Validation of Non-Traditional Approaches to Annual Average Daily Traffic (AADT) Volume Estimation

Publication No. FHWA-PL-21-033

September 2021

U.S. Department of Transportation
Federal Highway Administration

Notice

This document is disseminated under the sponsorship of the United States Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. This report does not constitute a standard, specification, or regulation.

The United States Government does not endorse products or manufacturers. Trade and manufacturers' names appear in this report only because they are considered essential to the object of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-PL-21-033		

4. Title and Subtitle	5. Report Date
Validation of Non-Traditional Approaches to Annual Average Daily Traffic (AADT) Volume Estimation	September 2021
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Joseph Fish, Stanley Young, Alana Wilson, Brennan Borlaug (NREL)	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
National Renewable Energy Laboratory 15013 Denver West Parkway Golden, CO 80401	11. Contract or Grant No.
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Highway Administration Office of Highway Policy Information 1200 New Jersey Avenue SE Washington, DC 20590	14. Sponsoring Agency Code

15. Supplementary Notes

Project performed in cooperation with the U.S. Department of Transportation, Federal Highway Administration, and a Technical Advisory Committee (TAC) consisting of Alaska DOT & Public Facilities, Caltrans, Colorado DOT, Idaho DOT, Illinois DOT, Georgia DOT, Maryland DOT, Minnesota DOT, Nebraska DOT, New Jersey DOT, North Carolina DOT, North Dakota DOT, Ohio DOT, Oregon DOT, Pennsylvania DOT, South Carolina DOT, Texas DOT, and Virginia DOT. Steven Jessberger served as FHWA's Task Manager.

16. Abstract

For this validation study, NREL compiled traffic volume data from permanent traffic counters at over 500 locations across the U.S. These data were used to validate annual average daily traffic (AADT) estimates developed by Streetlight Data using passive data sources. The analysis revealed a strong correlation between estimates derived from passive sources and the permanent counters. However, statistical differences were observed, and deviations from ground truth may be unacceptable for some applications. Toll locations were particularly prone to high errors, possibly due to complex geometry. While the validation sheds light on the utility of AADT estimation from passive sources, its findings were limited by a lack of permanent counter data. It is possible that an analysis using a larger sample size could result in more robust findings.

17. Key Words	18. Distribution Statement
Annual average daily traffic, AADT, probe data, validation	No restrictions. This document is available to the public.

19. Security Classif.(of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	42	

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

Acknowledgments

We would like to acknowledge the members of the Technical Advisory Committee for their guidance and feedback throughout the project.

Scott Vockeroth, Alaska DOT & Public Facilities
Afrid Sarker, Caltrans
Steve Abeyta, Colorado DOT
John Phillips, Idaho DOT
William (Bill) Morgan, Illinois DOT
Eric Conklin, Georgia DOT
Carole Delion, Maryland State Highway Administration
Gene Hicks, Minnesota DOT
David Schoenmaker, Nebraska DOT
Chris Zajac, New Jersey DOT
Kent Taylor, North Carolina DOT
Terry Woehl, North Dakota DOT
Dave Gardner, Ohio DOT
Josh Roll, Oregon DOT
Greg Dunmire, Pennsylvania DOT
Todd Anderson, South Carolina DOT
Chris Didear, Texas DOT
Hamlin Williams, Virginia DOT
Steven Jessberger, FHWA

Executive Summary

Traffic engineers and planners have historically used a variety of continuous and temporary counters in conjunction with traffic prediction models to estimate traffic volume throughout the roadway network. These systems are expensive to install and maintain, requiring significant staff or contractor time to achieve broad coverage. Over the last decade, new technologies and data have matured to the extent that several vendors have developed commercial traffic volume products based on passive data sources. These passive data may originate from various sources such as: vehicle-based sensors, smartphone-based GPS and location-based services (LBS) data, cell tower data, or Bluetooth detection. Additionally, improvements in cloud-based data storage and computing resources have enabled these technologies and data to be processed and made available at a scale that would have been unimaginable even a decade ago.

To promote further development and deployment of such advancements, the Federal Highway Administration (FHWA) organized a pooled fund effort TPF-5(384) with the objective of developing methods to check these new data sources that provide vehicle volume data and classification data using passively collected probe and other non-traditionally available data. Through the pooled fund effort noted above, FHWA awarded a contract to Streetlight Data (StL) to develop a methodology for estimating Annual Average Daily Traffic Volume (AADT) throughout the U.S. FHWA subsequently awarded contracts to two independent validation contractors, Cambridge Systematics with Texas Transportation Institute as a sub-contractor, and the National Renewable Energy Laboratory, or NREL. The validation teams were tasked with assessing the accuracy of the AADT estimates supplied by StL. This report documents the findings from NREL's independent validation.

The NREL validation team pursued several potential data sources that could serve as benchmark, or ground truth, data for testing. After extensive data discovery and quality control activities, 566 sites, including a combination of directional and bidirectional locations, were used in the validation. Among these, roughly two-thirds are from FHWA's Travel Monitoring Analysis System (TMAS), with the remaining coming from toll systems. The locations were roughly split between urban and rural contexts.

After compiling and cleaning the validation data, NREL compared ground truth AADT with StL estimates for the same locations. A variety of model accuracy metrics were used to assess the accuracy of StL's estimates. The key findings of the analysis are:

- StL's estimates are well correlated with ground-truth data for all sites and across the different groupings and volume ranges evaluated.
- There are statistical differences between the AADTs obtained from the ground truth locations and StL's estimates. At a significance level of 0.05, the difference is unlikely due to random variation. However, these differences are not unexpected since StL's estimates are modeled while ground truth data are observed directly.
- For all sites and groupings evaluated, StL's estimates are positively biased compared to ground truth AADT.
- Errors are considerably lower for rural sites than for urban sites.
- Compared to higher volume ranges, the StL estimates for low volume sites exhibit greater error.

- Based on the statistical hypothesis test results, it is concluded that differences between StL's estimates and ground truth data are not due to random variation.
- Errors at toll locations were generally higher and contribute to lower accuracy reported in various groupings. This may be due to complex geometries, different vehicle occupancy, or other factors associated with toll locations.

There are a few potential sources of error that may go a long way in explaining the discrepancies observed between StL and ground truth estimates. While the ground truth estimates are thought to be reliable, they are not expected to be completely error free. This type of error could also impact StL's training dataset. Another potential error source relates to the geospatial matching of validation counter locations with OSM segments. While extensive quality control was conducted to ensure the best possible match between counters and segments, the process of checking for alignment between counter locations and OSM segments still leaves some room for error. This is particularly true where counters are located at or near a ramp, and the OSM segmentation might not correspond exactly with the lanes counted.

Outside of the validation dataset, it seems likely that complex geometry might account for a lot of the observed error in StL's estimates. For example, toll locations often count vehicles adjacent to other through lanes, which could make it difficult to obtain an accurate tally from passive data sources.

