967	901
MPR (%)	4.4	4.4	4.9	5.1	4.6	4.6	4.8	5.1	5.3	5.0

Due to changes in operating conditions, maintenance issues, or work zone activity, not all count stations are available across the five years. For comparison of MPR across the years at a count station level, 12 common count station locations were compared from 2021 to 2024. The details of these 12 count stations are shown in Table 2. Highway or roadway details with approximate mile marker (MM) information is also provided in Table 2.

Table 2. Count station details.

Station No.	Count Station ID	Road Type	County	Location Description
1	952300	Interstate	Grant	I-69 RM 268.2
2	954300	Interstate	Laporte	I-94 MM 44.5
3	990206	Interstate	Huntington	I-69 NB MM 78.2
4	990371	Interstate	Marion	I-65 MM 121.5
5	954600	Non-interstate	Marshall	US-31 (SR 10)
6	954700	Non-interstate	Porter	SR-49 (N E. 600 N)
7	955200	Non-interstate	Ripley	US-50 (RD 175 W)
8	990202	Non-interstate	Elhart	US-6 EB RM 93.6
9	990305	Non-interstate	Marion	Binford Blvd
10	990502	Non-interstate	Morgon	SR-67 SB RM 80.6
11	990505	Non-interstate	Ripley	US-421 SB RM 29.2
12	990607	Non-interstate	Vanderburgh	US-41 NB RM 15.3

Figure 5 shows MPR values for each of the common 12 count stations from 2021 to 2024, colored by the roadway type. The dotted line represents overall MPR calculated using Equation (2) across all the available stations during that year. The relative position of the count stations remained the same over the years.

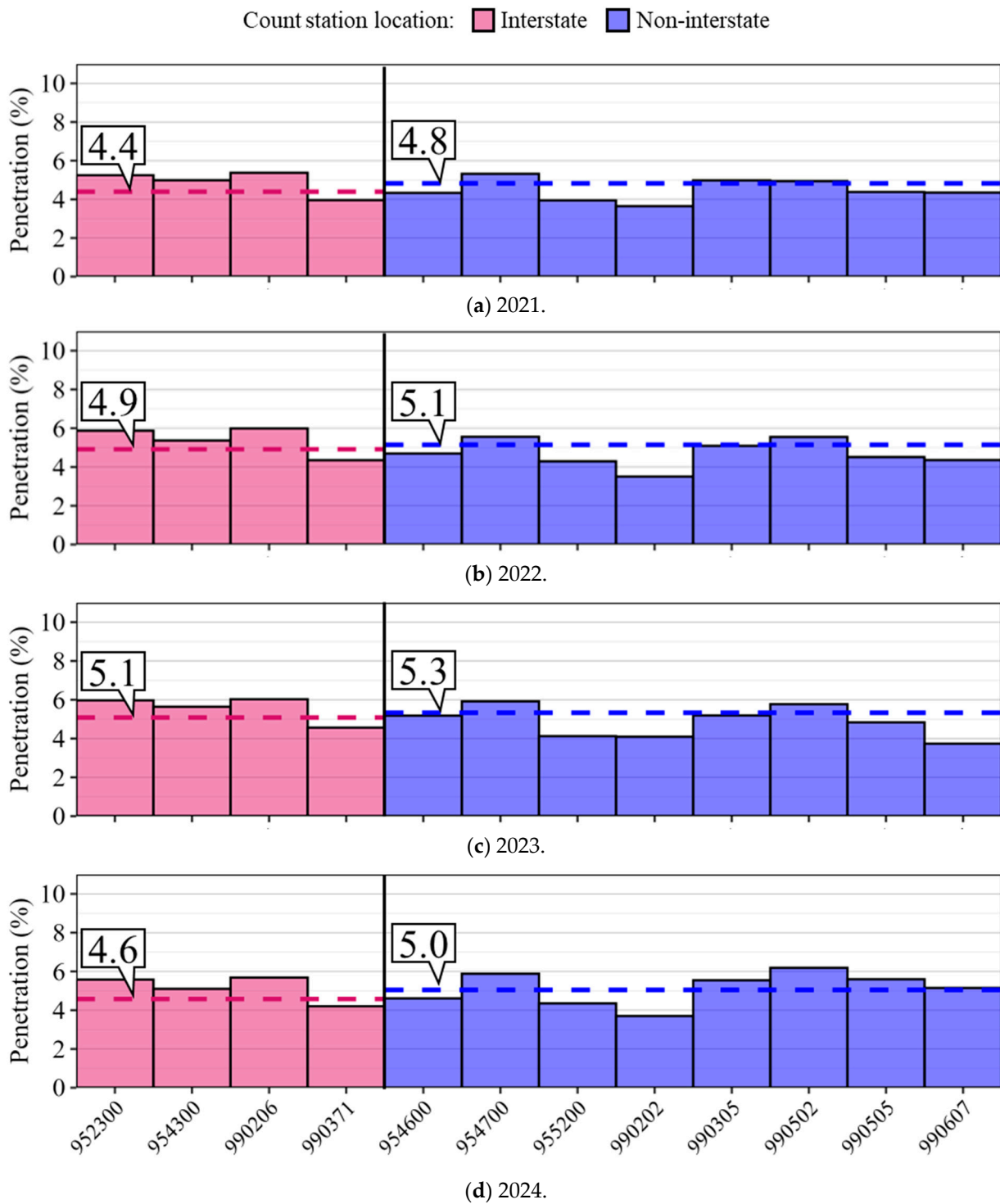


Figure 5. CV penetration comparison over the years across 12 common count stations.

3. Interstate, US Route, and State Route Coverage

Observing changes in CV data representativeness on interstates, US Routes, and State Routes is vital for agencies, as these road networks make up a majority of the roadway infrastructure maintained by a state. Knowing the representativeness of CV data will help identify opportunities for utilizing this CV data for continuous roadway mobility monitoring statewide, especially in locations with no existing sensor infrastructure. A total of twelve interstates (pink), four US Routes (blue), and three State Routes (orange) were

analyzed in this section, as shown in Figure 6. Interstate 80 (I-80) was excluded from the analysis owing to its full concurrency in the state of Indiana with routes I-94 and I-90.

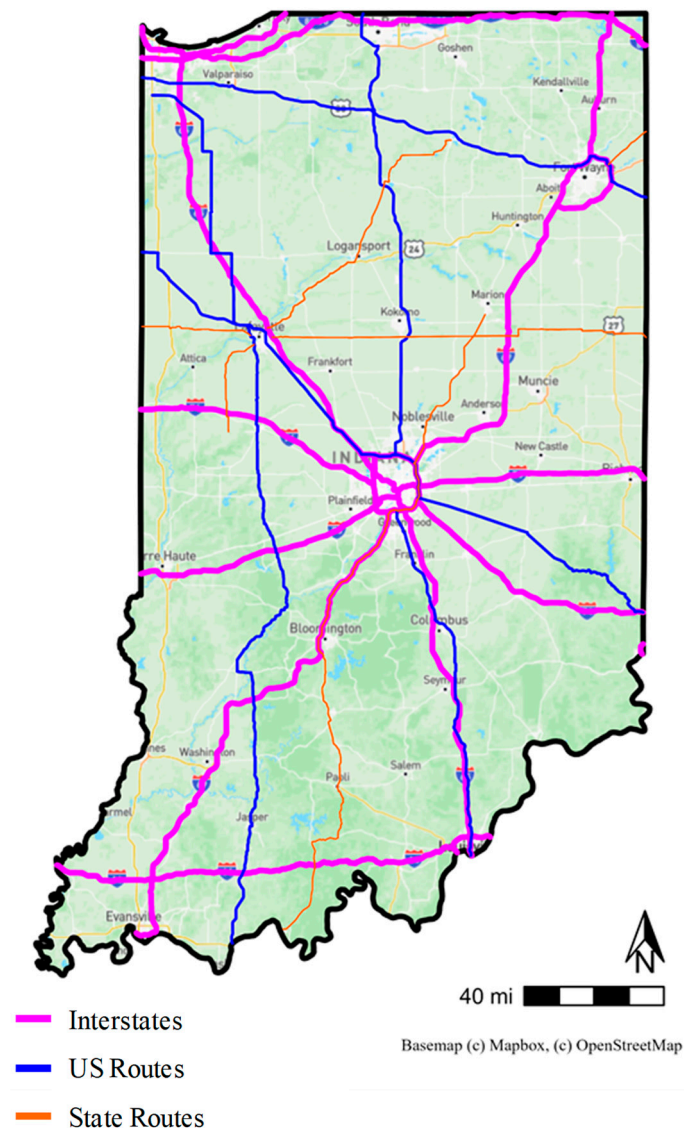


Figure 6. Map of Indiana showing routes analyzed—12 interstates, 4 US Routes, and 3 State Routes.

