APPENDIX A: Error Metrics used in Probe-Based Volume Estimates Validation

The table below summarizes the validation metrics used to quantify the discrepancy between the values reported by vendors and the 'reference' source. Given a paired dataset, the error at each location is calculated as the difference between vendor and reference volumes and may be expressed in the same units as the reported volume (vehicles) or as a percentage of the reference volume (percent error, often abbreviated as PE).

The mean and various percentiles of errors are calculated to communicate the central tendency and distribution (spread) of values across the dataset (or subset of the dataset), respectively. For example, the mean absolute percent error, referred to as MAPE, communicates the average absolute error expressed as a percentage of the reference volume, whereas a summary of absolute percent error (APE) percentiles communicates how much the values vary.

In most cases, results are typically grouped by volume bin, road class, or another data characteristic of interest. Percentage-based errors are sensitive to low volumes, so separating results by reference volume range can provide useful context for interpreting results.

Table A-1: Statical Metrics Used for Volume Validation

Given a set of *n* volume estimates ($\widehat{\boldsymbol{Y}}$) and corresponding reference volumes ($\boldsymbol{Y}$), the resulting signed errors, (***E***) are used as the basis for calculating various error metrics.

$$\mathbf{E} = \{e_1, e_2, \ldots, e_n\} = \{(\hat{y}_1 - y_1), (\hat{y}_2 - y_2), \ldots, (\hat{y}_n - y_n)\}$$

Metric	Formula	Explanation & Notes
Error Percentiles	$\mathrm{P}_X(\{e_1, e_2, \ldots, e_n\})$	The Xth percentile error represents the *signed* error value below which X% of observations lie (expressed as volume). • P_{50} is the median error (**MdE**) • P_{25} to P_{75} bracket middle 50% of observations
Percent Error Percentiles	$\mathrm{P}_X\left(100 \cdot \left\{\frac{e_1}{y_1}, \frac{e_2}{y_2}, \ldots, \frac{e_n}{y_n}\right\}\right)$	Similar to **Error Percentiles**, but errors are expressed as a percentage of reference volume. • P_{50} is the median error (**MdPE)**
Absolute Percent Error (APE) Percentiles	$\mathrm{P}_X\left(100 \cdot \left\{\left\lvert\frac{e_1}{y_1}\right\rvert, \left\lvert\frac{e_2}{y_2}\right\rvert, \ldots, \left\lvert\frac{e_n}{y_n}\right\rvert\right\}\right)$	Similar to **Error Percentiles**, but only the magnitude (not sign) of errors is considered, and errors are expressed as a percentage of reference volume. All values are positive. • P_{50} is the median error (**MdAPE)** • P_{90} represents 'tail risk'
Mean Absolute Percent Error (MAPE)	$\frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{e_i}{y_i}\right\rvert$	Avg. magnitude of error (as percent of reference volume)
Normalized Root Mean Square Error (NRMSE)	$100 \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i}$	Expresses avg. magnitude of error (as percent of average reference volume) but penalizes large errors more than MAPE.

APPENDIX B: Summary of FHWA Method

FHWA's report, titled "Guidelines for Obtaining AADT Estimates from Non-Traditional Sources, Revised" (FHWA-HPL-23-018), outlines a framework for evaluating the quality of "Non-Traditional" volume data sources (such as the probe-based volume estimates included in this study). The FHWA method includes a series of statistical tests that can be used to determine whether emerging probe-based volume estimates are suitable for Federal reporting purposes. Given agency interest in this particular use case, the validation team implemented this evaluation approach method to gain additional insight into vendor data quality. On thing to note about the FHWA analysis method is that it was developed and intended to be implemented on bi-directional traffic. As such, **directional volumes were aggregated to bi-directional volumes at each reference location to create a bi-directional dataset for application of the FHWA method.**

Before describing the FHWA evaluation approach, it is first helpful to understand how agencies currently estimate AADT at locations where there are not permanent counters. Aside from limited CCS sites where AADT can be computed directly, agencies typically estimate AADT by annualizing short duration counts. This process – often referred to as "factoring", involves (1) grouping similar CCS sites (often by road type and traffic characteristics), (2) developing factors to summarize how each group's counts vary by season and day-of-week, (3) assigning each short-term count location to the most similar factor group, and (4) multiplying daily counts by seasonality, day-of-week (and sometimes additional) factors to obtain estimates of annual average daily traffic. The FHWA guidelines for minimum accuracy of non-traditional sources is based on (referenced to) the accuracy of these short-term counts.

Accuracy and Precision Limits associated with factored Short-Term Counts

USDOT has previously benchmarked the accuracy and precision associated with the method of factoring 48-hour counts to annual averages (see Table B-1 below). The concept embedded in the FHWA guidelines is that if an agency wants to procure an AADT dataset (such as the vendor products included in this study), these products should be *at least as accurate and precise* as factored, short-term counts. Thus, the FHWA statistical method involves calculating the accuracy and precision of vendor provided volume estimates to see if they compare favorably with the values associated with the 'status quo' short-term counter approach.

Table B-1 (copied directly from Table 14 on page 49 of the FHWA report) summarizes the accuracy and precision acceptance limits based on USDOT's historical analysis of factored 48-hour counts, reporting limits separately for four different volume bins.

- **Accuracy Limit #1: *'Minimally 95% Probability, Traffic Count Error (TCE) Median Error Bias (%)'***: This metric provides a near-worst-case interval (middle 95%) of sample Median TCE values expected from factored AADTs, given a sample of N sites. At higher sample sizes the confidence interval becomes narrower, and the sample median more closely approximates the population

median. This accuracy metric is testing for bias (consistent over or under-estimation).

- **Accuracy Limit #2**: '**Minimally 95% Probability, MAPE Upper 95% CB (%)**': This metric reports the near-worst case MAPE (95th percentile) that would be expected from factored AADTs given a sample of N sites. At higher sample sizes this upper limit gets smaller, as large samples are less sensitive to being inflated by individual outliers.
- **Precision Limit**: ***Minimally 95% Probability, 95% TCE Population Error Range (%)'***:

  The precision limit describes how spread out (i.e., middle 95%) the site errors are when assessing factored AADTs. Unlike the accuracy limits above, this is a description of the underlying population of count site errors (not central tendency sample statistic like median PE and MAPE), so the limits do not depend on sample size. However, note that the testing procedure (described later) *does* depend on sample size.

The accuracy and precision limits as exhibited in Table B-1 from the FHWA guidance were obtained (from previous USDOT studies) by analyzing a sub-sample of 48-hour counts obtained from CCS. These sub-samples were treated as 48-hour short-term counts, factored appropriately, and then compared to the official AADT as derived from the full data set from their respective CCS. These fundamental studies of sub-samples of 48-hour synthesized STC, formed the basis for accuracy expectations for non-traditional volume estimates.

Table B-1: Accuracy and Precision Limits from Factored Short-Term Counts (from FHWA report)

AADT Volume Range†	Sample Size for Evaluation	Minimally 95% Probability, TCE Median Error (Bias) (%)	Minimally 95% Probability, MAPE Upper 95% CB (%)	Minimally 95% Probability, 95% TCE Population Error Range (%)
100 – 499 (very low)	N ≥ 1,000	± 2.7	18.8	± 64.4
	N = 200	± 4.1	19.9	
	N = 100	± 5.6	20.9	
	N = 50	± 7.6	22.3	
	N = 25	± 10.1	24.0	
500 – 4,999 (low)	N ≥ 1,000	± 1.6	13.0	± 43.4
	N = 200	± 3.1	14.5	
	N = 100	± 4.5	15.5	
	N = 50	± 6.4	16.7	
	N = 25	± 8.7	18.1	
5,000 – 54,999 (medium)	N ≥ 1,000	± 2.0	10.5	± 33.6
	N = 200	± 2.9	11.3	
	N = 100	± 3.8	11.9	
	N = 50	± 5.2	12.7	
	N = 25	± 6.9	13.6	
55,000 + (high)	N ≥ 1,000	± 2.7	8.2	± 22.0
	N = 200	± 4.0	8.9	
	N = 100	± 5.3	9.8	
	N = 50	± 7.1	11.2	
	N = 25	± 9.4	12.9	

†Roadways with AADT less than 100 have no pre-defined standards.

Evaluating Accuracy and Precision of Vendor AADT Estimates

The limits in Table 5 serve as the basis for assessing the accuracy and precision of vendor AADT estimates. Evaluating accuracy is straightforward; Median TCE (i.e., Percentage Error) and MAPE are computed for each volume bin and compared to the limits provided in Table 5, resulting in a pass/fail designation for each bin. Note that this table only reports accuracy limits for specific sample sizes (N=25, N=50). In practice, it is often useful to know the precise limits for sample sizes in between the values listed in the table. In such cases it is acceptable to either interpolate between records in the table or use a set of equations provided by FHWA (see Table 14a in FHWA report), for the specified sample size.

The precision test is also straightforward to execute but relies on more sophisticated underlying statistical quality control theory. To execute this test, the TCE (i.e., Percentage Error) is calculated for each record in the dataset and compared to the precision interval limit (last column in Table B-1). Each record is flagged as “in” or “out” of the precision interval limit, and the total number of observations that fall outside the limits are summed across the whole dataset (or subset to be included in the precision test). This total is then compared to the number of allowable failures based on specified sampling plan, which should be determined ahead of time. If the number of observations that fall outside the precision limits is less than or equal to the allowable limit, the precision test passes, otherwise it fails. This technique is often used in manufacturing to determine whether to accept or reject a production lot, based on inspecting a sample.

At a high level, the sampling plan seeks to determine the number 'failures' can be considered acceptable, while balancing risk to both the agency (a potential consumer of the data) and vendor (the producer). A sampling plan requires two binomial equations to be solved simultaneously:

$$1 - \alpha = \sum_{d=0}^{c} \frac{n!}{d!\,(n-d)!} p_1^d (1 - p_1)^{(n-d)}$$

$$\beta = \sum_{d=0}^{c} \frac{n!}{d!\,(n-d)!} p_2^d (1 - p_2)^{(n-d)}$$

- *n*: sample size
- *c*: number of allowable failures (**what we want to solve for)**
- *p1*: failure rate that is considered "good performance". Here failure rate is defined as the proportion of records that fall outside of precision intervals from Table 5.
- *p2*: failure rate that is considered "bad performance". Here failure rate is defined as the proportion of records that fall outside of precision intervals from Table 5.
- α: vendor risk (usually set to 0.05, or 5%). This is the probability that the test fails even though the actual failure rate (proportion of records that fall outside precision intervals) is less than *p1* (the threshold for "good performance")
- β: agency risk (usually set to 0.05, or 5%). This is the probability that the passes even though the actual failure rate (proportion of records that fall outside precision intervals) is greater than *p2* (the threshold for "bad performance")

Solving these equations is not trivial because some parameters – like sample size – are automatically determined based on the number of available validation locations and must be discrete values. The value for c (number of failures) is what we want to solve for and must also be a discrete value. Alpha and beta are often set to 0.05 (5%) to balance agency and vendor risk, which just leaves p1 and p2 (the definitions of "good" and "bad" failure rates, respectively). *Evaluators are free to come up with sampling plans that satisfy these constraints; however, they must determine the value of* ***c*** *(number of failures) before running the precision test.*

To make things easier, FHWA has prepared a table of reasonable sampling plans – each of which uses alpha = 0.05 and beta = 0.05 and tries to pick p1 and p2 failure rates that are as fair as possible. This is summarized in Table B-2 below. For clarity, we use the sampling plan in Table 6 for our implementation of the FHWA method. In cases where the sample size falls between observed records, we perform linear interpolation and round the value of c (number of allowed failures) up to the nearest integer – giving the benefit of the doubt to the vendors.

Table B-2: Acceptance sampling plan for precision test (from FHWA report)

n	*c*	p_1	p_2	α	β
25	0	0.002	0.113	0.05	0.05
50	1	0.007	0.091	0.05	0.05
80	2	0.01	0.077	0.05	0.05
100	4	0.02	0.089	0.05	0.05
200	7	0.02	0.065	0.05	0.05
400	16	0.027	0.060	0.05	0.05
1000	50	0.04	0.063	0.05	0.05

With this information in mind, the FHWA test is conducted as follows:

1. Create a paired dataset with reference and vendor volumes (bi-directional) and calculate Percentage Error (PE) for each record.
2. Assign each record a volume bin based on reference volume value. (Very low = 1-499, Low = 500-4,999, Medium = 5,000-54,999, High = 55,000+)
3. Group records by volume bin. For each bin:
   - **Accuracy Test #1:** Calculate Median PE (called TCE) across records and compare it to "*Minimally 95% Probability Traffic Count Error (TCE) Median Error Bias (%)"* limit.
     - If Median PE is within interval → **Pass**
     - If Median PE is outside interval → **Fail**
   - **Accuracy Test #2:** Calculate MAPE across records and compare it to *"Minimally 95% Probability, MAPE Upper 95% CB (%)"* limit.
     - If MAPE is less than or equal to limit → **Pass**
     - If MAPE exceeds the limit → **Fail**
4. **Precision Test:** For each record in the full dataset (or subset of data to be included in precision test), compare the PE value to '*Minimally 95% Probability, 95% TCE Population Error Range (%)'* and count the total observations outside the interval.
   - If number of 'failures' is less than or equal to limit from sampling plan → **Pass**
   - If number of 'failures' exceeds limit from sampling plan → **Fail**

APPENDIX C: North Carolina Volume Validation (2023)

The North Carolina Volume Validation, conducted in 2023, was the inaugural effort to comprehensively validate the industry's capacity to deliver probe-based volume estimates at scale.

Traditionally, the Transportation Data Marketplace (TDM) emphasized outsourced probe data for operational purposes. However, beginning with the third iteration of the TDM in 2022, the market evolved to support various applications, including planning and performance metrics. Since no vendor offered real-time volume estimates at the time, this validation focused on available volume aggregations.

The full report for this initial activity is available here: TDM Validation Activity: North Carolina Volume Study – Nov 2023.

Key Findings

The North Carolina validation focused on quantifying volume data accuracy from four TDM vendors: HERE, INRIX, Iteris, and StreetLight. The significant findings of this exercise included:

- **Four Vendors Successfully Delivered Volume Data:** This marked a major industry milestone. Pre-pandemic, probe-based volume estimates were primarily in research and development. By this validation, multiple vendors were capable of providing estimates across hundreds of continuous count stations (CCS) in a standardized format, with minimal manual interaction required for geo-location or conflation.
- **Standardized Evaluation Framework Emerged:** A standard set of visualizations, metrics, and benchmarks was established for evaluating the quality of probe-based volume estimates. This foundational toolset (including statistical measures, scatter diagrams, error distributions, and the FHWA AADT evaluation method) became the common practice replicated in future validations.
- **Early Product Maturity Demonstrated:** Vendor accuracy was quantifiable. While moderate differences were observed between vendors, performance was generally mixed across volume ranges, urban/rural settings, and time aggregations. There was no statistically significant difference in overall performance within the vendor pool.
- **High Interest in AADT:** Driven by the expanded MIRE reporting requirements taking effect in 2026, Coalition members were most interested in the Annual Average Daily Traffic (AADT) results. Even though AADT could not be completely blind-tested, it served as a vital benchmark for industry capability.

Validation Context and Data Sets

The validation spanned the state of North Carolina and included 263 Continuous Count Stations (CCS). Short-term count locations were excluded from this initial validation but were integrated into later studies (e.g., Georgia in 2024).

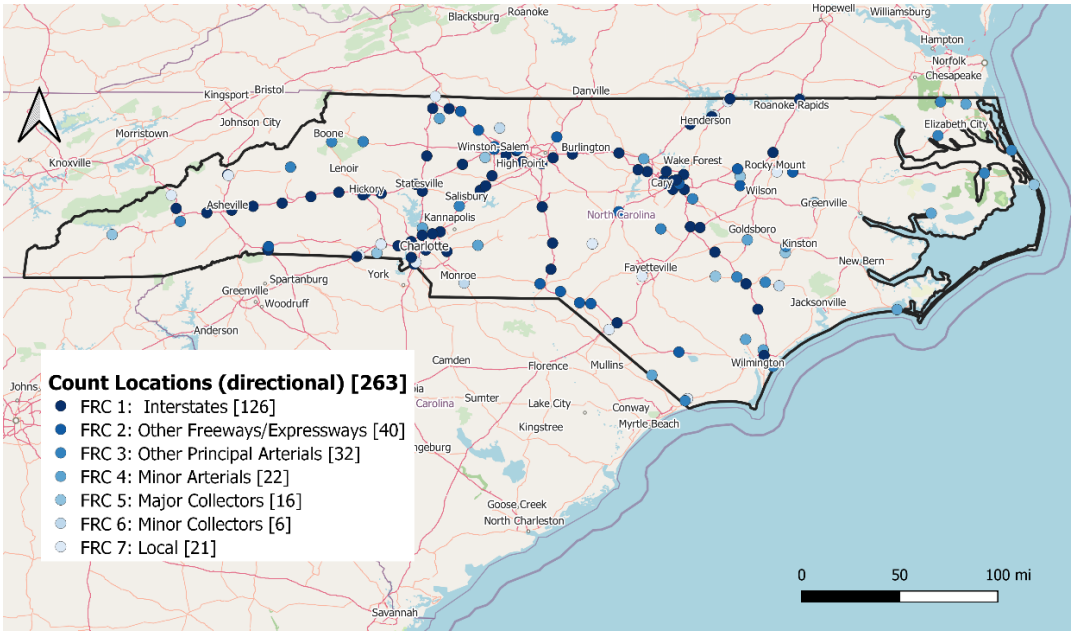


Figure C-1: Continuous Count State Locations in North Carolina - 2023

The methodology established three primary data sets:

1. **Calibration Data:** 17 months of historical traffic count data from CCS locations. This was provided to all vendors to train their machine learning and artificial intelligence models.
2. **Reference Data (Ground Truth):** Hourly traffic counts provided by NCDOT (Dec 17, 2022 – Jan 31, 2023) served as the benchmark for comparison for all measures except AADT (which utilized the full 2022 calendar year).
3. **Vendor Data:** Probe-based volume estimates provided by the four industry participants. To streamline evaluation and avoid conflation errors, vendors were required to submit data using a common format (the CATTWorks Geo-referencing Protocol, or CWGP).

Summary of Data Aggregations

The four reference data sets evaluated are summarized below in order of decreasing temporal granularity:

Table C-1: Summary of data items and study periods for the North Carolina Volume Validation

Data Item	Description	Study period	Reference Source
Hourly Volumes	Volume for a specific hourly period (e.g., 12/17/22 between 8-9am)	12/17/22 – 1/31/23	Continuous Count Stations (NCDOT)

Avg Hourly Daily Traffic (AHDT)	Avg hourly volume by day-of-week (e.g., Mondays between 8-9am)	1/1/23 – 1/31/23	Derived from CCS counts
Avg Daily Traffic (ADT)	Avg volume for a day during period	1/1/23 – 1/31/23	Derived from CCS counts
Avg Annual Daily Traffic (AADT)	Avg volume for a day during period	1/1/22 – 12/31/22	Derived from CCS counts

- **Hourly Volumes:** Included ~271,000 directional observations from over 250 CCS. The data showed many low-volume periods: 50% of hourly observations were under 350 vehicles per hour (vph), and 90% were under 1,000 vph.

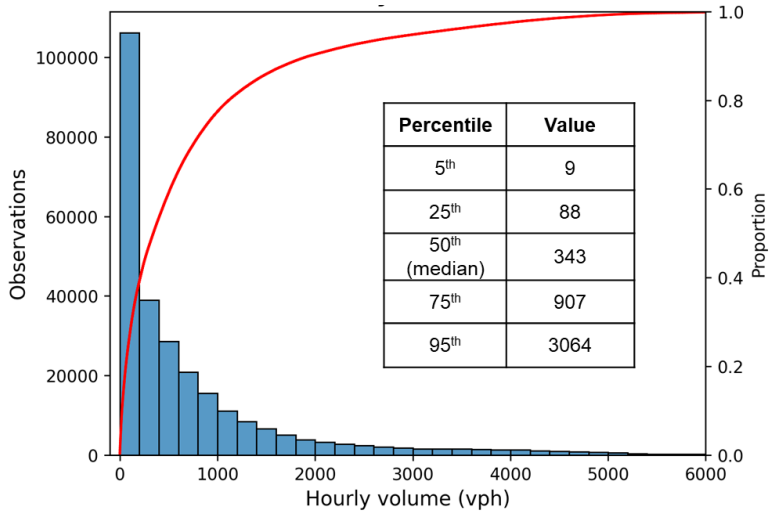


Figure C-2: Distribution of Hourly volumes in North Carolina Volume Validation

- **AHDT:** Yielded ~42,000 observations after aggregating by location, day of week, and hour (excluding holidays and other anomalous days).