There are many areas of opportunity for future research related to traffic volume estimation from passive data, including:

- Understanding and controlling for the influence of complex geometries or other roadway configurations on AADT estimation,
- Identifying additional factors that may influence probe penetration and associated accuracy of AADT, such as socio-demographics or varying levels of engagement with smartphones during trips,
- Identifying and improving the accuracy of passive data collection for other traffic measures, such as hourly or daily estimates, of for other modes of travel,
- Determining the appropriate number of ground truth traffic monitoring stations to support robust model calibration and validation for varying applications.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Project background

Traffic engineers and planners have historically used a variety of continuous and temporary counters in conjunction with traffic prediction models to estimate traffic volume throughout the roadway network. These systems are expensive to install and maintain, requiring significant staff or contractor time to achieve broad coverage. Additionally, the accuracy of modeled traffic volume at a given location is often unknown and likely to be poor in many cases. Locations that have not been counted for several years are also subject to a degree of error that may be unacceptable for certain applications. Where more reliable volume estimates are needed, such as for a corridor or intersection study, project-specific counts are routinely conducted.

Over the last decade, new technologies and data have matured to the extent that several vendors have developed commercial traffic volume products based on passive data sources. These passive data may originate from various sources such as: vehicle-based sensors, smartphone-based GPS and location-based services (LBS) data, cell tower data, or Bluetooth detection. Additionally, improvements in cloud-based data storage and computing resources have enabled these technologies and data to be processed and made available at a scale that would have been unimaginable even a decade ago. Data from permanent traffic sites are used to calibrate passive data sources.

To promote further development and deployment of such advancements, the Federal Highway Administration (FHWA) organized a pooled fund effort TPF-5(384) with the objective of developing methods to check these new data sources that provide vehicle volume data and classification data using passively collected probe and other non-traditionally available data.

If found to be accurate enough reliable traffic volume estimates obtained from passive data sources could reduce costs and improve efficiency for State Departments of Transportation (DOTs), Metropolitan Planning Organizations (MPOs), and local agencies. They could also reduce risks to employees and contractors who place sensor devices in and on the roadways by dramatically reducing the number of traffic counts needed.

Through the pooled fund effort noted above, FHWA awarded a contract to Streetlight Data (StL) to develop a methodology for estimating Annual Average Daily Traffic Volume (AADT) throughout the U.S. FHWA subsequently awarded contracts to two independent validation contractors, Cambridge Systematics with Texas Transportation Institute as a sub-contractor or CS/TTI, and the National Renewable Energy Laboratory, or NREL. The validation teams were tasked with assessing the accuracy of the AADT estimates supplied by StL. This report documents the findings from NREL's independent validation.

1.2 The Need for Independent Validation

As public agencies consider integrating passive data sources into their traffic monitoring programs, they need reliable information about the accuracy of these data. Commercial traffic volume vendors may provide information about expected accuracy based on their internal

assessments, but as with any commercial product, an objective validation is preferable. This objective validation is necessary to ensure the reported accuracy is not unduly influenced by financial considerations and that the data provides the accuracy and precision required to meet agencies' needs. It is also necessary to determine if and how the reported accuracy changes when the prediction model is applied to a new set of roadways, as discussed in greater detail below.

While vendors may take different approaches to estimating traffic volume, machine learning models are commonly deployed. These models use available traffic count data in conjunction with passive data and other data sources, such as contextual information, roadway type, or others, to develop traffic volume estimates. The traffic count data is a 'training' dataset, which provides a benchmark that is needed to train the machine learning algorithm. Once trained, the algorithm can be used to estimate volume at locations not included in the training dataset, based on the characteristics of those locations. Among these characteristics, the passively collected movement data is expected to play large role in predicting actual traffic volume.

Machine learning model accuracy is often reported based on the results of a k-fold cross validation. K-fold cross validation involves splitting the dataset into *k* subsets of training and testing data. For each subset, the training data is used to develop the model, and its accuracy is determined by how close its predictions match the testing data.[1] The average accuracy across these *k* models is then computed and reported as the accuracy for the machine learning model.

Since high-quality, continuous count data is relatively sparse, k-folds cross validation improves model performance as compared to a conventional 'holdout' approach, where a portion of the dataset is reserved exclusively for testing. However, the ability for any model to make predictions outside the range of the input dataset is inherently uncertain, and the k-folds cross validation does not address this challenge.

In order for public agencies to truly understand and assess the accuracy of traffic volume estimates derived from passive data sources, these estimates must be compared to data that is blind to the model. Blind validation answers the question of how well the model can predict traffic volume on roads where there is no available count data. It also sheds light on the reliability of the accuracy estimates derived from the k-folds cross-validation. In other words, if the results of the blind validation are consistent with the vendor-reported accuracy, agencies can be more confident that the accuracy of the estimates from passive data reported by the vendor are reliable for roads where data counts do not currently exist.

2 Data Collection

The NREL validation team pursued several potential data sources that could serve as benchmark, or ground truth, data for testing. Initially, the team hoped to identify several thousand locations with high-quality count data that would allow StL's AADT estimates to be assessed across factors such as roadway type, traffic volume range, context (urban/rural areas), and state, among others.

[1] Various metrics can be used to assess model accuracy in k-folds cross validation. Mean Absolute Percentage Error, or MAPE, and Root Mean Square Error (RMSE) are among those commonly used.

In order for a count location to be included in the testing dataset, it would need to meet the following conditions:

- Technology and deployed configuration provides accurate traffic volume,
- Includes sufficient data to calculate AADT using FHWA's preferred estimation method (this requires data for each day of the week for each month of the year),
- Not included in StL's training dataset.

Finding a suitable number of locations that meet these conditions proved more challenging than hoped. A summary of datasets pursued is provided in Table 1. Table 1 also indicates whether these datasets were included in the final testing dataset and the rationale for their inclusion or exclusion.

Table 1. Testing Datasets Pursued

Technology / Source	Included in Final Testing Dataset	Rationale for Including / Excluding
Travel Monitoring Analysis System (TMAS) continuous counts	Yes	Continuous counts are considered to be the most reliable traffic volume data source for most purposes
Toll Plazas (Ohio, Illinois, North Carolina, Maryland, Virginia)	Yes	Toll plazas found to be highly reliable data sources and relatively easy to obtain
Short-duration counts	No	FHWA guidance to exclude short-duration counts based accuracy concerns
Intelligent Transportation Systems (ITS), including signal detection and ramp metering locations	No	Some systems recently installed and did not include sufficient data; signal detection configurations often unclear based on available information; errors found during calibration; geo-spatial location issues
Magnetometers	No	Unable to obtain data from vendors
Road weather information system (RWIS)	No	Unable to obtain counts from these systems
Doppler Radar Speed Feedback Signs	No	Unreliable location information; significant errors found during video ground truth checks

StL's model training dataset includes continuous counts from many state DOTs. As a result, a large portion of the Travel Monitoring Analysis System (TMAS) counters were unavailable for use as testing data. Working with FHWA and the CS/TTI team, NREL was able to obtain some continuous count data that were withheld from StL. These counts were withheld either by state DOTs or by FHWA. StL also agreed to exclude some of the counts they had already obtained, so that they could be used by the validation teams for testing. A breakdown of sites withheld from StL and those excluded by StL is shown in Table 3.