Each route was divided into tenth-of-a-mile segments (0.1 miles) for the analysis and for consistent comparisons between multiple years of data. A corresponding geospatial polygon for each such 0.1-mile segment was created, and CV data were matched to these segments to determine the exact mile-marker location along a route through which a CV waypoint passed. Following this geospatial matching process, distinct counts of CV journeys (essentially unique journey identifiers) passing through each 0.1-mile segment were computed for two analysis weeks—namely, 7–13 May 2023 (historic, non-fuzzified), and 5–11 May 2024 (current, fuzzified). The 2024 dataset contains fuzzified records that cannot be considered because their GPS coordinates have been truncated and their geospatial representation does not indicate their actual location. Furthermore, their lack of heading information makes it difficult to assign them to specific directions of travel, and the lack of speed information makes them lose significant value for various freeway studies that rely on this characteristic.

The study location for this section is represented by a total of 54,284 0.1-mile segments of roadway in Indiana, with 26,838 of them being on interstates, 17,640 on selected US

Routes, and 9806 on selected State Routes. Table 3 shows that nearly 22% of interstate segments, 41% of US Route segments, and 55% of State Route segments observed an increase in CV journeys from 2023 to 2024. A very small percentage of segments showed no change, while an even lower percentage of segments could not be directly compared due to missing data in either of the analysis years.

Table 3. Summary statistics of change in number of journeys for route segments analyzed.

Type of Routes	Total 0.1-Mile Segments	Segments with Increase in Journeys	Segments with No Change in Journeys	Segments with Decrease in Journeys	Segments That Could Not Be Compared
Interstates	26,838	5835 (21.7%)	18 (0.1%)	20,976 (78.2%)	9 (0.0%)
US Routes	17,640	7174 (40.7%)	52 (0.3%)	10,403 (59.0%)	11 (0.0%)
State Routes	9806	5355 (54.6%)	58 (0.6%)	4391 (44.8%)	2 (0.0%)
Total	54,284	18,364 (33.8%)	128 (0.2%)	35,770 (65.9%)	22 (0.0%)

Figure 7 shows a box-and-whisker diagram of the network level percentage and absolute CV journey changes observed for the three types of routes analyzed. General trends show that the median percentage and absolute change in CV journeys for interstates is the highest, followed by slightly lower changes in US Routes and a median change of nearly 0% on State Route segments.

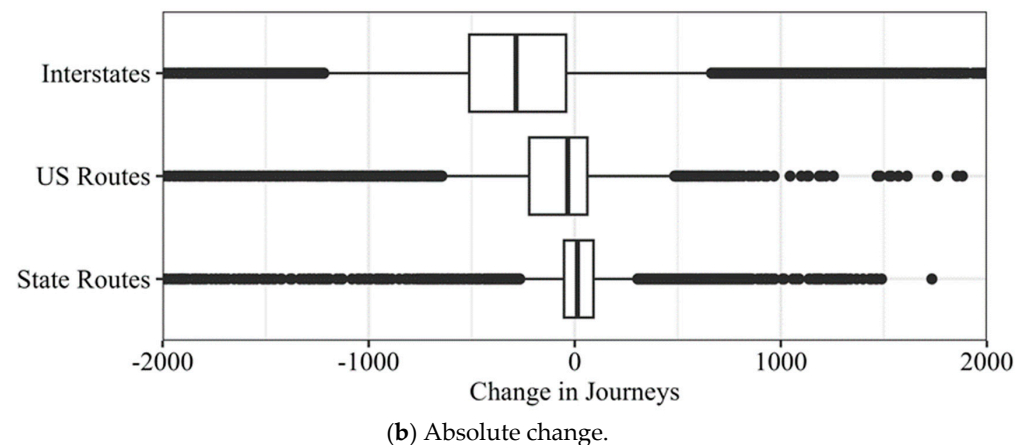
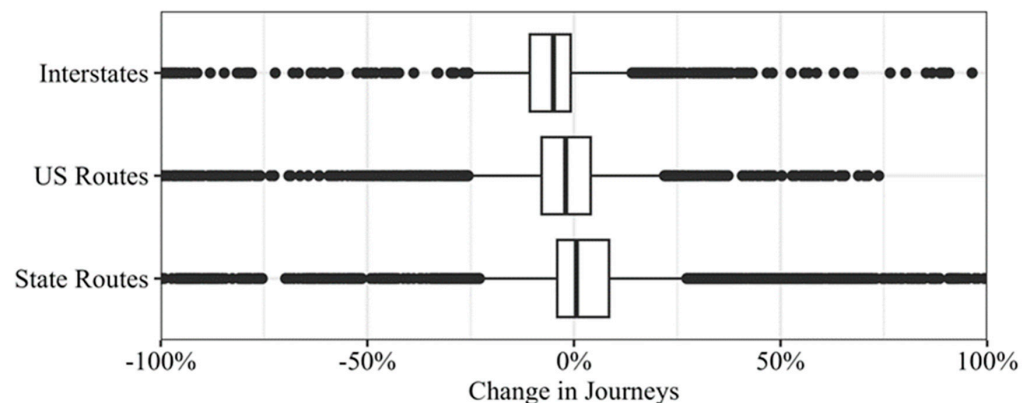
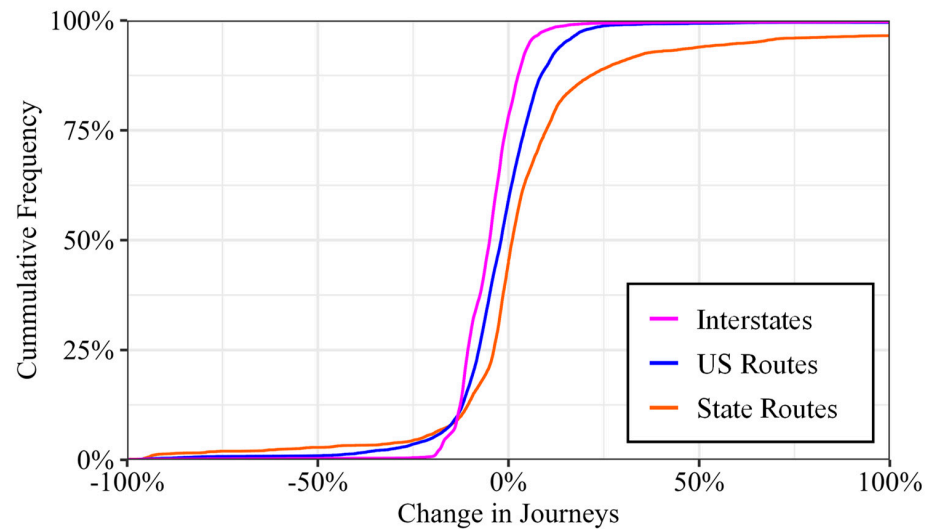
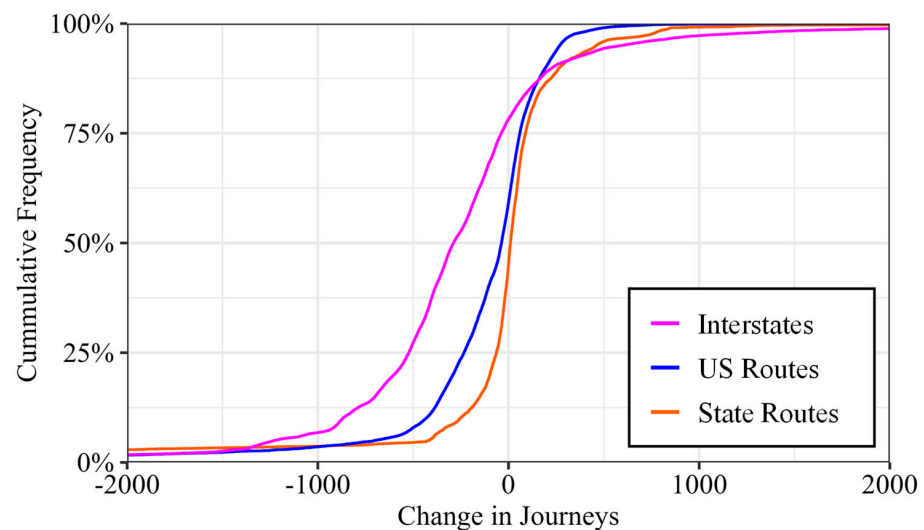


Figure 7. Box-and-whisker diagrams of the network level change on the number of journeys by route type and 0.1-mile segment.

A cumulative frequency distribution (CFD) of all 0.1-mile segments and their percentage and absolute changes in CV journeys between the two years is shown in Figures 8a and 8b, respectively. Median values for percentage change in journeys range from -4.95% for interstates to -1.96% for US Routes to $+0.98\%$ for State Routes. Correspondingly, median values for absolute change in journeys range from -289 for interstates to -37 for US Routes to $+10$ for State Routes. A number of segments showed reductions in CV journeys of more than 2000, possibly due to construction-related road closures significantly dropping volumes over those segments. Similarly, a number of segments showing increases in CV journeys of more than 2000 were a result of construction work being completed and converting an arterial into an interstate, thus leading to a significant rise in CV traffic passing through that segment.



(a) Percentage change.



(b) Absolute change.

Figure 8. CFD of the network level change on the number of journeys by route type and 0.1-mile segment.

