VTRC
Virginia Transportation Research Council

We Bring Innovation to Transportation

Developing a Standardized Acceptance Framework for Traffic Count Devices: Traffic Count and Classification Verification

https://vtrc.virginia.gov/media/vtrc/vtrc-pdf/vtrc-pdf/26-R53.pdf

CHIEN-LUN LAN, Ph.D.
Research Scientist

MO ZHAO, Ph.D., P.E.
Senior Research Scientist

Final Report VTRC 26-R53

VIRGINIA TRANSPORTATION RESEARCH COUNCIL
530 Edgemont Road, Charlottesville, VA 22903-2454
vtrc.virginia.gov

Standard Title Page - Report on Federally Funded Project

1. Report No.: FHWA/VTRC 26-R53	2. Government Accession No.:	3. Recipient's Catalog No.:

4. Title and Subtitle:	5. Report Date:
Developing a Standardized Acceptance Framework for Traffic Count Devices: Traffic Count and Classification Verification	June 2026
	6. Performing Organization Code:
7. Author(s): Chien-Lun Lan and Mo Zhao	8. Performing Organization Report No.: VTRC 26-R53
9. Performing Organization and Address: Virginia Transportation Research Council 530 Edgemont Road Charlottesville, VA 22903	10. Work Unit No. (TRAIS): 11. Contract or Grant No.: 126329
12. Sponsoring Agencies' Name and Address: Virginia Department of Transportation 1221 E. Broad Street Richmond, VA 23219 Federal Highway Administration 400 North 8th Street, Room 750 Richmond, VA 23219-4825	13. Type of Report and Period Covered: Final 14. Sponsoring Agency Code:

15. Supplementary Notes:
This is an SPR-B report.

16. Abstract:

As the Virginia Department of Transportation (VDOT) transitions from traditional in-pavement sensors to non-intrusive traffic monitoring technologies, a standardized and mathematically defensible verification process is required to ensure data integrity. This research establishes a comprehensive Device Acceptance Framework designed to evaluate the volumetric and classification accuracy of emerging traffic count devices.

The framework was developed through a multi-stage methodology. First, a user needs survey involving VDOT stakeholders established a 5% error threshold as the primary accuracy target for engineering and planning applications. Second, a robust operational baseline was established by calibrating existing continuous count stations against absolute, human-verified video ground truth data.

Third, the established baseline and parameters were integrated into a formal eight-step "fail-fast" protocol. This protocol subjects candidate devices to progressive evaluations, ranging from simple 24-hour volumetric screenings to rigorous classification tests under adverse environmental conditions (e.g., heavy rain, dense fog, and snow). A field demonstration of the framework on a candidate video-based detector illustrated its practical application and the effectiveness of the pass-fail criteria.

The implementation of this framework provides significant operational and financial benefits. By utilizing the pre-calibrated continuous count station baseline instead of manual video processing, VDOT can achieve a direct cost avoidance of approximately $321,300 per device certification request. Furthermore, the framework establishes a transparent roadmap for both the agency and technology vendors, ensuring that only resilient, high-accuracy tools are deployed across Virginia's traffic monitoring network. The study recommends developing a standardized traffic count device acceptance procedure, particularly the "fail-fast" framework to establish a consistent, resource-efficient vetting process for emerging non-intrusive traffic count devices for VDOT approval.

17. Key Words:	18. Distribution Statement:
Traffic Count Devices, Traffic Monitoring, Device Approval	No restrictions. This document is available to the public through NTIS, Springfield, VA 22161.

19. Security Classif. (of this report):	20. Security Classif. (of this page):	21. No. of Pages:	22. Price:
Unclassified	Unclassified	46	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

FINAL REPORT

DEVELOPING A STANDARDIZED ACCEPTANCE FRAMEWORK FOR TRAFFIC COUNT DEVICES: TRAFFIC COUNT AND CLASSIFICATION VERIFICATION

Chien-Lun Lan, Ph.D.
Research Scientist

Mo Zhao, Ph.D., P.E.
Senior Research Scientist

Virginia Transportation Research Council
(A partnership of the Virginia Department of Transportation
and the University of Virginia since 1948)

Charlottesville, Virginia

June 2026
VTRC 26-R53

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Virginia Department of Transportation, the Commonwealth Transportation Board, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation. Any inclusion of manufacturer names, trade names, or trademarks is for identification purposes only and is not to be considered an endorsement.

ABSTRACT

As the Virginia Department of Transportation (VDOT) transitions from traditional in-pavement sensors to non-intrusive traffic monitoring technologies, a standardized and mathematically defensible verification process is required to ensure data integrity. This research establishes a comprehensive Device Acceptance Framework designed to evaluate the volumetric and classification accuracy of emerging traffic count devices.

The framework was developed through a multi-stage methodology. First, a user needs survey involving VDOT stakeholders established a 5% error threshold as the primary accuracy target for engineering and planning applications. Second, a robust operational baseline was established by calibrating existing continuous count stations against absolute, human-verified video ground truth data.

Third, the established baseline and parameters were integrated into a formal eight-step "fail-fast" protocol. This protocol subjects candidate devices to progressive evaluations, ranging from simple 24-hour volumetric screenings to rigorous classification tests under adverse environmental conditions (e.g., heavy rain, dense fog, and snow). A field demonstration of the framework on a candidate video-based detector illustrated its practical application and the effectiveness of the pass-fail criteria.

The implementation of this framework provides significant operational and financial benefits. By utilizing the pre-calibrated continuous count station baseline instead of manual video processing, VDOT can achieve a direct cost avoidance of approximately $321,300 per device certification request. Furthermore, the framework establishes a transparent roadmap for both the agency and technology vendors, ensuring that only resilient, high-accuracy tools are deployed across Virginia's traffic monitoring network. The study recommends developing a standardized traffic count device acceptance procedure, particularly the "fail-fast" framework to establish a consistent, resource-efficient vetting process for emerging non-intrusive traffic count devices for VDOT approval.

FINAL REPORT

DEVELOPING A STANDARDIZED ACCEPTANCE FRAMEWORK FOR TRAFFIC COUNT DEVICES: TRAFFIC COUNT AND CLASSIFICATION VERIFICATION

Chien-Lun Lan, Ph.D.
Research Scientist

Mo Zhao, Ph.D., P.E.
Senior Research Scientist

INTRODUCTION

Infrastructure-based traffic counts are a critical component of traffic monitoring. Traffic count and vehicle classification data are essential not only for short-term highway engineering management but also for long-term traffic demand forecasts.

The Virginia Department of Transportation (VDOT) manages more than 500 continuous count stations (CCSs) across interstates and major arterials. Roughly 300 of these CCSs rely on traditional in-road (intrusive) sensors, like inductive loops and piezoelectric sensors. Although these traditional sensors have been the industry standard for decades, they come with clear operational downsides. Installation and routine maintenance are difficult, often requiring pavement cutting that degrades the roadway surface. More importantly, repairing these sensors exposes field crews to severe safety risks from live traffic and necessitates costly lane closures that disrupt traffic flow. Routine resurfacing projects also frequently destroy these in-road sensors, forcing repetitive and expensive replacement cycles.

Non-intrusive technologies solve many of these logistical and safety issues by keeping equipment entirely out of the travel lanes, typically mounting it on roadside poles or overhead structures. During the past decade, VDOT has successfully transitioned more than 200 count stations to radar-based, non-intrusive devices. Although radar is highly effective for basic volume counting, it has significant limitations in generating classification data. Because radar systems generally classify vehicles by overall length rather than physical axle configuration, they cannot generate the precise axle-based vehicle classification data needed by VDOT planners, engineers, and the Federal Highway Administration's (FHWA) Highway Performance Monitoring System.

To fill this data gap, the industry is seeing a surge in advanced video-based detectors powered by machine vision and artificial intelligence. Vendors are introducing smart camera systems capable of tracking vehicles and visually identifying specific axle configurations. These systems may be cost-effective ways to collect accurate, axle-based vehicle classification without requiring any in-pavement sensors.

To confidently deploy these emerging technologies, VDOT needs a standardized, reproducible way to test traffic count devices (TCDs), regardless of the type of detection technologies used. Relying on unverified traffic data can lead to flawed infrastructure planning and misallocated funding. Clear, practical guidelines must be established to verify device accuracy and reliability in the field, especially under challenging lighting and adverse weather conditions. Developing a robust acceptance framework helps ensure that the agency invests only in the most effective, resilient, and efficient tools for monitoring and managing traffic.

PURPOSE AND SCOPE

The purpose of this study was to develop a TCD acceptance framework so that a standardized procedure can be applied for all future TCDs that need approval for VDOT use.

METHODS

The following tasks were conducted to achieve the study objectives:

1. A literature review and state-of-practice synthesis.
2. A user needs survey.
3. The development of a defensible baseline reference.
4. The development of device acceptance framework.
5. The demonstration of device acceptance framework.

In addition, at the request of the project technical review panel, this study included a task that assessed speed measurements from a TCD against a selected baseline.

Literature Review and State-of-Practice Synthesis

The research team performed a comprehensive synthesis of existing literature to summarize the state of practice for TCD acceptance tests. This task involved a thorough review of federal policy frameworks, primarily the FHWA (2022) *Traffic Monitoring Guide*, and technical standards from ASTM International regarding device accuracy and testing environments. Because detailed state-level protocols are often not publicly available, the research team, in conjunction with the technical review panel, identified several peer state departments of transportation recognized as national leaders in this field for further consultation. The review focused on identifying standardized error tolerances, sample size requirements, and established "ground truth" methodologies. By reconciling technical standards with practical state-level acceptance criteria, this task provided a reference for establishing acceptance criteria and testing protocol.

User Needs Survey

The primary objective of this survey was to understand users' expected accuracies of traffic volume and classification data for their specific operational and analytical use cases, such as safety analysis, pavement design, and federal reporting. The survey was distributed to both internal VDOT stakeholders and external data consumers, including Metropolitan Planning

Organizations and state universities. The survey instrument, provided in the Appendix, was designed to capture granular details regarding how users currently access traffic volume and vehicle classification data, their primary analytical purposes, and their accuracy expectations (e.g., 90%, 95%, or 99%). Furthermore, the questionnaire investigated the required spatial granularity (e.g., lane-specific, directional, or full-link aggregation) and temporal aggregation levels (e.g., 15-minute, hourly, or daily intervals) necessary for these applications. A specific module of the survey was also dedicated to vehicle classification, inquiring whether the respondents adhered strictly to the standard FHWA 13-category classification system, shown in Figure 1, or used customized aggregated groupings tailored to their specific engineering or planning use cases.

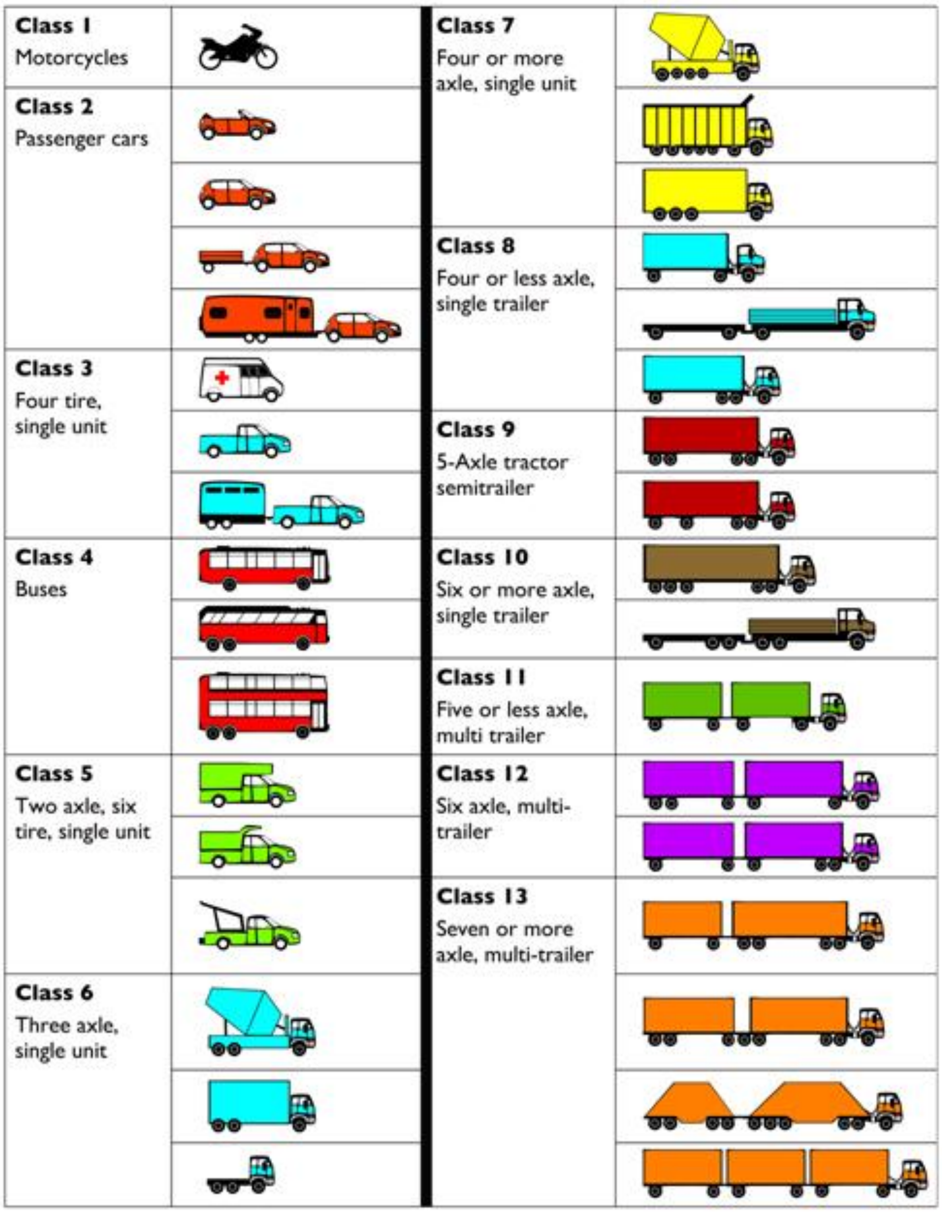


Figure 1. Federal Highway Administration 13 Vehicle Category Classification (FHWA, 2022)

Baseline Establishment

The fundamental prerequisite for evaluating any candidate TCD is establishing a baseline reference (or ground truth). Traditionally, evaluation frameworks have often depended directly on CCS as a baseline. Although existing CCSs are meticulously maintained, they still face fundamental constraints when resolving low-volume multi-axle commercial vehicles. Treating CCS output as an absolute, flawless standard without quantifying these inherent variances may introduce statistical risk. Specifically, it risks generating Type I errors, where highly capable candidate sensors are falsely rejected because of the limitations in the baseline. This task outlined a methodology to mitigate this risk and establish a defensible framework.

Continuous Count Station Site Selection

Selecting the proper CCS as a baseline is critical because it dictates testing feasibility and environmental validation integrity. To ensure the site provides a statistically viable environment satisfying the evaluation framework, the site should consider a number of aspects discussed in this section. Although a single CCS site meeting all aspects is ideal, the evaluation framework may leverage multiple sites to collectively satisfy the criteria.

Traffic Profile

The ideal CCS site should exhibit a diverse vehicle composition and carry moderate to high traffic to satisfy the sample size requirements. A historical moderate to high average annual daily traffic (AADT) with a significant proportion of Class 9+ vehicles is required. In addition, the site should ideally experience distinct Level of Service degradation during peak periods to evaluate candidate device against occlusion and stop-and-go shockwaves. Finally, to ensure a dependable baseline, the candidate CCS must consistently achieve high data quality—a "Class Quality" rating of 4 or 5 based on the current VDOT system.