3 Data Cleaning and Processing

Data checking, cleaning and processing accounted for a significant portion of the overall project effort. The key steps are documented in this section.

3.1 Data Organization and Standardization

Station and count data were received in different file formats (CSVs, excel files, shapefiles, and Microsoft Access database files) with a variety of schemas. Original count data also had varying spatial and temporal resolutions, depending on the agency and dataset. To facilitate analysis, these varying file formats and schemas were standardized into a common format agreed upon by the project team.

Within the TMAS dataset (all data from TMAS followed the 2001 TMG formats), some states, including Arizona, Ohio, and Virginia, report counts by both direction and lane. In these cases, all raw count data were aggregated by hour and direction. Count data from toll systems also required some manipulation and aggregation, such as where separate data were provided for different types of revenue collection (license plate reader, manual collection, etc.).

Among the various stations identified by the project team, only a small percentage were used in the eventual analysis (many of the identified stations were from short-duration count locations or sites using other technologies, which were later determined to be unusable). Station files were filtered to include only the stations of interest for analysis, and these stations were assigned a unique identifier.

Each pair of cleaned station and count data (e.g. Arizona TMAS station file and Arizona TMAS count file) were then merged based on their original station identifier (specific to each dataset), and further analysis was carried out, focusing on the established stations of interest.

3.2 TMAS Quality Control Checks

TMAS sites were checked in accordance with FHWA's guidance, and sites not meeting the quality control criteria were removed from the dataset.[2] More specifically, two conditions had to be met for a site to be included:

- At least one entire day of count data must have been available for each day of week in each of the 12 months of the year; and
- The data could not include more than six consecutive hours with a count of zero.

Sites that did not meet these criteria were eliminated from the dataset. Additionally, some TMAS records from 2019 included duplicate rows for the same day at the same station. These were handled by discarding duplicate entries with a difference of greater than 10% and averaging duplicate entries where the difference was less than 10%. Remaining sites were then checked again to ensure the weekday/month criteria mentioned above was still satisfied.

[2] https://www.fhwa.dot.gov/policyinformation/tmguide/tmg_2013/quality-control-checks.cfm

The original number of sites, the number of sites removed through this QC process, and the final remaining sites are in Table 2. These directional sites were later merged where appropriate resulting in the final counts used in Table 3.

Table 2. TMAS Sites Before and After Quality Control

State	Original Sites (Directional)	Sites Removed through QC	Sites Remaining
AL	46	4	42
AR	10	2	8
AZ	34	8	26
CA	33	7	26
CO	30	2	28
CT	4	0	4
DE	4	1	3
FL	69	3	66
GA	50	0	50
ID	49	7	42
IL	22	8	14
KY	16	0	16
MD	66	18	48
NC	24	8	16
ND	82	31	51
NE	12	0	12
NM	14	8	6
NV	22	8	14
OH	46	3	43
OK	26	0	26
PA	25	6	19
RI	8	4	4
SC	34	6	28
VA	128	7	121
Total	**854**	**141**	**713**

3.3 Spatial Review and Processing

Counter location information was provided in various formats, depending on the data source. The TMAS station files included latitude and longitude coordinates, but data from toll operators did not always include precise location information. For these datasets, the project team used Google Maps to manually geolocate and digitize each counter.

Once all counters were mapped, the team reviewed each location to ensure they were mapped to allow them to be joined to the appropriate roadway through an automated process. For example, some counters are set back from the highway such that they are closer to a nearby road, rather than the road being counted. Additionally, some counters needed to be duplicated and offset in order to account for each direction of travel.

The next step in the process was to perform a spatial join to associate testing locations with the appropriate Open Street Map (OSM) segment identifier. OSM was selected by both validation teams as the reference street network since StL associates its AADT estimates with OSM segments. Using OSM rather than another publicly available road network eliminated the potential for errors that might have resulted from network conflation. To improve the likelihood of achieving a successful match, attributes such as heading and directionality (one-way or two-way) were considered, along with distance to the matched segment. Counters that were more than 500 ft. from the nearest segment were not used.

3.4 Additional Quality Control Checks

Along with the quality control and data cleaning discussed above, additional reviews were performed where the StL estimates deviated substantially from validation data. Testing locations were removed at this point if it was determined that the counter was recording traffic volume on a different segment or lanes than what was represented in the associated OSM segment. Many locations with large discrepancies remained in the testing dataset, as there were no grounds for their removal.

3.5 Final Testing Dataset

The last step in the data cleaning process was to merge directional sites to obtain bidirectional AADTs, where appropriate. The NREL team had initially organized the data with the intent of evaluating each direction separately, but FHWA later directed the team to produce bidirectional AADT, consistent with FHWA and DOT conventions. Some directional sights remained in the final dataset, due to the location having only one direction of counts, or because one direction was removed through the QC processes described above.

Table 3 provides a breakdown of all testing sites by state and whether, for TMAS locations, they were withheld from StL or excluded by StL. Table 3 also indicates the total number of directional and bidirectional sites by state. Most of the directional testing sites were from toll systems, which do not always account for both directions.

Table 3. Final Testing Locations by State and Data Source

State	TMAS		Toll Systems	Total	Directional	Bidirectional
	Withheld from StL	Excluded by StL				
AL	0	22	0	**22**	2	20
AR	0	4	0	**4**	0	4
AZ	0	13	0	**13**	0	13
CA	0	14	0	**14**	2	12
CO	0	14	0	**14**	0	14
CT	0	2	0	**2**	0	2
DE	0	2	0	**2**	1	1
FL	0	32	0	**32**	10	22
GA	0	22	0	**22**	7	15
ID	0	20	0	**20**	2	18
IL	0	7	91	**98**	90	8
KY	0	8	0	**8**	2	6
MD	24	0	9	**33**	9	24
NC	8	0	34	**42**	35	7
ND	30	0	0	**30**	15	15
NE	0	7	0	**7**	2	5
NM	0	3	0	**3**	0	3
NV	0	7	0	**7**	0	7
OH	0	22	30	**52**	9	43
OK	0	13	0	**13**	0	13
PA	0	10	0	**10**	1	9
RI	0	1	0	**1**	1	0
SC	0	19	0	**19**	2	17
VA	0	78	20	**98**	73	25
Total	**62**	**320**	**184**	**566**	**263**	**303**

4 Summary of Testing Dataset

In this section of the report, a brief overview of the data is provided. Measures of frequency and normality are provided for context, and the results are presented in the next section.

4.1 Frequency Statistics

Figures 1 through 5 visually depict the frequency of testing sites by data source, urban/rural context, state, and AADT range.