Figure 9 documents percentage and absolute changes in CV journeys on the three types of routes as a Pareto-sorted column plot with each 0.1-mile segment represented. As evidenced by the zone of no-change gradually moving from left to right for interstates, US Routes, and State Routes, the highest percentage decrease in journeys is seen among interstate segments (78.2%), while the lowest percentage decrease in journeys is seen among State Route segments (44.8%).

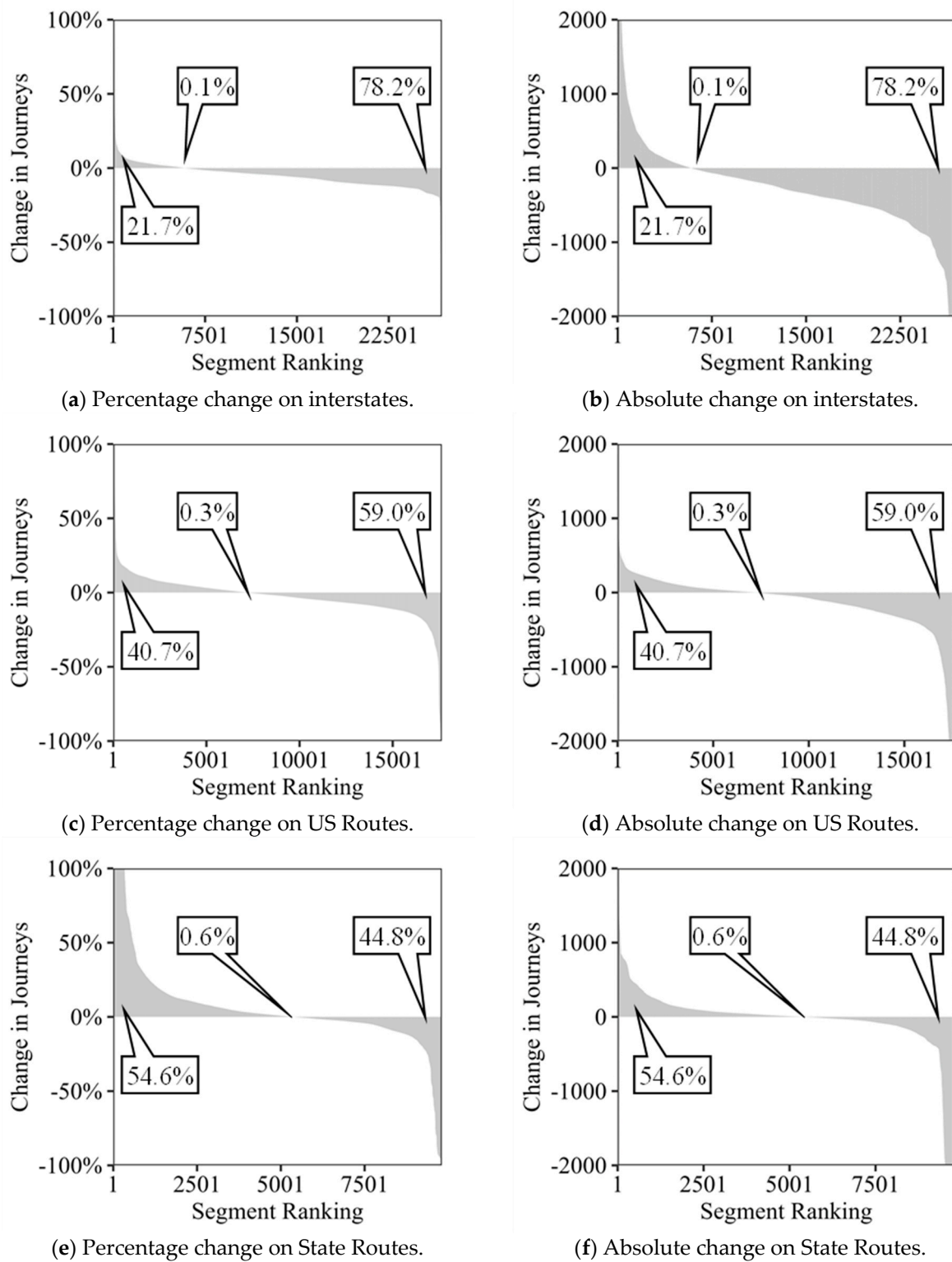
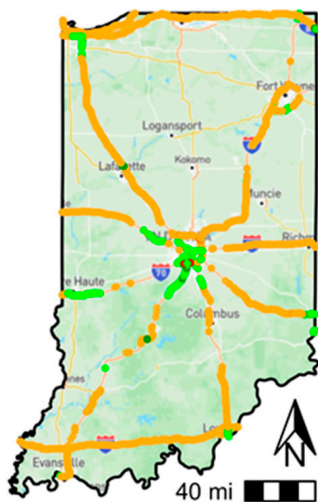


Figure 9. Pareto-sorted analyzed 0.1-mile route segments ranked by their change in journeys.

Figure 10 shows a map-based visualization of the change in the number of CV journeys across the three types of routes. To remove outliers or minor changes, any segments with percentage changes in CV journeys between -5 and 5 were ignored in Figure 10a–c and any segments with absolute changes in CV journeys between -100 and 100 were removed in Figure 10d–f.

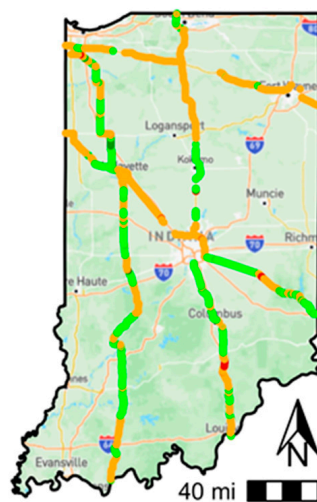
Figure 10a shows that a majority of rural interstate segments showed a decrease in journeys (−50% to −5%). Some segments at the Indiana–Ohio border on I-90 and I-74 showed increases in journeys, which may be attributable to construction projects in 2023 that were completed in 2024 leading to higher volumes or off-interstate construction projects causing additional diverting traffic to utilize the interstate. Segments highlighted in red in Figure 10d–f near the Indianapolis region in central Indiana are majorly attributable to a construction project in the northeast corner of Indianapolis that resulted in reduced or otherwise rerouted traffic through the area.

● < −50% ● −50% to −5% ● +5% to +50% ● ≥ +50%



Basemap (c) Mapbox, (c) OpenStreetMap

(a) Interstate percentage change.



Basemap (c) Mapbox, (c) OpenStreetMap

(b) US Route percentage change.



Basemap (c) Mapbox, (c) OpenStreetMap

(c) State Route percentage change.

● < −2000 ● −2000 to −100 ● +100 to +2000 ● ≥ +2000



Basemap (c) Mapbox, (c) OpenStreetMap

(d) Interstate absolute change.



Basemap (c) Mapbox, (c) OpenStreetMap

(e) US Route absolute change.



Basemap (c) Mapbox, (c) OpenStreetMap

(f) State Route absolute change.

Figure 10. Change in the number of journeys at the 0.1-mile route segment level in Indiana.

In general, the slightly lower MPR as documented in the preceding section, coupled with a smaller fleet of OEMs being represented in the CV data and the associated fuzzification can together be assumed to cause the reduction in observed CV journeys in 2024 compared to 2023 for interstate, US Route, and State Route segments. These visuals will be vital for agencies and practitioners in evaluating the usability of this novel form of CV data and identifying any significant changes that may bias year-by-year comparisons with CV data from 2020 to 2023.

4. Intersection Coverage

This section compared the number of vehicle trajectories available for analysis at 2827 signalized intersections, 158 roundabouts, and 304 all-way stops in Indiana between the historic 7–13 May 2023, dataset without fuzzified waypoints and the current 5–11 May 2024, CV dataset with fuzzified waypoints.

For a vehicle trajectory to be available for performance analysis, its movement at the intersection, that is, its direction of travel (i.e., northbound, eastbound, southbound, and westbound) and its turn type (i.e., right, through, and left), need to be identified. Therefore, the trajectory of each sampled vehicle contained in the historic and current CV datasets near the analyzed intersections was analyzed and, if possible, assigned an intersection movement [13].

Similar to the previous section, any fuzzified waypoint has truncated GPS coordinates with missing heading information, making it impossible to determine its trajectory's movement through the intersection. Therefore, fuzzified waypoints are not available for intersection movement performance analysis. The results of the differences between the two datasets at distinct analysis levels are presented as follows:

- First, the distribution of the change in available trajectories for analysis at the movement level is discussed.
- Then, the change in the number of vehicle trajectories by turn type is evaluated.
- Finally, a statewide qualitative analysis at the intersection level is provided.

4.1. Change by Movement

It is important to evaluate the amount of data available for analysis at the movement level. This is because movement level traffic signal performance measures provide practitioners with insights on the operational conditions in which each of the intersection's phases serve traffic. Depending on the performance results for all movements at an intersection, signal retiming [13] or capital investment [22] activities may be suggested to improve operations.