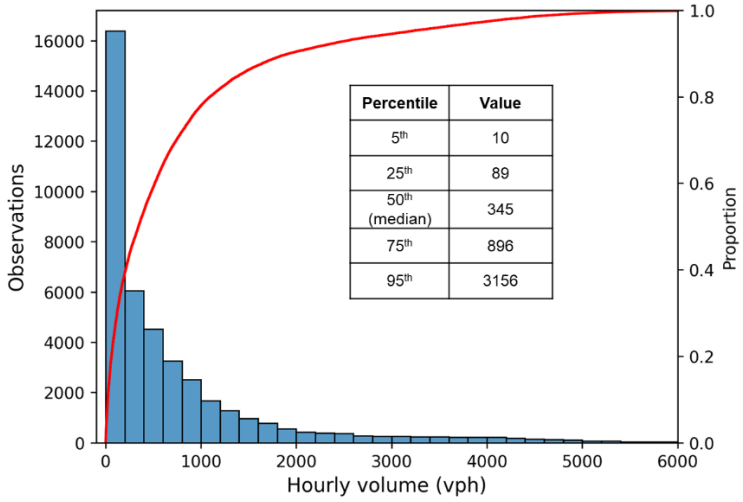


Figure C-3: Distribution of ADHT volumes in North Caroline Volume Validation

- **ADT:** Yielded 250 observations. ADT values were obtained by grouping the hourly counts in January 2023 by location and computing the volume corresponding to an average day, following the FHWA-recommended method from of the 2022 Traffic Monitoring Guide[1] (see MADT definition). Half of the locations had ADT values under 11,000 vehicles per day (vpd).

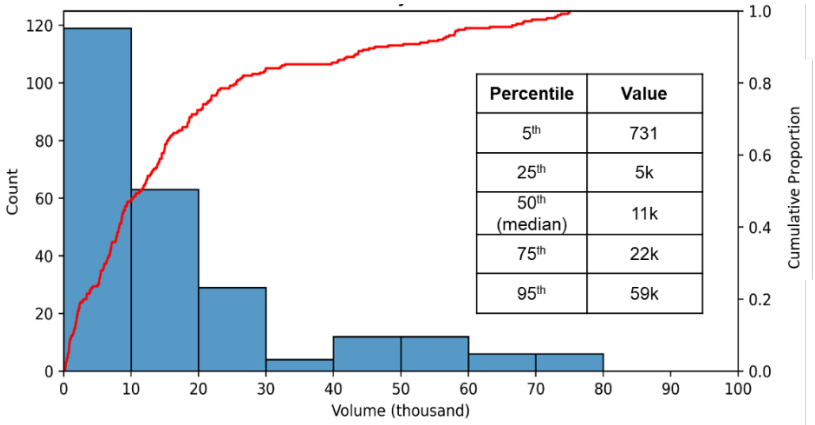


Figure C-4: Distribution of ADT volumes in North Carolina Volume Validation

- **AADT:** Yielded 261 observations for 2022 using the FHWA recommended method. Half of the locations had AADT values under 14,000 vpd.

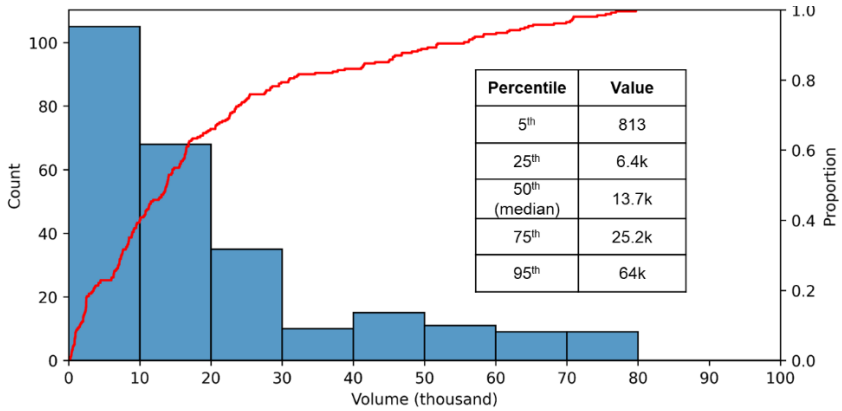


Figure C-5: Distribution of AADT volumes in North Carolina Volume Validation

Note on Blind Testing: NCDOT temporarily restricted public access to continuous count data during the winter evaluation period, al+lowing for a blind evaluation of hourly, AHDT, and ADT measures. However, a completely blind evaluation of AADT was not feasible due to the public availability of historical yearly counts.

Vendor Performance

All four vendors successfully engaged with the validation activity and delivered data in the specified CWGP format within the timeframe. This was a notable accomplishment

[1] 2022 FHWA Traffic Monitoring Guide (link)

given the major supply chain disruptions occurring at the time (e.g., mobile OS privacy policy changes and the loss of a leading connected vehicle data provider).

The accuracy results for AADT are reflected in the scatter plots shown if Figure C-6. The dashed red line represents points with zero error (i.e., vendor data volume estimates are equal to reference values), while the dotted black lines represent +/- 15% error. Figure C-6 reflects the capability of the industry to deliver accurate AADT volume estimates in early 2023. This visualization is used in subsequent validation activities (Georgia and Massachusetts).

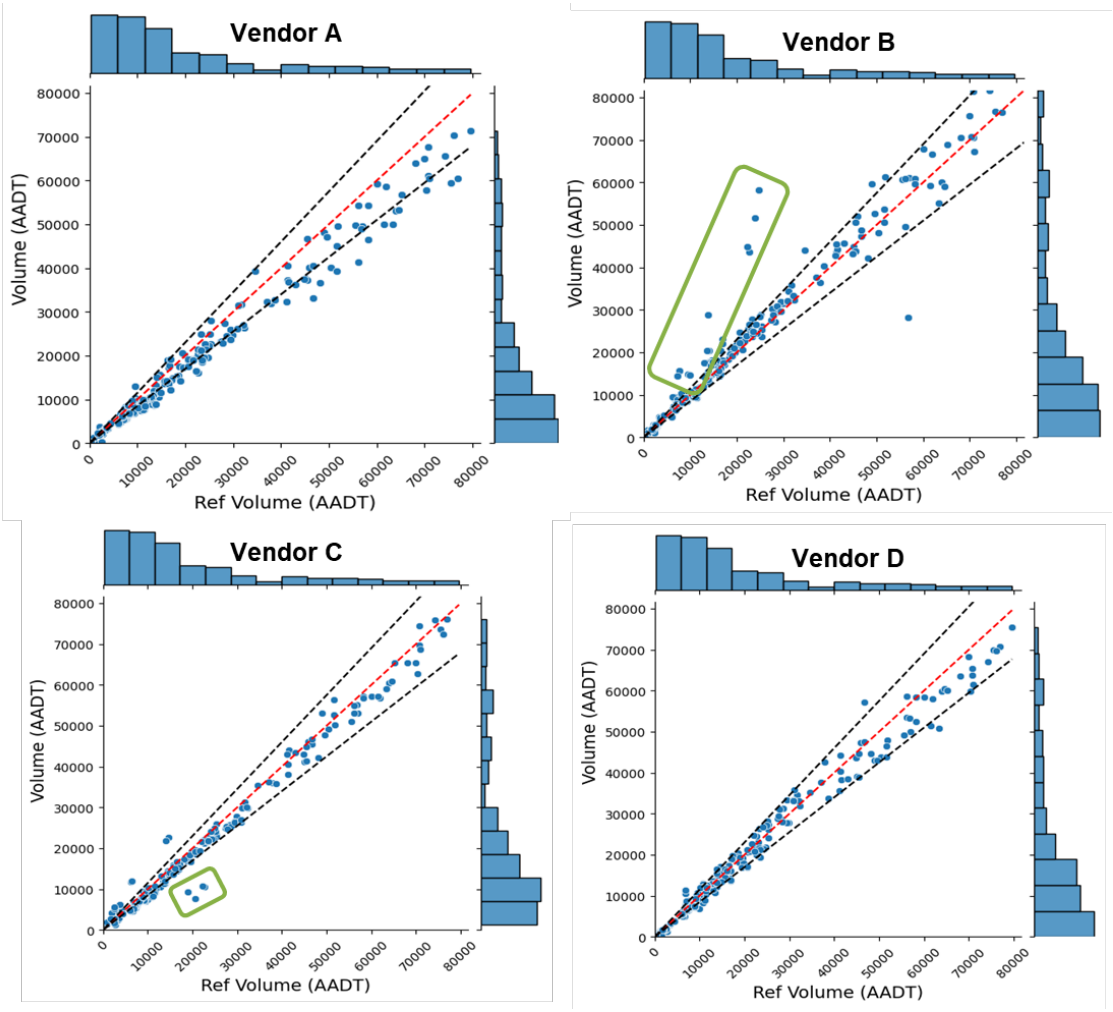


Figure C-6: AADT scatter plots for all vendors

Figure C-7 shows the error distribution across various volume bins. Th boxplots to depict error distributions. Figure C-7 summarizes the AADT percent error for the four vendors by volume bin, with the blue lines indicating +/- 15% error similar to the reference dashed lines in the scatter plots in

Figure C-6. The volume bins coincide with FHWA's volume bin definitions for assessing quality for non-traditional volume measures. Number of observations are not distributed equally across volumes ranges. The lowest volume bins have small sample sizes. Each boxplot summarizes the distribution of error, with the box showing the

middle 50% of data (encompassing data from the 1st to the 3rd quartiles – often referred to as the interquartile range), with the vertical line in the middle of the box representing the median value (50th percentile). The whiskers of each boxplot extend below the 25th percentile (Q1) and above 75th percentile (Q3) by 1.5 times the inter-quartile range (Q3-Q1) to represent the estimated non-outlier range, and red 'X's represent the mean value of the error distribution. Large offsets of the mean from the median indicate the presence of large outliers.

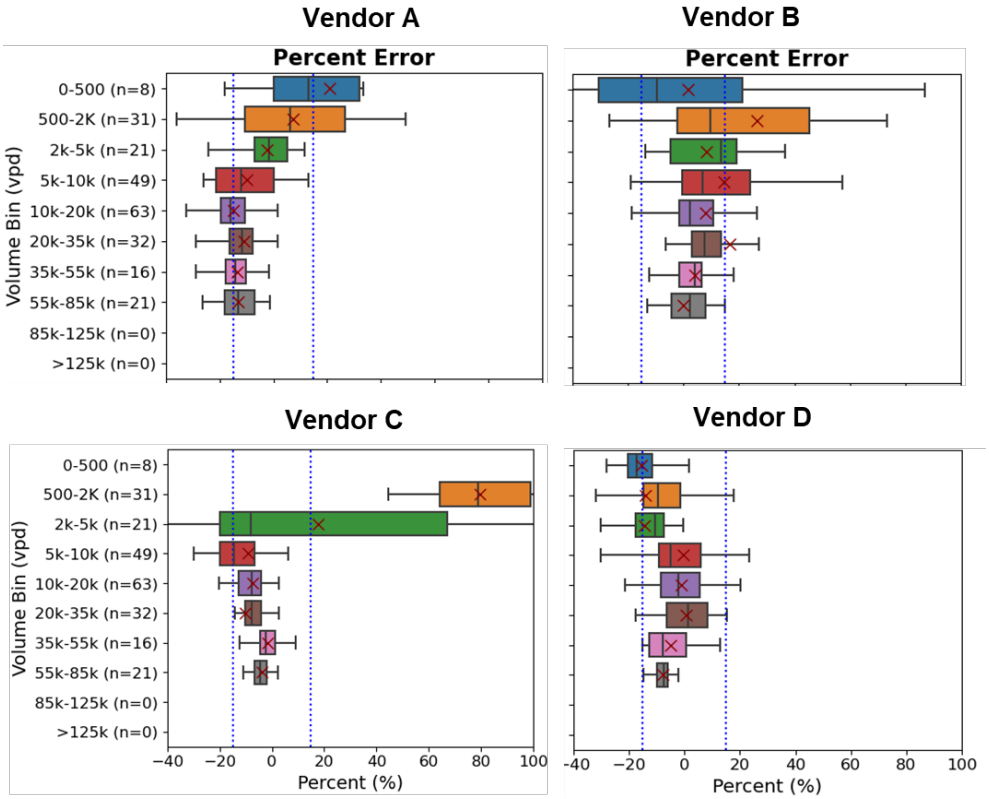


Figure C-7: Distribution of AADT percent error metrics by volume bin in NC

The general performance characteristics for AADT estimations among the four vendors are outlined below:

- **Vendor A:** Exhibited a systematic bias error, consistently underestimating volumes. This underestimation was most prevalent in volume bins starting at 5,000 vpd and higher.
- **Vendor B:** Showed less overall bias than Vendor A, with high accuracy in the higher volume bins. However, accuracy diminished in lower volume bins, and the vendor produced several prominent outliers that significantly overestimated volumes.
- **Vendor C:** Demonstrated the tightest accuracy overall, with most data points falling within a +/- 15% error margin. Despite a few notable outliers, the primary weakness was a significant drop in accuracy in the lowest volume bins (under 500 vpd).

- **Vendor D:** Displayed minimal bias, with most data points falling within the +/- 15% error margin. While its overall error distribution was not as tight as Vendor C, Vendor D did not exhibit instabilities at lower volume ranges (<5,000 vpd).

Figures C-6 and C-7 are standard visual representations of the accuracy of vendor provided volume estimates used through the TDM validation procedures from 2022 through 2026. The full North Carolina report provides these visualizations for all time aggregations (Hourly, AHDT, ADT, and AADT). These visualizations, as well as their accompanying core statistical metrics (mean, median, quartiles, and outliers in various volume bins) form the basis for recommended statistical visualizations in this guide.

The statistical results from FHWA method were also applied to North Carolina data set. These results are presented later alongside those for Georgia and Massachusetts as comparative of the progression of industry capability and maturity from 2022 to 2026.

Appendix D: Georgia Volume Validation (2024)

In 2024, the TETC conducted a volume validation exercise in Georgia, building upon the foundational 2023 North Carolina study. This validation evaluated both planning-level (AADT, ADT, AHDT) and granular (daily, hourly) volume estimates using reference data from continuous count stations (CCS) and temporary short-term count (STC) sites across all functional road classes (FRC 1-7).

The full report for this activity is available here: TDM Validation Activity: Georgia Volume Study.

Key Findings

Five vendors participated in the Georgia validation: HERE, INRIX, Iteris, StreetLight, and Replica. Significant takeaways included:

- **Measurable Industry Progression:** Despite ongoing data supply chain disruptions, vendor products showed noticeable improvement over the previous year. Coverage expanded to include lower-volume rural facilities (FRC 6-7), and several vendors successfully captured day-of-week, time-of-day, and atypical volume patterns.
- **Vendor Separation Emerged:** Vendors A, C, and D stood out as strong performers, nearing FHWA accuracy and precision benchmarks for AADT. Vendors C and D also demonstrated the ability to produce highly sensitive, operational-level daily and hourly estimates.
- **Multiple Sources of Error Identified:** The validation revealed that vendor modeling is only one piece of the puzzle. Significant errors were traced to map conflation and geo-referencing issues, anomalies within the agency's own reference data (e.g., counting only a subset of lanes), and misinterpretations of roadway directionality.

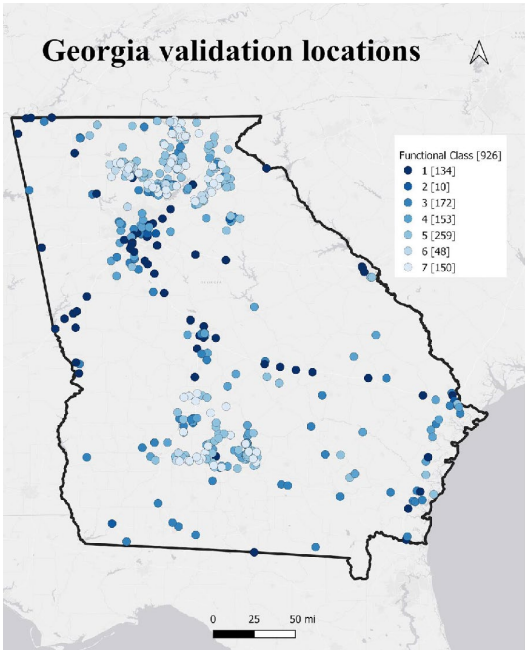


Figure D-1: Georgia Volume Validation Sites

Validation Context and Data Sets

Vendors submitted data using the standard CATTWorks Geo-referencing Protocol (CWGP). AADT estimates were evaluated for the entire 2023 calendar year, while the remaining aggregations (ADT, AHDT, daily, and hourly volumes) focused on April 2024.

During this time the industry was recovering from broad-based data supply chain disruptions with respect to security policy changes at Apple and Android impacting location-based service (LBS) data and the insolvency of the industry's primary connected vehicle data (CVD) provider. As in North Carolina, reference data were obtained from continuous count stations. In

Georgia, reference data were also collected at temporary coverage counts sites as illustrated to the right, covering all functional road classes (1-7) on the state highway system. The temporary coverage counts are typically 48 hours in duration and provide a larger, spatially diverse sample but for short durations.

Data Availability Threshold To ensure reference values were truly representative, a strict 75% data availability rule was applied to CCS locations. For AADT, this yielded 286 directional evaluation locations, half of which had volumes under 8,000 vehicles per day (vpd).

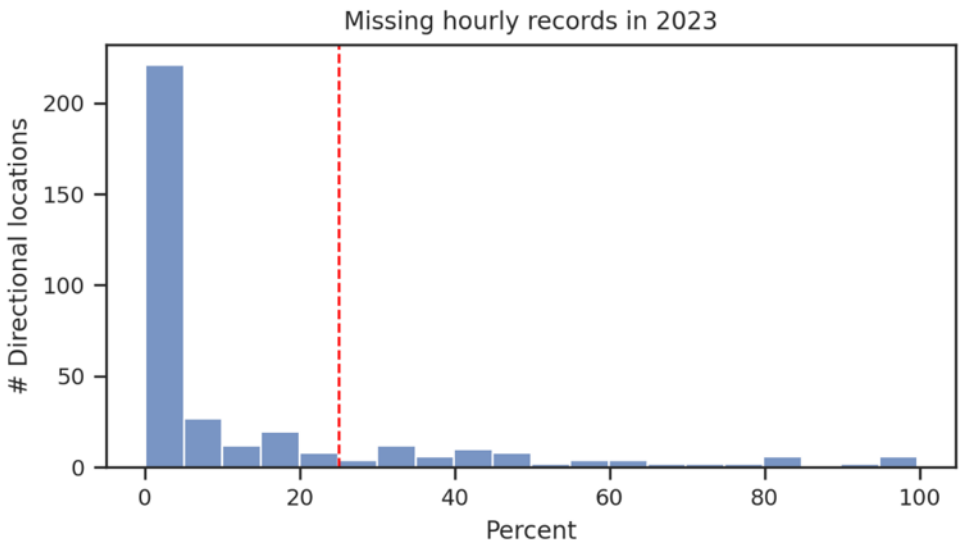


Figure D-2: Applying 75% data availability thresholds to GA data

Vendor Submission Summary

Three vendors delivered all requested data items, while two provided only the most aggregate product (AADT).