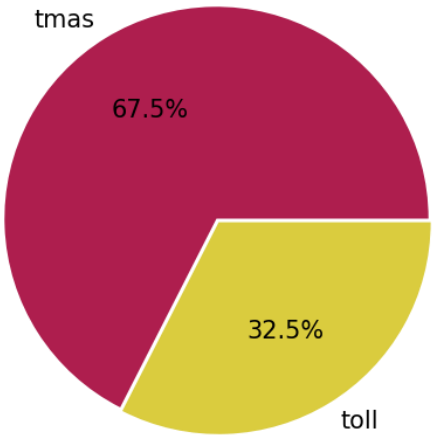

Figure 1. Percentage of Testing Sites by Data Source

Roughly two-thirds of all testing locations are from TMAS, with the remaining coming from toll systems.

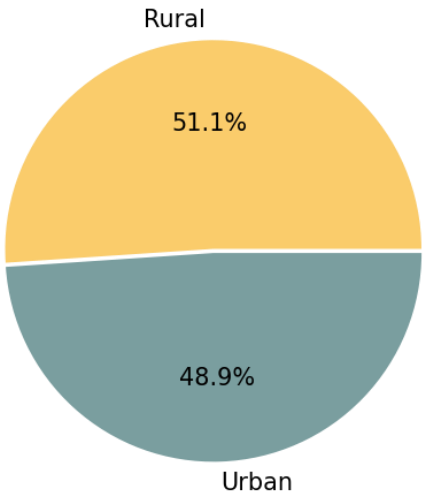

Figure 2. Percentage of Testing Sites by Urban/Rural Context

Testing sites are roughly split between rural areas (51 percent) and urban areas (49 percent).

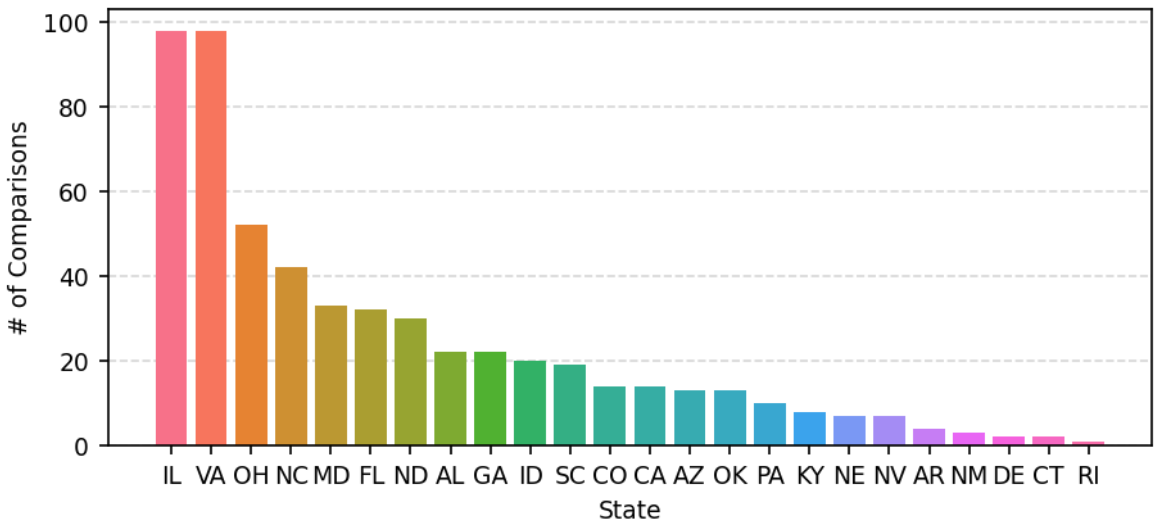

Figure 3. Number of Testing Sites by State

Illinois and Virginia stand out as having the largest number of testing locations, with close to 100 in each state. Many of the locations in Illinois are from the Illinois Tollway, while most in Virginia are from TMAS.

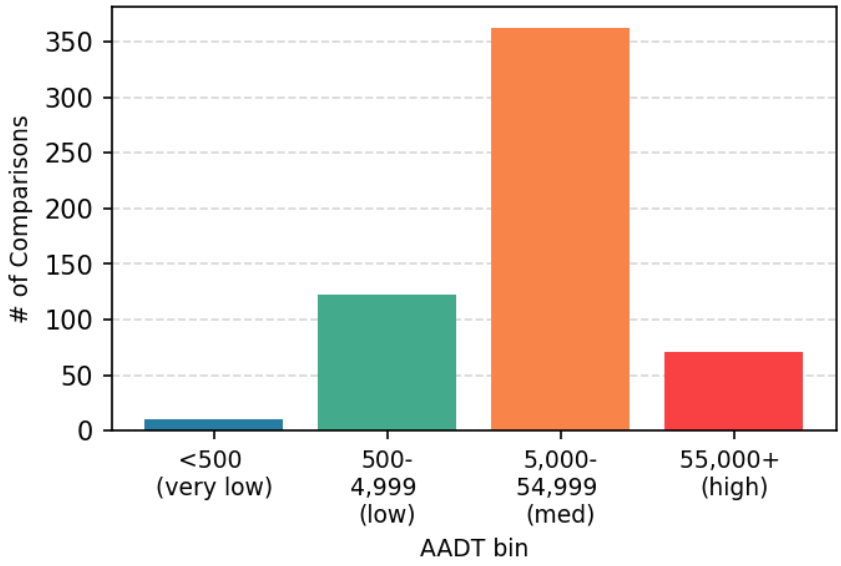

Figure 4. Number of Testing Sites by AADT bin (wide volume bins)

The vast majority of testing sites are in the 'medium' AADT bin; however, this accounts for a very wide traffic volume range.

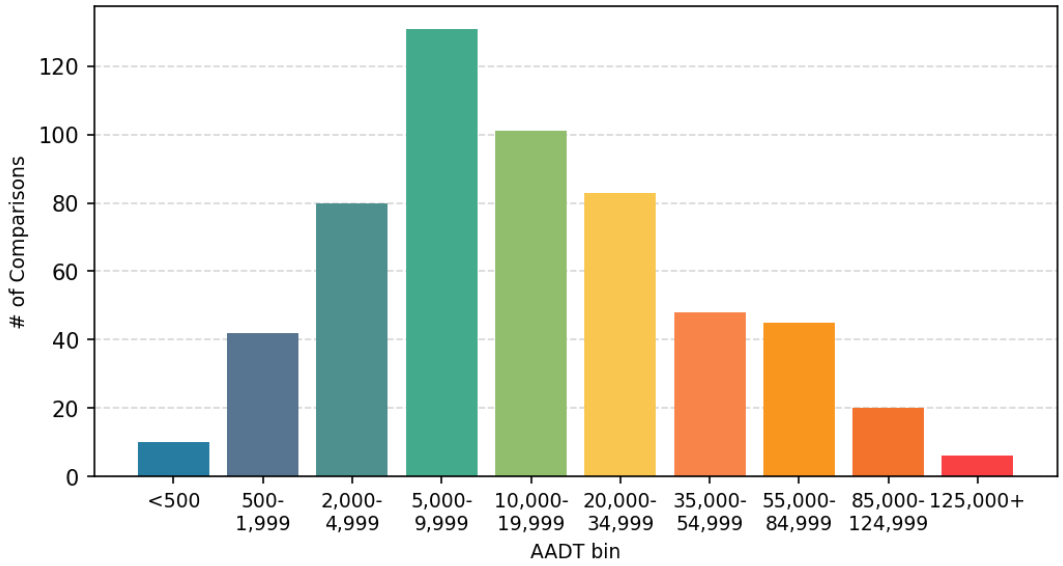

Figure 5. Number of Testing Sites by AADT bin (narrow volume bins)

Testing sites are well distributed throughout the range of the data.