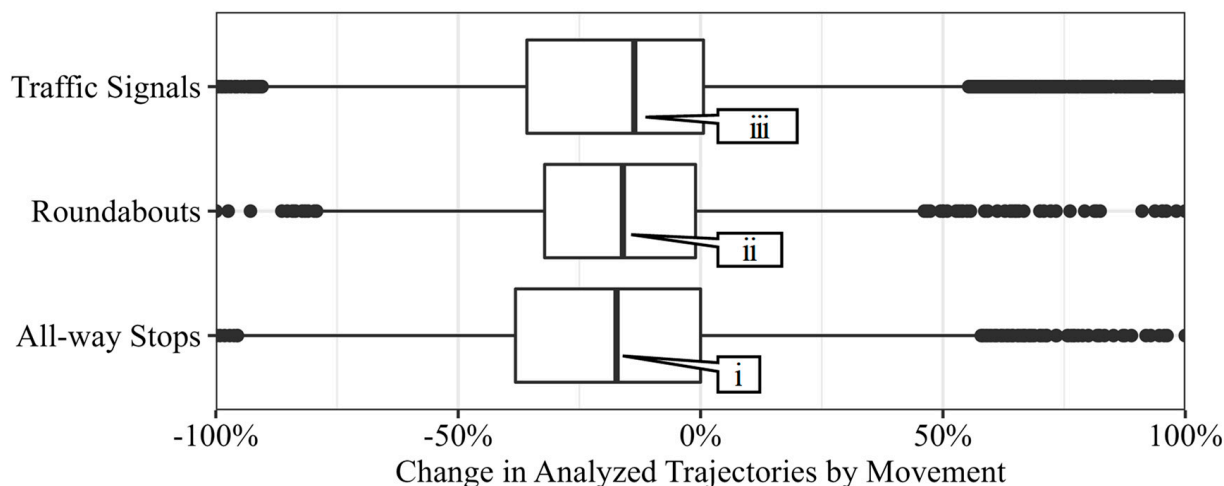
The change in the number of trajectories available for analysis (ΔT_{ijk}) at intersection i , direction of travel j , and turn type k was calculated as follows:

$$\Delta T_{ijk} = T_{ijk}^C - T_{ijk}^H \quad (3)$$

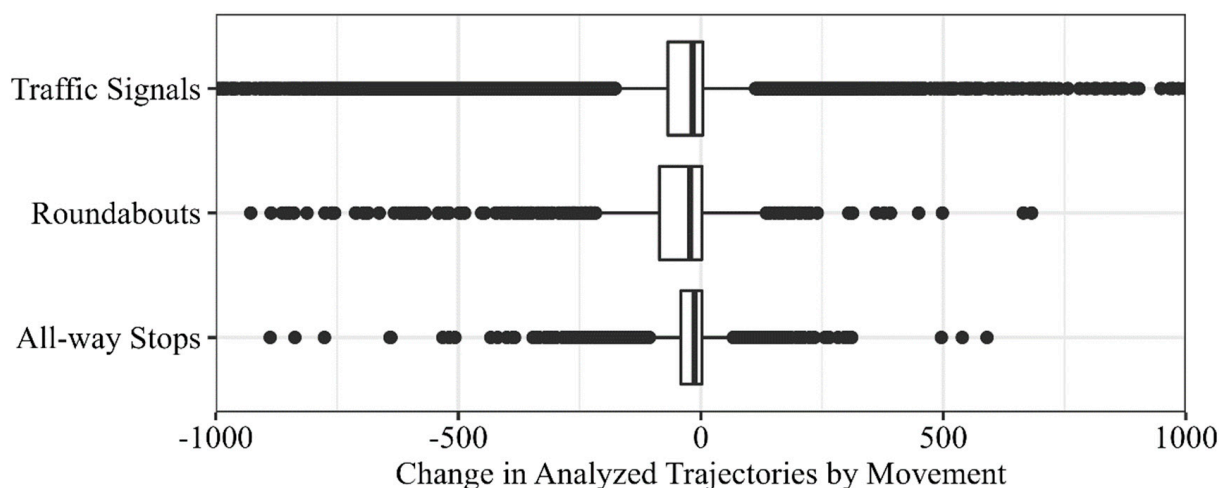
where T_{ijk}^H and T_{ijk}^C are the total number of trajectories assigned direction of travel j and turn type k at intersection i from the historic and current CV datasets, respectively. The percentage change $\% \Delta T_{ijk}$ was calculated as follows:

$$\% \Delta T_{ijk} = \left(\frac{\Delta T_{ijk}}{T_{ijk}^H} \right) * 100 \quad (4)$$

Figure 11 summarizes in box-and-whisker plots the distribution of the percentage (Figure 11a) and absolute (Figure 11b) trajectory count changes at the 26,291 movements of the intersections analyzed. All-way stops showed the largest reduction in median percentage change (callout i), followed by roundabouts (callout ii). Signalized intersections had the smallest median reduction (callout iii).



(a) Percentage change.

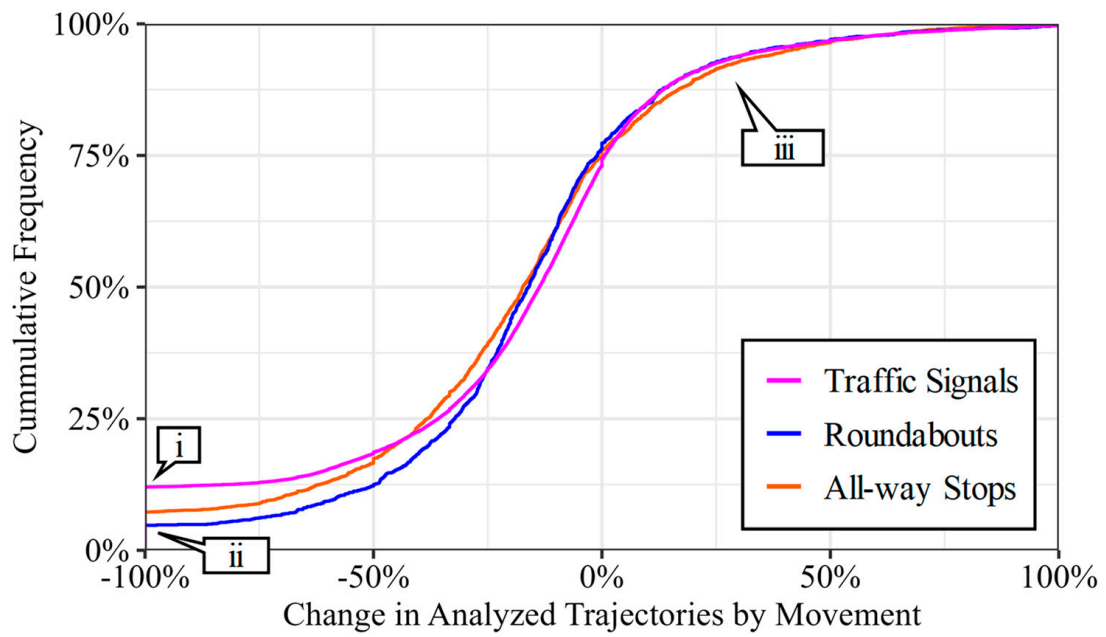


(b) Absolute change.

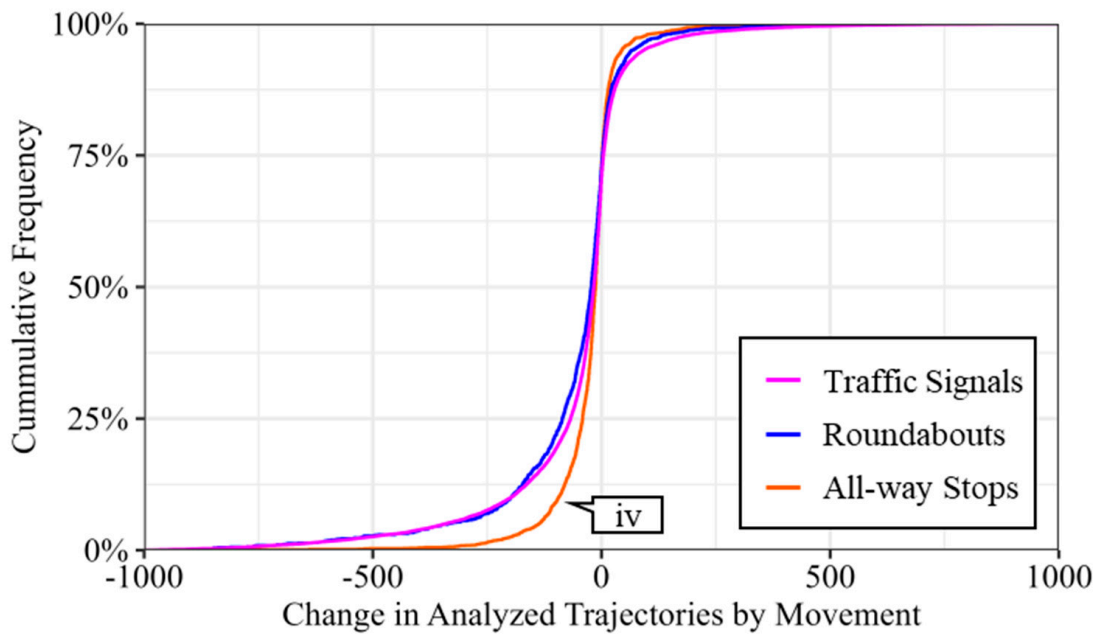
Figure 11. Box-and-whisker diagrams of the network level change on the number of analyzed vehicle trajectories by intersection and movement.