Road Weather Information Systems

To evaluate device resilience against specific weather variables (e.g., rain intensity greater than 0.25 in/hr and visibility less than 600 feet), the site requires accurate environmental information. Therefore, the CCS must be co-located with a Road Weather Information System (RWIS) to provide high-quality data and enable direct time-syncing of local precipitation and visibility metrics. In addition, the site requires a documented history of adverse weather, such as heavy rain, fog, and snow, ensuring empirical data availability.

Physical Infrastructure

To ensure a true side-by-side comparison, the testing environment must eliminate bias between the CCS and the test device. The site must feature existing infrastructure, such as sign gantries or dedicated Intelligent Transportation System (ITS) poles for mounting the testing device. The location must provide an unobstructed field of view of the exact detection zone the CCS monitors. It must also be sufficiently close to the in-pavement loops or piezoelectric sensors to minimize the possibility of vehicles changing lanes between the CCS and test device detection zone. In addition, accessible power and communication backhaul are required.

Ground Truth Data Collection

To accurately evaluate candidate TCDs, a baseline must be established. Although relying exclusively on manual counting is resource prohibitive, relying solely on fully automated computer vision models cannot guarantee the rigorous quality required for ground truth, introducing unacceptable baseline errors, particularly for complex FHWA 13 category classifications. To balance cost efficiency with absolute precision, this task centered on deploying a custom Python-based semi-automated vehicle tracking and classification system that fuses fast You Only Look Once-based deep learning object detection with rigorous human-in-the-loop verification.

The primary data source for this evaluation was video captured at the CCS site. To ensure comprehensive coverage of the site, two pan-tilt-zoom cameras were deployed to record continuous traffic across all monitored lanes. These recordings were archived as video files and transferred for laboratory testing, where they served as the input for the semi-automated vehicle tracking and classification system to extract the definitive ground truth dataset.

The workflow begins by using a graphical interface to establish the scene's spatial geometry. Virtual "tripwires" are mapped to individual travel lanes as spatial triggers, and exclusion zones are drawn over background areas to prevent false positive records. Once calibrated, the deep learning tracker traces moving vehicles frame by frame. When a tracked vehicle's center intersects a tripwire, the system pauses playback and presents a cropped image. A human reviewer then definitively assigns the FHWA classification and confirms the lane assignment.

To guarantee data integrity against edge cases, like severe truck occlusion or poor illumination, a manual failsafe allows analysts to force a capture using a localized bounding box tool. Every classified vehicle is systematically exported to a comma-separated values dataset, logging its unique identifier, verified FHWA class, lane, frame, and timestamp. Crucially, the system also saves a linked image of every cropped vehicle to facilitate rigorous future auditing.

Statistical Calibration of Continuous Count Station Baseline

CCS devices possess inherent hardware limitations, so treating the CCS output as ground truth mathematically risks generating Type I errors, falsely rejecting a capable candidate device because of underlying baseline noise. This task outlines the methodology to quantify the baseline uncertainty of the CCS by comparing it with human-verified, timestamped video logs extracted via a Python-based tool. The resulting calibration establishes the operational noise floor, ensuring candidate devices are evaluated against a statistically defensible standard.

Data Synchronization and Temporal Aggregation

The first step is to synchronize the highly granular video ground truth data with the CCS reporting format. This study extracted the ground truth with a timestamped log of every vehicle passing through the detection zone. Next, temporal aggregation is performed by binning both the manual video ground truth (***Y***) and the CCS sensor output (***X***) into discrete 15-minute temporal intervals, denoted by *t*. Through this variable assignment, for each 15-minute bin *t* and each of the 13 standard FHWA vehicle classes, denoted by *c*, the ground truth volume is defined as $Y_{c,t}$ and the CCS reported volume as $X_{c,t}$.

Vehicle Class Aggregation

Evaluating individual low-frequency classes (e.g., FHWA Class 13) often yields excessively wide confidence intervals and high variance. To ensure the baseline achieves a 95% statistical confidence level, the vehicle count of the 13 FHWA classes must be aggregated prior to error calculation into three strategic tiers determined based on literature, the user needs survey, and discussions with stakeholders: Group 1 (G_1), consisting of Classes 1–3 (Motorcycles, Passenger Cars, Light Trucks); Group 2 (G_2), containing Classes 4–8 (Buses and Single-Unit Trucks); and Group 3 (G_3), isolating Classes 9–13 (Multi-Axle and Multi-Trailer Heavy Trucks). For any given group *G* and time interval *t*, the aggregated group volumes are calculated as follows using Equations 1 and 2:

$$Y_{G,t} = \sum_{c \in G} Y_{c,t} \qquad \text{(Equation 1)}$$

$$X_{G,t} = \sum_{c \in G} X_{c,t} \qquad \text{(Equation 2)}$$

Calculation of the Weighted Absolute Percentage Error

Standard metrics like the mean absolute percentage error (MAPE) are highly vulnerable to zero-denominator scenarios and disproportionately penalize sensors during low-volume intervals. For instance, missing one vehicle when the true volume is 1 yields a 100% error rate. To calculate the baseline accuracy, this study computed the weighted absolute percentage error (WAPE). By summing the absolute errors across all time intervals before dividing by the total volume, WAPE neutralizes the zero-denominator issue and weights errors according to their actual effect on traffic operations. For a specific vehicle group *G* evaluated during a total testing duration of *T* intervals, the baseline error is calculated using Equation 3:

$$WAPE_G = \frac{\sum_{t=1}^{T} |X_{G,t} - Y_{G,t}|}{\sum_{t=1}^{T} Y_{G,t}} \qquad \text{(Equation 3)}$$

Deriving the Acceptance Threshold Parameters

The empirically calculated $WAPE_G$ values directly inform the acceptance parameters used in the acceptance framework evaluations (Task 4). First, researchers must define the base rate (T_{base}) because the established baseline variance becomes the minimum boundary for testing. If the CCS baseline variance against manual ground truth yields a specific error rate (e.g., 3%) for Group 3, the target evaluation threshold (T_{base}) cannot be logically set less than that baseline noise floor. Based on stakeholder consensus from the Task 2 user needs survey, the target threshold for ideal conditions will be selected. Second, researchers establish the absolute minimum (T_{min}) to prevent low-volume failure cascades. Missing a small number of vehicles can trigger a percentage failure, thus requiring an absolute minimum vehicle tolerance, T_{min} (e.g., 5 or 10 vehicles). Finally, the parameters derived from this calibration step are fed into the dynamic threshold equation to formulate the maximum allowable error (E_{max}) to govern device acceptance in Task 4. For any group *G*, the maximum allowable error boundary is calculated using Equation 4:

$$E_{max} = max\left(T_{base} \times V_{base,G}, T_{min}\right) \qquad \text{(Equation 4)}$$

Where:
$V_{base,G}$ is the base volume being evaluated at that specific moment (see Task 4).

Statistical Validation Across Environmental Scenarios

To ensure the overarching baseline calibration is robust across varying weather and day or night conditions, the error distributions of those scenarios (e.g., daytime clear or nighttime snow) must be statistically compared. Because the percentage-based WAPE data are strictly bounded between zero and one, they inherently violate the mathematical assumption of normal distribution required for parametric tests such as the analysis of variance (ANOVA). Consequently, the Kruskal-Wallis H test, a non-parametric rank-based method, is employed to determine statistical equivalence among the conditions. To apply this test, all observed WAPE values across all *k* environmental scenarios are pooled and ranked from lowest to highest, yielding a total of *N* observations. The sum of the ranks for each specific scenario *i*, containing observations n_i, is calculated and denoted as R_i. The Kruskal-Wallis test statistic (*H*) is then computed using Equation 5:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \qquad \text{(Equation 5)}$$

In the presence of tied ranks across the dataset, a correction factor *C* is applied, defined as $C = 1 - \sum(t^3 - t)/(N^3 - N)$, where *t* represents the number of tied observations in a given grouping, adjusting the final statistic to *H/C*. To determine statistical significance, the resulting *H* statistic is compared with the critical value from the Chi-square (χ^2) distribution. The degrees of freedom for this evaluation are determined by subtracting one from the total number of distinct environmental scenarios being compared (*df* = *k* - 1). If the calculated *H* exceeds the critical χ^2 value at the 95% confidence level ($\alpha = 0.05$), the null hypothesis is rejected, indicating a statistically significant divergence in sensor accuracy under specific conditions. Conversely, failing to reject the null hypothesis validates that the aggregated operational noise floor remains statistically stable regardless of external environmental factors.

Device Acceptance Framework

Building on the operational noise floor established in the previous section, this task developed a multi-phased acceptance framework for testing candidate TCDs. This framework translated the statistical baseline into actionable acceptance criteria. To optimize agency resources, this framework was designed following a “fail-fast” logic. If a candidate device fails an early, simple baseline test, the testing process is immediately terminated. Devices that pass the initial screening advance through increasingly rigorous scenarios. The evaluation was structured into three distinct phases. Phase I evaluates total volumetric screening, Phase II evaluates classification accuracy, and Phase III assesses environmental resilience.

Demonstration of Device Acceptance Framework

To provide a practical example of how to apply the Device Acceptance Framework, a demonstration was conducted using a candidate non-intrusive traffic count technology, referred to as “Device A” hereafter. Device A is a video-based detector utilizing proprietary classification algorithms. Device A was deployed at the established CCS site on southbound Interstate 77 (I-77) at mile marker 4.4 in Fancy Gap. The data collection period for Device A was from May 26 to October 24, 2025, and from November 25, 2025, to February 25, 2026. The demonstration followed the developed acceptance framework and processed data chronologically through the phases, verifying the error thresholds and the Temporal Reliability Factor requirements before advancing the device to the subsequent operational test.

Speed Comparison

This task presents a comparative analysis of speed measurements between the candidate device (Device A) and the CCS baseline. It is important to note that this speed comparison was conducted specifically at the request of the technical review panel to provide supplementary insights into the device’s operational capabilities. This analysis is entirely independent of the Device Acceptance Framework established in the preceding tasks.

Data Sources

- Speeds from Device A: Speed readings from Device A at the I-77 study site were available for approximately 189 days between July 15, 2025, and February 25, 2026. Speed of each observed vehicle was provided.
- Speeds from CCS: CCS speed records for the I-77 study site from July 2025 to February 2026 were obtained from the VDOT Traffic Operations Division Oracle database, the Commonwealth of Virginia Traffic Engineering Data Operations Portal. Direct speed measurements were unavailable. Instead, the following volume counts for each of the 15 speed bins were recorded at 15-minute intervals.
    - 0 to 15 mph.
    - 15 to 25 mph.
    - 25 to 30 mph.
    - 30 to 35 mph.
    - 35 to 40 mph.
    - 40 to 45 mph.
    - 45 to 50 mph.
    - 50 to 55 mph.
    - 55 to 60 mph.
    - 60 to 65 mph.

- 65 to 70 mph.
- 70 to 75 mph.
- 75 to 80 mph.
- 80 to 85 mph.
- 85 mph and up.

- Precipitation and visibility: Rain intensity (light, moderate, or heavy) and visibility (feet) data were obtained from VDOT's RWIS. Data were reported for about every 5- to 6-minute interval.
- Sunrise and sunset times: The times of sunrise and sunset for the study site were collected through the National Oceanic and Atmospheric Administration (NOAA) data services (NOAA, 2026).

Data Preparation

Traffic volume counts by speed bin from CCS were used to estimate speed using Equation 6.

$$Average\ speed = \frac{\sum_{i=1}^{n} S_i \times C_i}{\sum_{i=1}^{n} C_i} \qquad \text{(Equation 6)}$$

Where:
n = the total number of speed bins.
S_i = midpoint of speed bin i.
C_i = the number of vehicles in bin i.

Because CCS data were available at 15-minute intervals, all other data were aggregated to match this interval. To align with the CCS data, the per-vehicle speed readings from Device A were categorized into the 15 speed bins mentioned previously, and then an average speed was calculated using Equation 6.

From the RWIS data, the time intervals with heavy rain (greater than 0.3 inches per hour based on the RWIS device definition) or low visibility (less than 600 feet) were identified. If multiple readings occurred within a 15-minute interval, the highest rain intensity or the lowest visibility observed was used.

Because snowfall data were not available from RWIS, the hours with snowfall were identified by reviewing site video recordings. In addition, video recordings from January 2026 were reviewed to identify periods affected by sun glare at the study site. Researchers considered the data collected between 8:00 and 9:00 am in that month to be affected by sun glare.

Comparative Analysis

Aggregated speeds calculated from Device A speed readings were compared with the baseline of speeds from count stations. The comparisons were performed for different times of day (day, night, dawn, or dusk) and weather conditions (clear, snow, low visibility, heavy rain, and sun glare) combinations. For each combination, two-tailed paired *t*-tests and the Bland-Altman analyses were conducted if sample size was sufficient.

The two-tailed paired *t*-test was used to compare the speed samples from CCS and Device A, matched for each 15-minute interval, to determine whether their mean difference was statistically significant. The hypotheses of the *t*-test were:

- Null hypothesis (H_0): The mean difference between the paired speed observations is zero.
- Alternative hypothesis (H_a): The mean difference is not zero.

The Bland-Altman analysis was used to evaluate the agreement between the speed samples from the CCS and Device A. Rather than testing for a statistical difference, this analysis assessed how closely the two sets of speeds aligned by examining the bias and determining the range within which most differences fell, known as the limits of agreement (LoA). The range between the lower and upper LoA covers approximately 95% of the differences between the two speed datasets. The percentile-based LoA was used in this study because this approach is less influenced by tails of the bias distribution compared with methods relying on standard deviation, which assume normality. Key metrics were defined as:

- Bias = speed from Device A – speed from CCS.
- Upper LoA = 2.5th percentile of bias.
- Lower LoA = 97.5th percentile of bias.

In addition to comparing paired speed observations, the overall distributions of speeds from CCS and Device A were also analyzed. Percentile speeds (1st through 100th) were computed for both datasets and then compared using the two-sample Kolmogorov-Smirnov (K-S) test and the Wasserstein distance. The K-S test was used to determine whether the two speed distributions differ significantly, whereas the Wasserstein distance measured the overall separation between the two distributions. The hypotheses of the two-sample K-S test were:

- Null hypothesis (H_0): The cumulative distributions of speeds from CCS and Device A are identical.
- Alternative hypothesis (H_a): The cumulative distributions of speeds from CCS and Device A are different.

RESULTS AND DISCUSSION

Literature Review and State-of-Practice Synthesis

Permanent traffic monitoring programs traditionally rely on a combination of intrusive (in-pavement) and non-intrusive sensors. Intrusive sensors, such as inductive loop detectors and piezoelectric sensors, provide exceptional lane-by-lane discrimination and are largely immune to atmospheric weather conditions. Inductive loops detect the magnetic anomaly created by the conductive mass of a vehicle entering the detection zone, altering the loop’s inductance and triggering a call to the controller. Piezoelectric sensors, which can be constructed from polymer cables or precise quartz crystals, generate a proportional electrical charge when subjected to the mechanical strain of a passing tire, allowing for both axle counting and dynamic weight estimation.