Table D-1 Vendor Submission Summary for GA Volume Validation

Vendor	AADT	ADT	AHDT	Daily Volumes	Hourly Volumes
A	✓	✓	✓	✓	✓
B	✓				
C	✓	✓	✓	✓	✓
D	✓	✓	✓	✓	✓
E	✓				

Vendor Performance and Error Analysis

- **AADT Accuracy:** Vendors A, C, and D were the most consistent performers across most volume ranges. Vendor C demonstrated exceptionally tight accuracy in mid-to-high volume ranges but struggled with significant overestimations in the lowest volume bins (<500 vpd). Vendors A and D maintained consistent error distributions across all ranges.
- **FHWA Bi-Directional Testing:** Applying the FHWA's updated statistical method (which relies on bi-directional volumes), Vendors A and D passed 7 out of 9 possible accuracy and precision tests. Vendor C passed 5 of 9 tests, primarily held back by outliers in low and medium volume categories.
- **Urban vs. Rural Performance:** Segmenting the data by urban and rural classifications revealed no systematic degradation in performance between the two location types.
- **Spatial Consistency:** A conservation of flow test conducted at a specific interchange showed that four of the five vendors were internally consistent, with approach and departure volumes agreeing within roughly 3%. (Vendor C initially showed a 25% discrepancy due to a processing error, which was subsequently corrected).
- **Permanent vs. Short-Term Counters:** When evaluating daily volumes, errors were notably higher at temporary short-term count (STC) sites compared to permanent CCS locations. Performance degraded most significantly on FRC 6 and 7 roads. Because vendors rely heavily on publicly available CCS data to train their machine learning models, estimating volumes on lower-class roadways with minimal historical training data remains a challenge.

The Georgia results contain data from both continuous counters and temporary counters. The number and distribution of continuous counters versus temporary count sites are shown in Figure D-3. Continuous count stations tend to be located on higher class roadways (FRC 1 through 5) and less representative of lower volume roads (FRC

6 and 7). In contrast, temporary count sites are clustered in lower road classes (FRC 3 through 7). High volume roadway classes (FRC 1 and 2) are nearly exclusively covered by continuous counters, whereas extremely low volume roadways (FRC 6 and 7) are nearly exclusively counted by temporary count activities.

Calibration data is dominated by data from CCS. For this reason, better volume estimates would be expected in FRC 1 through 5. This is particularly relevant for AADT estimates from probe data for MIRE reporting purposes because the majority of the expansion occurs on lower class roadways.

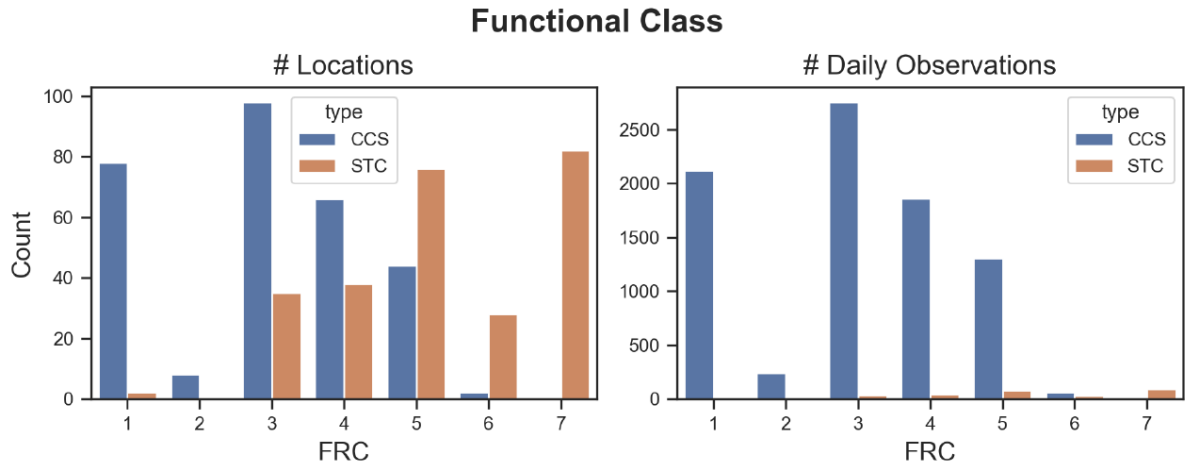


Figure D-4: Distribution of Continuous and Temporary Volume Counts by Roadway Classification for Georgia Volume Validation

Figure D-5 presents scatter plots of the vendor's AADT volume accuracy (similar format as that explained for the North Carolina validation). All vendors consistently produced estimates that frequently fell within the 15% bounds, with some notable outliers (e.g. overestimates by Vendor C at a handful of locations) observed.

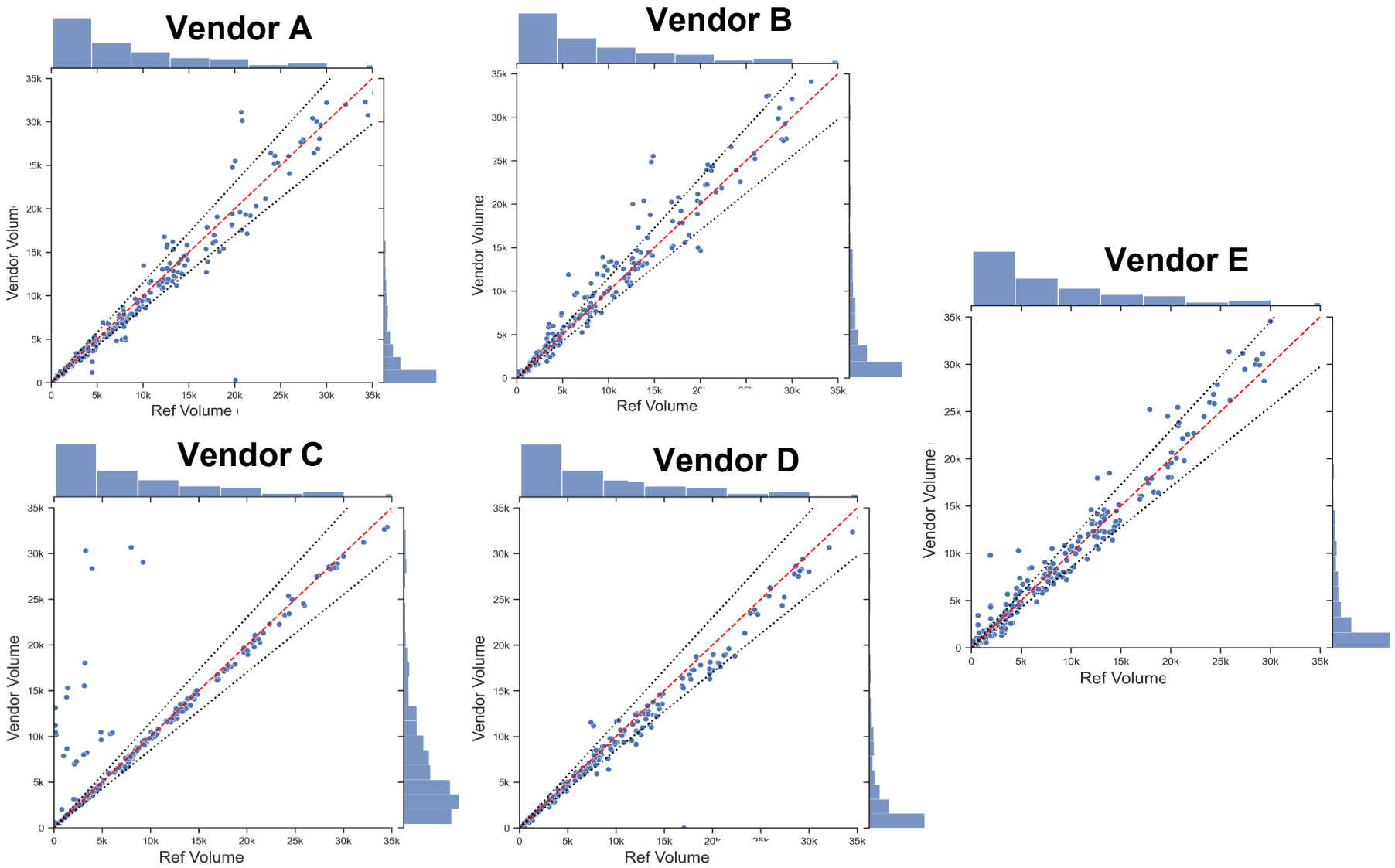


Figure D-5: AADT scatter plots by vendor for Georgia Volume Validation.

Figure D-6 shows the error distribution of AADT error percentages for each vendor via boxplots, separated into volume bins. Vendors A, C and D stand out as consistent performers across most volume ranges, with Vendor C showing very high accuracy across mid-to-high volume ranges, and Vendors D and A showing consistent performance across all volume ranges.

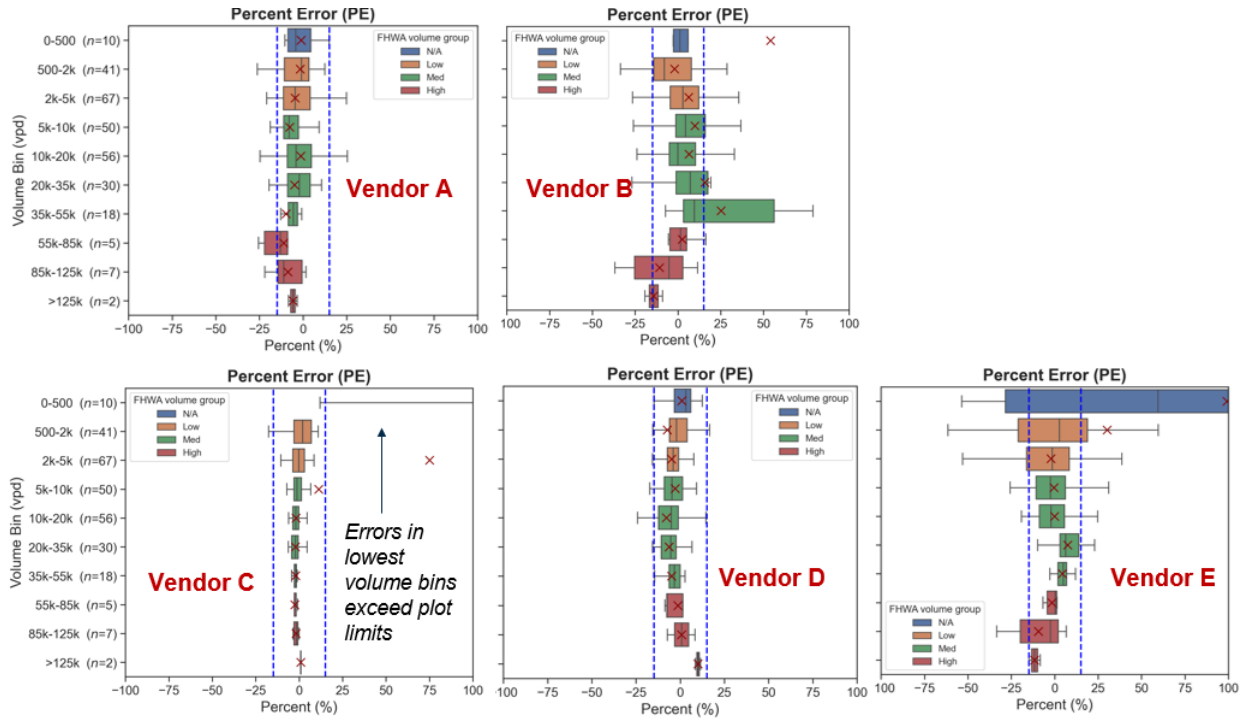


Figure D-6: AADT Error Distribution Plots in Georgia Volume Validation.

In comparison to North Carolina results, many of the bias issues reflective of consistent over or under estimation of volumes appear to be resolved. Error distributions appear to have improved in most volume ranges with low and extremely flow volumes still problematic for some vendors.

Assessing Performance in Urban vs Rural Locations

At the direction of the TDM Technical Advisory Committee, results were separated into rural versus urban to determine whether performance is consistent across location types. Using GDOT's classification for each count station, Figure D-7 shows the results separated by urban/rural location for each vendor, with urban results on the left and rural results on the right. Based on the plots in Figure D-7, no systematic evidence exists to suggest degradation in performance between urban and rural location types.

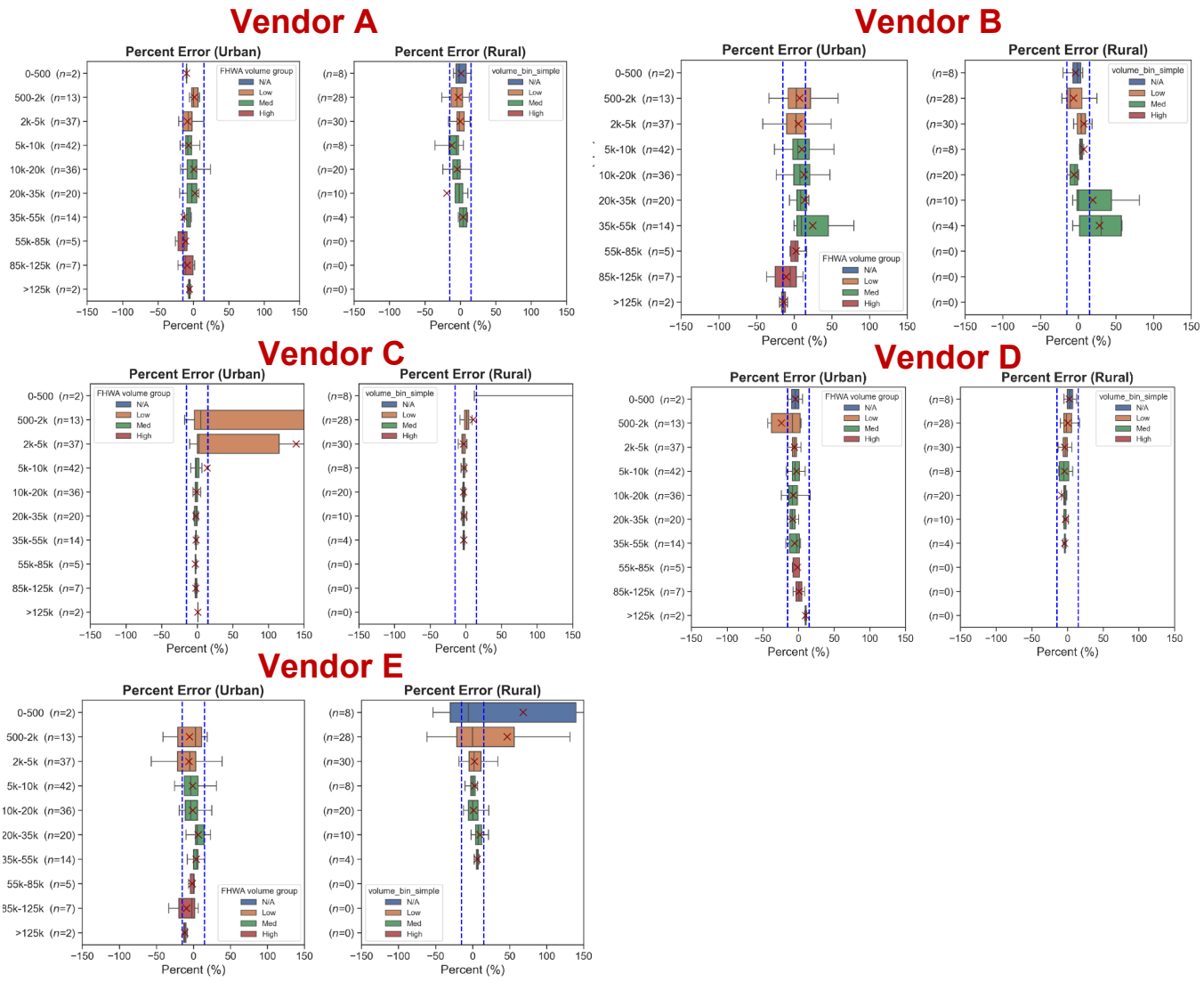


Figure D-7: AADT Error Distribution Plot broken down by Urban vs Rural locations

Spatial Consistency Test

A spatial consistency test was introduced to check whether vendor volumes estimates are internally consistent at an intersection. This spatial consistency test can be run even if reference data is not available to quantify accuracy measures at a specific location. Figure D-8 shows one such interchange where the spatial consistency test was executed. Eight validation reference locations were defined at this interchange, one for every approach and departure. Vendor volume estimates were tested for conservation of flow (i.e., volume from four legs entering the interchange should roughly equal volumes from the four legs leaving the intersection). This test was executed for all temporal aggregations, including AADT.

Four of the five vendors showed basic conservation of flow, with volumes into and out of the interchange agreeing within approximately 3%. The remaining vendor (Vendor C) had about a 25% difference at this location. After bringing this to the vendor's attention, Vendor C discovered an error in data processing for one of the locations. After addressing this issue, their performance was consistent with other vendors.

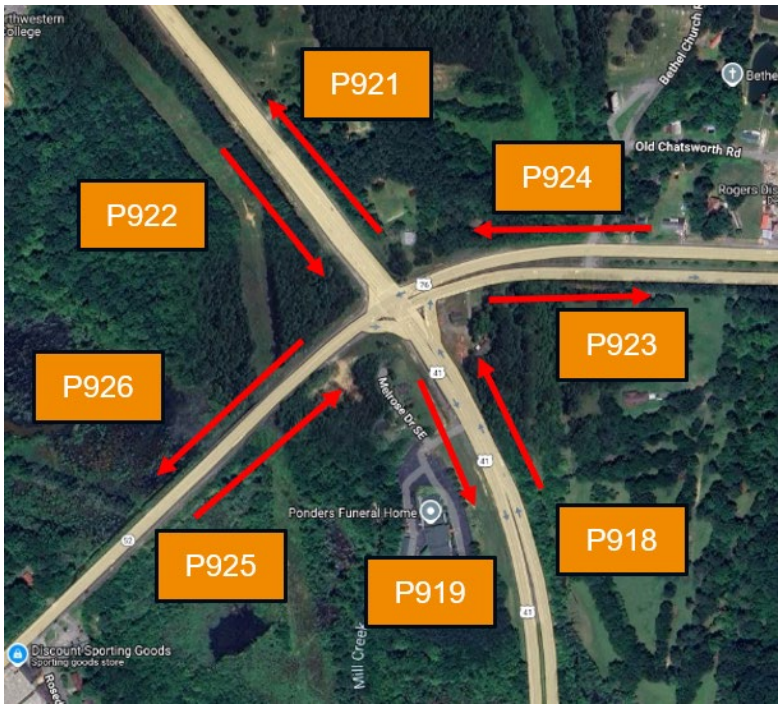


Figure D-8: Spatial Consistency Test Location Diagram

Applying the FHWA Method in Georgia Volume Validation

The FHWA method establishes accuracy and precision benchmarks in three broad volume ranges (Low: 500-4999 vpd, Medium: 5000-54999 vpd, High: 55000+ vpd). These benchmarks reflect the expected precision/accuracy of factoring 48-hour counts to AADT values and are based on prior FHWA research. The FHWA test calculates statistical measures for each vendor's AADT product and compares those with FHWA's benchmarks for 48-hour counts. There are three statistical benchmarks, two for accuracy and one for precision. Each test is applied in each of the three volume bins, resulting in 9 opportunities for evaluation. (Beginning with the Massachusetts volume validation, the precision test was applied across all three volume bins due to low statistically low sampling, bringing the total number of tests to 7 rather than 9.)

Accuracy result for previously displayed metrics have been explicitly for *directional* traffic volumes, regardless of facility type. FHWA's statistical method is based on bi-directional volumes, using the total volume associated with both directions of traffic flow at a count site. To obtain bi-directional volumes, volumes in both directions of travel were summed at each site, both for reference and vendor data sets. When directional volumes are obtained from bi-directional flow two effects occur: (1) the sample size is halved and (2) the observation may shift to a higher volume bin because volumes roughly double at each location.

Table D-2 summarizes the results for the FHWA test using *bi-directional volumes*. Results are presented for each of the three tests and three volume bins for each vendor, resulting in nine tests per vendor. Benchmark limits are also shown to provide context for how close each vendor's result was to the pass/fail threshold. A tally of the total number of tests passed by each vendor appears to the right. Table D-2 presents accuracy and precision results in all three volume bins, with the sample size for each bin designated. The 'High Volume' bin has 20 observations, which is less than the 25 minimum recommended observations. All vendors passed at least 4 of the 9 tests. Two vendors, Vendors D and A, passed 7 of the 9 tests.