Figure 12 supplements the distribution analysis from Figure 11 by showing the change in available trajectories by movement as CFD diagrams. Traffic signals had the largest proportion of movements, around 12.5%, where the current CV dataset did not provide any trajectory available for analysis (callout i). This phenomenon only occurred for less than 6.25% of roundabout movements (callout ii). Overall, the increase in the availability of trajectories for analysis followed a similar distribution for all intersection types (callout iii).

It is important to note that most absolute changes were within the ± 500 trajectory count range (Figure 12b). All-way stops presented the smallest decrease in trajectories available for analysis (callout iv), likely because this type of intersection usually serves fewer vehicles than the others. The information provided in Figure 12 provides valuable insights on the expected number of trajectories available for analysis by intersection type for a week of data.



(a) Percentage change.



(b) Absolute change.

Figure 12. CFD of the network level change on the number of analyzed vehicle trajectories by intersection and movement.

Figure 13 shows Pareto-sorted bar graphs displaying the percentage and absolute changes on available trajectories for analysis for each evaluated movement. Callouts indicate the percentage of movements where the available trajectories increased (~29%), stayed the same (~1%), and decreased (~70%). The next subsection discusses the changes at the turn type level.

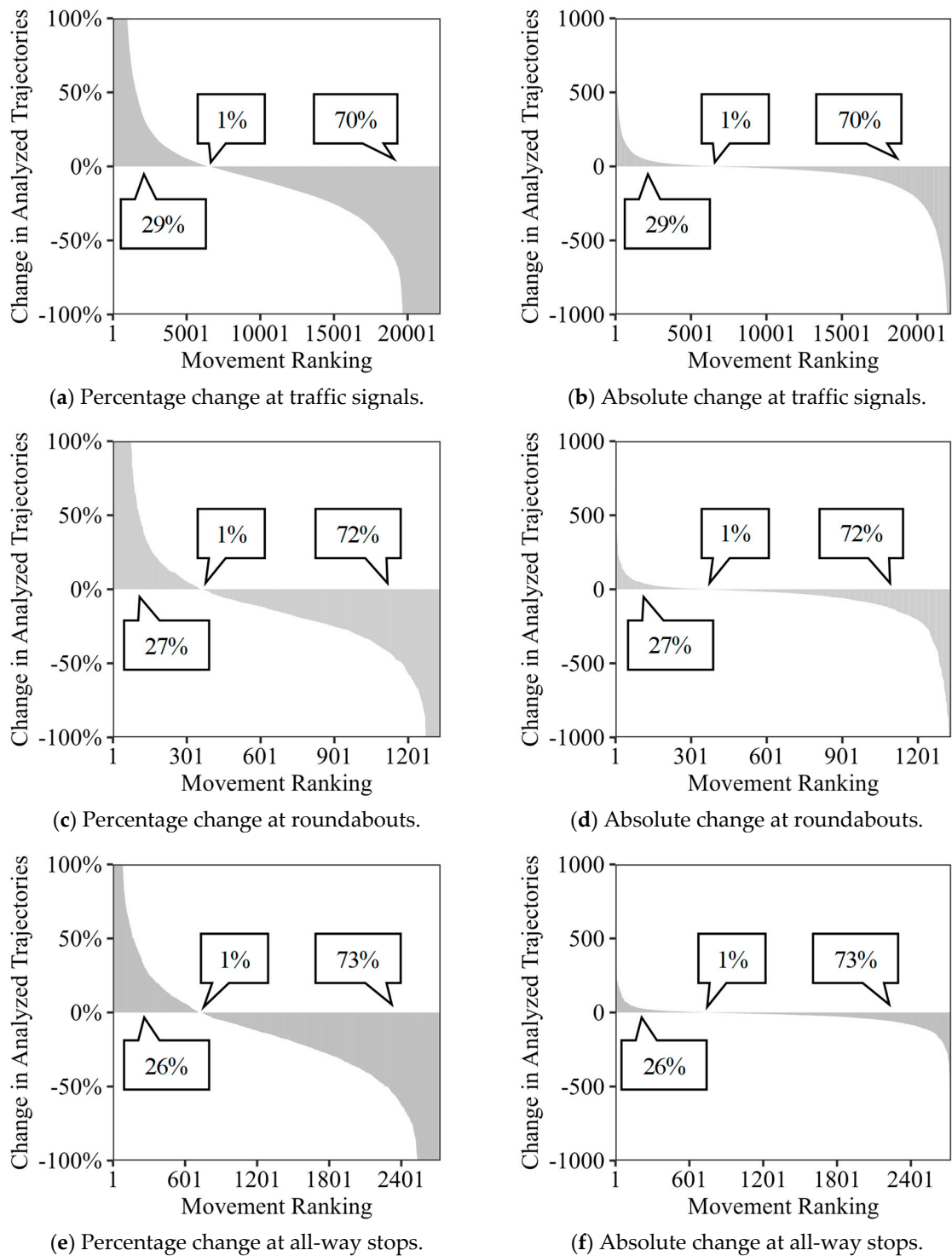


Figure 13. Pareto-sorted analyzed movements ranked by their change in analyzed vehicle trajectories.

4.2. Change by Turn Type

In addition to the movement level analysis, it is important to evaluate the change in the number of available trajectories for analysis at the turn level (i.e., right, through, and left). The change in the number of trajectories available for analysis (ΔT_k) that followed turn type k was calculated as follows:

$$\Delta T_k = T_k^C - T_k^H \tag{5}$$

where T_k^H and T_k^C are the total number of trajectories assigned turn type k from the historic and current CV datasets, respectively. The percentage change $\% \Delta T_k$ was calculated as follows:

$$\% \Delta T_k = \left(\frac{\Delta T_k}{T_k^H} \right) * 100 \quad (6)$$

Table 4 summarizes the change in the number of trajectories available for analysis by turn type. Every category shows a reduction in available trajectories as a consequence of waypoint fuzzification (Figure 1) and MPR reduction (Table 1). Considering all turn types, signalized intersections had the smallest percentage decrease (−11.22%) and roundabouts the largest (−17.66%). In particular, through trips at signalized intersections had the smallest percentage reduction (−10.73%), while through trips at roundabouts had the largest (−19.46%). The following subsection presents the changes at the intersection level.

Table 4. Number of vehicle trajectories analyzed by intersection and turn type.

Intersection Type	Turn Type (k)	T_k^H	T_k^C	ΔT_k	$\% \Delta T_k$
Traffic signal	All	13,774,803	12,228,874	−1,545,929	−11.2%
	Right	948,878	802,500	−146,378	−15.4%
	Through	11,138,846	9,943,666	−1,195,180	−10.7%
	Left	1,687,079	1,482,708	−204,371	−12.1%
Roundabout	All	562,707	463,349	−99,358	−17.7%
	Right	83,080	72,518	−10,562	−12.7%
	Through	389,827	313,975	−75,852	−19.5%
	Left	89,800	76,856	−12,944	−14.4%
All-way stop	All	498,109	426,220	−71,889	−14.4%
	Right	102,935	85,202	−17,733	−17.2%
	Through	292,083	254,043	−38,040	−13.0%
	Left	103,091	86,975	−16,116	−15.6%

4.3. Change by Intersection

An evaluation of the change in the number of available trajectories for analysis at the intersection level can help identify overall trends and statewide conditions. The change in the number of trajectories available for analysis (ΔT_i) at intersection i was calculated as follows:

$$\Delta T_i = T_i^C - T_i^H \quad (7)$$

where T_i^H and T_i^C are the total number of trajectories available for analysis at intersection i from the historic and current CV datasets, respectively. The percentage change $\% \Delta T_i$ was calculated as follows:

$$\% \Delta T_i = \left(\frac{\Delta T_i}{T_i^H} \right) * 100 \quad (8)$$

Figure 14 qualitatively shows the percentage (Figure 14a–c) and absolute (Figure 14d–f) changes at each analyzed intersection. In general, rural areas displayed less pronounced differences between the datasets. This is likely because those intersections may serve more commuting traffic than urban intersections, which can reduce the impact of fuzzification as fewer vehicles have rural areas as destination. Additionally, most intersections fall within the $\geq -50\%$ to $< +50\%$ trajectory change categories, which indicates general moderate differences at the intersection level.