Non-intrusive sensors represent a rapidly expanding segment of the permanent monitoring landscape because of their ability to be installed and maintained without subjecting personnel to hazardous lane closures. Microwave radar sensors transmit electromagnetic energy and analyze the reflected Doppler shift or frequency modulation to calculate vehicle presence, speed, and length. Video image processing systems utilize advanced machine vision algorithms to establish virtual tripwires and tracking zones, identifying vehicles based on pixel contrast and object geometry. However, because non-intrusive devices monitor the roadway from a distance, they are highly susceptible to occlusion (when a large vehicle blocks the sensor's view of a smaller vehicle in an adjacent lane) and atmospheric interference. Consequently, the approval processes for these devices may require stringent environmental and spatial testing requirements.

The systematic development of an evaluation and acceptance process for TCD is contingent on a hierarchical structure of federal guidance and industry-specific technical standards. At the national level, the FHWA *Traffic Monitoring Guide* serves as the foundational policy document, establishing requirements for continuous versus short-duration data collection, quality assurance and control, and statistical precision targets, such as plus or minus 10% precision at 95% confidence for AADT estimation. The *Traffic Monitoring Guide* is defined as a set of recommendations designed to improve business processes and technological integration rather than a rigid federal standard. To align the intensity of monitoring efforts with the quality of data collected, the *Traffic Monitoring Guide* utilizes a "nested" sample design. In this hierarchy, sessions for weighing trucks represent a specialized subset within vehicle classification sessions, which are contained within broader volume counting sessions (FHWA, 2022).

Technical Specifications of ASTM E2532-09(2024)

ASTM E2532-09(2024), titled *Standard Test Methods for Evaluating Performance of Highway Traffic Monitoring Devices*, provides a practical framework for assessing the data accuracy and operational performance of TCDs through standardized testing protocols (ASTM International, 2024). This standard defines two primary categories of acceptance tests based on the lifecycle stage of the hardware. *Type approval tests* serve as a rigorous, comprehensive evaluation intended for new TCD models to verify all functional features and baseline accuracy across a mix of anticipated vehicle classes. Conversely, *onsite verification tests* are applied to specific production units that have already cleared type approval but are being commissioned at a new location or following major repairs. This procedure ensures that localized environmental factors or installation nuances have not compromised the unit's intended performance. To achieve a statistically valid result, the standard requires that performance be determined by comparing device outputs with accepted reference values, typically established by two or more human observers analyzing synchronized video imagery. To ensure the integrity of the ground truth, the user and seller shall agree on the data-measuring accuracy requirement for all reference equipment, with the standard recommending that reference equipment possesses an accuracy at least one order of magnitude greater than the accuracy specified for the device under test, where possible (ASTM International, 2024).

ASTM E2532 establishes a protocol for field evaluations, ensuring that devices are tested against the complexities of real-world traffic dynamics rather than ideal laboratory conditions. The standard mandates several rigorous field-testing parameters:

- **Operational Flow States:** Devices must be evaluated under vehicle flow rates that include vehicle-to-vehicle gaps of both less than and greater than 1 second. This particular benchmark evaluates how effectively a sensor distinguishes between a preceding vehicle's rear and the subsequent vehicle's front, which is a common failure point for radar and inductive logic.
- **Environmental Matrix:** Testing must encompass day, night, dawn, and dusk lighting, including periods where artifacts such as sun glare or shadows occur. Sun glint is particularly problematic for optical sensors because it can "blind" cameras, leading to missed detections or locked detection zones. Users are responsible for specifying localized environmental conditions, such as temperature extremes and precipitation rates, to ensure the device is "hardened" for the local climate.
- **Reference Accuracy Requirements:** Equipment used to establish reference values should possess an accuracy at least one order of magnitude greater than the accuracy specified for the TCD under test. Reference values for counts and classifications are established by two or more human observers analyzing synchronized video. If the discrepancy between observers exceeds 10% of the specified device tolerance, reference observations must be repeated.
- **Speed Reference Standards:** Reference values for vehicle speed are obtained using either matched axle-detecting sensors at known distances (creating a precise speed trap) or calibrated radar or laser speed guns.

Temporal and Sampling Variables

A robust evaluation process must systematically test a device against the specific variables that cause sensor failure, algorithmic confusion, or data degradation. TCD accuracy is not a static coefficient. It fluctuates dynamically based on the time of day, the specific sample size of the test, and the physical characteristics of the vehicles being tracked.

Temporal Variables: Minimum Hours of Data Collection

The duration of data collection during an acceptance test is a critical parameter that must be calibrated to capture sufficient variability in traffic flow and ambient lighting conditions while remaining economically feasible. Testing duration typically falls into three stratified categories based on sensor criticality:

1. **Short-Duration Visual Comparison (15–60 minutes):** Can be used for routine maintenance, portable deployments, or immediate post-installation verification. For instance, the Florida Department of Transportation (FDOT) mandates a manual verification count of 50 vehicles or a 15-minute observation period per travel lane, whichever occurs first, to ensure the hardware is counting and classifying correctly during initial site commissioning (New York State Department of Transportation [NYSDOT], unpublished data, 2024).
2. **Daily Cycle Test (24 hours):** Can be used for permanent CCSs or critical signal actuation sensors. Agencies such as the Georgia Department of Transportation (GDOT) require 24 hours of continuous ground truth to evaluate devices across morning peaks, midday, evening peaks, and low-volume nighttime hours (GDOT, 2022a). This duration

is a statistical necessity to test the sensor's resilience to changing lighting (critical for video-based detectors) and shifting car-to-truck ratios (FHWA, 2000).

3. **Extended Verification (48–72 hours):** Can be used for Weigh-In-Motion (WIM) systems or critical freeway ITS deployments. A longer verification is necessary to see a full range of loads, axle configurations, speeds, and weather conditions. For example, WIM systems are often evaluated for multiple days because thermal cycles significantly affect the structural response of the pavement and the strain readings of bending plates and piezoelectric sensors (FHWA, 2000).

Vehicle Classification and Sample Size Thresholds

Achieving statistical confidence in an acceptance test requires adequate sample sizes for each of the 13 standard FHWA vehicle classes (FHWA, 2006). The unique physical and behavioral characteristics of specific vehicle types often complicate achieving these sample quotas.

- **Motorcycle (Class 1):** Motorcycles represent a persistent challenge due to their small physical footprint, low metal mass, and tendency to travel in staggered configurations (FHWA, 2022). Standard loop detectors often overcount them as passenger cars or miss them entirely. Specialized loop configurations and test samples that actively include groups of motorcycles are recommended to ensure sensitivity without triggering false positives (FHWA, 2006).
- **Heavy Vehicle (Class 10–13):** Because heavy vehicles constitute a small percentage of total traffic, short-duration tests often fail to capture enough samples (e.g., Class 13 multi-axle trucks) to evaluate class-specific accuracy. Acceptance test plans should explicitly state minimum sample size requirements for heavy classes, often necessitating an extension of the test window if quotas are not met (FHWA, 2006).

Environmental and Traffic State Variables

Acceptance testing must subject devices to a matrix of traffic states and environmental conditions. A sensor performing well in free-flow traffic may fail during a nighttime snowstorm in stop-and-go congestion. The following variables are reported in the literature.

Traffic State Variables

1. **Free-Flow Conditions:** Baseline state in which vehicles are separated by adequate headways, allowing for easy identification and accurate calculation of speed across speed traps (FHWA, 2010).
2. **Congested and Stop-and-Go Conditions:** High traffic density often leads to "occlusion" or "tailgating" errors, when the detection logic fails to distinguish the gap between vehicles and incorrectly aggregates multiple passenger cars into a single, larger vehicle class. Furthermore, rolling shockwaves in congestion violate the assumption of constant velocity, potentially making travel time estimation from detectors inaccurate (Wisconsin Department of Transportation, 2005).

Environmental Variables

Atmospheric and lighting conditions can significantly degrade sensor accuracy by obscuring optical paths or introducing physical interference. Table 1 provides a summary synthesized from existing literature and industry standards, detailing the specific environmental variables, their operational effects, and the corresponding metrics used in the ASTM standard.

Table 1. Environmental Variables for Traffic Count Device Acceptance Testing

Environmental Variable	Effect on Sensor	ASTM E2532 Testing Metrics (ASTM International, 2024)
Ambient Lighting	Video contrast loss at dawn/dusk; headlight "blooms" at night; daytime shadows cause double counts.	Dawn, dusk, and nighttime ambient lighting levels.
Solar Glare	Direct sun (sun glare) blinds cameras, causing false negatives or locking detection zones.	Sun glint angles and light intensity.
Rain and Precipitation	Splashing obscures lenses; radar signal scattering; tire hiss overwhelms acoustic sensors.	Rain rate (intensity measured in mm/hr).
Fog and Dust	Reduces video contrast; scatters laser beams, leading to truncated vehicle length estimates.	Human visual range (visibility distance).
Snow and Ice	Obscures video; snowplows damage intrusive road tubes and surface-mounted loops.	Snowfall rate and flake size.

Peer State Approval Protocols for Video-Based Traffic Count Devices

Before a permanent traffic counting device can be deployed on state highways, it typically undergoes a formal acceptance process (the type approval), before it can be placed on an Approved Products List, Qualified Products List, or Authorized Materials List. This prequalification phase ensures that commercial off-the-shelf devices meet baseline hardware, electrical, communication, and safety specifications before any field testing is authorized. This section examines type approval practices for video-based TCDs. The following case studies illustrate how peer state departments of transportation structure these evaluations.

Florida Department of Transportation Traffic Engineering Research Laboratory

FDOT maintains centralized control over all traffic monitoring equipment through its Traffic Engineering Research Lab. Under Florida Statutes Section 316.0745, the Traffic Engineering Research Lab must evaluate and certify all traffic control and monitoring devices before they can be utilized within the state. Manufacturers initiate this process by submitting formal applications through the Product Application Tracking History system. The Traffic Engineering Research Lab evaluation process assesses product compliance against FDOT Standard Specifications and Developmental Specifications. Technical experts from various disciplines, including structures, materials, and traffic operations, review the device's technical documentation, structural integrity, and communication protocols. On successful completion of laboratory inspections and administrative reviews, the device receives a unique Approved

Products List designation. For type approval, FDOT mandates minimum accuracy thresholds of 95% for volume, 90% for occupancy, and 90% for speed across all monitored lanes (FDOT, 2025).

New York State Department of Transportation Video Review Standards

NYSDOT has established rigorous standards for the acceptance of video-based traffic count technologies (NYSDOT, unpublished data, 2024). The field certification protocol mandates system deployment at two distinct New York State locations characterized by a high volume of Class 3 vehicles and heavy trucks. To ensure reliability across operational variables, NYSDOT requires a minimum of two non-consecutive hours of video for daytime, nighttime (9 pm to 4 pm), and adverse weather conditions (rain, fog, or snow). Furthermore, the certification is hardware-specific; any approval granted is tied to the exact number of lanes tested and the specific lighting configuration used during the evaluation.

Technical acceptance is determined by a tiered accuracy matrix that varies based on the classification scheme (6 bin versus 13 bin) and environmental conditions. For a standard 13-bin daytime count, the system must achieve at least 95% accuracy for total volume, 90% for classification bins recording 30 or more vehicles per hour, and 75% for bins with fewer than 30 vehicles per hour. Acceptance thresholds are adjusted for nighttime or adverse weather, reducing the high-volume bin requirement to 80%. Failure in any single classification bin results in overall test failure, requiring system recalibration and a complete reevaluation.

The evaluation framework depends on manually generated "ground truth" data, which was generated through manual human observation. All test videos are reviewed by a minimum of two people to verify the counts. This human-generated count serves as the baseline against the candidate technology. Recognizing that human observation introduces variability, particularly under suboptimal lighting or low-resolution conditions, NYSDOT implements a strict quality control protocol for manual verification. Counts failing internal quality control thresholds are rejected and require full recounting. The department notes that this labor-intensive validation process imposes significant economic costs on the type approval workflow.

Georgia Department of Transportation Video Review Standards

GDOT has established standards for the acceptance of video-based traffic count technologies through its Product Evaluation Program and Qualified Products List. The field certification protocol mandates that applicants submit a third-party accuracy test demonstrating compliance with GDOT Supplemental Specification Section 937 accuracy requirements for volume and speed (GDOT, 2022b, 2025). If a third-party test is unavailable, GDOT requires the submission of a comprehensive Accuracy Test Plan to the Qualified Products List Products Evaluation Committee. To ensure reliability across operational variables, GDOT requires the test site to include both free-flowing (normal freeway speed) and congested (low freeway speed) traffic conditions. Furthermore, the certification requires a robust sample size; a minimum of 1,000 data points must be collected for each lighting, pavement, and temperature condition.

Technical acceptance is determined by comparing the device's output with ground truth data. For ITS freeway and CCS applications, the system must achieve at least 90% accuracy for vehicle volume and 90% accuracy for vehicle speed compared with the actual count of vehicular traffic. The deviation represents the absolute value of both over and undercounting actuation events.

The evaluation framework depends on ground truth data, which can be generated through manual human observation. The applicant must provide contact information for the ground truth manual data collection vendor. Alternatively, in-ground loops can be used to compare device results, provided the applicant submits verification that the loops are maintained, tested, and highly accurate.

Summary of Findings

A critical synthesis of these findings highlights the tension between standardized rigor and practical economic feasibility within TCD validation. Although protocols successfully address sensor modality limitations, such as occlusion and environmental sensitivity in video-based sensors, achieving statistically robust representation for low-frequency vehicle classes (e.g., Class 13 heavy trucks) often necessitates extended testing windows. Furthermore, the reliance on labor-intensive manual observation to establish ground truth, as mandated by several peer state departments of transportation, introduces significant operational costs and inherent variability into the validation workflow. Therefore, this review establishes a clear technical gap: Current acceptance frameworks are optimized for comprehensive accuracy but lack an integrated mechanism to maintain high statistical confidence across all traffic regimes while simultaneously ensuring economic viability.

User Needs Survey

The survey yielded detailed operational insights from 26 participants who collectively submitted 72 specific use-case requirements for traffic count data. Analysis of the required accuracy thresholds revealed that 38% of the reported use cases (27 of 72) demanded a 95% accuracy threshold. These applications predominantly require 15-minute or hourly data aggregation across specific lanes or directions, largely supporting critical operational functions, such as dynamic signal timing optimization, corridor performance measurement, queue estimation, and complex delay studies. A 99% accuracy threshold was required by 25% of the use cases (18 of 72), which largely consisted of per-vehicle record (PVR) extractions for speed studies, as well as granular interval data for localized benefit-cost analyses. Conversely, only 20% of the use cases (14 of 72) expected a 90% accuracy threshold, which users deemed acceptable for broader planning activities, such as work zone scheduling, alternative route screening, and average daily traffic (ADT) forecasting. Although 15-minute temporal aggregation emerged as a frequent request for micro-level traffic operations, hourly aggregation was commonly expected for broader corridor studies and performance reporting.

Responses regarding vehicle classification data included 40 use cases submitted by 23 participants. The largest proportion of classification data use cases, representing 43% of the submitted use cases (17 of 40), requested an accuracy of 95%. These applications required hourly aggregated data for structural engineering and corridor modeling tasks, including mechanistic-empirical pavement design, simulation model calibration, and large-scale freight planning. In pavement design, for instance, accurately differentiating between medium and heavy truck classes is vital because misclassifying heavy, multi-axle vehicles can significantly underestimate equivalent single-axle loads (ESALs) and predicted pavement wear. A 90% accuracy requirement was captured in 28% of the classification use cases (11 of 40), utilized primarily for detour capacity analysis, work zone queuing projections, and macroscopic safety modeling. The highest accuracy tier of 99% was necessary for 18% of the classification use cases (7 of 40). This high accuracy was typically required for federal reporting (e.g., Highway Performance Monitoring System), speed studies, and heavy truck restriction evaluations.