Table D-2: FHWA Method Bi-Directional Results for Georgia Volume Validation.

			Accuracy Test #1			Accuracy Test # 2			Precision Test			
Vendor	Volume Category	Number of Locations	Absolute Median TCE	Limit	Result	MAPE	Limit	Result	# Failures	Limit	Result	# Tests Passed
A	Low	27	0.5	8.7	Passed	8.9	18.1	Passed	0	0	Passed	7
	Medium	93	4.2	5.2	Passed	10.7	12.7	Passed	5	2	Failed	
	High	20*	6.1	9.4	Passed	10.4	12.9	Passed	2	0	Failed	
B	Low	27	8.1	8.7	Passed	16.8	18.1	Passed	1	0	Failed	4
	Medium	93	3.1	5.2	Passed	15.0	12.7	Failed	11	2	Failed	
	High	20*	3.1	9.4	Passed	19.8	12.9	Failed	5	0	Failed	
C	Low	27	1.4	8.7	Passed	149.5	18.1	Failed	5	0	Failed	5
	Medium	93	1.7	5.2	Passed	21.0	12.7	Failed	6	2	Failed	
	High	20*	2.0	9.4	Passed	2.2	12.9	Passed	0	0	Passed	
D	Low	27	3.0	8.7	Passed	10.1	18.1	Passed	1	0	Failed	7
	Medium	93	5.5	5.2	Failed	7.7	12.7	Passed	2	2	Passed	
	High	20*	2.3	9.4	Passed	4.7	12.9	Passed	0	0	Passed	
E	Low	27	1.3	8.7	Passed	44.7	18.1	Failed	6	0	Failed	5
	Medium	93	0.3	5.2	Passed	12.3	12.7	Passed	8	2	Failed	
	High	20*	2.5	9.4	Passed	7.0	12.9	Passed	1	0	Failed	

Note: Results are reported for 'High' volume category although sample size is less than 25.

Comparison with Prior Results

The Noth Carolina datasets were scored for bi-directional volumes, similar to Table D-2 above. Additional data filtering (using the same data availability threshold of 75% as in GA) was implemented when re-running the FHWA analysis.

Tables D-3 summarizes the bi-directional results for NC. Performance is are fairly consistent across all four vendors, ranging from 4-6 of the 9 tests passed. Based on these FHWA tests, several vendors improved performance form NC to the Georgia validation one year later, with three vendors passing at least 7 of the 9 tests.

Table D-3: FHWA Method Results for North Carolina Volume Validation.

Vendor	Volume Category	Number of Locations	Accuracy Test #1			Accuracy Test # 2			Precision Test			# Tests Passed
			Absolute Median TCE	Limit	Result	MAPE	Limit	Result	# Failures	Limit	Result	
A	Low	20*	16.1	8.7	Failed	33.0	18.1	Failed	5	0	Failed	6
	Medium	75	0.7	5.2	Passed	10.1	12.7	Passed	1	1	Passed	
	High	31	4.0	9.4	Passed	8.6	12.9	Passed	1	0	Failed	
B	Low	23*	9.5	8.7	Failed	16.7	18.1	Passed	2	0	Failed	4
	Medium	76	7.0	5.2	Failed	11.5	12.7	Passed	3	1	Failed	
	High	31	2.5	9.4	Passed	7.0	12.9	Passed	2	0	Failed	
C	Low	2*										5
	Medium	68	0.5	5.2	Passed	7.9	12.7	Passed	2	1	Failed	
	High	30	2.4	9.4	Passed	6.4	12.9	Passed	0	0	Passed	
D	Low	26	5.5	8.7	Passed	11.7	18.1	Passed	2	0	Failed	6
	Medium	76	9.2	5.2	Failed	12.3	12.7	Passed	2	1	Failed	
	High	31	0.0	9.4	Passed	8.0	12.9	Passed	0	0	Passed	

**Results are reported for volume category although sample size is less than 25.*

One caveat is that AADT evaluation was not a blind study. Historical CCS data was provided to each vendor for calibrating their models for the period for which AADT was evaluated. Higher temporal aggregations, such as ADT, provides an opportunity to compare performance when reference data is not available to the vendor. The accuracy results for ADT were consistent with the assessed AADT accuracy providing higher confidence in the AADT results.

Performance on Reference Data not used for Training

Volume accuracy temporary counter locations test's a vendors' ability to accurately report volumes on roadways beyond that which was used for model training. Validation performance was separated into continuous count stations where vendors had access to historic data to train the models and temporary count stations where it was unlikely that vendors had access to training data for the models. This could be applied to Daily volume estimates, as temporary counts are typically 48 hours in duration.

Figure D-9 compares error distributions for Daily Volumes at CCS locations versus temporary counter locations for vendors that provided Daily Volume estimates. In all cases the performance on the temporary count sites were worse than at continuous count sites, with performance degrading with lower roadway class. Vendors A and D performance degradations were within expected limits, while the performance degradation of Vendor 3 was substantially greater, suggesting over learning of data training data from CCS locations.

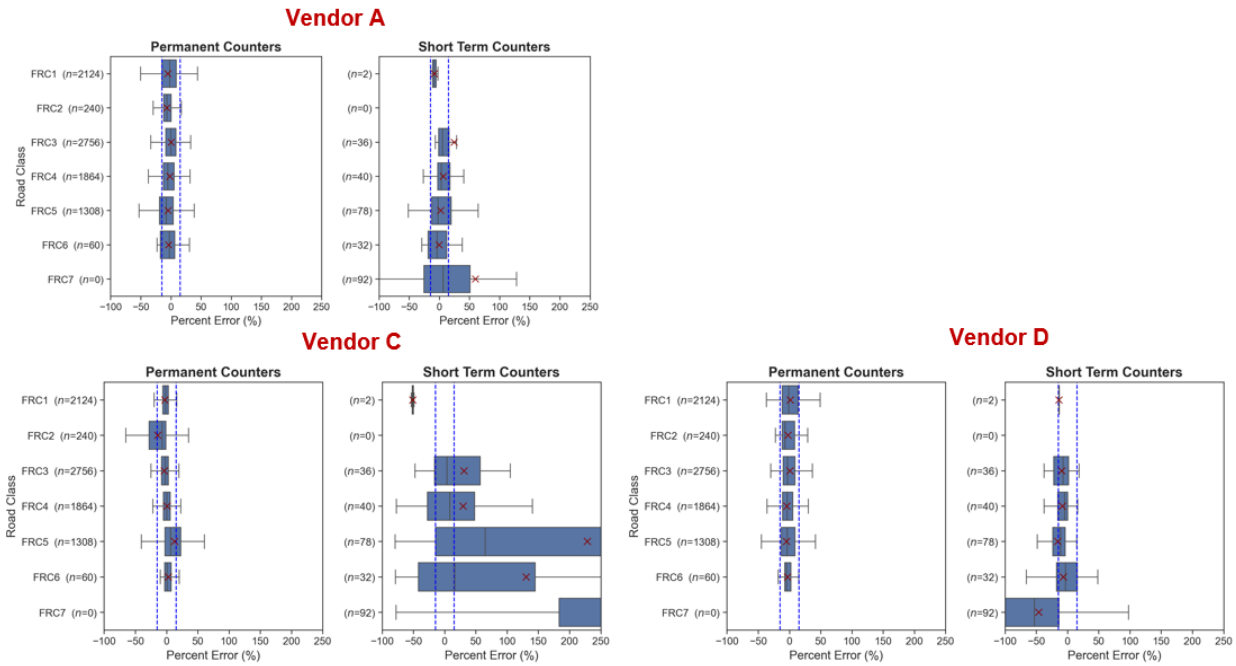


Figure D-9: Daily Volume Percent Error for permanent vs short-term count sites

APPENDIX E: Massachusetts Volume Validation (2025)

The TETC performed a validation of industry volume estimates in the state of Massachusetts in collaboration with the Massachusetts DOT (MassDOT) in 2025. As with Georgia, five TDM vendors participated: Bentley (HERE), INRIX, Iteris, StreetLight, and Replica. Four of the five vendors delivered volume estimates at all five temporal aggregations: annual average daily traffic (AADT), average daily traffic (ADT), average hourly daily traffic (AHDT), daily volumes, and hourly volumes, while one vendor provided AADT only. Volume estimates were provided at over 1,000 evaluation locations across the state.

The original intent of this study was to characterize vendor performance across all five temporal aggregations, using metrics, statistical tests, and visualizations consistent with prior studies in Georgia and North Carolina. However, the initial AADT-level analysis uncovered several issues that warranted further investigation. At the time of the release of this guidebook, the initial MassDOT volume validation report (Part I) that focused exclusively on the analysis of AADT volume estimates has been published. A second edition (Part II) is being prepared that covers the other measures of ADT, AHDT, daily, and hourly volume estimates.

The full Part I report for this activity is available here: TDM Validation Activity: Massachusetts Volume Study Part I (AADT) .

The MassDOT validation report (Part I) uses updated guidance from the FHWA for assessing non-traditional volume sources. Compliance with the updated FHWA methodology is an important benchmark for states seeking to use probe-based volume products for federal reporting purposes. Prior reports included a variation of this method, as FHWA's updated guidance occurred after the Georgia validation. The validation team included the updated methodology in the MassDOT volume validation and also retroactively applied it to previous validation results (North Carolina and Georgia) for direct comparison.

Initial analysis of AADT volume estimates revealed high errors and major outliers. These results were unexpected based on vendor performance in prior validations. As such, the validation team spent considerable effort tracking down the cause of these discrepancies. This process uncovered several different sources of error (beyond just vendor inaccuracy), a theme first exposed in the Georgia validation. These errors are important to understand because they produced a disproportionate impact on resulting validation metrics and distorted apparent industry performance, including initial results on the FHWA statistical test. It was discovered that some erroneous data was included in the calibration dataset provided to the vendors. For this reason, vendors were allowed to resubmit results after suspect locations were identified and tagged. All suspect reference data locations were excluded from the final analysis.

Key Findings

Several critical observations and takeaways emerged from the MassDOT volume validation study:

- **Diverse Counter Technology:** Agency counter technology and deployment locations were more diverse in Massachusetts than in prior validations. While primarily utilizing traditional inductive loop counters for its Continuous Count

Station (CCS) network, Massachusetts also monitors tolled facilities and utilizes a sizeable inventory of video count cameras at signal-controlled intersections. This mix of technologies and deployment contexts introduced additional complexity and resulted in varying levels of volume estimation performance.

- **Multiple Sources of Error Identified:** Although some outliers were attributed to vendor estimation error, many significant errors stemmed from other sources, such as miscoded site locations, ambiguity in georeferencing, and reference data errors. These non-vendor errors were equally or more impactful to the initial validation results than vendor inaccuracy.
- **Caution with Video Counting Technology:** Video cameras, deployed primarily for traffic control at intersections, exhibited greater volume estimation errors than traditional CCS and Automated Electronic Toll (AET) count sites. Because the data quality of these cameras is less validated for volume reporting applications, accuracy results were reported both with and without these camera locations to provide a more accurate comparison to previous validations.
- **Improved Vendor Accuracy:** After removing suspect locations and processing final submissions, four of the five vendors produced volume estimates that agreed well with reference AADT data, particularly when focusing on non-camera locations.
- **Strong FHWA Statistical Test Performance:** Implementing the updated FHWA statistical tests for bi-directional AADT volume accuracy revealed measurable industry progression. When evaluating accuracy against continuous counters and toll road counters (excluding video cameras), two vendors passed all components of the test, with two others nearing the benchmark.

The overarching insight from this validation exercise is the significant impact of data quality issues originating outside of vendor estimation models. Because performance evaluation statistics—including visualizations, error metrics, and the FHWA test—are highly sensitive to outliers, understanding and mitigating these external sources of uncertainty is critical. Identifying these issues early is essential for structuring future validations and fairly evaluating vendor products. Furthermore, these insights extend beyond validation procedures; they are highly relevant to any application requiring industry-sourced data to be conflated or geo-referenced to an agency base map. The necessity of managing various error sources to successfully apply outsourced vendor data was initially observed in the Georgia validation and strongly amplified during the MassDOT volume validation.

Validation Context and Data Sets

The validation encompassed 1,044 directional roadway locations across Massachusetts, covering functional road classes 1 through 7 (with the exception of FRC 6). See Figure E-1. These locations correspond to MassDOT count sites, including continuous count stations (CCS) that collect data 24/7 and temporary short-term count (STC) sites. AADT estimates were evaluated using 2024 calendar year data, while the remaining temporal aggregations (ADT, AHDT, daily, and hourly volumes) focused on May 2025.

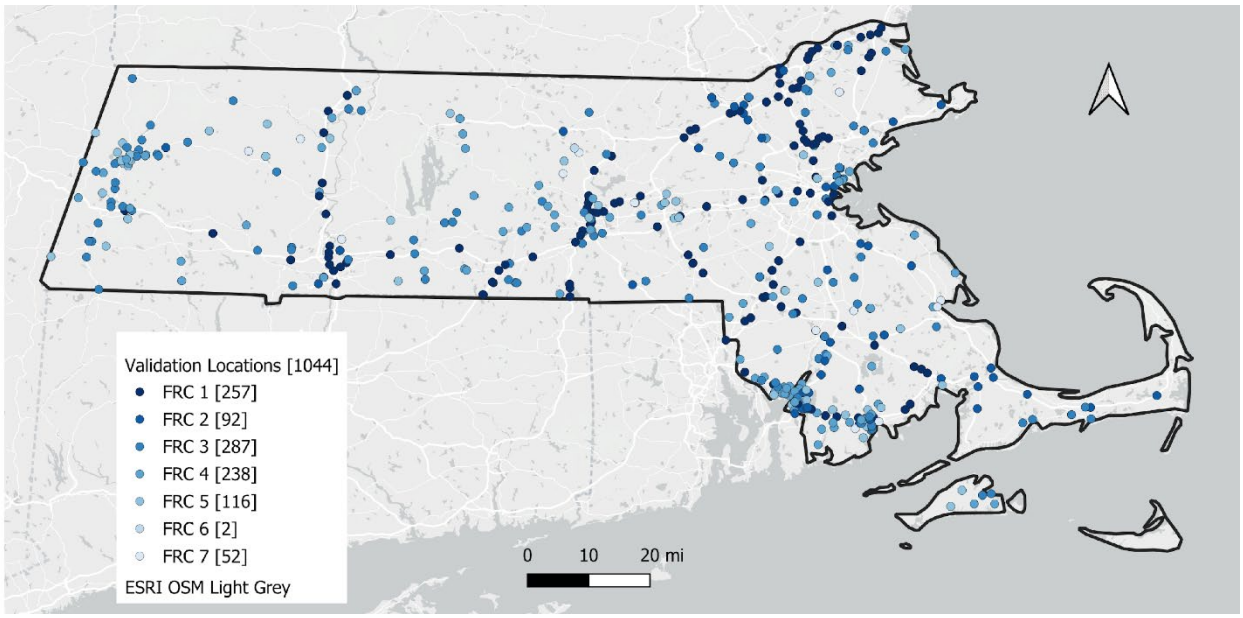


Figure E-1: Massachusetts Evaluation locations

Vendors submitted data using the standard CATTWorks Geo-referencing Protocol (CWGP). MassDOT provided 12 months of historical count data (January–December 2024) from the continuous count sites for model calibration. Unlike previous validation exercises, the MassDOT network included a diverse mix of counter types and deployment contexts: traditional inductive loop counters on public roadways, inductive loops at automated electronic tolling (AET) facilities, and video cameras typically deployed at signalized intersections.

Reference Data:

- **Data from Continuous Count Sites (CCS)**

  Data Availability Threshold: To ensure reference values were truly representative, a strict 75% data availability rule was applied to CCS locations, consistent with the methodology used in Georgia and North Carolina. After applying this filter for AADT, 486 directional CCS locations were retained for analysis, while 131 were omitted due to insufficient data.

- **AADT from Factored Short-Term Counts**

  MassDOT provided directional AADT estimates based on factored, short-term counts. The TETC validation team performed no further processing or filtering on this data set.

Figure E-2 shows the distribution of directional locations by (a) volume and (b) road class, with values reported separately for Continuous (blue) and Short-Term (orange) sites. In Massachusetts there are no sites classified as FRC 6 (minor collectors). All other functional classes are represented.

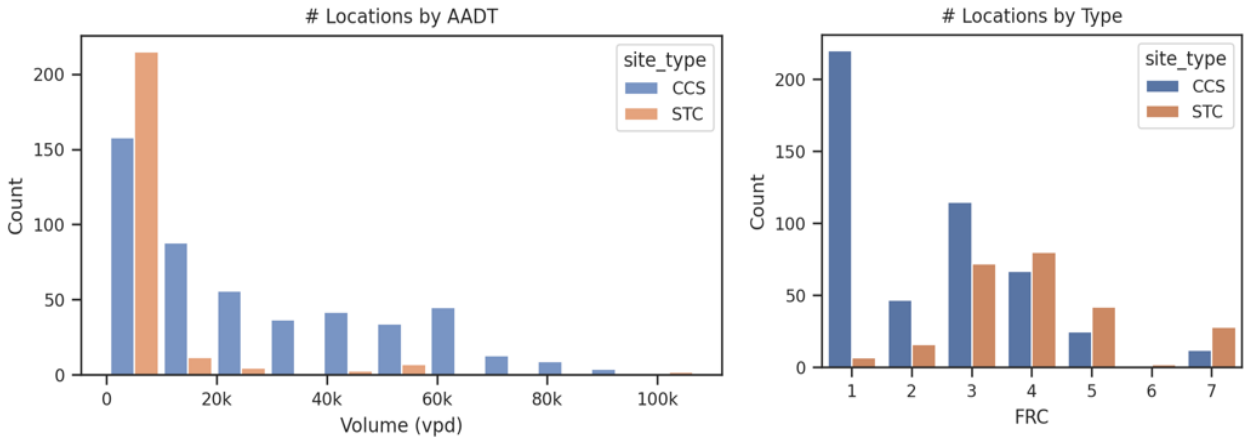


Figure E-2: Distribution of AADT volumes across volume and functional class in Massachusetts Volume Validation.

Unlike previous validation exercises, MassDOT's continuous count sites encompassed multiple types of counters and counter locations, broadly categorized as follows: (1) traditional inductive loop counters on public roadways, (2) inductive loops at automated electronic tolling (AET) facilities, and (3) video cameras, typically deployed at signalized intersections. The inductive-loop counters at AET facilities are typically located near merge/diverge points in proximity to automated toll collection gantries. Figure E-3 summarizes the distribution of the three types of counter locations, in addition to short-term count sites.

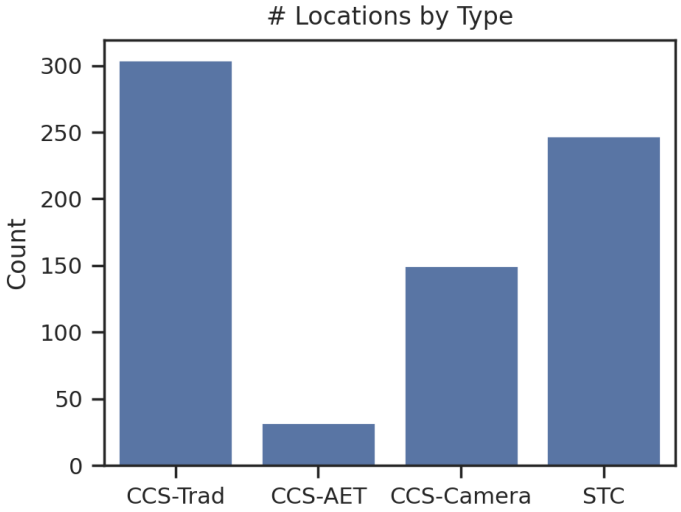


Figure E-3: Distribution of count locations by counter classification in Massachusetts Volume Validation.

Vendor Data:

The validation team requested data from the vendors in five time aggregations: AADT, ADT, AHDT, Daily, and Hourly volumes. Table E-1 summarizes the vendor submissions for each volume product and highlights that all vendors submitted AADT estimates.