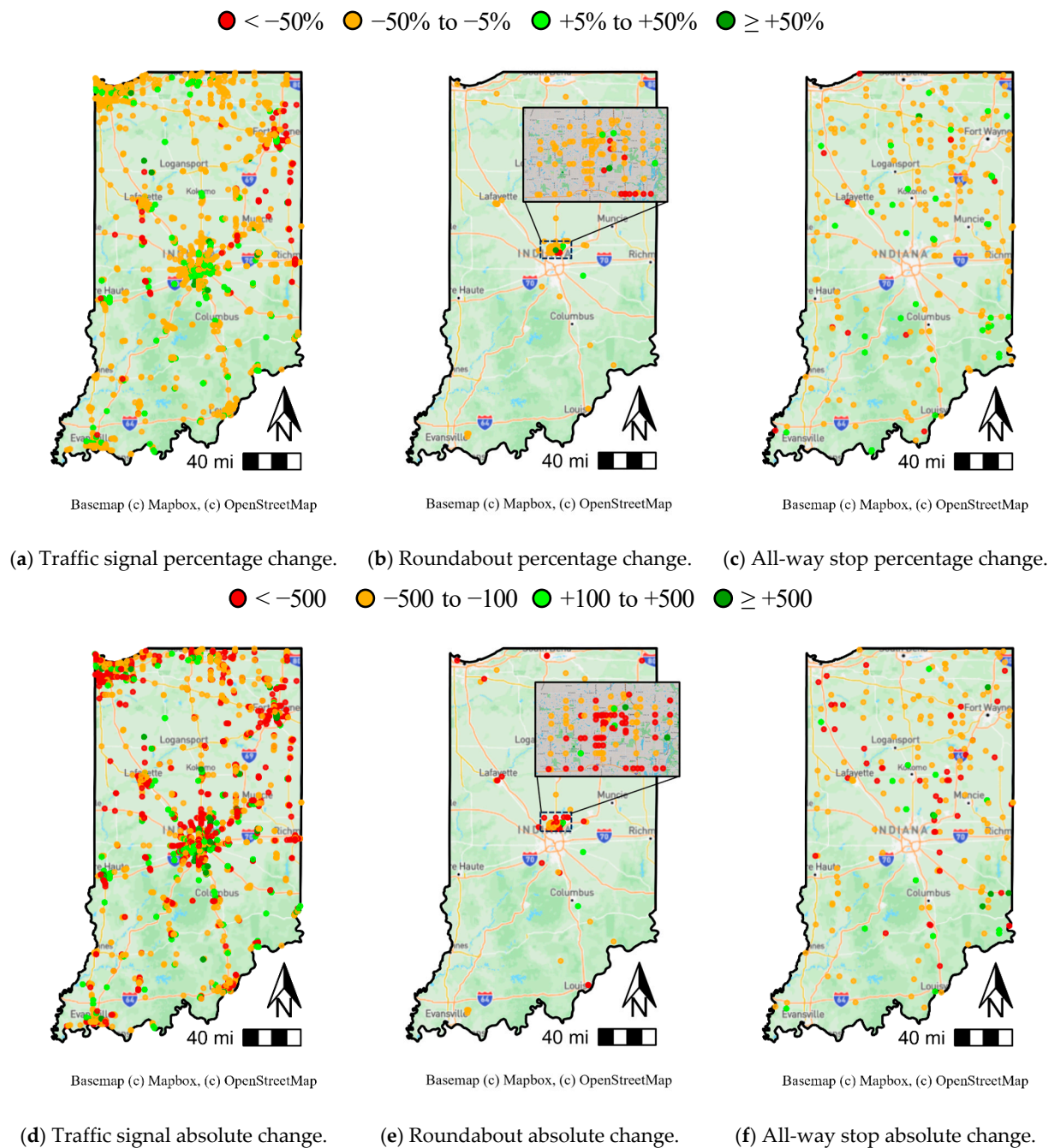


Figure 14. Change in the number of vehicle trajectories analyzed at the intersection level in Indiana.

5. Discussion

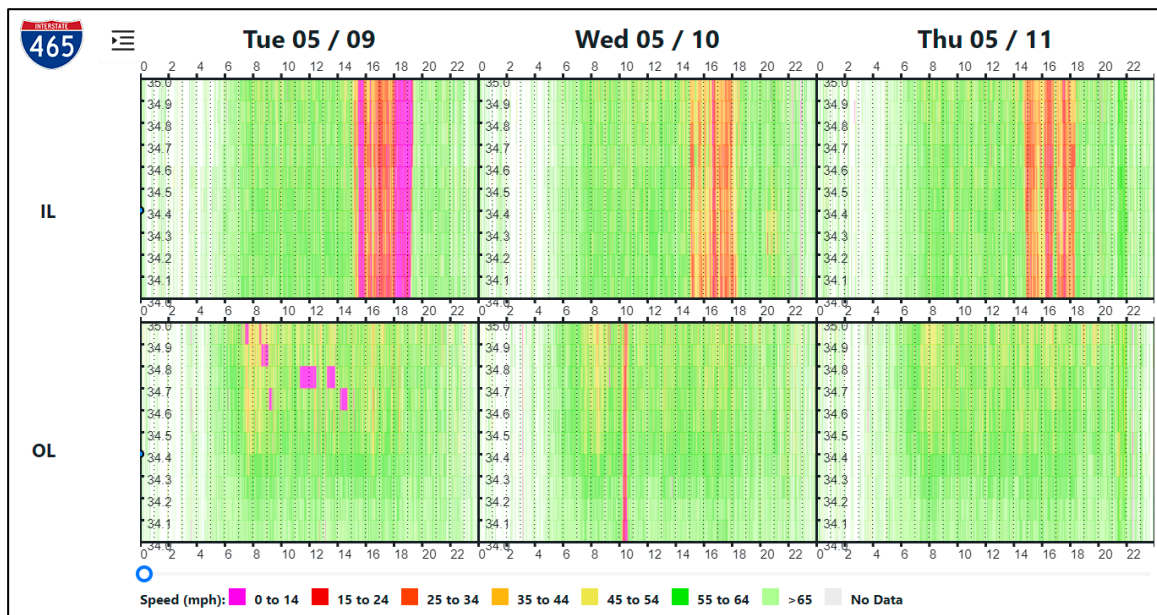
CV trajectory data have demonstrated to be a versatile dataset that can be used to assess mobility under a wide range of scenarios and scales. This study presented a high-level analysis on the impact that data fuzzification and fleet reduction (Figure 1) had on a current commercial CV dataset. An evaluation of the current and historic estimated MPRs showed reductions of 0.5% and 0.3% along interstate and non-interstate roadways, respectively (Table 1). The summary impact at freeways and intersections is summarized below.

5.1. Impact at Freeways

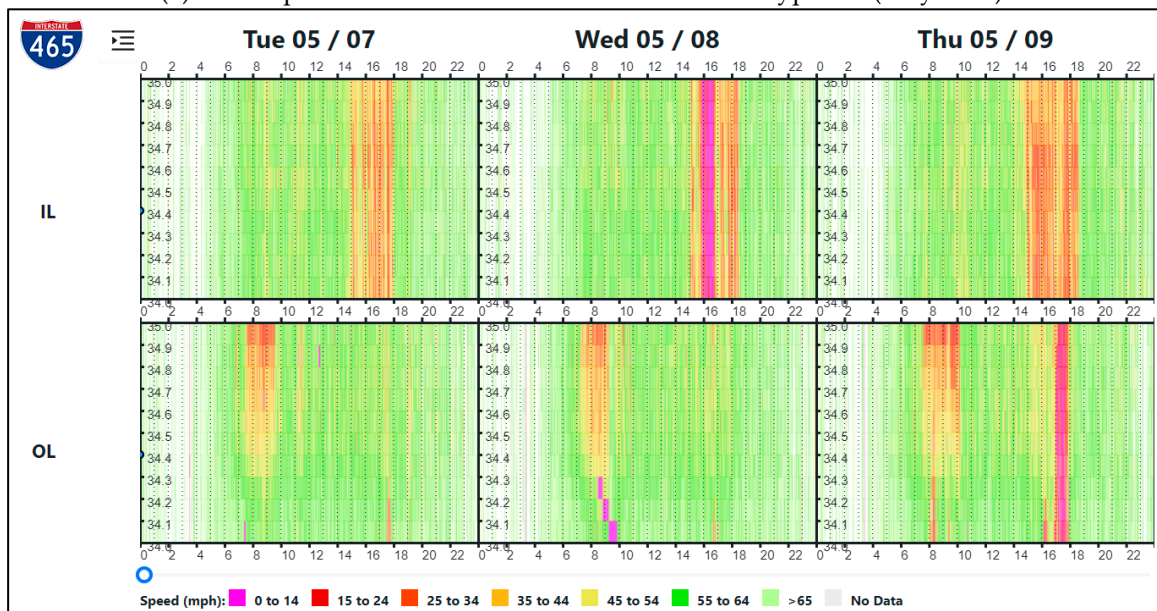
A comparison of the CV data available on 54,284 individual 0.1-mile segments of interstates, US Routes, and State Routes between a week of historic non-fuzzified and a

week of current fuzzified data showed that the number of journeys available for analysis increased for 33.8% of segments and decreased for 65.9%.