When evaluating categorical parameters for classification data, a divergence in classification needs was observed among the 27 participants. Of the 27 users, 18 (67%) directly utilize the standard FHWA 13-category classification scheme for their analyses; the remaining 9 (33%) apply customized vehicle class aggregations specifically tailored to their use cases. For example, VDOT’s Structure and Bridge Division extracts AADT and specific truck percentages to calculate the fatigue categories of steel structures. This task often requires grouped classification by precise axle weight specifications because bridge fatigue life is hyper-sensitive to heavy axle configurations rather than general vehicle shape. Similarly, environmental teams utilize heavy vehicle counts as a direct surrogate for corridor noise level modeling, isolating commercial truck volumes because of their distinct acoustic and emission profiles. Transportation planning units frequently aggregate the data into broad operational bins, such as passenger vehicles (Classes 1–3, 15), buses (Class 4), and heavy vehicles (Classes 5–13), to efficiently estimate passenger car equivalents for capacity analyses. By synthesizing these diverse user requirements, it was determined, in conjunction with the stakeholders, that an overall acceptance threshold (ε) of 95% is a balanced choice.

Baseline Establishment

Continuous Count Station Site Selection

Based on established spatial and operational criteria, the CCS on southbound I-77 at mile marker 4.4 in Fancy Gap, Virginia (Station ID 120027), was selected as the baseline testing site. Although I-77 does not typically experience the severe congestion or Level of Service degradation seen on urban corridors, this minor compromise is heavily outweighed by the exceptional infrastructural and environmental attributes of the site. Given the objective to develop a robust statewide acceptance process, access to localized weather data and existing camera infrastructure was essential.

The location provides a moderate traffic volume for classification, maintaining daily traffic volumes between 20,000 and 30,000 vehicles. Figure 2 illustrates the daily volume profile calculated as an average of 15-minute intervals from May to October 2025. The profile demonstrates complex diurnal and lane-specific dynamics. Although averaged data yield a peak volume of roughly 1,000 vehicles per hour per lane, raw empirical records document actual peak surges exceeding 2,000 vehicles per hour per lane.

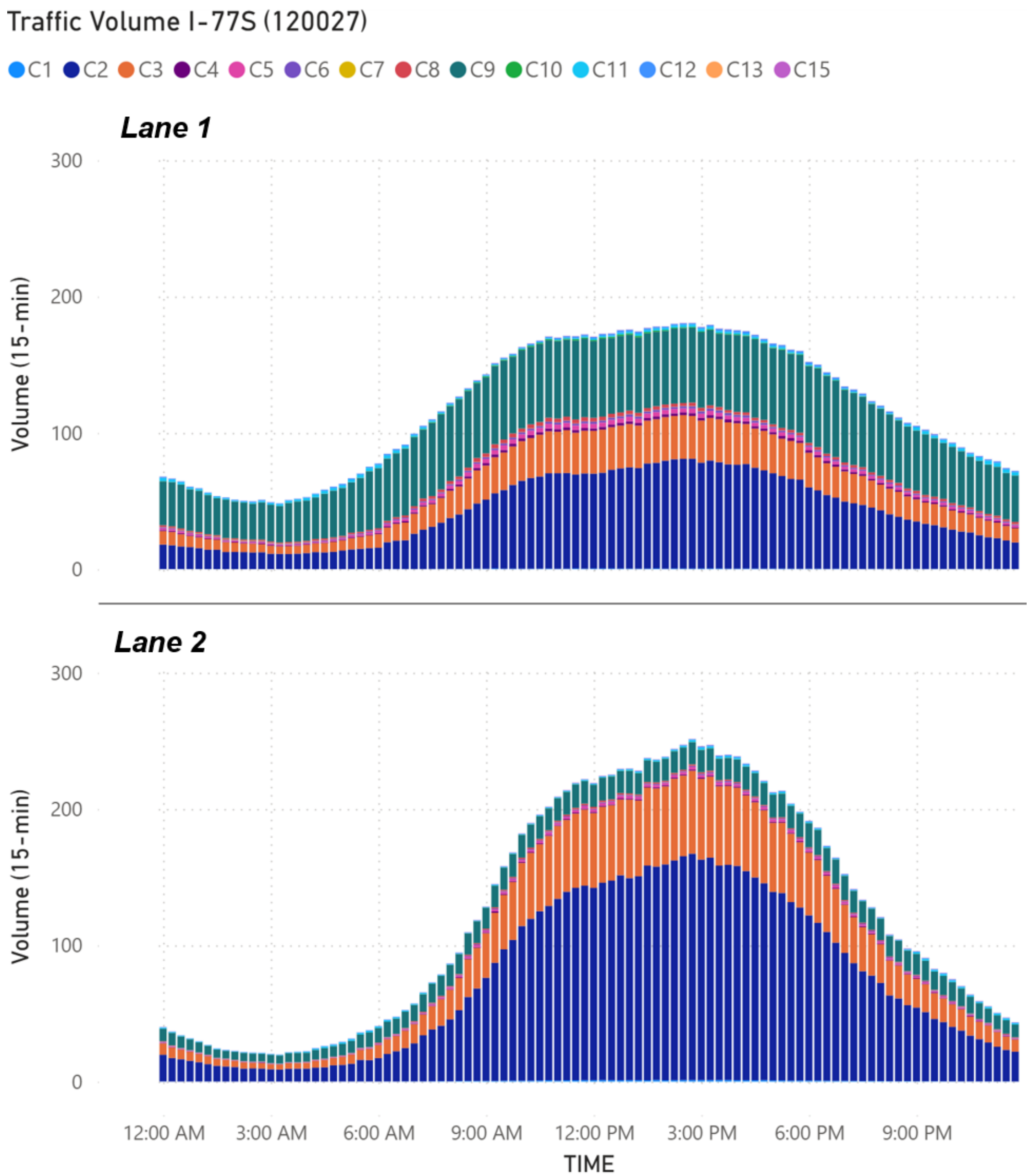


Figure 2. Average Daily 15-Minute Traffic Volume Profile by Class and Lane from May to October 2025. C1 through C13 represent Federal Highway Administration's Vehicle Classes 1 through 13; C15 represents unclassifiable vehicles from the continuous count station. Lane 1 = right lane; Lane 2 = left lane.

The site satisfies heavy vehicle composition requirements through distinct lane utilization patterns. Lane 1 carries a high volume of commercial vehicles, averaging 30% trucks during the daytime and rising to roughly 65% overnight. This concentration continuously challenges test devices to classify complex multi-axle configurations under varying lighting conditions. Conversely, passenger vehicles dominate Lane 2. This asymmetry of heavy vehicles in Lane 1 alongside higher speed passenger traffic in Lane 2 creates an ideal testing ground for evaluating a sensor's ability to resist occlusion and separate adjacent disparate vehicle signatures. In addition, the studied CCS consistently maintains historical Class Quality ratings of 4 and 5, guaranteeing a reliable comparative baseline.

To establish visual ground truth, existing infrastructure supported a dual pan-tilt-zoom camera configuration with supplemental infrared illumination for night operations (Figure 3). This dual-angle approach, utilizing forward (Figure 4) and side (Figure 5) views, was engineered to mitigate truck occlusion, where smaller vehicles are hidden from a single sensor by adjacent heavy trucks.

Figure 3. Camera and Infrared Illumination Mounting Location. PTZ = pan-tilt-zoom camera.

Figure 4. Camera Forward Viewing Angle

Figure 5. Camera Side Viewing Angle

Furthermore, close alignment with an active RWIS station enabled a comprehensive multi-season data collection strategy for inclement weather condition testing. The deployment captured baseline data across distinct temporal and meteorological windows, including a summer and fall monitoring period (May 26 to October 24, 2025) and a winter observation period (November 25, 2025, to February 25, 2026). This winter deployment captured a targeted snow event subset in late January 2026. Leveraging RWIS data, researchers also identified roughly 1,500 fog events with visibility below 600 feet. Capturing this high frequency of degraded visibility and snow provides the precise environmental variance needed to evaluate sensor resilience under adverse weather.

Ground Truth Data Collection

The deployment of the semi-automated vehicle tracking and classification system on the video feeds captured from the study site successfully generated a high-fidelity ground truth dataset. The graphical user interface allowed researchers to overlay the detection tripwires (labeled L1 and L2) directly onto the southbound travel lanes while successfully masking the peripheral infrastructure and opposing traffic utilizing the red exclusion zone functionality (Figure 6).

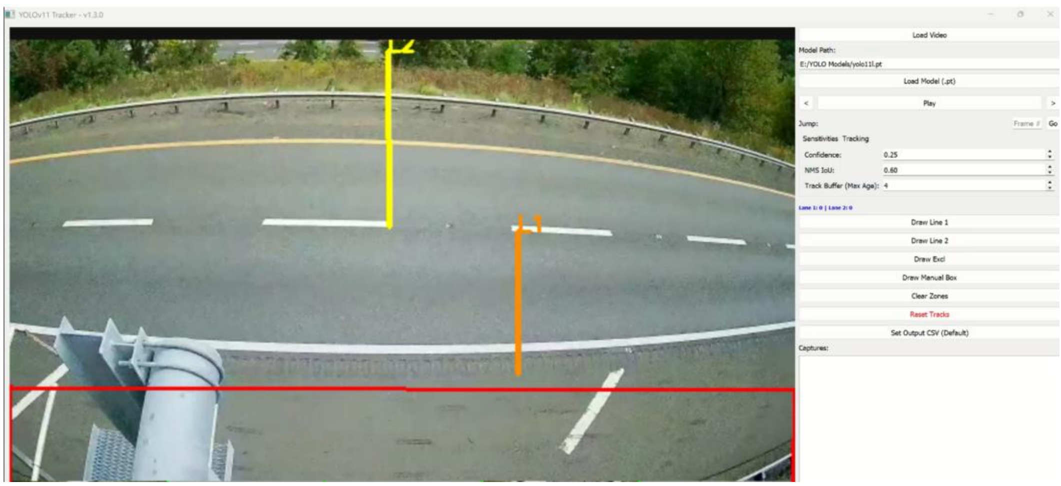


Figure 6. Semi-Automated Vehicle Tracking and Classification System Interface

By processing the continuous footage through this interface, the research team effectively eliminated the baseline classification errors that typically plague fully automated vision systems. The human-in-the-loop validation process ensured that subtle morphological differences, such as distinguishing a heavy Class 5 two-axle, six-tire, single-unit truck from a lighter Class 3 four-tire pickup truck, were accurately captured. When a vehicle triggers the lane tripwire, the system forces a definitive classification via a dedicated user dialog, presenting a cropped night vision or optical capture alongside the FHWA 13-category reference matrix to ensure rigorous adherence to federal reporting standards (Figure 7).

This level of accuracy was mathematically necessary to isolate the error rate of the candidate device from the inherent uncertainty of the baseline. Beyond automatic classification, the software's manual review capabilities proved critical for resolving complex spatial traffic dynamics, notably severe geometric occlusion. For example, when a large commercial tractor-trailer in the right lane physically shadows a smaller passenger vehicle in the passing lane, the automated tracker may temporarily lose the smaller vehicle's silhouette. The human reviewer can intervene using manual override tools to bound and log the obscured vehicle, maintaining the absolute integrity of the classification count.

Figure 7. Human-in-the-Loop Federal Highway Administration Classification Dialog

Similarly, the developed software successfully maintained tracking fidelity during periods of high-density, stop-and-go traffic. Under congested conditions, vehicles follow close headways, frequently causing standard, unsupervised computer vision models to double count them or to merge two distinct vehicles into a single bounding box. The application's playback control and manual capture features allowed analysts to visually separate and individually classify vehicles clustered within dense platoons.

The critical value of this hybrid human-artificial intelligence workflow became most distinctly evident when processing video data captured during inclement weather conditions. The continuous camera feeds successfully captured traffic operations across distinct diurnal cycles, including challenging nighttime snowing conditions. During the intensive snow events in January 2026 and the approximately 88 hours of low-visibility events in which human visual range dropped below the 600-foot threshold, the automated tracker naturally experienced localized degradation due to heavily obscured vehicle silhouettes, severe headlight blooming, and precipitation artifacts on the camera lenses.

However, the program's manual capture override allowed researchers to step in, manually bound the obscured commercial vehicles traversing the corridor in severe weather, and definitively log their exact timestamps and classifications. This hybrid capability ensured that the baseline ground truth remained unbroken and statistically viable even under the most severe atmospheric degradation, providing a solid dataset. Furthermore, the generation of the timestamped comma-separated values, or CSV, dataset and the corresponding cropped vehicle image fulfills the rigorous data quality standards required for an evaluation framework.

Statistical Calibration of Continuous Count Station Baseline

The Python-based Computer Vision tool processed more than 40 hours of video ground truth data, synchronized to 15-minute intervals against the CCS output at the study site. This extraction period was strategically stratified to encompass continuous diurnal cycles and severe meteorological shifts, ensuring at least 4 hours of validated ground truth data were collected for each specific factor combination, such as daytime clear skies, nighttime clear skies, daytime snow, and nighttime snow events. This sampling framework yielded a robust dataset of comparative intervals. As outlined in the methodology, to determine if the baseline accuracy fluctuated significantly under these specific environmental or diurnal conditions, the Kruskal-Wallis H test was applied to the stratified scenarios. The calculated H statistic fell below the critical Chi-square threshold. Thus, the test confirmed that no statistically significant difference appeared in the baseline accuracy among the various scenarios. Because the baseline error remained statistically stable across conditions, the aggregate WAPE was successfully calculated for the three defined vehicle groups to establish the operational noise floor.

The quantitative calibration revealed expected values based on vehicle morphology. Group 1 (Classes 1–3) exhibited a highly accurate baseline WAPE of 1.92%, reflecting the relative consistency with which both CCS and ground truth identify standard light-duty profiles. Group 2 (Classes 4–8) demonstrated a baseline WAPE of 3.4%. The highest baseline variance was observed within the complex freight configurations of Group 3 (Classes 9–13), which yielded a WAPE of 4.7%.

Because this maximum operational noise floor was less than the targeted ideal evaluation threshold of 5%, as stakeholders predominantly requested, it was confirmed as mathematically sound and statistically defensible. Setting the evaluation threshold less than WAPE would risk evaluating the candidate devices against a margin tighter than the baseline uncertainty. In addition, analysis of historical interval data, calculated by comparing CCS volumes to manual video ground truth for each period, established a Temporal Reliability Factor (R_T) of approximately 75%. That is, the CCS stayed within the 5% error threshold for approximately 75% of the time.

Finally, acknowledging the effect of severe atmospheric variables during the proposed Phase III testing discussed in the next section, the ideal threshold was deemed empirically restrictive for adverse environments. Extreme weather conditions, such as nighttime snow, account for only a small proportion of the entire year and are typically accompanied by significantly lower traffic volumes. Because these events are infrequent and affect fewer vehicles overall, accepting a lower accuracy during these periods is operationally practical. Therefore, the necessity for a relaxed adverse weather threshold ($T_{adverse}$) may be established. Rather than prescribing a rigid value at this calibration stage, the framework empowers stakeholders to select an appropriate expanded tolerance (anticipated to be 10% or 15%) during the acceptance procedure development in Task 4. This selectable threshold accommodates the inherent physical limitations of field sensors during severe rain, snow, or fog conditions while still maintaining a sufficiently rigorous boundary to reject candidate devices that experience total classification failure.