4.2 Normality

Many statistical methods rely on an assumption that the data follow a normal distribution. If this assumption is violated, test results may be interpreted incorrectly. As such, determining whether StL's errors are normally distributed was key to deciding which statistical tests are most appropriate. This is especially important because errors arising from traffic volume estimation processes may be non-normally distributed.

There are multiple ways of assessing whether data follows a normal distribution, each with its pros and cons. One of the more common approaches is through visual inspection of the data distribution in a histogram. Figure 6 shows the distribution of AADT errors for all testing sites. Overall, the data appears to be somewhat normally distributed (it generally follows a bell-shaped curve), though it is positively skewed, as indicated by its skewness measure of 1.86.

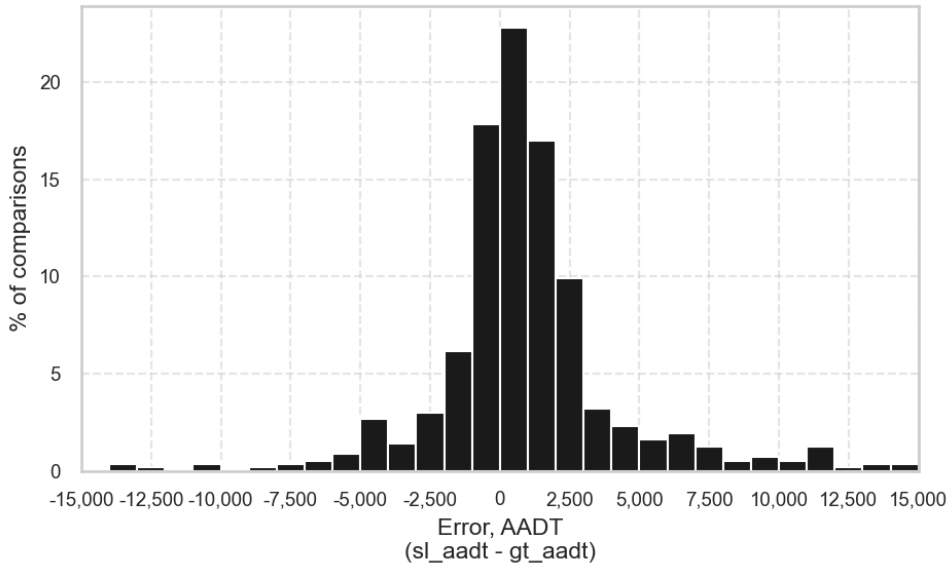

Figure 6. AADT Error Distribution for All Sites

The AADT error distribution is slightly positively skewed, but generally follows the shape of a normal distribution.

Another approach for assessing normality is through the application of formal hypothesis tests. The Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) tests can be used for this purpose. These tests check whether the subject data was drawn from a normal distribution. Both tests were performed, and the results indicated the AADT errors are not normally distributed. The p-values for the KS and SW tests were 3.6e-24 and 4.8e-24, respectively.

Despite the formality of the KS and SW tests, it is worth noting that these do not necessarily provide definitive proof that the AADT errors are not normally distributed. These tests are known to be conservative and tend to reject the hypothesis that the data are normally distributed if the sample size is large, as is the case here.[3]

The results presented below generally do not hinge on whether the data is normally distributed, as most of the validation metrics are not based on hypothesis testing. For audiences that might be interested in a formal hypothesis test approach, results from alternate versions are provided, with one assuming normally distributed data (paired sample t-test), and another not requiring this assumption to be met (Wilcoxon signed ranks test).

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/

5 Results

To assess the accuracy of the StL volume estimates relative to the ground truth counts, several validation metrics were calculated. The validation team believes that a single metric should not be used to assess model accuracy, but instead a variety of metrics should be considered, along with supporting visualizations. The metrics used throughout this section are described in Table 4.

Table 4. Validation Metrics.

Validation Metric	Interpretation	Notes
AADT error percentiles	The AADT error below which occurs the percent of observations indicated by the percentile. E.g., 50 percent of observations fall below the 50th percentile AADT error	The 16 to 84 percentile AADT error range accounts for 68% of observations while the 2.5 to 97.5 percentile range accounts for 95% of observations
95th Percentile Error Range	The upper and lower bounds of error within which 95 percent of all observations occur, expressed in percentage terms.	
Median percentage error (bias)	Low values indicate more accurate estimates; sign (+/-) indicates direction of bias	Can mask large errors if the error distribution is symmetric
Median Absolute Percentage Error (MdAPE)	Low values indicate more accurate estimates; metric suggests a typical error without regard to direction	Less sensitive to outliers than MAPE
Mean Absolute Percentage Error (MAPE)	Low values indicate more accurate estimates; metric suggests a typical error without regard to direction	More sensitive to outliers than MdAPE; constrained in negative direction to -100%, but not in positive direction
Normalized Root Mean Square Error (NRMSE)	Low values indicate a better model fit (ranges from 0 to 1, expressed as a percentage)	Primarily used to compare models; there is no accepted NRMSE threshold
Spearman's rho	Values range from -1 to 1 with 0 implying no correlation. Correlations of -1 or 1 imply an exact monotonic relationship	Correlation coefficient for non-parametric data based on ranked estimates and ground truth data
Paired sample t-test	p values > 0.05 indicate the test failed to reject the assumption that mean difference between the estimates and ground truth datasets is zero	For normal distributed data
Wilcoxon matched-pairs signed-ranks test	p values > 0.05 indicate the test failed to reject the assumption that there is not a statistically significant difference between the median ranks of the estimates and ground truth datasets	For non-normal distributed data; not reliable for low sample size

5.1 Summary of Findings

The accuracy of StL's predictions were evaluated across several groupings. The results are summarized in Table 5 and described further in subsequent sections. Additional metrics and visualizations, along with supporting data and analysis code may be found at: https://github.com/NREL/fhwa-streetlight-aadt-validation

Table 5. Summary of Model Validation Results.