Table E-1: Vendor Submission Summary in Massachusetts Volume Validation

Vendor	AADT	ADT	AHDT	Daily Volumes	Hourly Volumes
A	✓	✓	✓	✓	✓
B	✓	✓	✓	✓	✓
C	✓	✓	✓	✓	✓
D	✓	✓	✓	✓	✓
E	✓				

Changes to FHWA Method Test Implementation

The Massachusetts volume validation applied a slightly modified FHWA test methodology compared to previous validations. The key adjustments include:

- **Focus on Bi-Directional Volumes:** The analysis exclusively used total (bi-directional) volumes rather than directional volumes. This aligns with USDOT research and FHWA pooled fund study recommendations, which base accuracy and precision limits on total volume.
- **Updated Accuracy Limits:** Benchmark limits were determined using FHWA-provided equations based on sample size. While slightly more stringent, this update better accommodates intermediate sample sizes with minimal overall impact on the results.
- **Updated Precision Limits:** Previously, precision tests were applied to individual volume bins, resulting in statistically small sample sizes where a single outlier could trigger a failure. In this validation, precision limits were applied to the entire pooled dataset. This maximized the sample size, reduced sensitivity to isolated outliers, and provided a fairer evaluation of vendor performance. Minor adjustments were also made to the calculation of acceptable failures (detailed in Appendix B).

(Note: A complete summary of the FHWA methods implemented during the MassDOT validation is available in Appendix B.)

Summary of AADT Volume Validation Results in Massachusetts

Preliminary evaluation results reflect vendor performance prior to identifying several non-vendor sources of error. Figure E-4 displays initial scatter plots for all five participating vendors in a format consistent with prior validation studies in Georgia and North Carolina. The x-axis spans 0 to 50,000 vehicles per day (vpd), capturing approximately 85% of all evaluation locations. Data points are color-coded to distinguish Continuous Count Stations (CCS, blue) from Short-Term Count sites (STC, orange).

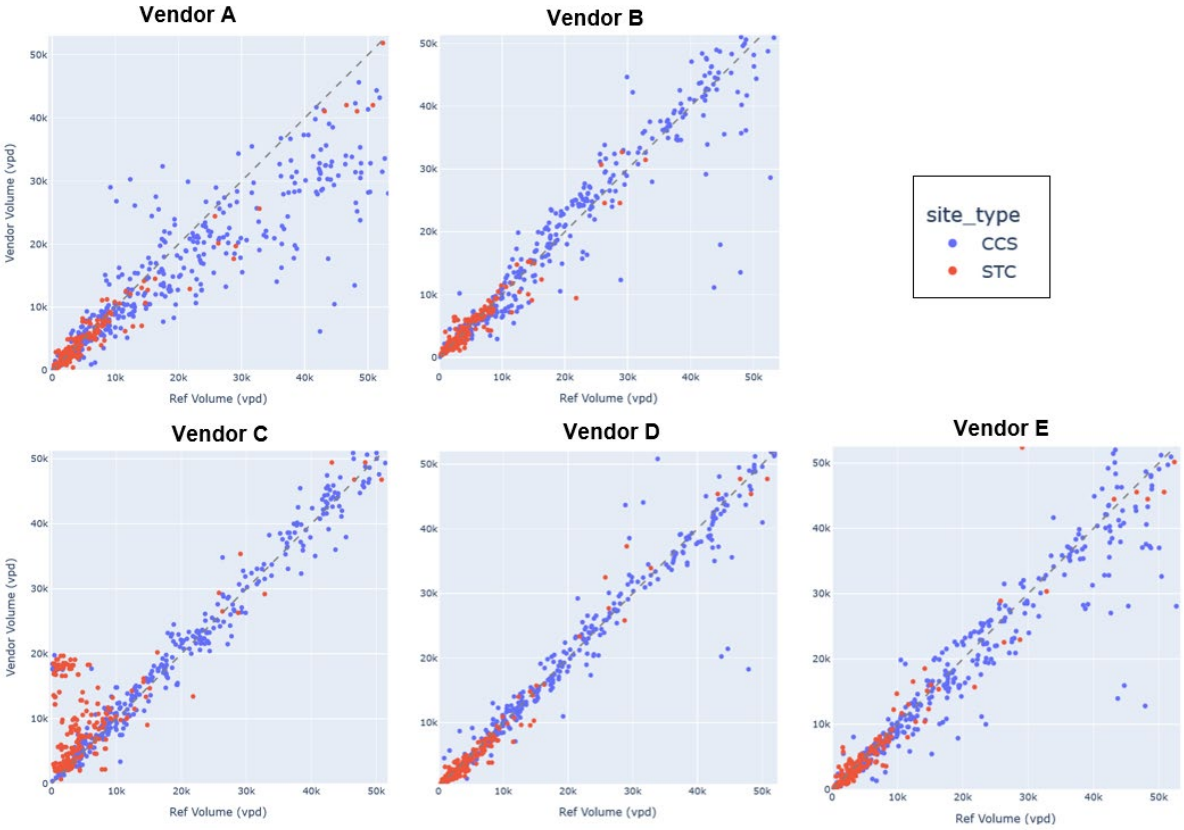


Figure E-4: AADT scatterplot for vendors' initial submissions

The primary finding from these initial submissions was the unexpected frequency and magnitude of severe outliers across all five vendors:

- **Systematic bias:** Vendor A exhibited consistent underestimation across volume ranges, while Vendor C showed marked overestimation on lower-volume facilities.
- **Sporadic outliers:** All vendors exhibited individual data points deviating from agency reference counts by several thousand vehicles per day.

Because vendor performance had demonstrated consistency in earlier validation efforts, the validation team reasoned that broad industry modeling capability had not deteriorated. Consequently, the team launched a dedicated outlier investigation to isolate the root causes of these discrepancies.

Outlier Investigation

The validation team audited the ten most significant outliers per vendor dataset. This involved reviewing counter deployment types, verifying exact georeferencing, assessing cross-vendor consistency at identical stations, and comparing directional flow balances. Because annual average traffic on opposing directions is generally balanced, pronounced directional asymmetry often helped pinpoint data anomalies.

The root causes of major outliers fell into four primary categories:

1. **Miscoded Site Locations in Agency Records:** Manual review revealed that MassDOT's database contained erroneous coordinates for 39 count sites (representing 52 directional locations). MassDOT independently identified and resolved these database errors after validation activities had commenced. Because vendors received flawed initial coordinates, their estimates reflected roadway segments that differed substantially from the actual physical counter locations. All 52 affected locations were inductive loop CCS sites.
2. **Location Referencing and Conflation Ambiguities:** Site locations were distributed to vendors using the standard CATTWorks Geo-referencing Protocol (CWGP), which specifies latitude, longitude, bearing/heading, road name, and functional class. In complex geometric configurations—such as freeway interchanges, ramp terminals, and closely spaced intersections—these attributes proved insufficient to define the intended count target unambiguously.
    - *Example (I-90 Toll Gantry):* As illustrated in Figure E-5, coordinates alone left ambiguity as to whether Westbound estimates should encompass through lanes only or include adjacent exit ramp lanes past the gore point. MassDOT loop assignment schematics confirmed that reference counts included both through and ramp traffic, whereas most vendors modeled through traffic only.
3. **Reference Counter Occlusion and Functional Limitations:** Several recurring anomalies coincided with intersection video detection cameras. Although deployed primarily for traffic signal actuation and turning-movement detection, aggregated counts from these cameras had been incorporated into the agency count database.
    - *Example (Signalized Intersection Slip Lane):* As illustrated in Figure E-6, reference data for the northern leg showed southbound AADT at three times the volume of the northbound leg. An audit revealed that a channelized right-turn slip lane fell outside the camera's field of view, causing a significant portion of entering volume to be systematically omitted from reference counts.
4. **Vendor Estimation Errors:** A subset of errors originated directly from vendor modeling workflows. These errors were typically unique to a single vendor at a specific station, with other vendors aligning closely with reference data.

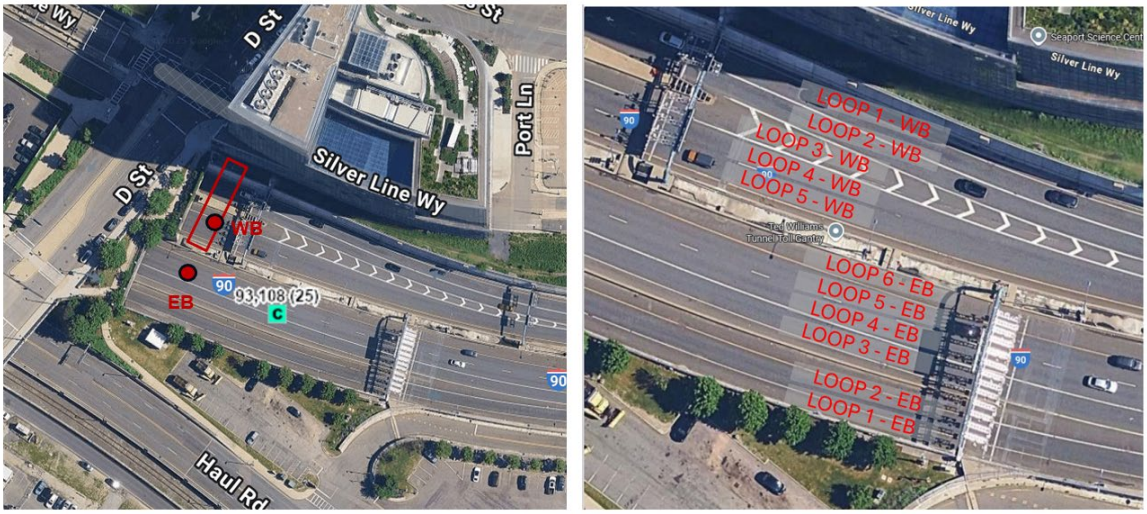


Figure E-5: Example of georeferencing and lane-configuration ambiguity at an I-90 toll counter

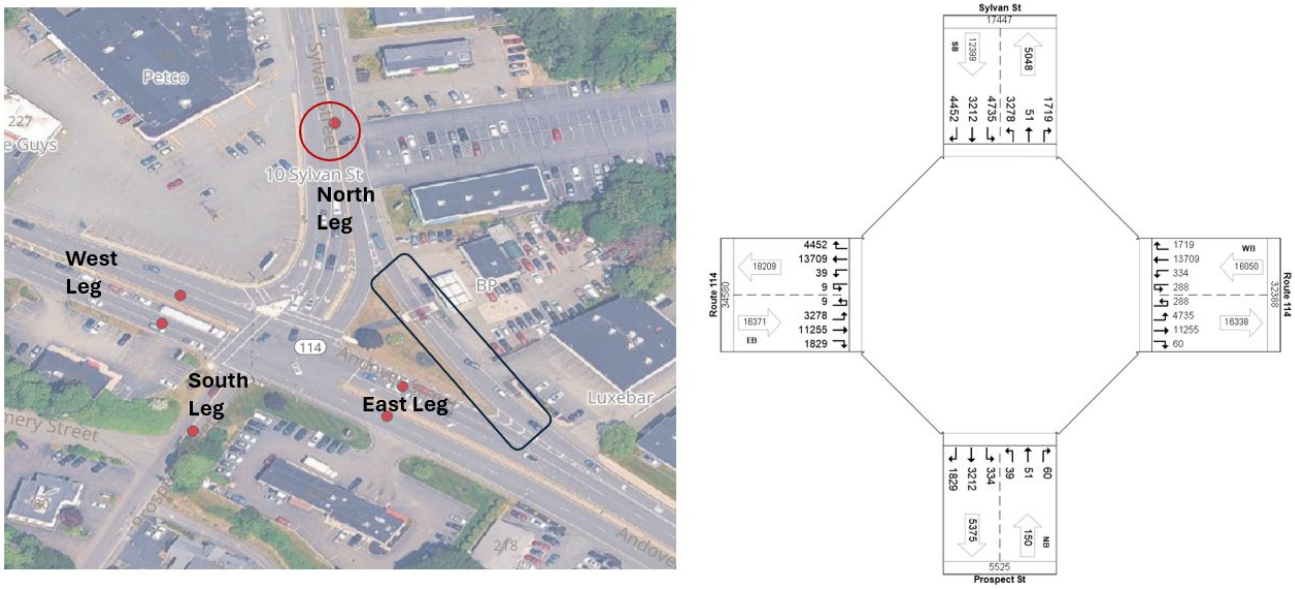


Figure E-6: Example of reference data undercounting caused by camera occlusion at a signalized intersection slip lane

Performance Stratified by Counter Technology

To understand performance variations across sensing environments, the validation team analyzed results across three distinct CCS technologies: Traditional Inductive Loops, Automated Electronic Tolling (AET) loops, and Intersection Video Cameras (Figures E-7 through E-9). To isolate counter effects independent from systematic vendor bias, this comparison focused on Vendors B, D, and E only (vendors without systematic bias errors). Each scatter plot is color coded to flag 'suspect locations' in red, referring to outliers caused by error other than vendor error and thus unrelated to vendor performance. The validation team later omitted suspect locations from performance results.

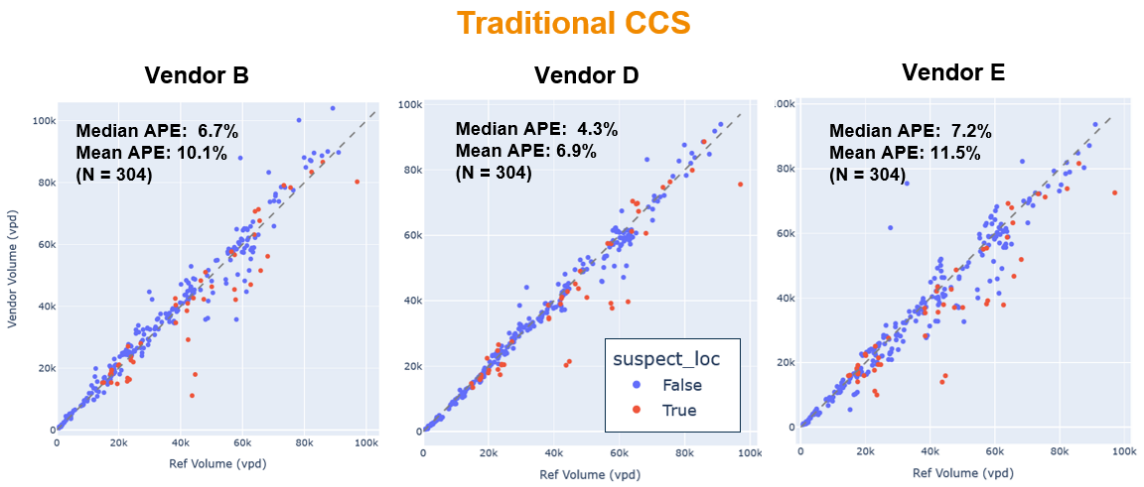


Figure E-7: Performance at Traditional CCS locations (Vendors B, D, and E)

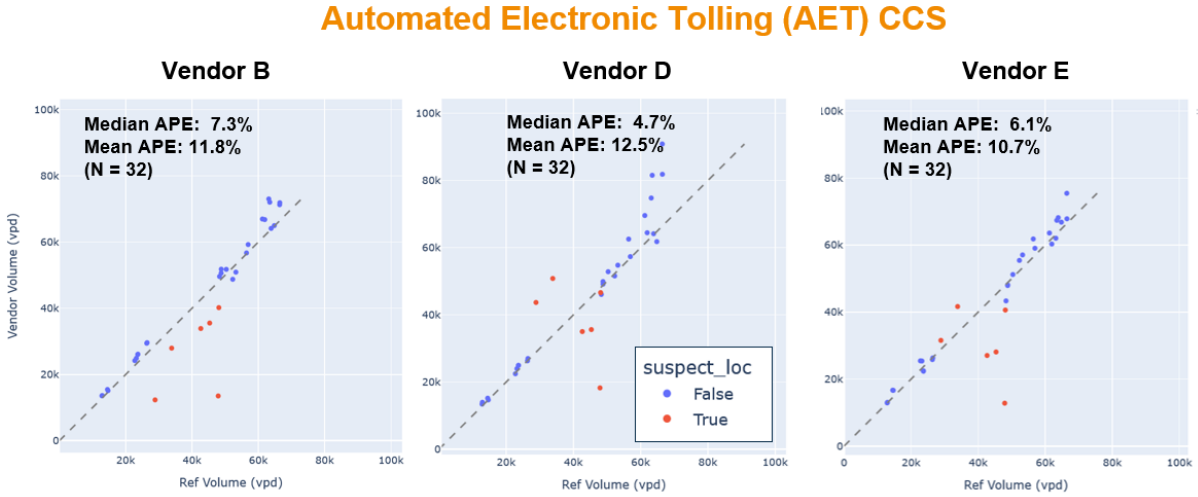


Figure E-8: Performance at AET CCS locations ((Vendors B, D, and E)

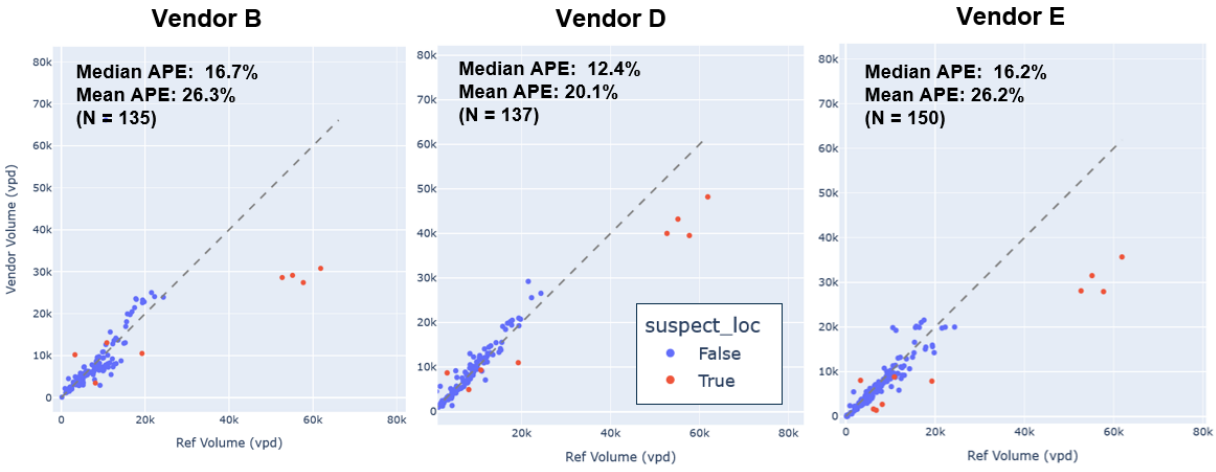


Figure E-9: Scatter plots showing Camera CCS locations (Vendors B, D, and E)

Impact of Network and Site Complexity

The high error rates observed at camera locations prompted an assessment of spatial roadway complexity. The validation team classified each CCS station into **Low**, **Medium**, or **High** complexity based on the count of OpenStreetMap (OSM) roadway segments within a 30-meter (~100-foot) radius (Figure E-10):

- **Low Complexity (1–2 candidate segments):** Mainline uninterrupted highway sections with minimal conflation ambiguity.
- **Medium Complexity (3–5 candidate segments):** Divided arterials, grade-separated ramps, or simple surface intersections.
- **High Complexity (6+ candidate segments):** Dense urban intersections, multi-leg junctions, slip ramps, and frontage road configurations (Figure E-11).