Device Acceptance Framework

The research team developed this framework by synthesizing findings from the literature review, incorporating feedback from the technical review panel, and performing statistical analysis of baseline data. The primary objective was to establish a clear and concise step-by-step testing method for VDOT to evaluate proprietary TCDs for approval. The following eight steps, grouped in three phases, outline the procedures for testing. This protocol operates on a fail-fast basis. If a candidate device fails to meet the pass requirement at any individual step, the testing process is terminated, and the device will not move forward to subsequent tests.

Phase I: Volumetric Screening

Step 1: 24-Hour Ideal Condition Screening

- **Condition:** Ensure the weather is clear, and baseline CCS Data Quality Code for all evaluated timestamps is 4 (Acceptable for Use in AADT Calculation) or 5 (Acceptable for all Traffic Monitoring System Use).
- **Action:** Calculate the total volume of vehicles for a full 24-hour period for each specific travel lane and direction. Compare the total volume (both per-direction and per-lane) reported by the test device against the baseline volume.
- **Pass Requirement:** The volume difference between the test device and the baseline must be less than 5%.

Step 2: Diurnal Variation (Light Sensitivity)

- **Condition:** Collect data for 1 full month (30 days). Derive data for dawn, dusk, and night periods using the NOAA Sunrise/Sunset Calculator (https://gml.noaa.gov/grad/solcalc/sunrise.html).
- **Action:**
    - Define *dawn* as the 1-hour period before sunrise, using the NOAA Sunrise/Sunset Calculator to determine the exact time, rounded to the nearest 15 minutes *before* sunrise. For example, if the sunrise is at 5:25 am, round to 5:15 am, making the dawn test period from 4:15 am to 5:15 am.
    - Define *dusk* as the 1-hour period after sunset, using the NOAA Sunrise/Sunset Calculator to determine the exact time, rounded to the nearest 15 minutes *after* sunset. For example, if the sunset is at 6:25 pm, round to 6:30 pm, making the dusk test period from 6:30 pm to 7:30 pm.
    - Define *night* as a 4-hour period per day, selected during the middle of the night (e.g., 10:00 pm to 2:00 am).
    - Calculate the hourly volume differences for these periods between the test device and the baseline.
- **Pass Requirement:** The test device volume must be within 5% of the baseline volume for a preset target number of hours based on the 75% Temporal Reliability Factor (that is, passing 135 out of the 180 tested hours).
- **Note:** If a device needs infrared or extra lighting to pass this test, it is only certified to be used with that specific equipment.

Step 3: High-Volume Test

- **Condition:** Collect data for 30 weekdays, excluding holiday traffic identified by the Traffic Operations Division, and baseline CCS Data Quality Code for all evaluated timestamps is 4 or 5. Identify the 3 hours with the highest traffic volume each day to represent peak-hour traffic, yielding 90 hours of data total.
- **Action:** Calculate the hourly volume (per-direction and per-lane) difference between the test device and the baseline during these specific peak hours.

- **Pass Requirement:** The test device volume must be within 5% of the baseline volume for a set target number of hours based on the 75% Temporal Reliability Factor (that is, passing 67 out of the 90 tested hours).

Phase II: Classification Accuracy

Step 4: 24-Hour Average Class Accuracy

- **Condition:** Ensure the weather is clear, and baseline CCS Data Quality for all evaluated timestamps is 4 or 5.
- **Action:** Calculate the classification accuracy for a 24-hour period for each lane and per direction. Use Equation 7 to find the total error across all 13 standard vehicle categories:

$$Classification\ Error = \frac{\sum_{c=1}^{13} \left| V_{TEST,c} - V_{CCS,c} \right|}{\sum_{c=1}^{13} V_{CCS,c}} \qquad \text{(Equation 7)}$$

- **Pass Requirement:** The calculated Classification Error must be less than 5%.

Step 5: 24-Hour Grouped Class Accuracy

- **Condition:** Ensure the weather is clear, and baseline CCS Data Quality for all evaluated timestamps is 4 or 5.
- **Action:**
  - Combine the FHWA 13 vehicle classes into the three established groups.
    - Group 1: Classes 1–3.
    - Group 2: Classes 4–8.
    - Group 3: Classes 9–13.
  - Calculate the 24-hour volume using Equation 8 for each of these three groups for each lane and per direction to find the group error:

$$Group\ Error_G = \frac{\left| \sum_{c \in G} V_{TEST,c} - \sum_{c \in G} V_{CCS,c} \right|}{\sum_{c \in G} V_{CCS,c}} \qquad \text{(Equation 8)}$$

- **Pass Requirement:** The group error for every group must individually be less than 5%.

Step 6: Grouped Class Accuracy with Diurnal Variation

- **Condition:** Collect data for 1 full month (30 days), and baseline CCS Data Quality Code for all evaluated timestamps is 4 or 5. Derive data for dawn, dusk, and night periods using the NOAA Sunrise/Sunset Calculator (https://gml.noaa.gov/grad/solcalc/sunrise.html).
- **Action:**
  - Determine dawn, dusk, and night evaluation time periods using the same rules as Step 2.
  - Calculate the hourly group error for all three groups for each lane and per direction using Equation 8 in Step 5.
- **Pass Requirement:** The hourly volumes for each group must be within 5% of the baseline volume for a set target number of hours based on the 75% Temporal Reliability Factor (that is, passing 135 out of the 180 tested hours).

Step 7: Grouped Class Accuracy Under High Traffic Volume

- **Condition:** Collect data for 30 weekdays, excluding holiday traffic identified by the Traffic Operations Division, and baseline CCS Data Quality Code for all evaluated timestamps is 4 or 5. Identify the 3 hours with the highest traffic volume each day to represent peak hour traffic, yielding 90 hours of data total.

- **Action:** Calculate the hourly group error for all three groups for each lane and per direction using Equation 8 in Step 5.
- **Pass Requirement:** The hourly volumes for each group must be within 5% of the baseline volume for a set target number of hours based on the 75% Temporal Reliability Factor (that is, passing 67 out of the 90 tested hours).

Phase III: Environmental Resilience

Step 8: Adverse Weather Evaluation

- **Condition:** Identify periods with adverse weather events, including rain heavier than 0.25 inches per hour, fog with visibility below 600 feet, or any measurable snow.
- **Action:**
  - Collect a minimum of 4 hours of data for each specific weather condition. At least 2 of these hours must be during the night.
  - Group the class volume into the three groups and calculate the directional volume.
- **Pass Requirement:** The grouped class volumes must meet the adverse weather threshold ($T_{adverse}$).
- **Note:** Devices should not be penalized if vehicles drift out of their lanes, because adverse weather (especially snow accumulation) often causes natural lane wandering by drivers. Thresholds may be relaxed slightly depending on specific stakeholder needs.

Demonstration of Device Acceptance Framework

The evaluation of Device A demonstrated the practical application of the acceptance framework. The following section details the specific actions the evaluator took at each step of the framework and the corresponding performance of the candidate device. Please note that the quantitative results, pass rates, and performance metrics presented in this section are for demonstration purposes to illustrate the execution of the valuation protocols; they may not exactly reflect the actual empirical data.

Phase I: Volumetric Screening

Step 1: 24-Hour Ideal Condition Screening

On May 1, 2026, the evaluator extracted full 24 hours of data from Device A corresponding to the known clear-weather day (June 1, 2025). The evaluator aggregated the raw timestamped outputs to calculate the total directional volume and the individual lane-specific volumes. On comparing these totals to the CCS baseline, the evaluator calculated a directional volume error of -0.1% and a lane-specific volume error of -0.2%. Because these volumetric differences fell below the 5% threshold, the evaluator certified Step 1 as a “Pass” and advanced the device to Step 2.

Step 2: Diurnal Variation (Light Sensitivity)

The evaluator expanded the queried dataset to encompass a full continuous month of clear and mostly clear weather (June 2025). Using the NOAA Improved Sunrise/Sunset Calculator, the evaluator precisely identified the 1-hour dawn and dusk intervals for each specific day, alongside the 4-hour night periods. After calculating the hourly volume differences between Device A and the baseline, Device A maintained a volume error of less than 5% for 153 out of the 180 tested transitional hours (85%). Because this result exceeded the required 75% Temporal Reliability Factor target of 135 passing hours, the evaluator certified Step 2 as a "Pass."

Step 3: High-Volume Test

The evaluator isolated 30 non-holiday weekdays (from September and October 2025) from the dataset by excluding holiday travel days identified by the Traffic Operations Division. The evaluator identified the 3 hours with the highest traffic volume for each of those days, yielding a testing pool of 90 peak hours. The evaluator calculated the hourly volume differences between the test device and the CCS baseline. Device A maintained its volumetric accuracy within the 5% error threshold for 85 of the 90 tested peak hours (94.4%). This result surpassed the required threshold of 68 passing hours (75% reliability target). Consequently, the evaluator passed Device A through Phase I and advanced it to the Phase II classification tests.

Phase II: Classification Accuracy

Step 4: 24-Hour Average Class Accuracy

Extracting a new 24-hour clear weather dataset from July 12, 2025, the evaluator verified that the baseline CCS reported a Class Quality rating of 4 or 5 for all evaluated timestamps. The evaluator then applied the classification error Equation 7, evaluating Device A across all 13 standard FHWA categories simultaneously. The total absolute classification error across all individual classes was calculated at 3.8%. Because this cumulative error remained under the 5% boundary, the evaluator certified Step 4 as a "Pass."

Step 5: 24-Hour Grouped Class Accuracy

Using the same 24-hour dataset, the evaluator executed the designated grouping strategy, mathematically aggregating the 13 FHWA categories into Group 1 (Classes 1–3), Group 2 (Classes 4–8), and Group 3 (Classes 9–13). Applying the group error Equation 8, the evaluator calculated a directional accuracy error of 2.0% for Group 1, 4.4% for Group 2, and 3.0% for Group 3. The evaluator then repeated this grouping logic on a per-lane basis, yielding similar error distributions. Because the volume error for every single group remained within the required 5% threshold, the evaluator marked Step 5 as a "Pass."

Step 6: Grouped Class Accuracy with Diurnal Variation

The evaluator expanded the queried dataset to encompass a full continuous month of clear or mostly clear weather (September 2025). The observed vehicles were grouped into the three required categories, and the hourly volume difference was calculated for each group for the dawn, dusk, and night periods. Device A met the 5% group classification threshold for 147 out of the 180 tested hours (81.7%). Having met the 135-hour success parameter, the evaluator passed the device to Step 7.

Step 7: Grouped Class Accuracy Under High-Traffic Volume

The evaluator isolated 30 non-holiday weekdays (from September and October 2025) from the dataset by excluding holiday travel days identified by the Traffic Operations Division. The evaluator calculated the hourly volume difference for the three specific vehicle groups. Device A maintained its grouped class volumetric accuracy within the 5% error threshold for 82 of the 90 tested peak hours (91.1%). Having exceeded the 68-hour success target, the evaluator certified Phase II as complete and advanced Device A to the Phase III environmental phase.

Phase III: Environmental Resilience

Step 8: Adverse Weather Evaluation

The evaluator queried the RWIS logs to isolate adverse weather events. Extracting data between November 2025 and February 2026, the evaluator analyzed periods of heavy rain (greater than 0.25 in/hr) and fog (less than 600-feet visibility). Device A maintained group classification errors below 5%.

On extracting 3 hours of data from a nighttime snow event in January 2026, the directional volume error was 12%. The grouped class accuracy errors for Groups 2 and 3 were 15%.

Final Acceptance Test Results

Device A cleared all volumetric and classification protocols outlined in Steps 1 through 7 under ideal, diurnal, and high-volume constraints. However, the final testing of the device stopped entirely on the Step 8 Adverse Weather evaluation. Because Device A exhibited a 12% volume error and a 15% classification error during nighttime snow events, its ultimate pass-fail status is determined by the specific stakeholder threshold selected for adverse conditions ($T_{adverse}$). If the operational stakeholders selected a relaxed adverse weather tolerance of 15% or 20%, Device A is certified. Conversely, if stakeholders mandated a strict 10% maximum error limit, the evaluator must fail the device at Step 8, concluding the evaluation.

Speed Comparison

The aggregated speeds from CCS and Device A, along with the distributions of the two speed datasets, were compared through visualizations and statistical tests. Whether the device was calibrated for speed estimation at the study site is unclear. As a result, the findings may not reflect Device A's performance under optimal configuration. The results presented in this section should therefore be interpreted as exploratory rather than definitive. They apply only to the specific site and time studied and should not be generalized to other locations or operational conditions without further validation.

Comparison of Aggregated Speeds

Figure 8 shows the distribution of bias between Device A and CCS speeds (Device A speed – CCS speed) under clear weather conditions. The mean bias was -11 mph, and distribution was skewed with a long left tail. A Shapiro-Wilk test was conducted to assess normality. The result showed that the bias was not normally distributed (Shapiro-Wilk test statistic = 0.7952, p-value < 0.001). The skewed bias distribution suggested that agreement between CCS and Device A speeds was not consistent across the range of measurements, certain speed ranges or conditions may exhibit larger or more frequent discrepancies, and large discrepancies occurred more frequently than expected under normality, particularly in the negative direction.