		AADT Error @ Percentile					Median % Error	Absolute % Error		NRMSE	rho	Wilcoxon Signed Ranks Test p-value	Paired Samples t-test p-value
	N	2.5th	16th	50th	84th	97.5th		Mean	Median				
All comparisons	566	-5,768	-1,068	650	3,032	13,285	5.7	18.8	11.8	9%	0.98	0.000	0.000
Rural	289	-5,703	-837	199	1,979	5,264	3.6	16.3	11.6	8%	0.99	0.000	0.045
Urban	277	-5,631	-1,244	1,249	6,157	22,301	8.8	21.4	12.4	12%	0.98	0.000	0.000
< 500 AADT	10	-9	27	93	199	220	27.1	30.3	27.1	57%	0.66	0.006	0.003
500-4,999 AADT	122	-769	-303	243	1,365	6,117	12.6	32.3	23.2	13%	0.88	0.000	0.000
5,000-54,999 AADT	363	-4,693	-1,203	893	3,085	8,577	5.3	15.6	10.9	10%	0.96	0.000	0.000
55,000+ AADT	71	-14,129	-5,256	1,055	12,116	26,510	1.1	10.1	7.3	21%	0.90	0.063	0.014
TMAS	382	-6,834	-1,685	43	1,814	10,468	0.4	11.7	7.2	8%	0.99	0.461	0.286
Toll	184	-872	821	1,995	5,971	22,658	25.2	33.7	25.2	15%	0.95	0.000	0.000
Withheld by StL	320	-6,957	-1,671	58	1,838	10,273	0.7	10.9	7.1	8%	0.99	0.300	0.289
Blind Validation Sites	246	-4,040	-12	1,555	4,334	21,796	20.4	29.0	21.8	14%	0.97	0.000	0.000

5.2 All Sites (n=566)

NREL's validation covered 566 sites, including a combination of directional and bidirectional locations, as discussed in previous sections. The results from the evaluation of these sites are shown in Table 6.

Table 6. Summary of Model Validation Results for All Sites.

Validation Metric	Result
AADT error percentiles	2.5th: -5,768 16th: -1,068 50th: +650 84th: +3,032 97.5th: +13,285
95th percentile error range	-28% to +64%
Median percentage error (bias)	+5.7%
Median Absolute Percentage Error (MdAPE)	11.8%
Mean Absolute Percentage Error (MAPE)	18.8%
Normalized Root Mean Square Error	9%
Spearman's rho	0.98
Paired sample t-test	0.000
Wilcoxon matched-pairs signed-ranks test	0.000

Key Takeaways (All Sites)

- 68% of observations fall within an error range of -1,068 to +3,031 AADT
- 95% of observations fall within an error range of -5,768 to + 19,053 AADT
- Compared to ground truth AADT, StL's estimates are positively biased, with a median error of 5.7%.
- The MdAPE is considerably lower than the MAPE, indicating the mean error is influenced by outliers.
- The results suggest the StL and ground truth estimates are statistically different, regardless of normality assumptions.

Visualizations provided in Figure 7 through Figure 9 provide additional insight into the comparison of StL and ground truth estimates for all sites.

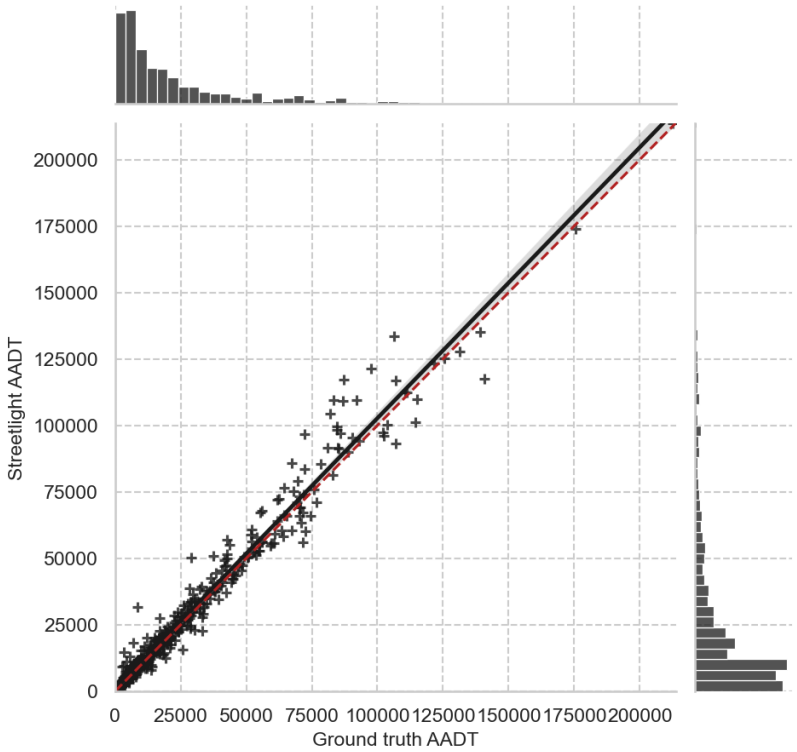

Figure 7.Scatterplot of Ground Truth and Streetlight AADT for All Sites[4]

The StL AADT estimates match ground truth AADT reasonably well. The dashed red line represents the line of equality, i.e., a perfect estimator. The black regression line approximates the line of equality. The 95% confidence interval is shaded. There is a slight tendency for StL to overestimate AADT, particularly in the 75,000 to 100,000 AADT range. However, this range accounts for a small portion of observations. The majority of observations are below 25,000 AADT.

[4] In this and similar scatterplots throughout the report, the red dashed line indicates a perfect correlation, while the black line indicates the observed correlation.

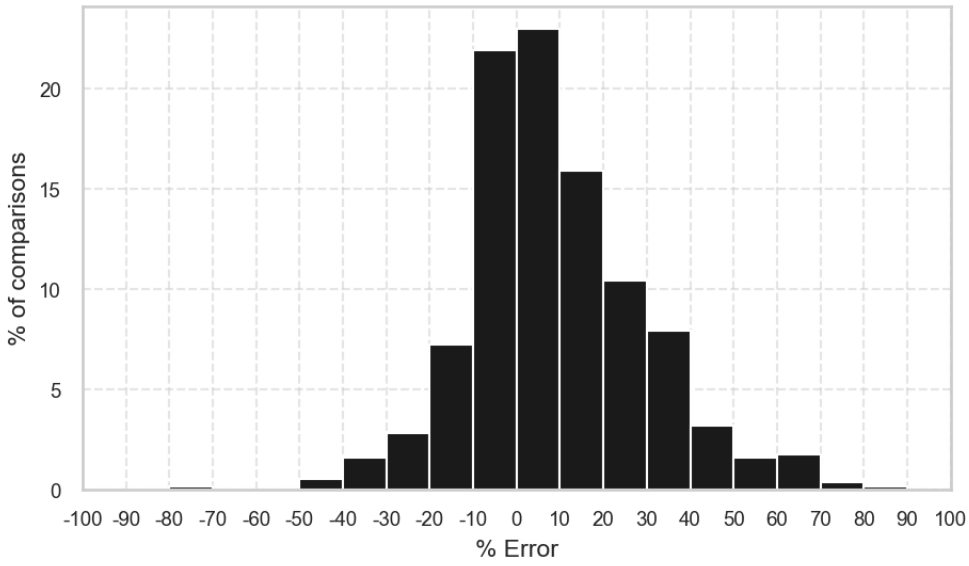

Figure 8. Percentage Error for All Sites

The observed percentage error for the StL AADT estimates is positively skewed.