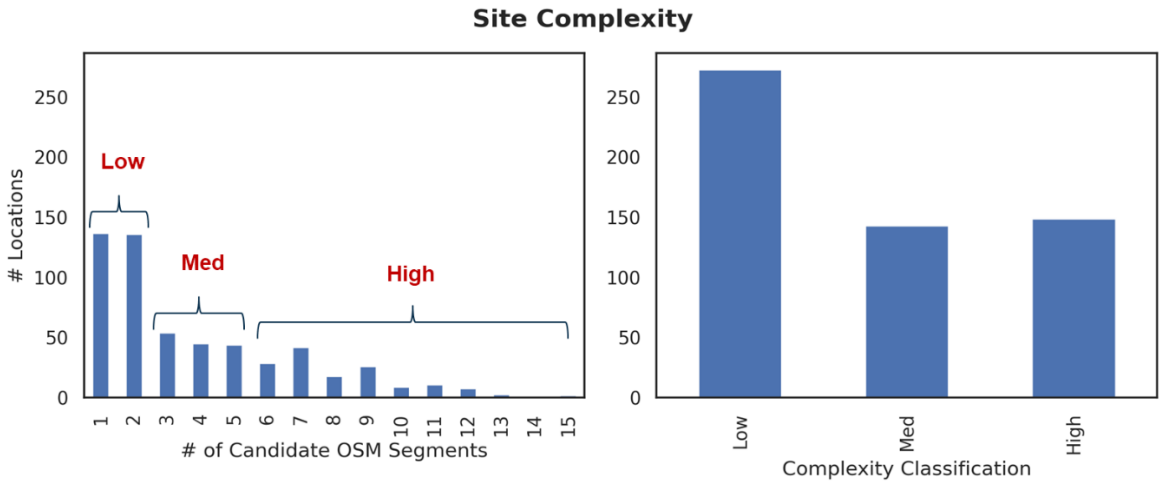
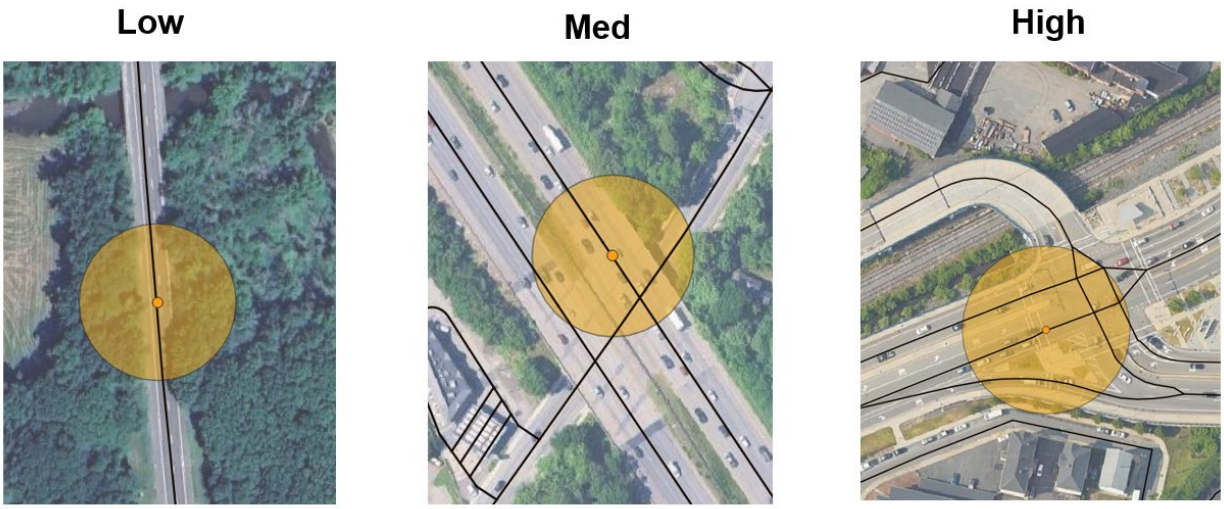


Figure E-10: Classification thresholds and distribution of candidate OSM segments.



Loc P0007 – 1 candidate segment

Loc P0213 – 5 candidate segments (underpass)

Loc P0563 – 15 candidate segments

Figure E-11: Aerial examples of Low, Medium, and High complexity count environments

As detailed in Figure E-12, counter technology strongly correlates with geometric complexity:

- **Traditional loops** reside almost exclusively on low-complexity mainline corridors.
- **AET loops** are split between low- and medium-complexity segments near toll plazas.
- **Video cameras** are concentrated in high-complexity intersection environments.

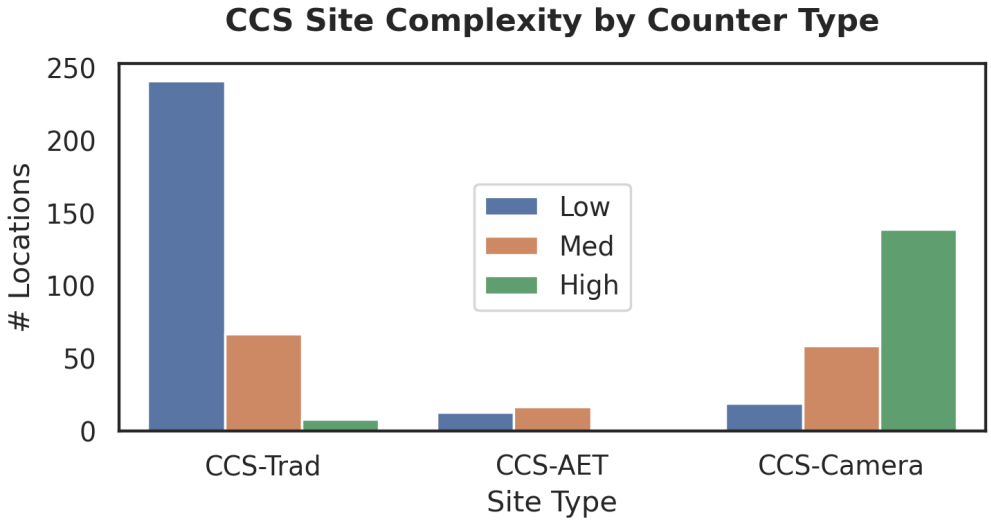


Figure E-12: Distribution of site complexity classifications across CCS counter types

Consequently, higher error rates at camera stations reflect both sensor-specific limitations (occlusion, signal-actuation tuning) and the heightened geometric difficulty of conflating probe data onto multi-lane turning facilities.

Outlier Mitigation and Final Resubmissions

Based on the diagnostic investigation, the validation team designated **72 directional locations** as suspect data points and excluded them from further evaluation. These comprised miscoded database coordinates, verified camera occlusion errors, and unresolved conflation ambiguities.

Because suspect reference data had originally been supplied in the baseline training and calibration package, vendors were given the option to recalibrate their models and submit revised estimates. Four of the five vendors provided updated datasets; for the remaining vendor, the 72 suspect stations were removed from their initial submission.

Due to lingering data-quality concerns surrounding intersection cameras, the final performance results are reported under two distinct frameworks:

1. **All CCS Locations (inclusive of video cameras)**
2. **Non-Camera CCS Locations (Traditional loops and AET toll stations only)**

The non-camera evaluation provides direct parity with the baseline count conditions evaluated in Georgia and North Carolina.

Final AADT Volume Validation Performance

Figure E-13 presents final scatter plots across all five vendors for volumes up to 50,000 vpd. Compared to initial submissions, data points align much closer to the line of equality ($y = x$), with a sharp reduction in outlier frequency. Vendor C eliminated its low-volume bias, while Vendor A reduced its systematic underestimation.

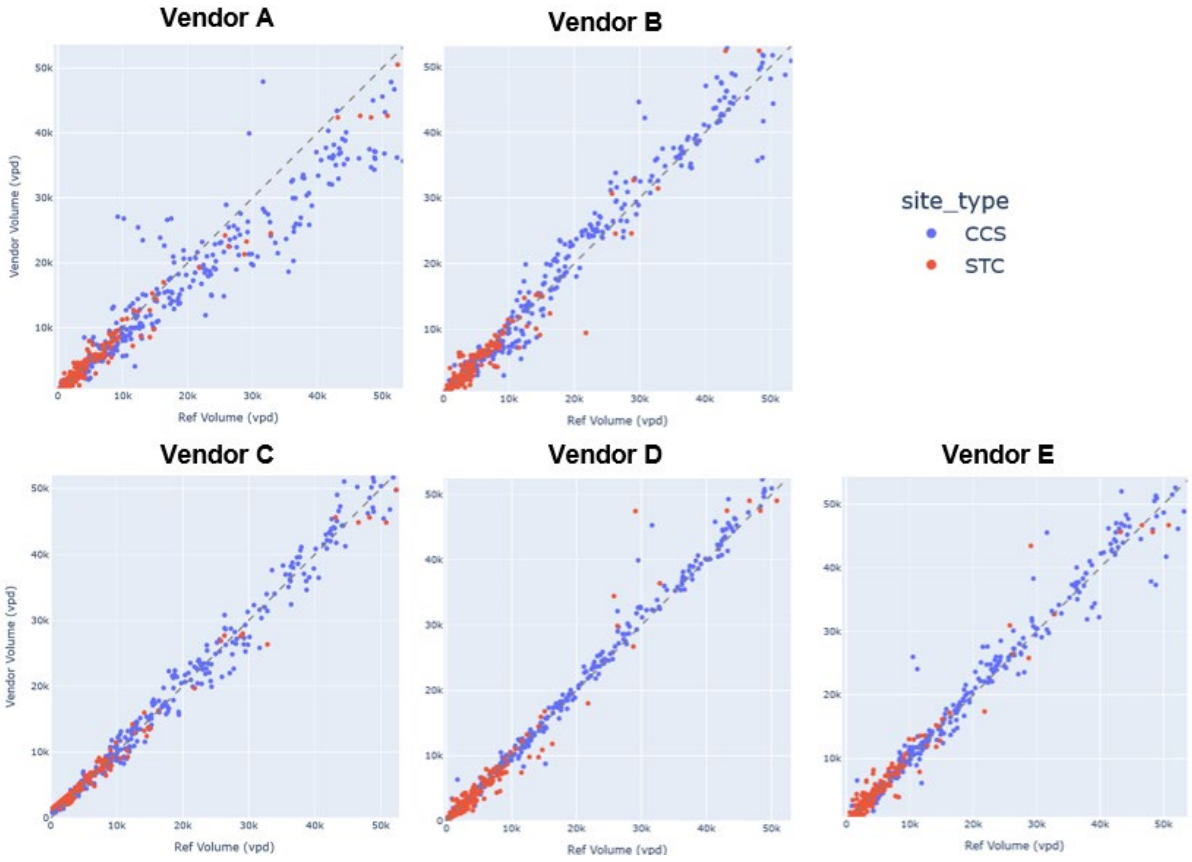


Figure E-13: Final AADT scatter plots across all vendors (0–50k vpd)

Figures 3.4.4.2 through 3.4.4.6 show the individual scatterplots for each vendor, with separate plots and summary metrics (median and mean APE) reported for CCS and STC. Accompanying boxplots summarize the distribution of errors in units of volume (vehicles).

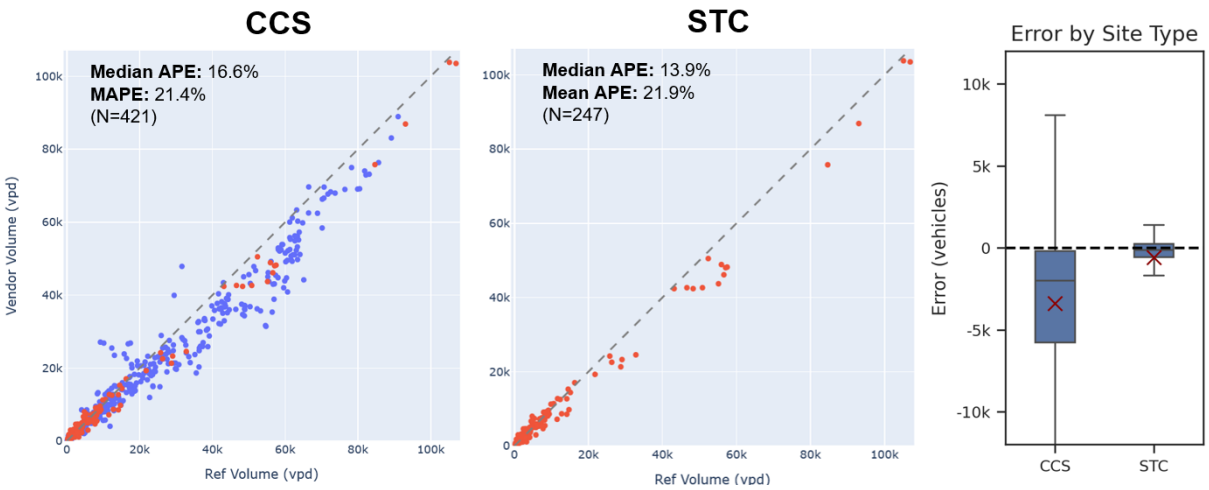


Figure E-14: Final individual vendor scatter plots and error distributions for Vendor A

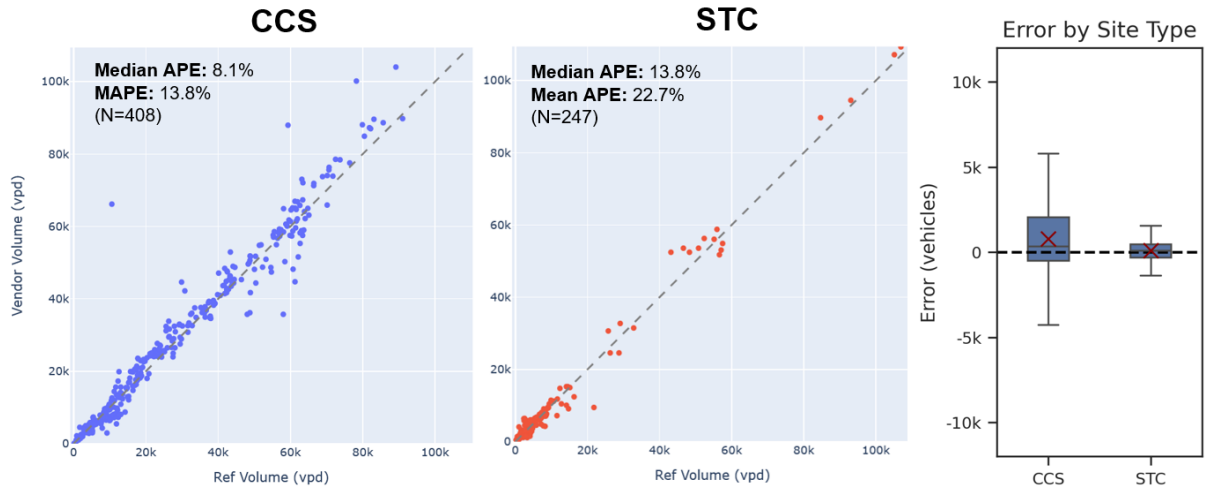


Figure E-15: Final individual vendor scatter plots and error distributions for Vendor B

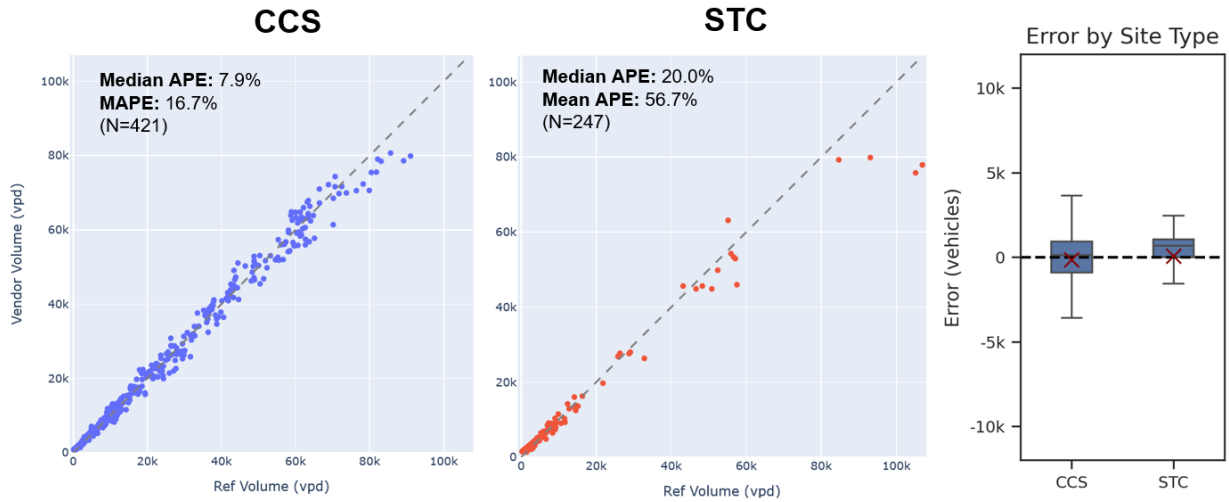


Figure E-16: Final individual vendor scatter plots and error distributions for Vendor C

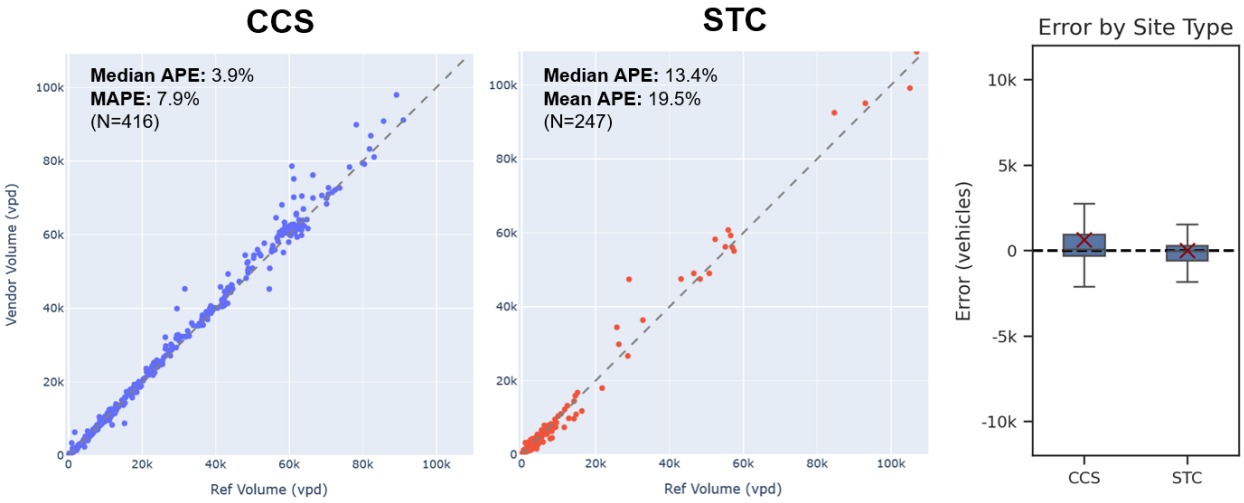


Figure E-17: Final individual vendor scatter plots and error distributions for Vendor D

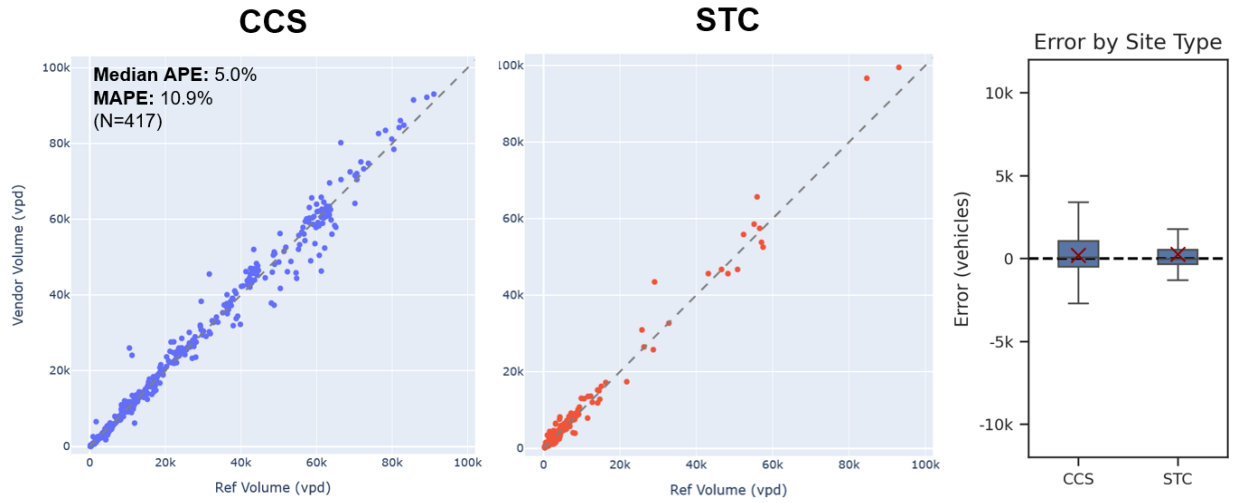


Figure E-18: Final individual vendor scatter plots and error distributions for Vendor E

Performance by Volume Bin

Figures 3.4.4.7 and 3.4.4.8 show the error distribution for each vendor aggregated by volume bins. Figure 3.4.4.7 plots results for all CCS locations, and Figure 3.4.4.8 focuses on non-camera sites (Traditional and AET locations). The validation team chose the volume bins based on FHWA's volume bin definitions. The validation team divided FHWA bi-directional volume bins in half to create a directional version of bin thresholds.