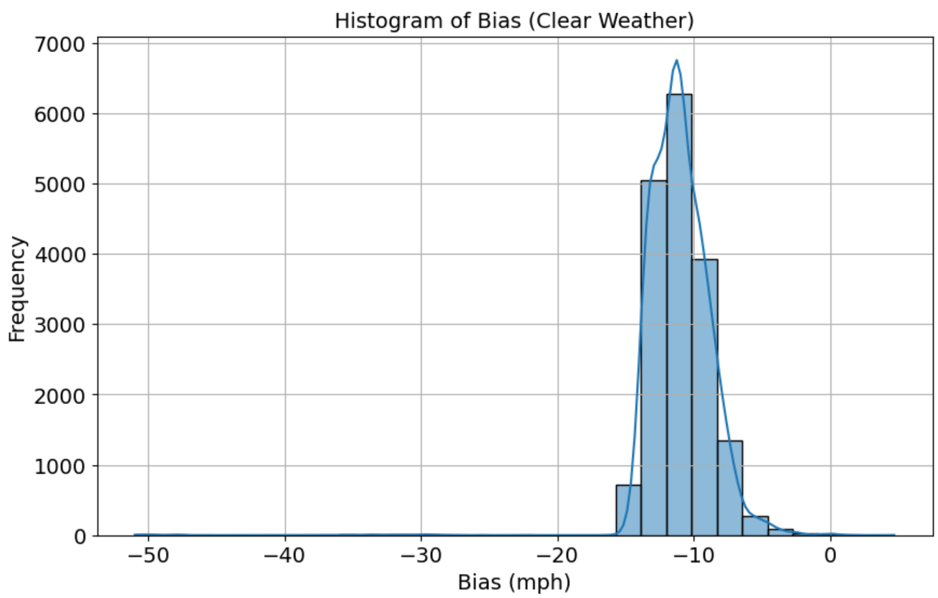
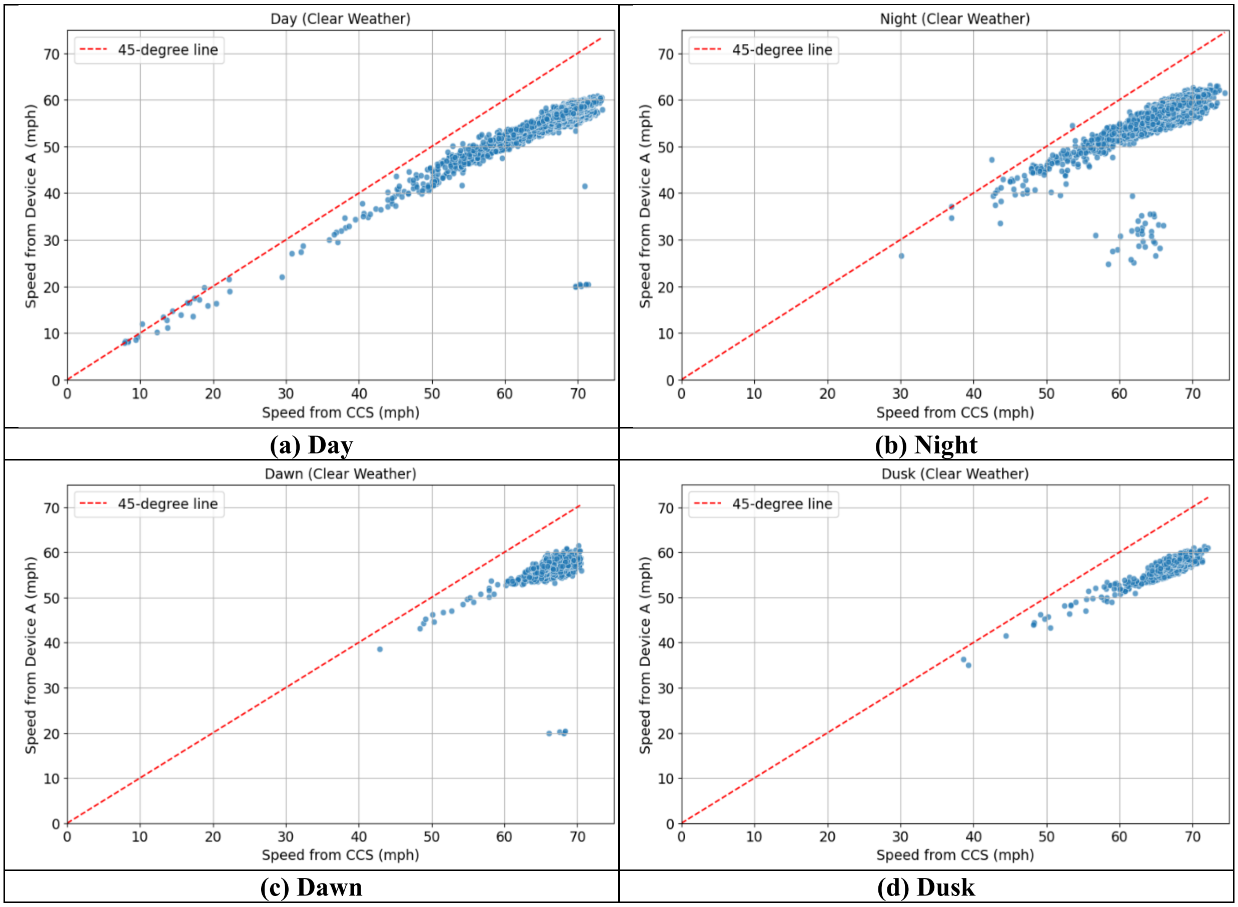


Figure 8. Distribution of Speed Differences Under Clear Weather Conditions

Figure 9 shows scatter plots comparing speeds from CCS and Device A under clear weather conditions. These plots illustrate the magnitude and consistency of the differences between the two speed measurements. Across all times of day (night, dawn, day, and dusk) under clear weather conditions, speeds from Device A were lower than those from CCS, as the data points generally falling below the 45-degree reference line on the plots indicate. Although the overall pattern of Device A reporting lower speeds is consistent, the magnitude of difference varied. The spread of data points around the trend line suggested that the discrepancy might vary by time of day and speed bin. The night period data showed a wider spread of differences for speeds between 60 and 70 mph, indicating greater variability in Device A's performance under those conditions.

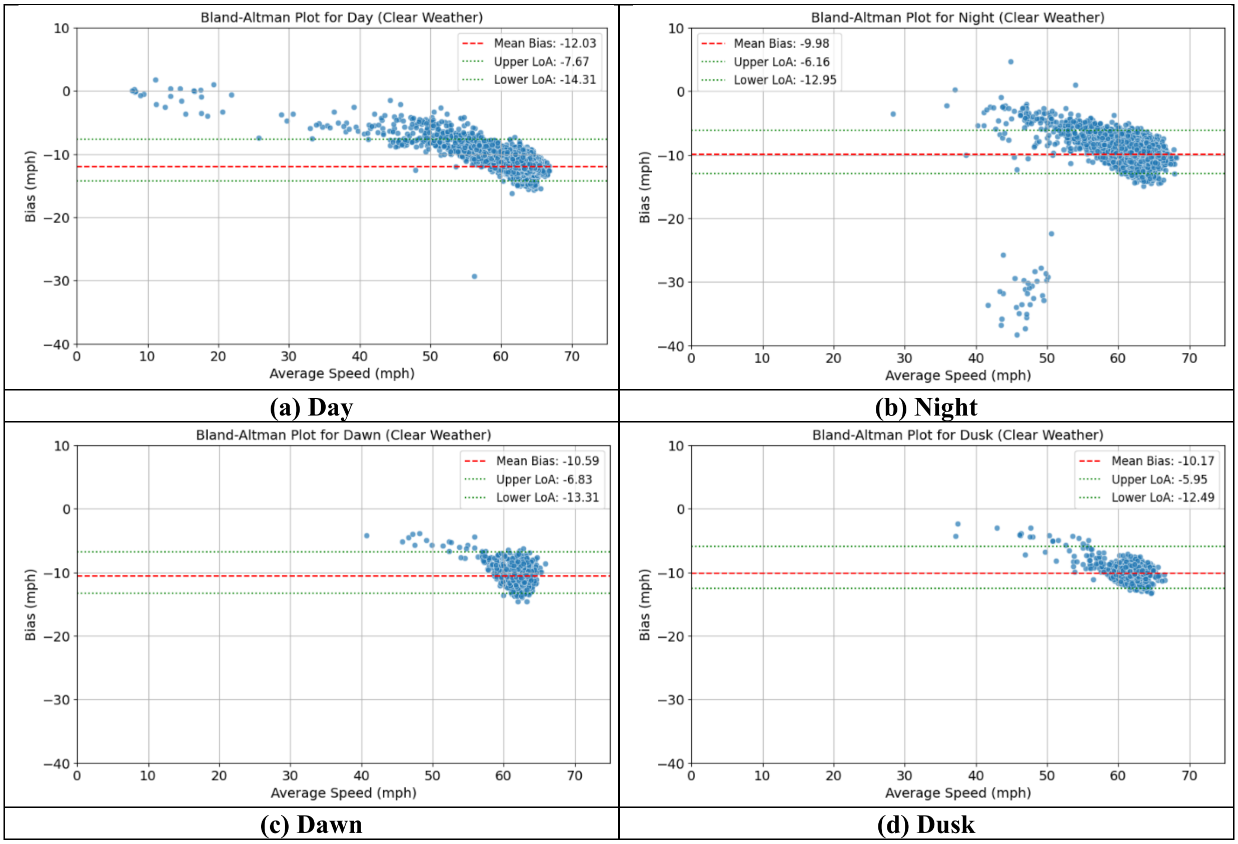


(a) Day **(b) Night**

(c) Dawn **(d) Dusk**

Figure 9. Scatter Plots of Speeds by Time of Day Under Clear Weather Conditions. CCS = continuous count station.

The observation that Device A reported lower speeds than CCS is further supported by the negative bias values shown in the Bland-Altman plots in Figure 10. These plots illustrate the agreement between the two speed datasets by showing the bias on the y-axis against the average of the two measurements on the x-axis. Under clear weather conditions, mean biases ranged from -9.98 mph at night to -14.31 mph during the day, with most bias value points falling below the zero bias line and within the calculated LoA. These patterns confirm that speeds from Device A were consistently lower compared with those from CCS across all times of day. For the daytime period, the mean bias was -12.03 mph, indicating that speeds from Device A were 12.03 mph lower on average than those from CCS. The corresponding upper and lower LoA showed that 95% of the speed difference fell between -7.67 and -14.31 mph. For all times of day, the LoA ranges were around 6.5 mph wide, reflecting moderate variability. The magnitude of biases generally increased as the average of CCS and Device A speed increased. During the night period, some large biases exceeding -30 mph were observed for average speeds between 40 and 50 mph, which may indicate that the reduced lighting affected the sensing accuracy of Device A.



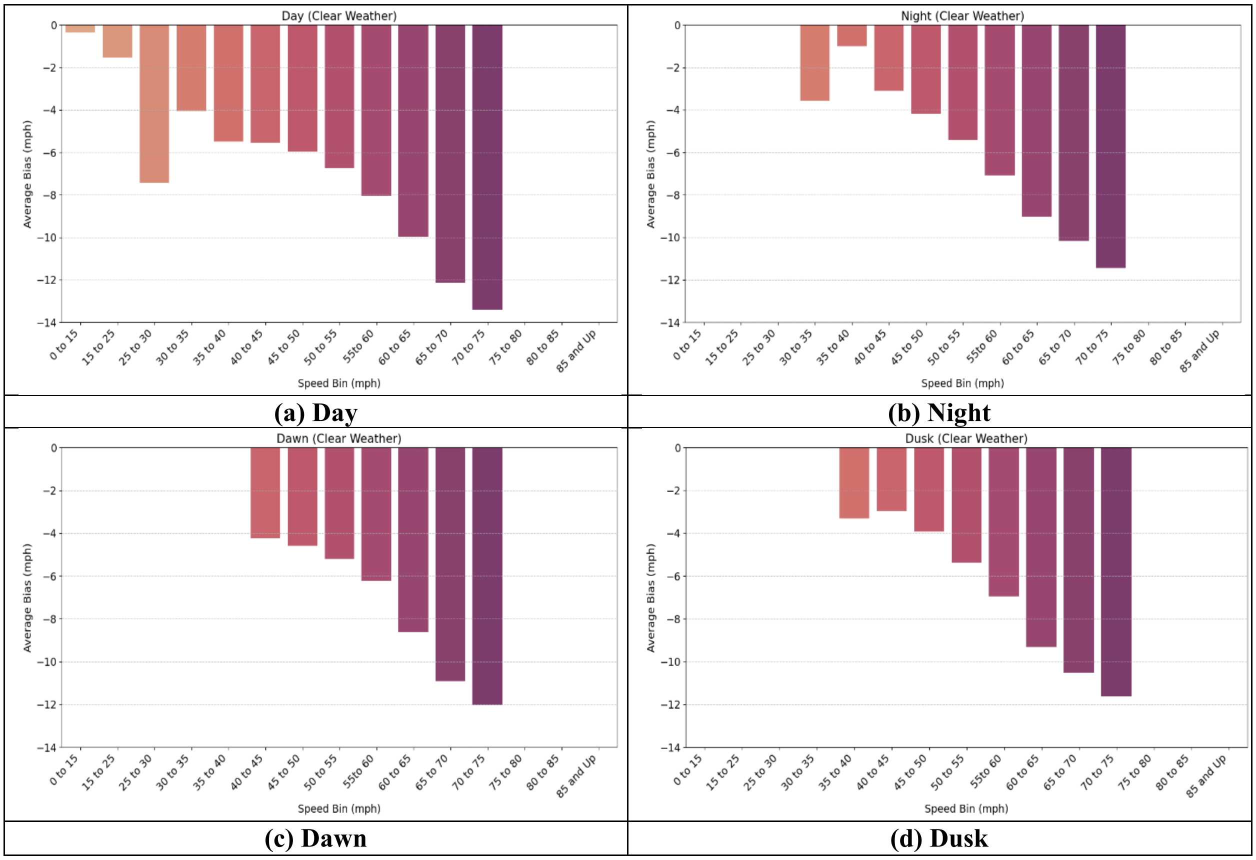
(a) Day **(b) Night**

(c) Dawn **(d) Dusk**

Figure 10. Bland-Altman Plots by Time of Day Under Clear Weather Conditions. LoA = limits of agreement.

Figure 11 illustrates how the average bias varies across different speed bins for each time of day, with speed bins defined using CCS speeds. Consistent with the Bland-Altman results in Figure 10, the average bias generally became more negative as CCS speeds increased, with only a few deviations. The largest average bias, -12.81 mph, occurred in the 70–75 mph bin during the day. Across all time periods, the magnitude of bias exceeded 10 mph for baseline speeds of 65 mph or higher, indicating increasingly underestimation by Device A at higher speeds. The magnitude of bias was relatively smaller during the night and dusk periods compared with the day and dawn periods, possibly because traffic speed was relatively lower during the night and dusk periods.

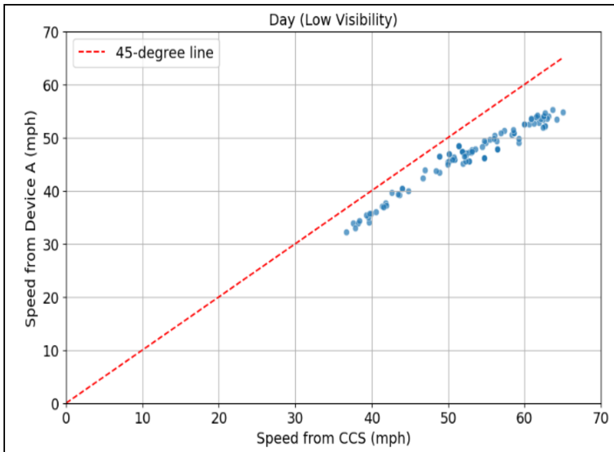
Figures 12 and 13 show scatter plots and Bland-Altman plots for low visibility and snow conditions, respectively. These analyses were conducted only for day and night periods because the sample sizes for dawn and dusk periods were limited (less than 20 samples). Similar to the patterns observed under weather, Device A consistently reported lower speeds than CCS in both conditions, as evidenced by data points falling below the 45-degree reference line in the scatter plots and by the negative bias and LoA in the Bland-Altman plots. Under low-visibility conditions, the average bias was -6.11 mph during the day and -4.31 mph at night. Under snow conditions, the average bias was -4.46 mph during the day and -3.83 mph at night. These biases and their corresponding LoA were smaller in magnitude than those observed under clear weather, probably because of lower traffic speeds in these conditions, particularly at night, when most speeds were below 60 mph. During the night period specifically, Device A underestimated speeds by an average of 4.31 mph in low-visibility (fog) conditions and 3.83 mph in snow, indicating more moderate differences between the CCS and Device A speeds in adverse weather compared with clear conditions. At the study site, speeds from Device A for vehicles in the left lane generally showed a higher bias than those in the right lane. Under low-visibility or snow conditions, more vehicles tended to use the right lane, which may partially explain the relatively lower average speed differences observed.