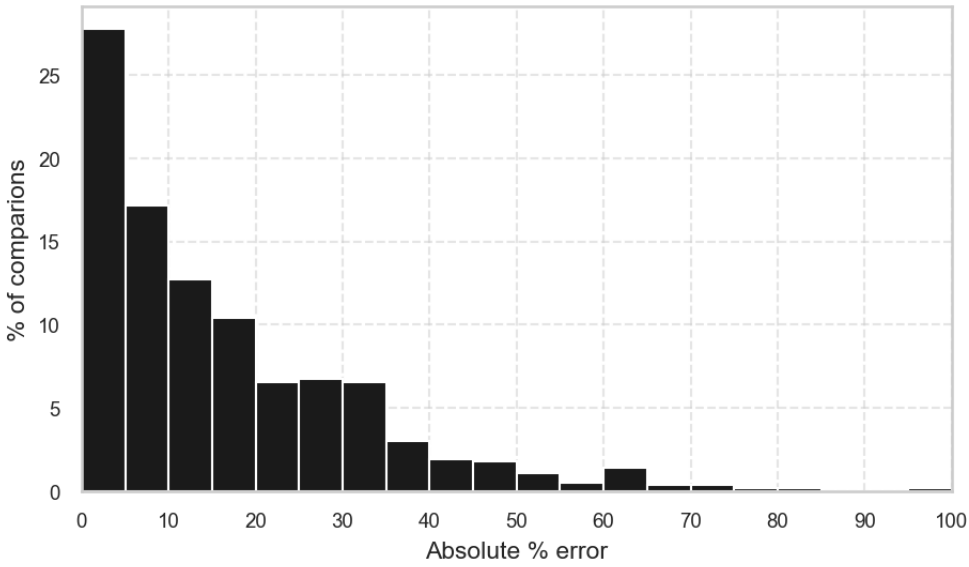

Figure 9. Absolute Percentage Error for All Sites

Half of all observations have an absolute percentage error (MdAPE) less than 11.8%.

5.3 Rural Sites (n=289)

Locations were flagged as either urban or rural, on the basis of census urbanized area boundaries. Results from the rural sites are summarized in this section. StL estimates generally performed better in rural areas than urban areas.

Table 7. Summary of Model Validation Results for Rural Sites

Validation Metric	Result
AADT error percentiles	2.5th: -5,703 16th: -837 50th: +199 84th: +1,979 97.5th: +5,264
95th percentile error range	-32% to +59%
Median percentage error (bias)	+3.6%
Median Absolute Percentage Error (MdAPE)	11.6%
Mean Absolute Percentage Error (MAPE)	16.3%
Normalized Root Mean Square Error	8%
Spearman's rho	0.99
Paired sample t-test	0.045
Wilcoxon matched-pairs signed-ranks test	0.000

Key Takeaways (Rural Sites)

- 68% of observations fall within an error range of -837 to +1,979 AADT
- 95% of observations fall within an error range of -5,703 to +5,264 AADT
- While results for rural sites are positively biased, errors are considerably lower (closer to zero) than for all sites and urban sites.
- As for all sites, the MdAPE for rural sites is considerably lower than the MAPE, indicating the mean error is influenced by outliers.
- The results suggest the StL and ground truth estimates are statistically different. However, less stringent hypothesis testing criteria such as a 90% confidence level would suggest a different interpretation, as the p-value for the paired sample t-test is 0.045, which is quite close to the threshold value of 0.05. The Wilcoxon signed-ranks hypothesis test indicates the StL and ground truth estimates are statistically different, but given the relatively large sample size, the t-test may be reliable for this comparison.

Visualizations provided in Figure 10 through Figure 12 provide additional insight into the comparison of StL and ground truth estimates for rural sites.

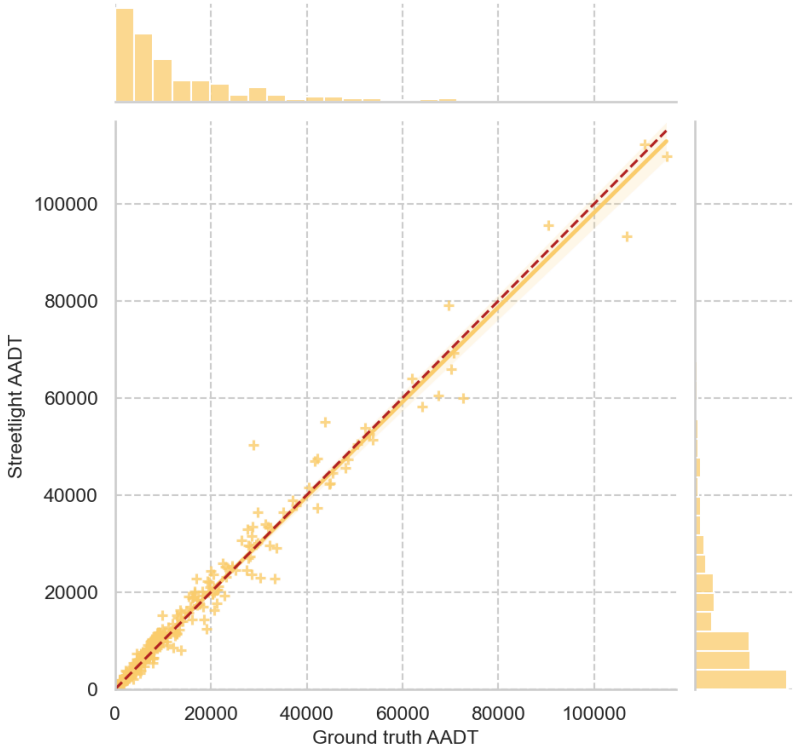

Figure 10. Scatterplot of Ground Truth and Streetlight AADT for Rural Sites

The StL AADT estimates for rural sites are strongly correlated with ground truth AADT.