- **Vendor A** consistently underestimates volume across all volume ranges, independent of whether camera locations are included (i.e., Fig 3.4.4.7 vs 3.4.4.8). The level of underestimation improved from the preliminary results, but still perceivable.
- **Vendor B**'s errors are typically within +/- 15%, but the sign of the error switches across volume bins when all count locations are included (Fig 3.4.4.7). When

camera locations are excluded (Fig 3.4.4.8), the magnitude of Vendor B's error decreases slightly and a slight positive bias emerges.

- **Vendor C** shows strong agreement with reference data across mid-to-high volume ranges (especially above 5k vpd) but has high error in the lowest volume ranges when cameras are included (Fig 3.4.4.7). If camera count locations are excluded, the performance improves and becomes more consistent.
- **Vendor D** shows high accuracy with minimal bias across all volume bins in Fig 3.4.4.7, and the magnitude of the error drops when excluding cameras count locations (Fig 3.4.4.8). Vendor D AADT estimates consistently matches the reference AADT values across all volume ranges.
- **Vendor E** shows strong agreement across mid-to-high volume ranges, with a slight tendency to underestimate at lower volumes when all locations are considered (Fig 3.4.4.7). When camera locations are excluded, performance improves.

Performance by Road Class

Figures 3.4.4.9 and 3.4.4.10 show the distribution of percent errors aggregated by road class for each vendor. Figure 3.4.4.9 shows results for all CCS locations, while Figure 3.4.4.10 focuses on non-camera sites (Traditional and AET). The results presented by road class correlates with volume bins, as higher road classes have higher volume. In Figure 3.4.4.9, FRC 3 (Other Principal Arterials) stands out as having notable error/bias across all vendors when camara locations are included. This bias disappears when camera locations are omitted. Road class 7 is exclusively covered by camera count locations. When cameras are excluded, no reference data is available for comparison in that class.

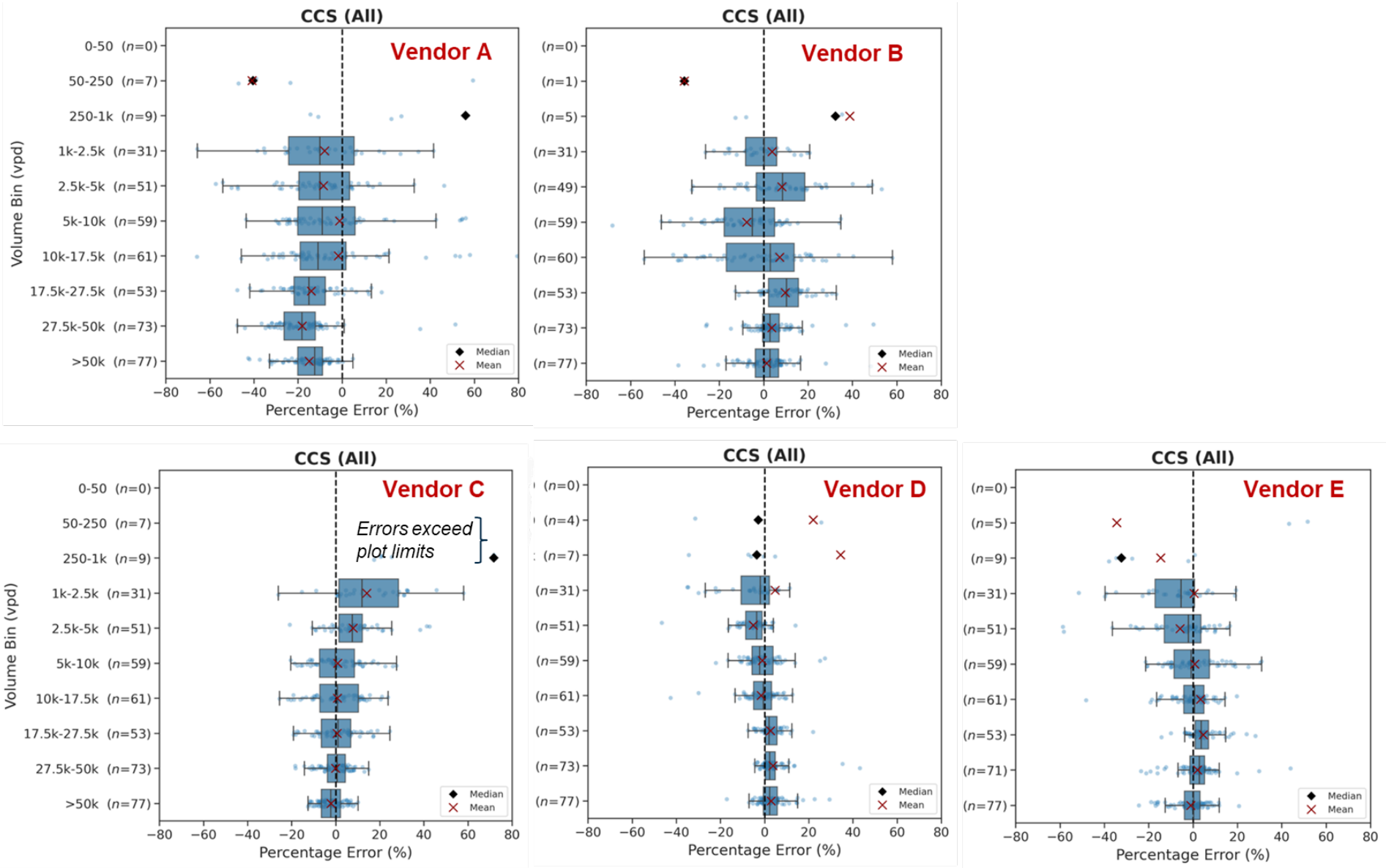


Figure 3.4.4.7: Error Distribution plots by volume bin - All CCS locations in Massachusetts

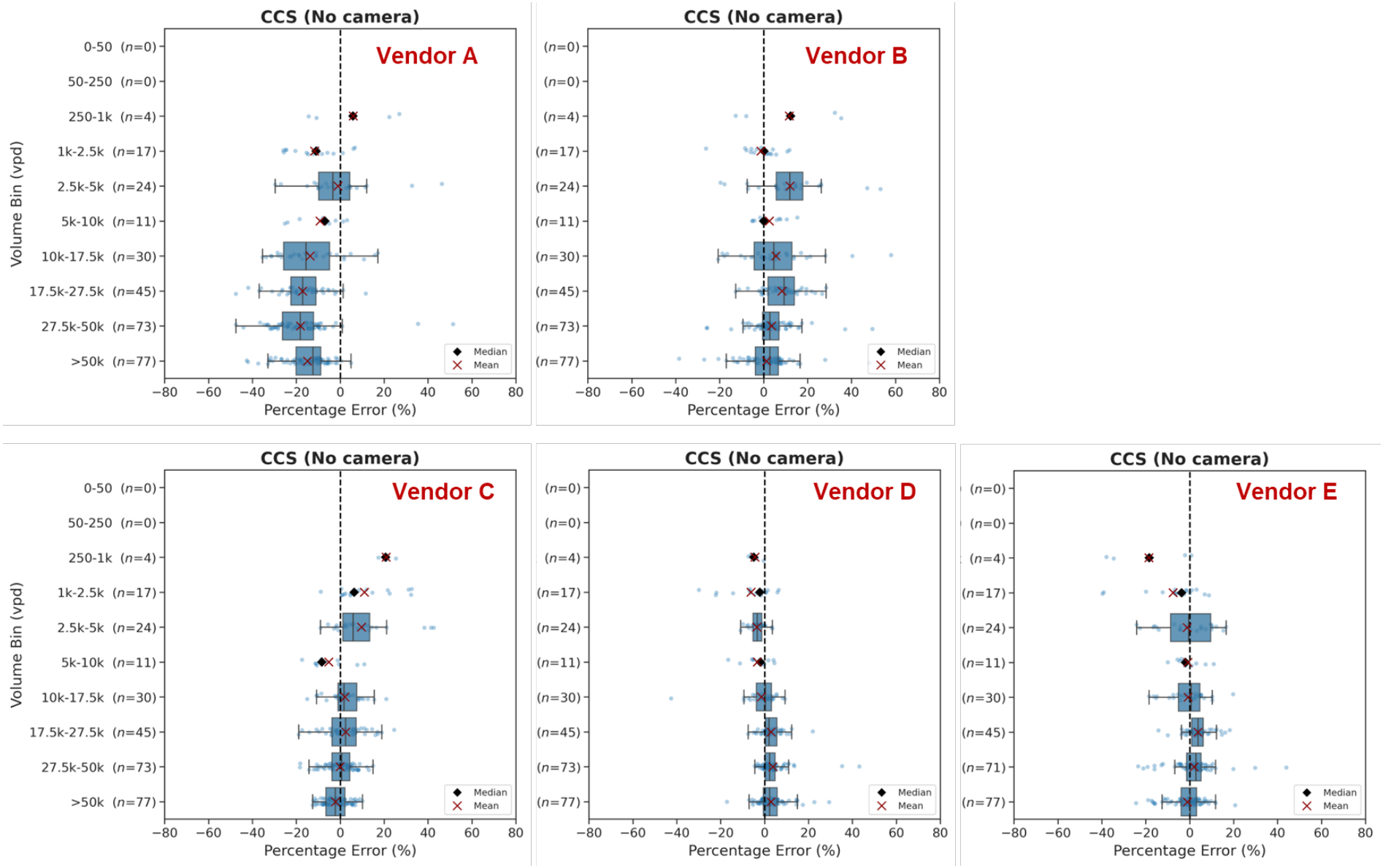


Figure 3.4.4.8: Error Distribution plots by volume bin – omitting Camera locations in Massachusetts

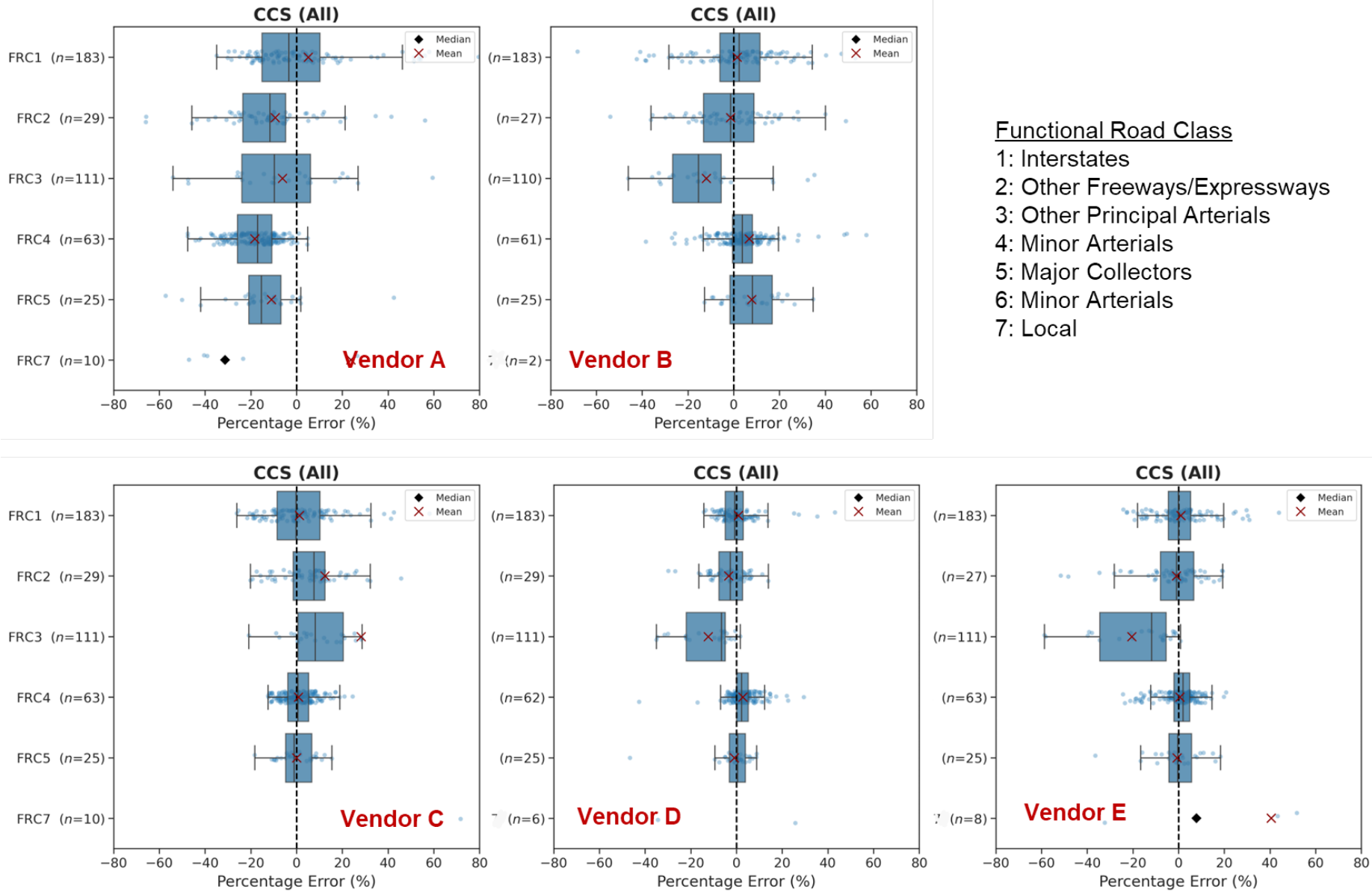


Figure 3.4.4.9: Error Distribution plots by functional roadway classification – All CCS locations in Massachusetts

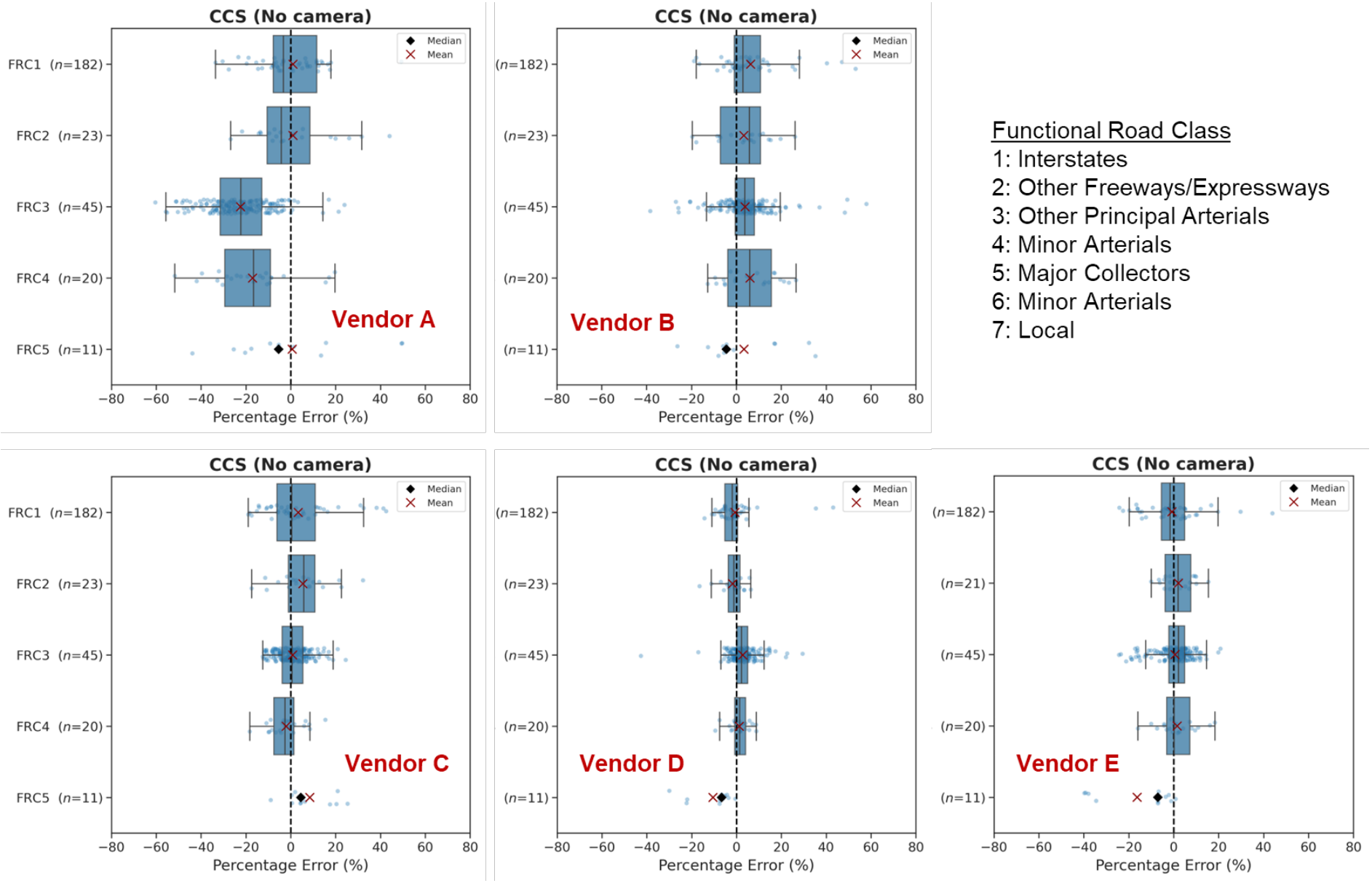


Figure 3.4.4.10: Error Distribution plots by functional roadway classification – omitting cameras location in Massachusetts

3.4.5 FHWA Test Methodology Results in Massachusetts

Table 3.4.5.1 summarizes the results using *bi-directional volumes for all CCS locations*, while Table 3.4.5.2 summarizes the results for *bi-directional volumes for non-camera CCS locations.* In both cases, the validation team conducts 7 statistical tests: two accuracy metrics applied to three volume bins (low, medium, high), and one overall precision test applied across all volume bins (see discussion in Appendix B). Note that the validation team omitted the "very low" volume bin due to insufficient observations in this volume range. Table 3.4.5.1 and 3.4.5.2 also displays benchmark limits in the table to provide context for how close each vendor's results were to the pass/fail threshold. The validation team provides a tally of the total number of tests passed to the right of the table.

All CCS Locations

Tables 3.4.5.1 depicts the result based on all CCS sites (including Camera sites). This report presents the results for all three volume bins, but the reader should note the 'Low' volume bin has only 20 observations, while a minimum of 25 observations is recommended. Overall, three vendors passed at least 5 of the 7 tests; Vendors D and E passed 6 of 7 tests (5/6 accuracy, 1/1 precision), with Vendor C passing 5 of 7 (4/6 accuracy, 1/1 precision). All failures from these three vendors occurred in the lowest volume bin where most camera count sites exist. Vendor B passed 4 of 7 tests (4/6 accuracy, 0/1 precision), with several outliers causing a failure on the precision test (limit = 8, observed outliers = 16). Vendor A did not pass any tests due to the issue with systematic underestimation.

Traditional CCS and AET Locations Only

Table 3.4.5.2 depicts the results if the validation team removes camera locations from the data set. The validation team deems these results most comparable to the GA and NC validation. This report presents results for all three volume bins, although the 'Low' volume bin has only 11 observations, while a minimum of 25 observations is recommended. With camera CCS locations removed, **Vendors D and E passed all 7 tests** (6/6 accuracy, 1/1 precision), while **Vendor C passed 6 of 7 tests** (5/6 accuracy, 1/1 precision) and Vendor B passed 5/7 (5/6 accuracy, 0/1 precision). Vendor A passed 1/7 (1/6 accuracy, 0/1 precision) again due to the systematic bias issues. The validation team deems these results, with Camera locations removed, to be most comparable to the methodology used in previous validations from GA and NC, and representative of industry capacity to provide accurate AADT volume estimates as of mid 2026.