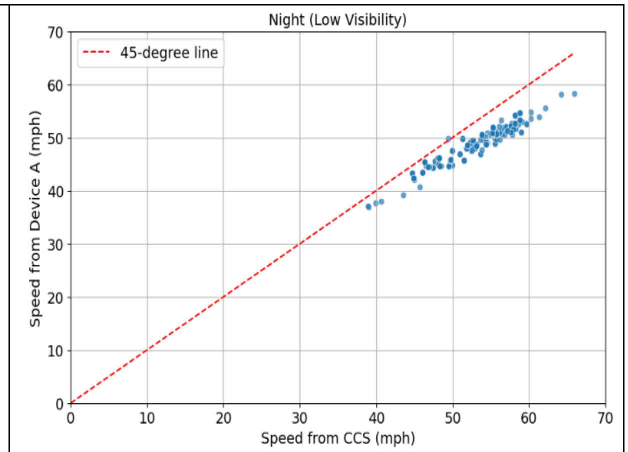
(a) Day | (b) Night

(c) Dawn | (d) Dusk

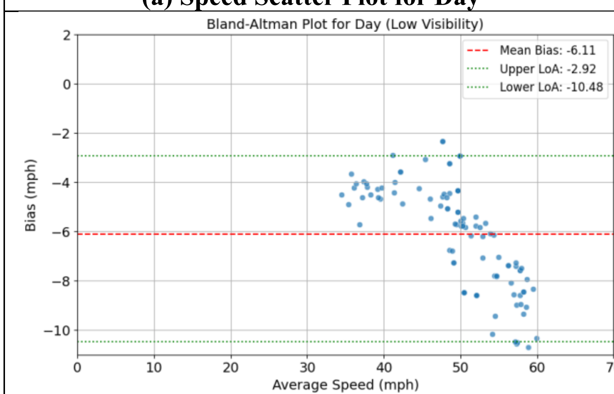
Figure 11. Average Bias by Speed Bin and Time of Day Under Clear Weather Conditions



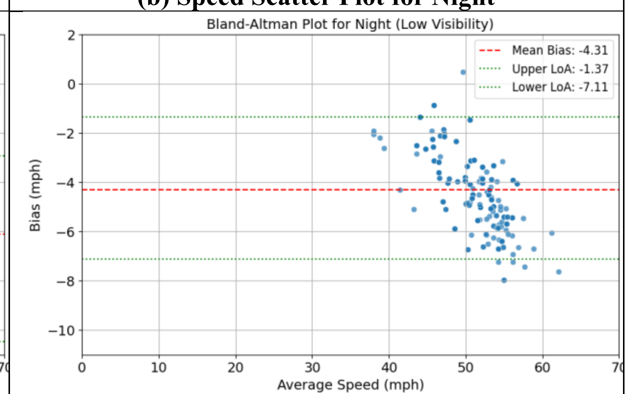
(a) Speed Scatter Plot for Day



(b) Speed Scatter Plot for Night

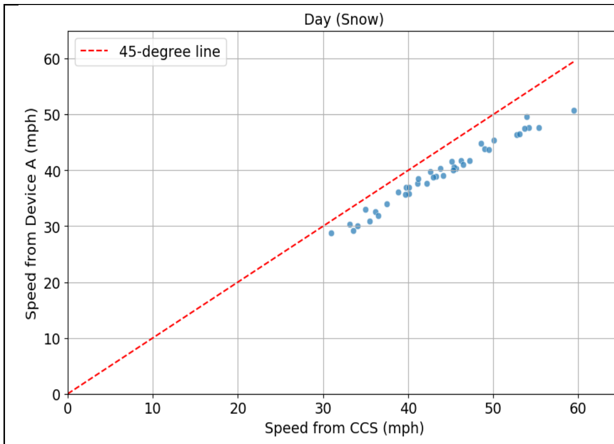


(c) Bland-Altman Plot for Day

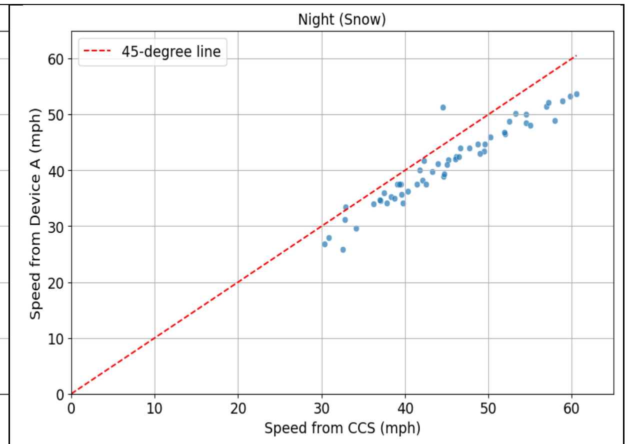


(d) Bland-Altman Plot for Night

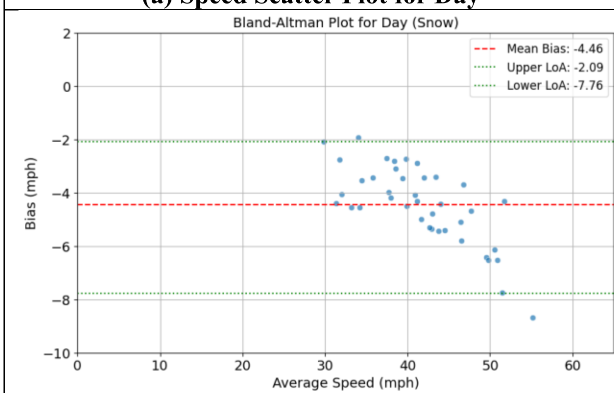
Figure 12. Scatter Plots of Speeds by Time of Day Under Low Visibility Conditions. LoA = limits of agreement.



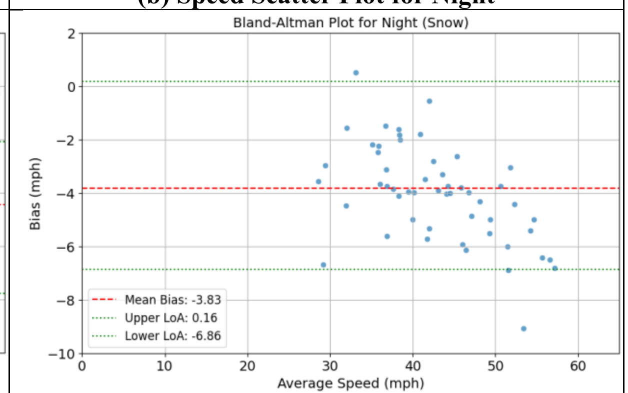
(a) Speed Scatter Plot for Day



(b) Speed Scatter Plot for Night



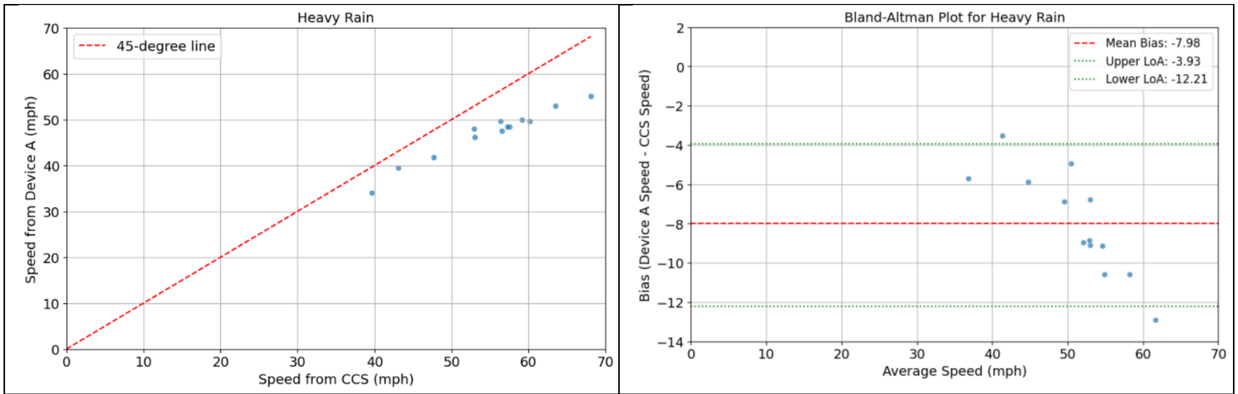
(c) Bland-Altman Plot for Day



(d) Bland-Altman Plot for Night

Figure 13. Scatter Plots of Speeds and Band-Altman Plots by Time of Day Under Snow Conditions. CCS = continuous count station. LoA = limits of agreement.

Because only 13 samples were available for the heavy-rain condition, the speed scatter plot and Bland-Altman plot in Figure 14 were generated using data from all time periods combined. In every case, Device A reported lower speeds than the corresponding CCS measurements. The average bias was -7.98 mph, and the LoA indicated that 95% of the bias was expected to fall between -12.21 and -3.93 mph. However, due to the small sample size, both the estimated bias and the LoA should be interpreted with caution, because they may be unstable and may not fully represent Device A's performance under heavy-rain conditions.



(a) Speed Scatter Plot | **(b) Bland-Altman Plot**

Figure 14. Scatter Plot of Speeds and Bland-Altman Plot for Heavy Rain Condition. CCS = continuous count station. LoA = limits of agreement.

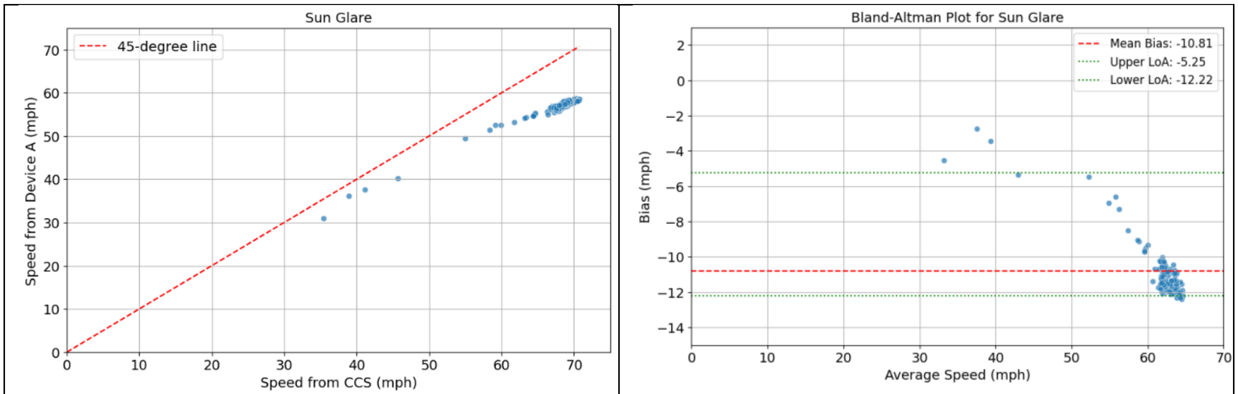
Table 2 summarizes the results of the two-tailed paired *t*-tests and Bland-Altman analyses for all weather conditions. In all the *t*-tests, the null hypothesis was rejected at the 0.001 significance level, indicating that speeds measured by Device A and CCS differ statistically across all tested weather and time of day conditions. The consistently negative *t*-statistics, *p*-values near zero, and negative mean biases all point to Device A systematic underestimating speed. The largest mean percentage difference, 17.66%, occurred during the day period in clear weather, whereas the smallest percentage difference, 9.03%, was observed at night in snow conditions. These results confirm that the underestimation of speeds by Device A was persistent but varied in magnitude, depending on weather and time-of-day conditions.

Table 2. Two-Tailed Paired *t*-Tests and the Bland-Altman Analyses Results

Weather	Time of Day	Sample Size	*T*-Statistic	*P*-Value	Mean Bias	MAPD	Lower LoA	Upper LoA
Clear	Day	8,527	– 524.97	0.0000	– 12.03	17.66	– 14.13	– 7.67
	Night	7,554	– 383.35	0.0000	– 9.98	14.9	– 12.95	– 6.16
	Dawn	929	– 100.02	0.0000	– 10.59	15.83	– 16.91	– 4.26
	Dusk	739	– 174.07	0.0000	– 10.17	15.26	– 12.49	– 7.06
Low Visibility	Day	102	– 29.17	0.0000	– 6.11	11.41	– 10.48	– 2.92
	Night	218	– 41.24	0.0000	– 4.31	7.99	– 7.11	– 1.37
	Dawn	14			– 4.89	8.54	– 6.40	– 3.22
	Dusk	18			– 2.69	6.09	– 3.94	– 1.76
Snow	Day	40	– 19.19	0.0000	– 4.46	10.03	– 7.76	– 2.09
	Night	54	– 12.19	0.0000	– 3.83	9.03	– 6.86	0.16
	Dawn	5			– 3.99	9.93	– 5.51	– 2.56
	Dusk	1			– 4.1	10.44		
Heavy Rain	Day	7			– 9.32	16.21	– 12.56	– 6.18
	Night	4			– 6.22	11.6	– 10.23	– 3.62
	Dawn	0						
	Dusk	2			– 6.83	10.44	– 6.88	– 6.77

LoA= limit of agreement. MAPD = mean absolute percentage difference.

To evaluate the effect of sun glare on Device A's speed estimation performance, data collected between 8:00 and 9:00 am in January 2026 under clear weather conditions were analyzed. A total of 112 samples were included. Figure 15 presents the corresponding speed scatter plots and Bland-Altman results. As observed in all other conditions, Device A consistently reported lower speeds than CCS. The average bias was -10.81 mph, which is approximately 2 mph smaller in magnitude than the average bias calculated for all day hours under clear weather. The LoA ranged from -12.22 to -5.25 mph, narrower than that for all daytime hours (-14.13 to -7.67 mph). Most bias values fell between -12.22 and -10.81 mph for average CCS and Device A speeds in the 60–70 mph range, which was consistent with the pattern observed for all daytime hours. These results suggest that sun glare did not meaningfully degrade Device A's speed measurement beyond what was typically observed under clear daytime conditions.



(a) Speed Scatter Plot	**(b) Bland-Altman Plot**

Figure 15. Scatter Plot of Speeds and Bland-Altman Plot for Sun Glare Condition. CCS = continuous count station. LoA = limits of agreement.

In summary, Device A consistently underestimated vehicle speeds relative to the CCS system, regardless of the time of day and weather conditions. This systematic underreporting was evident across all analyses, including scatter plots, paired *t*-tests, and Bland-Altman assessments. Although the magnitude of the bias varies somewhat by speed range and lighting and visibility conditions, the direction of the difference remained consistent. These findings indicate a persistent measurement bias in Device A that should be considered when interpreting or applying its speed measurements in traffic speed analyses.

Comparison of Speed Distributions

The distributions of speeds from Device A and CCS were compared across different times of day using data from all weather conditions. Figure 16 shows the cumulative distribution functions (CDFs) for the day, night, dawn, and dusk periods. Clear separation is visible between the CDFs for CCS and Device A for all time periods, highlighting the inconsistencies between the two sets of speed distributions across all time-of-day conditions. The CCS curves consistently lie above those of Device A, indicating that Device A reported lower speeds across the full distribution. The largest overall differences between the two sets of speed distributions were observed during the day period. For CCS, the 50th, 80th, 85th, and 95th percentile speeds were highest during the day, consistent with generally higher traffic speeds during that time. In contrast, the CDFs for Device A showed that the 50th, 80th, 85th, and 95th percentile speeds in night periods were generally higher than those in other periods.