Although the reductions are non-trivial, for simply computing average segment speeds, the authors do not believe this will significantly impact estimations. For example, Figure 15 shows trajectory-derived heatmaps of sampled vehicle speeds over one mile of I-465 in Indianapolis, Indiana. The vertical axis represents the location on the road and the horizontal axis the DOW and TOD; displayed trajectory segments color-coded based on their speed portray mobility conditions [15]. A similar coverage can be qualitatively concluded by comparing the heatmap generated from a non-fuzzified historic dataset (Figure 15a) and the heatmap generated from the fuzzified current dataset (Figure 15b).



(a) With a previous CV dataset without fuzzified waypoints (May 2023).



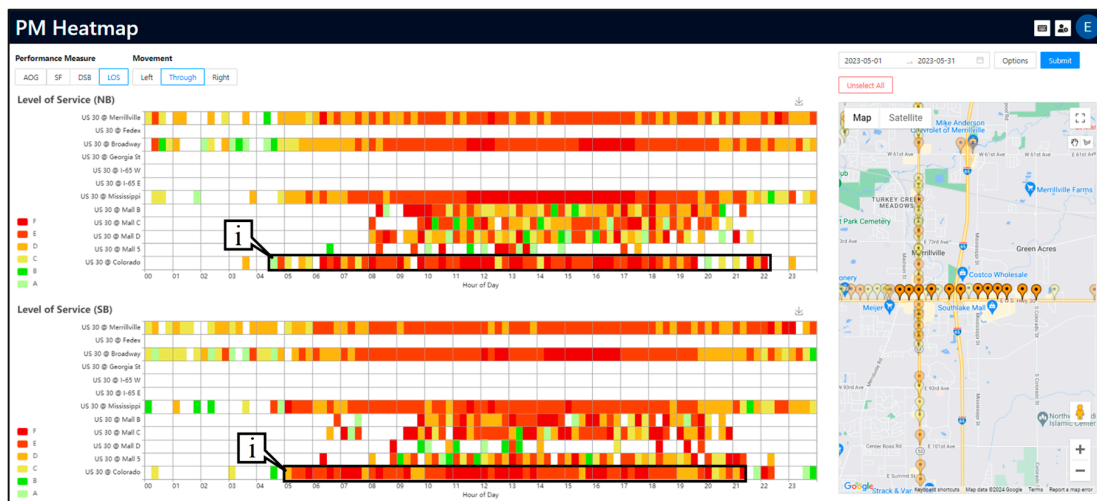
(b) With the current CV dataset with fuzzified waypoints (May 2024).

Figure 15. Speed analysis over one mile of I-465 around Indianapolis, Indiana (Note: IL = inner loop, OL = outer loop).

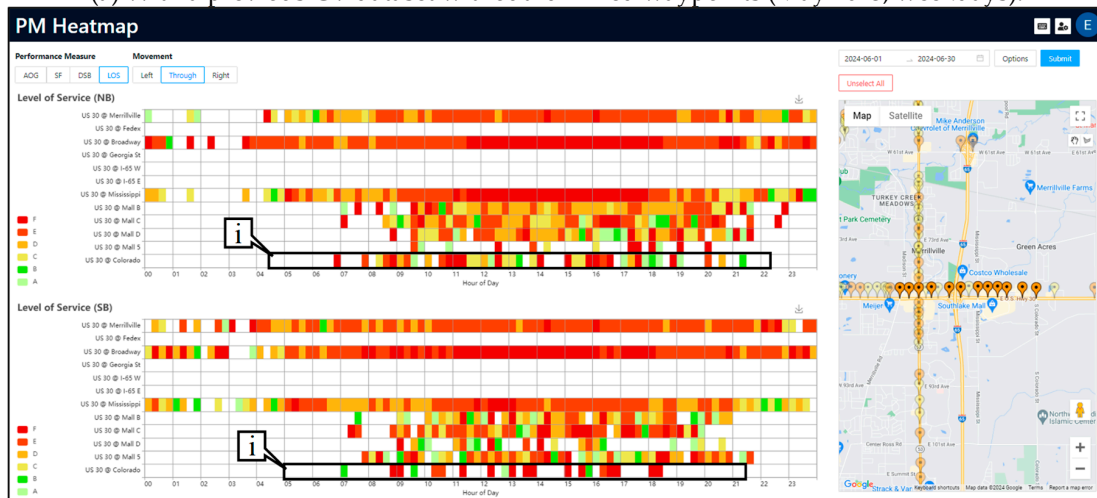
5.2. Impact at Intersections

A comparison of the number of CV trajectories available for movement level analysis at 3289 intersections between a week of historic non-fuzzified and a week of current fuzzified data showed an 11.6% overall reduction (Table 4). Of the 26,291 studied individual movements, the number of trajectories available for analysis increased at 28.3% and decreased at 70.4%. This large and non-uniform reduction in available trajectories for analysis may substantially decrease sampling on particular movements and perhaps bias intersection analysis. This change is a concern, because, in many cases, the movements that are fuzzified going to or from residential or commercial areas are of a significant interest. This is because signal retiming [13] and capital investment [22] opportunity identification algorithms rely on the unbiased sampling of vehicle trajectories at all movements at any intersection.

For example, Figure 16 shows trajectory-derived heatmaps of the estimated Highway Capacity Manual (HCM) level of service (LOS) [13,34] at the minor through movements of a corridor segment with 12 signalized intersections. The vertical axis represents each analyzed intersection, and the horizontal axis represents the TOD. It is evident how data coverage is reduced for the heatmap generated with the current fuzzified dataset (Figure 16b, callout i) when comparing it to the heatmap derived from the historic dataset without fuzzified waypoints (Figure 16a, callout i). Improvement opportunities will now be more difficult to identify for the intersection with decreased coverage (callout i).



(a) With a previous CV dataset without fuzzified waypoints (May 2023, weekdays).



(b) With the current CV dataset with fuzzified waypoints (June 2024, weekdays).

Figure 16. LOS estimation for the minor movements at a 12-intersection corridor.

5.3. Future Research

This study focused on providing a high-level overview of the impacts that CV data fuzzification and fleet composition changes have on data availability for transportation studies. Future research should provide a more in-depth investigation of the effects that these changes have on specific mobility analyses by type of roadway infrastructure. Subsequent efforts should provide special attention to the impact that data fuzzification has on movement level traffic signal performance estimations, as this granularity of analysis is critical to identify retiming [13] and infrastructure [22] improvement opportunities.

Furthermore, improved alternative data fuzzification techniques may be derived from an intentional dialog between transportation agencies, industry, and academia to further protect motorist privacy while minimizing adverse effects on data availability. Different fuzzification approaches should be extensively evaluated and best practices defined in future studies.

6. Conclusions

Even though there is a general reduction in the amount of data available for analysis in comparison with historic datasets, with MPR values $\geq 4.6\%$, the authors believe that the current CV dataset accounts for enough representativeness to derive most relevant mobility performance measures. Furthermore, as MPR follows an upward trend, it is just a matter of time before the current dataset contains more usable data than previous versions.

Of the infrastructure CV coverage analyzed, intersection performance will likely be the most affected by data fuzzification (Figures 1 and 16). This is because the current data filtering approach may induce unintended bias on the amount of data available for analysis by movement at intersections near popular commercial areas or large residential zones.

Although the results of this study are based on data from Indiana, the authors believe that similar trends are likely to be observed in other locations. This is because if the same fuzzification approach and fleet modifications are implemented elsewhere, proportional changes in data availability are expected to be produced.

Author Contributions: Conceptualization, E.D.S.-C., R.S.S., J.D., J.K.M. and D.M.B.; methodology, E.D.S.-C., R.S.S., J.D., J.K.M. and A.J.S.; software, E.D.S.-C., R.S.S., J.D., J.K.M. and A.J.S.; validation, E.D.S.-C., R.S.S. and J.D.; formal analysis, E.D.S.-C., R.S.S., and J.D.; investigation, E.D.S.-C., R.S.S. and J.D.; resources, E.D.S.-C., R.S.S., J.D., J.K.M. and D.M.B.; data curation, E.D.S.-C., R.S.S., J.D., J.K.M., A.J.S. and J.M.; writing—original draft preparation, E.D.S.-C., R.S.S. and J.D.; writing—review and editing, J.K.M. and D.M.B.; visualization, E.D.S.-C., R.S.S., J.D., J.K.M. and A.J.S.; supervision, D.M.B.; project administration, D.M.B.; funding acquisition, D.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Joint Transportation Research Program and Pooled Fund Study (TPF-5(519)) led by the Indiana Department of Transportation (INDOT) and supported by the state transportation agencies of California, Connecticut, Georgia, Minnesota, Mississippi, North Carolina, Ohio, Pennsylvania, Texas, Utah, and Wisconsin, and the Federal Highway Administration (FHWA) Operations Technical Services Team. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views or policies of the sponsoring organizations. These contents do not constitute a standard, specification, or regulation.