Table 3.4.5.1: FHWA method results for bi-directional AADT in - MassDOT all CCS Locations

			Accuracy Test #1			Accuracy Test # 2			Precision Test			
Vendor	Volume Category	# Obs	Absolute Median TCE	Limit	Result	MAPE	Limit	Result	# Failures	Limit	Result	Tests Passed
A	Low	20*	9.1	8.8	Failed	33.0	18.1	Failed	4			
	Med	121	12.3	3.5	Failed	19.6	11.7	Failed	15	8	Failed	0/7
	High	78	14.9	5.9	Failed	17.3	10.3	Failed	22			
B	Low	18*	0.8	8.8	Passed	19.7	18.1	Failed	1			
	Med	120	3.2	3.5	Passed	18.8	11.7	Failed	11	8	Failed	4/7
	High	78	2.4	5.9	Passed	7.4	10.3	Passed	4			
C	Low	20*	16.1	8.8	Failed	32.4	18.1	Failed	3			
	Med	121	2.8	3.5	Passed	8.3	11.7	Passed	1	8	Passed	5/7
	High	78	1.3	5.9	Passed	4.8	10.3	Passed	0			
D	Low	18*	3.0	8.8	Passed	25.2	18.1	Failed	1			
	Med	121	0.7	3.5	Passed	4.6	11.7	Passed	0	8	Passed	6/7
	High	78	2.6	5.9	Passed	4.4	10.3	Passed	2			
E	Low	20*	5.8	8.8	Passed	29.7	18.1	Failed	2			
	Med	121	0.9	3.5	Passed	8.5	11.7	Passed	2	8	Passed	6/7
	High	77	1.6	6.0	Passed	5.5	10.3	Passed	2			

Results are reported for 'Low' category although sample size is less than recommended 25

Table 3.4.5.2: FHWA method results for bi-directional AADT - MassDOT CCS, cameras omitted

			Accuracy Test #1			Accuracy Test # 2			Precision Test			
Vendor	Volume Category	# Obs	Absolute Median TCE	Limit	Result	MAPE	Limit	Result	# Failures	Limit	Result	Tests Passed
A	Low	11*	12.4	8.8	Failed	14.6	18.1	Passed	0			
	Med	61	13.1	4.8	Failed	15.8	12.4	Failed	4	6	Failed	1/7
	High	78	14.9	5.9	Failed	17.3	10.3	Failed	22			
B	Low	11*	1.5	8.8	Passed	10.0	18.1	Passed	0			
	Med	61	7.4	4.8	Failed	11.5	12.4	Passed	3	6	Failed	5/7
	High	78	2.4	5.9	Passed	7.4	10.3	Passed	4			
C	Low	11*	13.3	8.8	Failed	13.6	18.1	Passed	0			
	Med	61	2.4	4.8	Passed	7.1	12.4	Passed	1	6	Passed	6/7
	High	78	1.3	5.9	Passed	4.8	10.3	Passed	0			
D	Low	11*	4.3	8.8	Passed	8.1	18.1	Passed	0			
	Med	61	0.1	4.8	Passed	3.8	12.4	Passed	0	6	Passed	7/7
	High	78	2.6	5.9	Passed	4.4	10.3	Passed	2			
E	Low	11*	5.1	8.8	Passed	11.0	18.1	Passed	0			
	Med	61	1.6	4.8	Passed	6.2	12.4	Passed	0	6	Passed	7/7
	High	77	1.6	6.0	Passed	5.5	10.3	Passed	2			

Results are reported for 'Low' category although sample size is less than recommended 25

For direct comparison with prior results in Georgia, the Georgia vendor performance data was recompiled with using the revisions to the FHWA test implemented with the MassDOT validation. Table 3.4.5.3 summarizes the bi-directional results based on CCS locations in Georgia for 2023 AADTs. Note that there were no camera counters in Georgia, so these results are most comparable to Table 3.4.5.2 (non-camera CCS locations in Massachusetts). These results showing improvement from Georgia to MassDOT over the course of approximately one year reinforce the conclusion that industry continues to mature and provide volume data products of increasing and value.

Table 3.4.5.3 – FHWA method results for bi-directional AADT - Georgia

			Accuracy Test #1			Accuracy Test # 2			Precision Test			
Scenario	Volume Category	# Obs	Absolute Median TCE	Limit	Result	MAPE	Limit	Result	# Failures	Limit	Result	Tests Passed
Vendor A	Low	27	0.5	8.5	Passed	8.9	18.0	Passed	0	6	Failed	5/7
	Medium	93	4.2	4.0	Failed	10.7	12.0	Passed	5			
	High	20*	6.1	9.4	Passed	10.4	12.9	Passed	2			
Vendor B	Low	27	8.1	8.5	Passed	16.8	18.0	Passed	1	6	Failed	4/7
	Medium	92	3.1	4.0	Passed	14.4	12.0	Failed	10			
	High	20*	3.1	9.4	Passed	19.8	12.9	Failed	5			
Vendor C	Low	27	1.4	8.5	Passed	149.5	18.0	Failed	5	6	Failed	4/7
	Medium	93	1.7	4.0	Passed	21.0	12.0	Failed	6			
	High	20*	2.0	9.4	Passed	2.2	12.9	Passed	0			
Vendor D	Low	27	3.0	8.5	Passed	10.1	18.0	Passed	1	6	Passed	6/7
	Medium	93	5.5	4.0	Failed	7.7	12.0	Passed	2			
	High	20*	2.3	9.4	Passed	4.7	12.9	Passed	0			
Vendor E	Low	27	1.3	8.5	Passed	44.7	18.0	Failed	6	6	Failed	4/7
	Medium	93	0.3	4.0	Passed	12.3	12.0	Failed	8			
	High	20*	2.5	9.4	Passed	7.0	12.9	Passed	1			

Results are reported for 'High' category although sample size is less than recommended 25

3.4.6 MassDOT Volume Validation Lessons Learned and Recommendations

The validation experience in Massachusetts revealed several sources of error beyond those related to vendor modeling for volume estimation. It is critical to understand the different sources of error, such as:

- Ambiguity in georeferencing/conflation that correlates with network complexity
- Location error within an agency database
- Varying accuracy of different types of count technologies.

All of these are unrelated to quality of a vendors' volume estimation processes. This is a critical finding not only for validating volume estimates but also for application and integration of such data. The results from the MassDOT validation resulted in a number of recommendations for future validation activities. These recommendations included:

- Validation data collection (including calibration, reference, and vendor data) should continue to be collected/distributed as directional volume counts. Even though the FHWA method is based on bi-directional counts, directional counts at

the same location assists in identifying and understanding anomalies in both agency and vendor data.

- Cross-checking results across multiple vendors is highly valuable. If the vendor data shows agreement internally across multiple vendors, but disagrees with agency reference data, it is probable that the source of the anomalous data lies with the agency or ambiguities in geo-referencing.
- If validation results indicate the prevalence of outliers, operators should systematically evaluate the data to assess if additional error sources (other than vendor error) are present in the analysis. Investigating the top 10 anomalies or outliers is a starting rule-of-thumb.
- Assess performance across different types of count technologies and contexts. Massachusetts introduced toll road counts and camera counts, both of which had performance peculiarities not previously encountered. Observing results for rural/urban, functional road classification, or level of complexity provides, at a minimum, better understanding of performance, and may assist in isolating various data errors.
- FHWA Guidelines should be applied using bi-directional volume counts. Unless sample sizes are ample, apply FHWA precision metric to the entire data set (not specific volume bins).

The TETC will publish and make accessible the full MassDOT volume data validation on the TETC web site later in 2026. The report will contain additional details, charts, graphs, etc. The authors have summarized the MassDOT report here to the extent that it contributes to the recommendations for best practices and guidelines for using probe-based volume estimates for HPMS reporting.

APPENDIX F: Minnesota DOT Case Study (2021)

As an additional case study, a summary of the Minnesota DOT (MnDOT) *Mobile Source Data for AADT Implementation Plan* and its associated results are provided below. Three documents detail the MnDOT experience:

- **MnDOT Implementation Plan for Mobile Source Data for AADT:** The original plan (circa 2021) to qualify third-party volume estimates for submission to HPMS.
- **MnDOT Mobile Source Data for AADT Implementation Analysis:** This document (circa 2021) contains the initial analysis of third-party data for use on high-volume roadways.
- **MnDOT 2022 Mobile Source Data for AADT Analysis (6/5/2023):** This document contains the analysis results for lower-volume roadways.

All three documents are available for review here. Combined, they form a comprehensive case study of one state's effort to qualify a volume estimate vendor (StreetLight Inc.) for HPMS submission.

MnDOT Testing Plan

The first document, *MnDOT Mobile Source Data for AADT Implementation Plan*, outlines a structured framework created by MnDOT to integrate mobile-source data into their standard AADT monitoring. The plan details the technical methodologies, validation benchmarks, and step-by-step implementation protocols required to augment physical traffic counters with probe-based volume estimates sourced from the industry.

The focus of the statistical comparison is to evaluate the accuracy of these estimates. For the validation metrics, MnDOT tested the null hypothesis that the true mean difference between the paired samples (probe estimate versus counter estimate) is zero, using either a paired sample t-test or the Wilcoxon matched-pairs signed-rank test if the data was not normally distributed. This intent is similar to the FHWA test.

The initial year of implementation focused on roadway segments with an AADT above 20,000. Data was collected from over 10% of the count stations and compared with industry-provided volume estimates. MnDOT has 650 traffic segments in this higher volume range. The plan stated that if the data for this range proved acceptable, MnDOT would test lower AADT ranges in subsequent years. For the second year (data collection year 2022), MnDOT planned to lower the threshold to segments with AADTs greater than 10,000 vehicles, continuing to lower the range in following years if the validation results remained acceptable.

MnDOT Initial Results for High Volume Roadways

The second document, *MnDOT Mobile Source Data for AADT Implementation Analysis*, conveys the results of the initial analysis, which concentrated on performance on high-volume roads exceeding 20,000 AADT. The testing plan was approved in December 2021. The results compare traditional Federal Highway Administration (FHWA) AADT data from count stations with StreetLight Data's mobile-source metrics, evaluating how

closely big-data passive traffic metrics align with traditional highway performance monitoring systems in terms of accuracy and scalability.

The reference data used for comparison was AADT from 2021. MnDOT prepared a GIS file with "gates" (short lines bisecting the count location) and delivered this to the vendor. MnDOT performed quality control on the CCS count data similar to the TETC validation efforts, such as ensuring directional counts were within 90%.

After filtering, 272 count locations were used for comparison on roadways with greater than 20,000 AADT. The mean absolute percent error (MAPE) between paired volume estimates was 9.64%, with a standard deviation of 11.35%. Approximately 62% of sites exhibited a negative error, indicating a slight negative bias in the probe-based volume estimates for this range. The median difference was 1,620.1 vehicles. Visualizations of the comparisons are shown in the following figures.

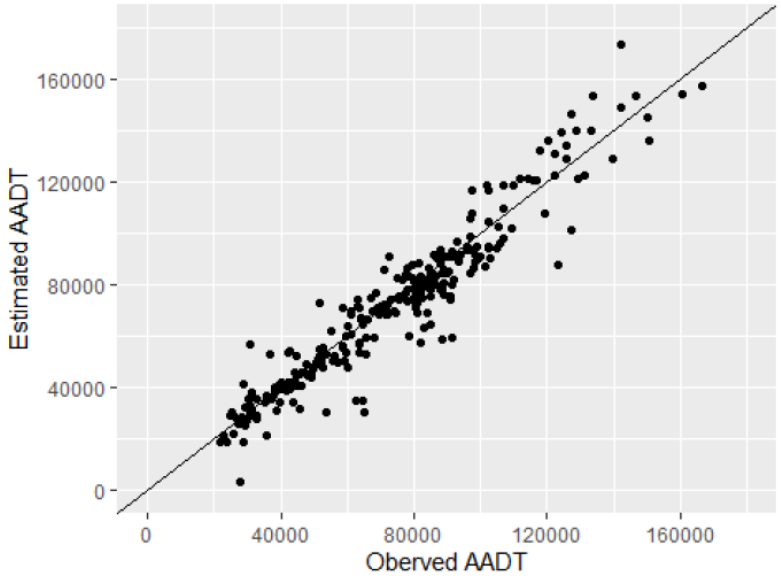


Figure F-1: Scatter plot of MnDOT Results

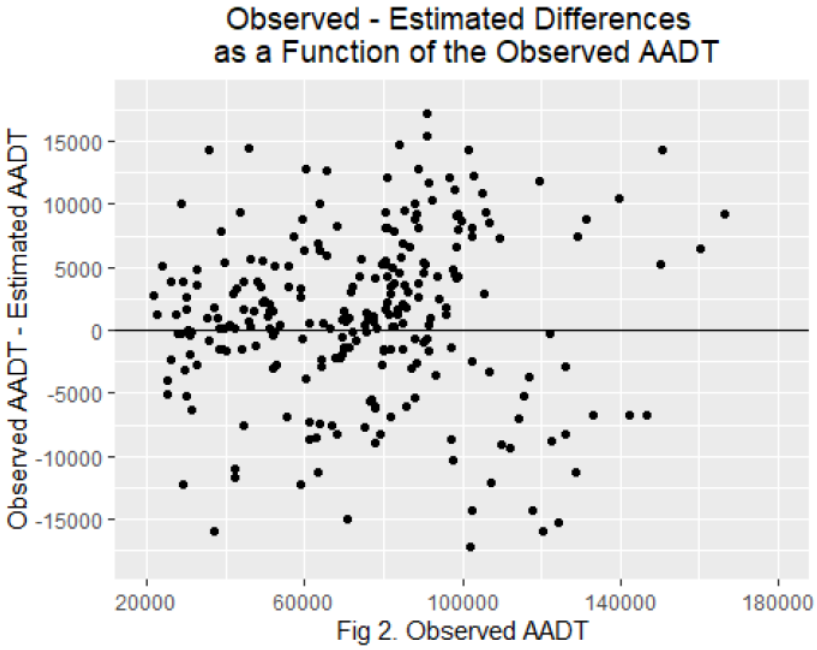


Figure F-2: Observed differences by Observed AADT

MnDOT further filtered the data by applying a percent tolerance criterion, specifying the acceptable percent AADT change from a previous count (comparing against the 2019–2020 average AADT to account for the strong COVID-related traffic downturn in 2020 and subsequent rebound in 2021). Tolerance criterion for this filter were chosen as ±15% at sites with a past official AADT of 5,000–49,999, and ±10% at site with volumes of 50,000+. These thresholds were consistent with count acceptance criteria already in effect at MnDOT as given in Table F-1.

Table F-1 – AADT Acceptable Percent Change

Past Official AADT Range	Acceptable % Change
0 – 99	No Recounts
100 – 499	±30%
500 – 1,499	±25%
1,500 – 4,999	±20%
5,000 – 49,999	±15%
>50,000	±10%

By imitating this operational process and only including sample data that met the acceptable change margins, MnDOT estimated the median percent difference between estimation methods in practical use. After applying this QC measure, the MAPE improved to approximately 5.4%, and the standard deviation lowered to 5.18%. The estimated median error improved to 730.5 vehicles. A comparison of the full dataset and the QC-filtered dataset is provided below.

MnDOT further filtered the data as described …

> *AADT estimates must pass a percent tolerance criteria specifying the acceptable percent AADT change from a previous count (This year, we compared against the 2019, 2020 average AADT in consideration of the strong COVID-related downturn in traffic in 2020 and subsequent 2021 rebound seen at our continuous sites). For sites with a past official AADT within the range of 5,000 – 49,999 the acceptable change is +/- 15%, for volumes of 50,000+ the acceptable change is +/- 10%. If the adjusted volume exceeds the acceptable percent change margin, a recount is requested. If we imitated this process with our sample data and only include data meeting acceptable change margins, we can get an estimate of the median percent difference between estimation methods in practical use.*

After further implementing this QC measure on the results (using only test sites that pass the filter previously described), the MAPE improved to approximately 5.4% and the standard deviation lowered to 5.18 %. The estimated median error improved to 730.5 vehicles. The comparison of the Full Data Set and the QC'd data set are provided in the Table below.

Table F-2: Comparison of Performance on Full Dataset vs. QC Dataset

Data	Mean Absolute Percent Error	Signed Absolute Percent Error
Full Dataset	9.64%	2.4%
QC Dataset	5.44%	0.5%

Based on this level of accuracy, and considering the improved personnel safety for traffic employees by reducing field collections, MnDOT concluded that the results warranted the use of mobile source data for roadways with a volume of 20,000 or greater.

Retest in 2023

The final document, dated June 5, 2023, details a targeted analysis expanding into lower-volume roads following the successful adoption of third-party data for roadways above 20,000 AADT. In 2023, MnDOT evaluated the 10,000–19,999 AADT range alongside the previously qualified 20,000+ AADT range.

This second analysis uncovered a significant difference in accuracy and bias between the two AADT range groups. In the 10,000–19,999 AADT range, the vendor's estimates tended to underestimate higher AADTs by around 5.4% and overestimate the lower AADT sites by an average of 23%. These results are visualized in the figure below. Notably, these biases persisted

even after excluding sites that exceeded MnDOT's acceptable percent change criteria.

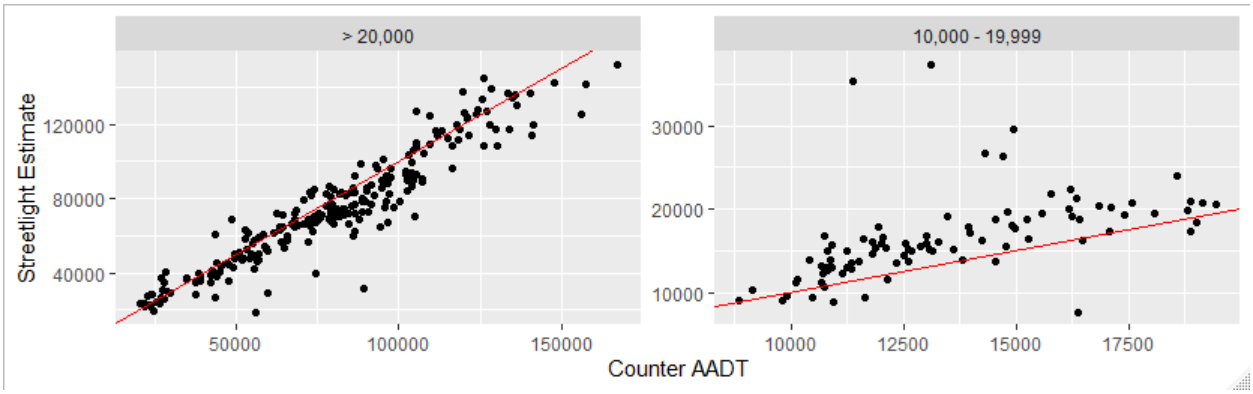


Figure F-3: Visual results of 2022 retest of Probe-Based Data

A comparison of the initial results (using 2021 Reference Data) with the subsequent year's analysis (using 2022 Reference Data) is shown in Table F-3. The vendor's estimates for 2022 failed to fall within the acceptable parameters for both the lower (10,000–19,999 AADT) and higher (20,000+ AADT) volume ranges.

Table F-3: Comparison of Initial Results (2021) with Subsequent Analysis (2022)

		2022		2021	
		Mean Absolute Percent Error	Mean Signed Percent Error	Mean Absolute Percent Error	Mean Signed Percent Error
Full Dataset	AADT >= 20,000	10.9%	5.4%	9.6%	2.4%
	AADT 10,000 - 19,999	26.1%	-23.2%		
Data within MnDOT Acceptable percent change criteria	AADT >= 20,000	6.7%	3.9%	5.4%	0.5%
	AADT 10,000 - 19,999	9.4%	-5.7%		

Together, the MnDOT documents form a comprehensive case study representing one of the first state-level attempts to qualify probe-based volume estimates for HPMS submission. This material assists states in understanding how agencies can plan, validate, and execute the shift from traditional physical infrastructure counting toward modern probe-based traffic volume estimates for official reporting.