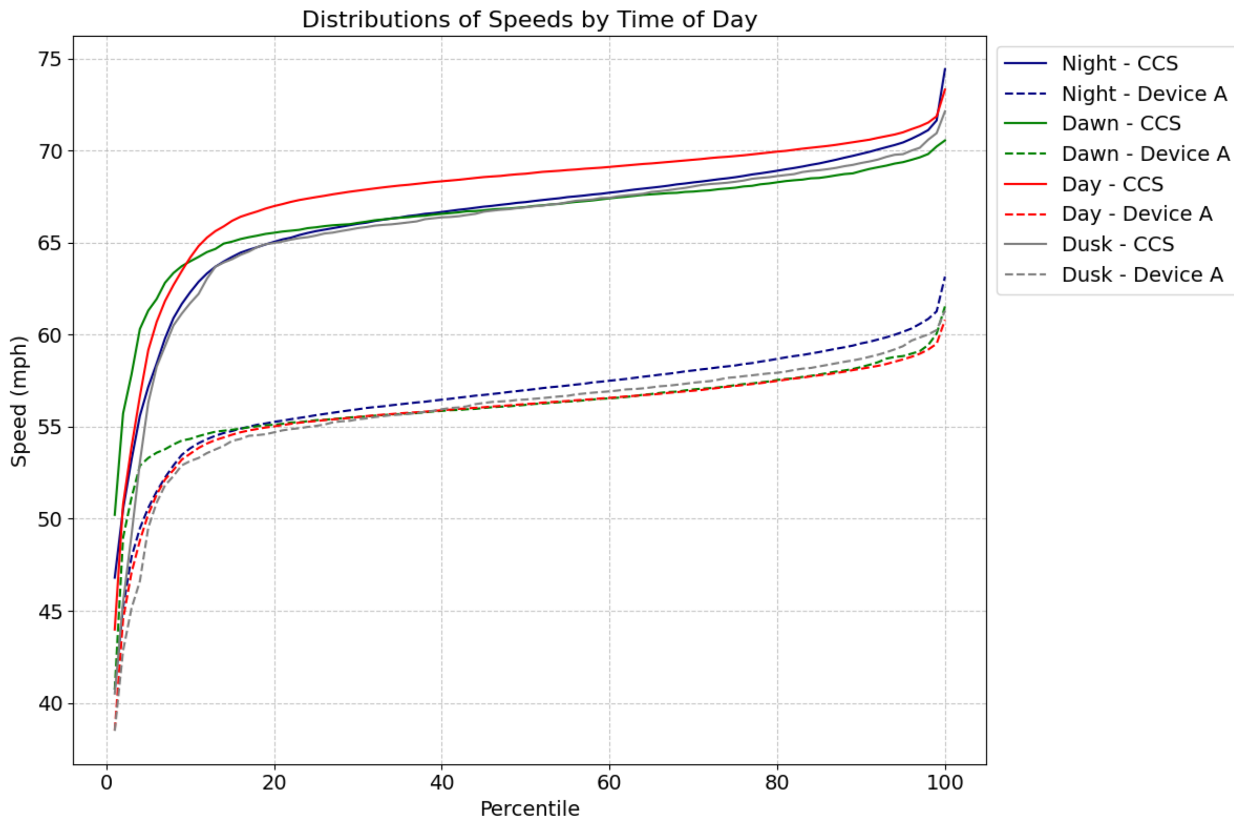


Figure 16. Speed Distributions by Time of Day. CCS = continuous count station.

Table 3 summarizes the K-S test results and the Wasserstein distances comparing the percentile distributions of speeds from Device A and CCS. For all time-of-day categories, the K-S tests produced very large statistics (close to 1) and extremely small *p*-values, indicating that the distributions of speeds from Device A and CCS were statistically different across all tested conditions. The Wasserstein distances, which quantify the "cost" to transform one distribution into the other, ranged from 9.80 to 11.95 mph. These distances capture the shifts in location, spread, and overall shape of the distributions, with a larger value suggesting a greater difference between the distributions. The highest Wasserstein distance (11.95) was observed during the day period, indicating the most significant difference in speed distributions compared with all other periods. A simplified interpretation is that values in Device A speed distribution are about 11 mph below on average from where they need to be to match the CCS speed distribution. These K-S tests and Wasserstein distance results demonstrated that speeds from Device A did not accurately replicate the full speed distribution measured by CCS, particularly during the day period.

Differences in traffic patterns, roadway geometry, environmental factors, and sensor deployment configurations may lead to substantially different performance outcomes. In addition, sensor technologies evolve over time. As such, these findings regarding speed measurements pertain solely to the I-77 site and time studied and should not be considered definitive.

Table 3. K-S Test Results and Wasserstein Distances

Time of Day	*K-S* Statistic	*P*-Value	Wasserstein Distance
Day	0.94	0.0000	11.95
Night	0.91	0.0000	9.80
Dawn	0.96	0.0000	10.39
Dusk	0.92	0.0000	10.02

K-S = Kolmogorov-Smirnov.

Framework Applicability

The TCD Acceptance Framework is designed for broad operational deployment by aggregating the 13 FHWA classes into three operational groups. This grouping prevents extreme statistical variance caused by low-frequency classes, making the framework viable. However, agencies requiring accurate classification for each of the FHWA 13 classes should supplement this evaluation with class-specific accuracy thresholds to meet their needs.

In addition, to ensure practical scalability, the framework evaluates candidate devices against baseline CCS rather than intensive human-verified video ground truth. This approach efficiently verified whether a candidate device performs at least as accurately as the baseline. Although highly practical, this relative comparison introduces a tradeoff regarding compounding variance. For example, if the baseline CCS undercounts a vehicle group by 3% relative to ground truth, and the candidate device undercounts by an additional 3% relative to the CCS, then the device passes the 5% framework threshold despite a potential 6% deviation from the ground truth. Agencies must account for this relative accuracy baseline when interpreting acceptance test results for highly sensitive data applications.

CONCLUSIONS

- *A 5% error threshold balances the precision required for VDOT's engineering and planning applications with the practical realities of field sensor deployment*. Based on the results of the literature review, user needs survey, and stakeholder discussions, it was determined that this 5% margin satisfies the diverse data requirements of VDOT stakeholders. Consequently, this threshold was established as the primary accuracy target for all volumetric and classification evaluations.

- *The developed acceptance framework with "fail-fast" logic provides a statistically sound and resource-efficient mechanism to certify emerging traffic count devices*. To create a defensible testing baseline, CCSs were verified against a rigorous human-in-the-loop video ground truth. This verification provided that aggregating the 13 standard FHWA vehicle classes into three broader operational groups maintains statistical robustness during testing. By requiring candidate devices to progressively pass volumetric screenings, classification tests, and adverse weather evaluations and immediately terminating the test if a device fails to meet the threshold at any step, the framework prevents wasted personnel hours on underperforming hardware.

RECOMMENDATION

1. *VDOT's Traffic Operations Division should develop and implement a standardized traffic count device acceptance procedure, utilizing the framework established in this study as its foundation*. By adopting the eight-step, "fail-fast" evaluation protocol and the vehicle grouping strategy demonstrated in this study, TOD can develop a consistent resource-efficient vetting process for emerging non-intrusive traffic count sensors for VDOT approval. This established framework can also provide vendors with clear, statistically defensible performance expectations

IMPLEMENTATION AND BENEFITS

The researcher and the technical review panel (listed in the Acknowledgments) for the project collaborate to craft a plan to implement the study recommendations and determine the benefits of doing so. This process is to ensure that the implementation plan is developed and approved with the participation and support of those involved with VDOT operations. The implementation plan and the accompanying benefits are provided here.

Implementation

Regarding Recommendation 1, within 2 years of the publication of this report, the Assistant Division Administrator for Operations Planning and Performance at the Traffic Operations Division will organize a technical meeting with relevant stakeholders to review their current traffic count device acceptance process and incorporate the framework proposed in this study into VDOT's Traffic Monitoring Collection Device Product Evaluation Guide. The effort will focus on developing a standardized acceptance procedure for evaluating emerging non-intrusive sensors for VDOT approval. Topics to be discussed may include the integration of the eight-step evaluation framework, the adoption of the three-tier vehicle class grouping strategy, and the establishment of specific acceptable tolerances for adverse weather scenarios.

Benefits

Traditionally, to achieve a statistically defensible evaluation, manual ground truth data collection and processing are required. The process includes deploying field cameras, capturing continuous video, and paying human analysts to manually review the footage to count and classify every vehicle. By using the CCS baseline data against which new devices are measured, VDOT eliminates the need for expensive, third-party video data processing during standard device certifications.

Based on a recent quote derived from a third-party data processing vendor, manually processing count and classification video data costs approximately $21,000 per 20 hours of footage, which equates to a rate of $1,050 per hour.

If VDOT were to execute the developed Device Acceptance Framework, a candidate device must undergo 306 hours of distinct evaluation: volumetric and classification screening (24 hours), diurnal variation testing (180 hours), high-volume conditions (90 hours), and adverse weather resilience (12 hours). This evaluation translates into $321,300 of ground truth data collection. Although a full certification cycle represents a maximum cost of $321,300, the developed "fail-fast" logic disqualifies underperforming devices at early stages, and the full expenditure is reserved only for those devices that successfully navigate the entire certification process. Please note this figure represents an ideal scenario for a full certification cycle; actual expenditures may be higher because of the operational challenges of capturing specific conditions, such as the required 12 hours of adverse weather.

The proposed acceptance process provides VDOT with a "fail-fast" mechanism, reducing the effort required to evaluate an underperforming device. The standardized framework establishes a transparent and predictable environment for the industry. Both VDOT and its vendors know exactly what performance metrics are required for certification. This clarity reduces the time and cost associated with ambiguous trial-and error deployment, ensuring a level playing field for all providers.

ACKNOWLEDGMENTS

The authors are grateful to the following individuals who served on the technical review panel for this study: Ron Hobson, Project Champion and Program Manager for Traffic Monitoring Construction & Contract Management, Traffic Operations Division; Justice Appiah, Associate Principal Research Scientist, Virginia Transportation Research Council; Mena Lockwood, Assistant State Traffic Operations Engineer, Traffic Operations Division, VDOT; Chris Medina, Program Admin Specialist III, Traffic Operations Division, VDOT; and Hamlin Williams, Traffic Monitoring System Program Manager, Traffic Operations Division, VDOT. The authors also appreciate the assistance of Marielle Bocaner of the University of Virginia for her work in data reduction from video files.

REFERENCES

ASTM International. *ASTM E2532-09(2024): Standard Test Methods for Evaluating Performance of Highway Traffic Monitoring Devices*. West Conshohocken, PA, 2024.

Federal Highway Administration. *A Summary of Vehicle Detection and Surveillance Technologies used in ITS*. Washington, DC, 2000. https://www.fhwa.dot.gov/ohim/tvtw/vdstits.pdf.

Federal Highway Administration. *Traffic Detector Handbook: Third Edition—Volume II*. FHWA-HRT-06-139. Washington, DC, 2006.

Federal Highway Administration. *Traffic Monitoring: A Guidebook*. FHWA-WFL/TD-10-002. Washington, DC, 2010.

Federal Highway Administration. *Traffic Monitoring Guide*. Washington, DC, 2022.

Florida Department of Transportation. Product Compliance Matrices and Supplemental Requirements, 2025. https://www.fdot.gov/traffic/traf-sys/product-specifications.shtm. Accessed May 12, 2026.

Georgia Department of Transportation. *Detection Systems Standard Operating Guide – Field Evaluation Process for Traffic Signal and TIS Vehicle Detection Systems*. Atlanta, GA, 2022a. https://www.dot.ga.gov/PartnerSmart/Materials/Documents/qplproducts/Detection_Systems_SOG.pdf.

Georgia Department of Transportation. *Detection Systems Standard Operating Guide – Field Evaluation Process for Traffic Signal and ITS Vehicle Detection Systems*. Atlanta, GA, 2022b.

Georgia Department of Transportation. Supplemental Specification Section 937 – Detection Systems, 2025.

Wisconsin Department of Transportation. *Traffic Engineering, Operations & Safety Manual*. Madison, WI, 2025. https://wisconsindot.gov/Pages/doing-bus/local-gov/traffic-ops/manuals-and-standards/teops/default.aspx.

APPENDIX

Understanding the Use of Traffic Volume Data at VDOT

Introduction

The Virginia Transportation Research Council (VTRC) is conducting a research project titled "Development of an Evaluation and Acceptance Process for Traffic Data Collection Devices." The goal of this project is to establish a standard verification and acceptance procedure for all traffic data collection devices seeking approval for use by the Virginia Department of Transportation (VDOT). To support this effort, we seek to better understand how VDOT utilizes data products from traffic data collection devices, including **traffic volume** and **vehicle classification**. We invite stakeholders to participate in this survey by answering the following questions. Your input will help shape the evaluation and acceptance framework for these devices.

Section A — About You

1. **District / Division: _____________**
2. **Job Title: ______________**
3. **How often do you work with traffic volume/classification data?** (Daily; Weekly; Monthly; Ad hoc)

Section B — Where you access data

4. **Traffic volume data sources you use** (check all that apply):
   • VDOT's TMS Database
   • TOD's PowerBI (Volume)
   • TransportationDataHub
   • On-Call Contractual Resources (special study)
   • Other (please specify)

5. **Vehicle classification (count) data sources you use** (check all that apply):
   • VDOT's TMS Database
   • TOD's PowerBI (Class)
   • TransportationDataHub
   • On-Call Contractual Resources (special study)
   • Other (please specify)

Section C — Key definitions

To keep answers consistent across districts, systems, and devices, we're standardizing a few key terms up front. VDOT teams often use similar words slightly differently (e.g., "directional volume" vs. "link volume"), and vendors/tools may aggregate or classify data in their own ways. The following definitions aims to ensure that everyone is answering with the same ground.

- **Spatial granularity:**
  • *By link (both directions)*; *By direction*; *By lane*; *Per-vehicle record (PVR)* (i.e., raw per-vehicle observations).
- **Temporal aggregation:** 1-min; 5-min; 15-min; 30-min; 60-min; daily; other (text).

- **Vehicle classes:** “All vehicles (combined)”, *FHWA 13-class*, or *User-defined* groupings.
- **Accuracy expectations (e.g., % within tolerance, MAPE):** target overall accuracy 99%, 95%, 90%, 85%, or specify in a common language.

Please keep these definitions in mind when answering Section D (per-application needs)

Section D — Applications detail (repeat block per application)
Instruction: *Please add one entry for each distinct application where you use traffic volumes/classification (e.g., signal timing, WZ planning, safety analysis, design volumes/AADT, HPMS/FHWA reporting, corridor performance, detour planning, freight planning, program reporting, other).*
(All fields below refer to this application only.)

6. **Application name / description** (short text)
7. **Which data do you need for this application?**
   (check all) Traffic volume; Vehicle classification; Both
8. **Spatial granularity needed** (check all that apply; state your **preferred** one in Others):
   □ By link (both directions) □ By direction □ By lane □ Per-vehicle record (PVR) □ Other: ______
9. **Temporal aggregation needed** (check all that apply; state your **preferred** one in Others):
   □ 1-min □ 5-min □ 15-min □ 30-min □ 60-min □ Daily □ Other: ______

10. **Vehicle classification for this application**
    ○ Not needed (use all vehicles combined)
    ○ FHWA 13-class
    ○ User-defined grouping (please list the classes and why you group them that way)

11. **Accuracy target for this application**
    a) Target level (choose one or specify): 99% / 95% / 90% / 85% / 80% / Other: ____ %
    b) **Metric used** (choose one): % within tolerance; MAPE; RMSE; Other (specify)
    c) **If tolerance-based**, what tolerance? (e.g., within ±5% of ground truth at 15-min)

12. **For classification (if used): accuracy requirements by grouping**
    • Overall class accuracy target: ____ %
    • Which classes or groups are most critical (e.g., buses, heavy trucks), and their targets? (text)

13. **Typical data sources you rely on for this application** (check all):
    □ TMS DB □ TOD PowerBI □ TransportationDataHub □ On-Call / study □ Other

14. **Conditions that matter for acceptance** (check all):
    □ Peak vs off-peak □ Congestion/queues □ Work zones □ Weather (rain/fog/snow)
    □ Nighttime/low-light □ Mixed facility types (urban freeway/arterial) □ Other: _____

15. **If your preferred granularity isn’t available, what’s acceptable?**
    (e.g., lane→direction; 1-min→15-min; FHWA-13→user group; none)

16. **Additional details for this application** (open text)

Section E — Global practices & preferences

17. **Do you commonly adjust FHWA 13 classes?**
    ○ No, we use FHWA-13 as is
    ○ Yes → *Describe your grouping and rationale* (examples encouraged)

18. **Anything else we should know?** (open text)