Data Availability Statement: The datasets presented in this article are not readily available because of commercial restrictions.

Acknowledgments: CV trajectory data from 2020 to 2023 used in this study was provided by Wejo Data Services, Inc. CV trajectory data from 5–11 May 2024, was provided by StreetLight Data, Inc.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American Society of Civil Engineers. A Comprehensive Assessment of America's Infrastructure: 2021 Report Card for America's Infrastructure. 2021. Available online: <https://www.infrastructurereportcard.org/> (accessed on 17 July 2024).
2. FHWA. December 2023 Traffic Volume Trends. Available online: https://www.fhwa.dot.gov/policyinformation/travel_monitoring/23dectvt/ (accessed on 17 July 2024).
3. FHWA. December 2003 Traffic Volume Trends. Available online: <https://www.fhwa.dot.gov/ohim/tvtw/03dectvt/index.htm> (accessed on 17 July 2024).
4. Chu, J.; Radow, L. Behind the Scenes at TMCs. FHWA. Available online: <https://highways.dot.gov/public-roads/julyaugust-2012/behind-scenes-tmcs> (accessed on 17 July 2024).
5. Tantillo, M.; Smith, K.; Packard, C.; Lomax, T.; Dhuri, S. *Transportation Management Center Performance Dashboards*; Department of Transportation: Washington, DC, USA, 2021.
6. Wuertz, S.C. *Indiana Department of Transportation Traffic Management Strategic Deployment Plan Version 2.4*; Indiana Department of Transportation: Indianapolis, IN, USA, 2008.
7. ITE; NOCoE. 2019 Traffic Signal Benchmarking and State of the Practice Report. 2020. Available online: <https://transportationops.org/trafficsignals/benchmarkingreport> (accessed on 1 January 2024).
8. National Transportation Operations Coalition. *National Traffic Signal Report Card*; National Transportation Operations Coalition: Washington, DC, USA, 2012. Available online: <https://transportationops.org/publications/2012-national-traffic-signal-report-card#downloads> (accessed on 1 January 2024).
9. Denney, R.W.; Head, L.; Spencer, K. *Signal Timing Under Saturated Conditions*; United States Federal Highway Administration: Washington, DC, USA, 2008.
10. Day, C.M.; Bullock, D.M.; Li, H.; Remias, S.M.; Hainen, A.M.; Freije, R.S.; Stevens, A.L.; Sturdevant, J.R.; Brennan, T.M. *Performance Measures for Traffic Signal Systems: An Outcome-Oriented Approach*; Purdue University: West Lafayette, IN, USA, 2014. [CrossRef]
11. Lattimer, C. Automated Traffic Signals Performance Measures. 2020. Available online: <https://ops.fhwa.dot.gov/publications/fhwahop20002/fhwahop20002.pdf> (accessed on 24 October 2022).
12. Guadamuz, R.; Tang, H.; Yu, Z.; Guler, S.I.; Gayah, V.V. Green time usage metrics on signalized intersections and arterials using high-resolution traffic data. *Int. J. Transp. Sci. Technol.* **2021**, *11*, 509–521. [CrossRef]
13. Saldivar-Carranza, E.D.; Li, H.; Mathew, J.K.; Desai, J.; Platte, T.; Gayen, S.; Sturdevant, J.; Taylor, M.; Fisher, C.; Bullock, D.M. *Next Generation Traffic Signal Performance Measures: Leveraging Connected Vehicle Data*; Purdue University: West Lafayette, IN, USA, 2023. [CrossRef]
14. Emtenan, A.M.T.; Day, C.M. Impact of Detector Configuration on Performance Measurement and Signal Operations. *Transp. Res. Rec. J. Transp. Res. Board* **2020**, *2674*, 300–313. [CrossRef]
15. Sakhare, R.S.; Desai, J.; Mathew, J.K.; McGregor, J.; Kachler, M.; Bullock, D.M. *Measuring and Visualizing Freeway Traffic Conditions: Using Connected Vehicle Data*; Purdue University: West Lafayette, IN, USA, 2024. [CrossRef]
16. Waddell, J.M.; Remias, S.M.; Kirsch, J.N. Characterizing Traffic-Signal Performance and Corridor Reliability Using Crowd-Sourced Probe Vehicle Trajectories. *J. Transp. Eng. A Syst.* **2020**, *146*, 04020053. [CrossRef]
17. Mahmud, S.; Day, C.M. Evaluation of Arterial Signal Coordination with Commercial Connected Vehicle Data: Empirical Traffic Flow Visualization and Performance Measurement. *J. Transp. Technol.* **2023**, *13*, 327–352. [CrossRef]
18. Desai, J.; Mathew, J.K.; Li, H.; Sakhare, R.S.; Horton, D.; Bullock, D.M. *National Mobility Analysis for All Interstate Routes in the United States: December 2022*; Purdue University: West Lafayette, IN, USA, 2023. [CrossRef]
19. ITSdigest. 470 Million Connected Vehicles On the Road by 2025. Available online: <https://www.itsdigest.com/470-million-connected-vehicles-road-2025> (accessed on 2 February 2023).
20. Sakhare, R.S.; Hunter, M.; Mukai, J.; Li, H.; Bullock, D.M. Truck and Passenger Car Connected Vehicle Penetration on Indiana Roadways. *J. Transp. Technol.* **2022**, *12*, 578–599. [CrossRef]
21. FHWA. Every Day Counts: Innovation for a Nation on the Move. 2023. Available online: https://www.fhwa.dot.gov/innovation/everydaycounts/edc_6/ (accessed on 18 July 2024).
22. Gayen, S. *Statewide Identification and Ranking of Signalized Intersections Needing Capacity Improvements*; Purdue University: West Lafayette, IN, USA, 2024.
23. Desai, J.; Mathew, J.; Li, H.; Bullock, D. Using connected vehicle data for assessing electric vehicle charging infrastructure usage and investment opportunities. *Inst. Transp. Eng. ITE J.* **2022**, *92*, 22–31. Available online: https://drive.google.com/file/d/1-0dHQmbqZk_npqEtyPb1-v8b1OdhEUrU/view (accessed on 17 March 2024).
24. Alsahfi, T.; Almotairi, M.; Elmasri, R.; Alshemaimri, B. Road Map Generation and Feature Extraction from GPS Trajectories Data. In Proceedings of the 12th ACM SIGSPATIAL International Workshop on Computational Transportation Science, Chicago, IL, USA, 5–8 November 2019; ACM: New York, NY, USA, 2019; pp. 1–10. [CrossRef]
25. Zhan, X.; Zheng, Y.; Yi, X.; Ukkusuri, S.V. Citywide Traffic Volume Estimation Using Trajectory Data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 272–285. [CrossRef]
26. Fan, J.; Fu, C.; Stewart, K.; Zhang, L. Using big GPS trajectory data analytics for vehicle miles traveled estimation. *Transp. Res. Part C Emerg. Technol.* **2019**, *103*, 298–307. [CrossRef]
27. Ugan, J.; Abdel-Aty, M.; Islam, Z. Using Connected Vehicle Trajectory Data to Evaluate the Effects of Speeding. *IEEE Open J. Intell. Transp. Syst.* **2024**, *5*, 16–28. [CrossRef]

28. Islam, Z.; Abdel-Aty, M.; Ugan, J. Signal Phasing and Timing Prediction Using Connected Vehicle Data. *Transp. Res. Rec. J. Transp. Res. Board* **2024**, *2678*, 662–673. [[CrossRef](#)]
29. Du, Z.; Yan, X.; Zhu, J.; Sun, W. Signal Timing Parameters Estimation for Intersections using Floating Car Data. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2673*, 189–201. [[CrossRef](#)]
30. Klein, L.A.; Mills, M.K.; Gibson, D.R.P. *Traffic Detector Handbook*, 3rd ed.; Turner-Fairbank Highway Research Center: McLean, VA, USA, 2006; Volume I. Available online: <https://rosap.ntl.bts.gov/view/dot/954> (accessed on 23 July 2024).
31. Chen, C.; Petty, K.; Skabardonis, A.; Varaiya, P.; Jia, Z. Freeway Performance Measurement System: Mining Loop Detector Data. *Transp. Res. Rec. J. Transp. Res. Board* **2001**, *1748*, 96–102. [[CrossRef](#)]
32. Oh, S.; Ritchie, S.G.; Oh, C. Real-Time Traffic Measurement from Single Loop Inductive Signatures. *Transp. Res. Rec. J. Transp. Res. Board* **2002**, *1804*, 98–106. [[CrossRef](#)]
33. Hunter, M.; Mathew, J.K.; Cox, E.; Blackwell, M.; Bullock, D.M. *Estimation of Connected Vehicle Penetration Rate on Indiana Roadways*; TRP Affiliated Reports—Paper 37; Joint Transportation Research Program: West Lafayette, IN, USA, 2021. [[CrossRef](#)]
34. Transportation Research Board. *Highway Capacity Manual 2010*; National Research Council (NRC): Washington, DC, USA, 2010.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.