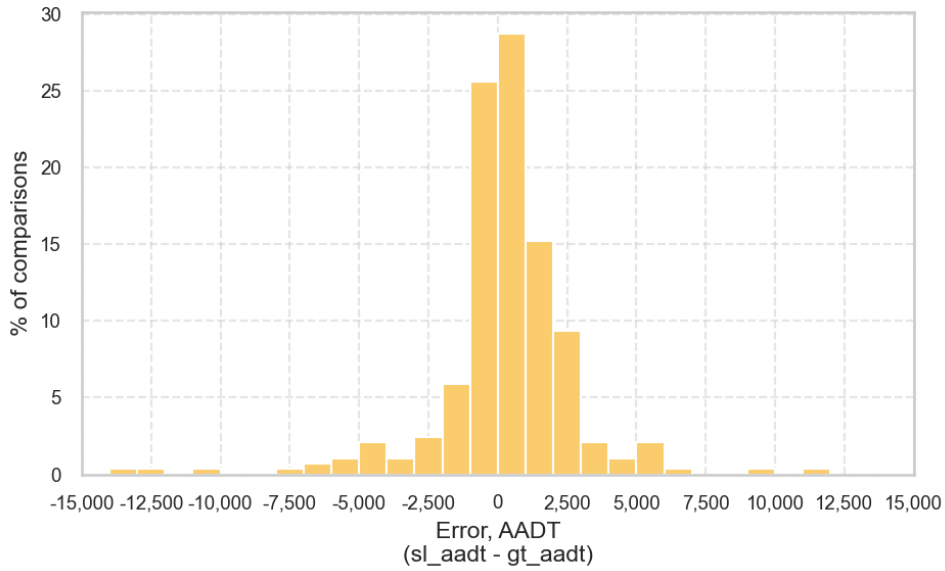

Figure 11. AADT Error for Rural Sites

The observed error for the StL AADT estimates is slightly positively skewed, but there are a few strong negative outliers.

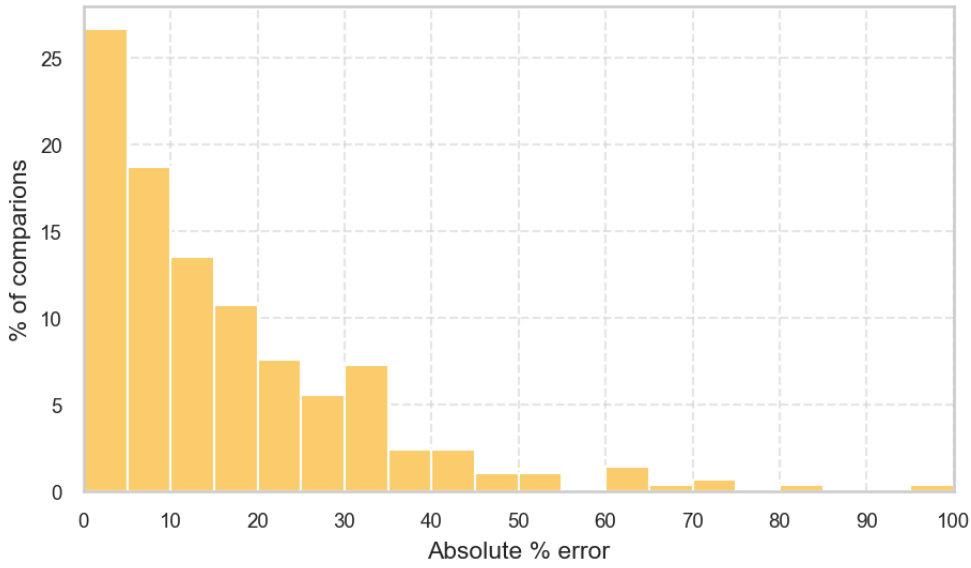

Figure 12. Absolute Percentage Error for Rural Sites

Half of all observations have an absolute percentage error less than 11.6%.

5.4 Urban Sites (n=277)

Compared to rural sites, StL estimates for urban sites tended to be slightly less accurate. In particular, large positive AADT errors in the upper volume range resulted in considerably higher mean and median errors.

Table 8. Summary of Model Validation Results for Urban Sites

Validation Metric	Result
AADT error percentiles	2.5th: -5,631 16th: -1,244 50th: +1,249 84th: +6,157 97.5th: +22,301
95th percentile error range	-22% to +83%
Median percentage error (bias)	+8.8%
Median Absolute Percentage Error (MdAPE)	12.4%
Mean Absolute Percentage Error (MAPE)	21.4%
Normalized Root Mean Square Error	12%
Spearman's rho	0.98
Paired sample t-test	0.000
Wilcoxon matched-pairs signed-ranks test	0.000

Key Takeaways (Urban Sites)

- 68% of observations fall within an error range of -1,244 to +6,157 AADT
- 95% of observations fall within an error range of -5,631 to +22,301 AADT
- Compared to rural sites, the StL estimates for urban sites exhibit higher error, both as a percentage and in actual AADT.
- The MdAPE for urban sites is considerably lower than the MAPE, indicating the mean error is influence by outliers.
- The results suggest the StL and ground truth estimates for urban sites are statistically different, regardless of normality assumptions.

Visualizations provided in Figure 13 through Figure 15 provide additional insight into the comparison of StL and ground truth estimates for urban sites.

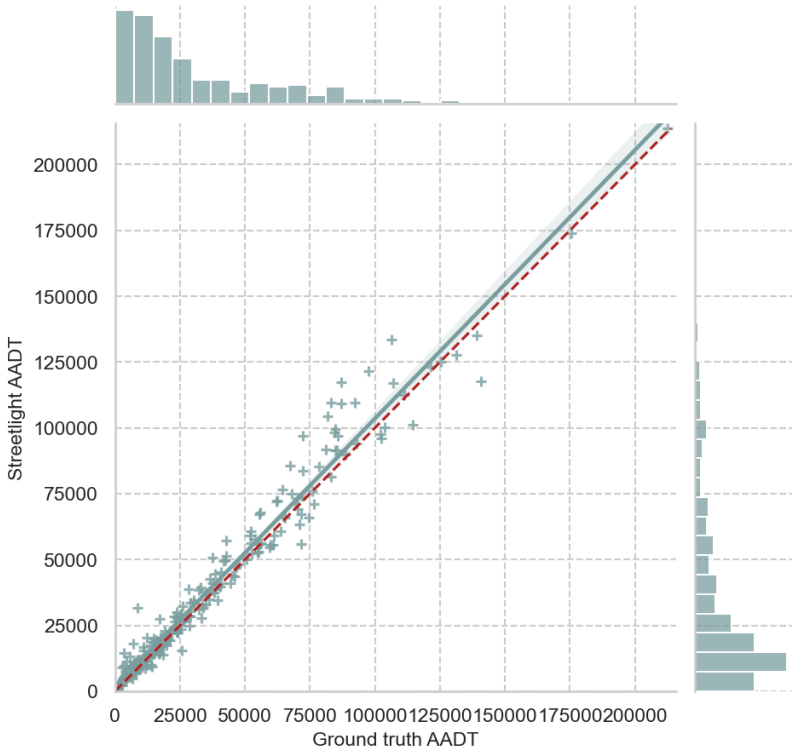

Figure 13. Scatterplot of Ground Truth and Streetlight AADT for Urban Sites

The StL AADT estimates for urban sites are generally correlated with ground truth AADT. However, StL estimates have a greater tendency toward overestimation for urban sites than for rural sites, particularly above 75,000 AADT.

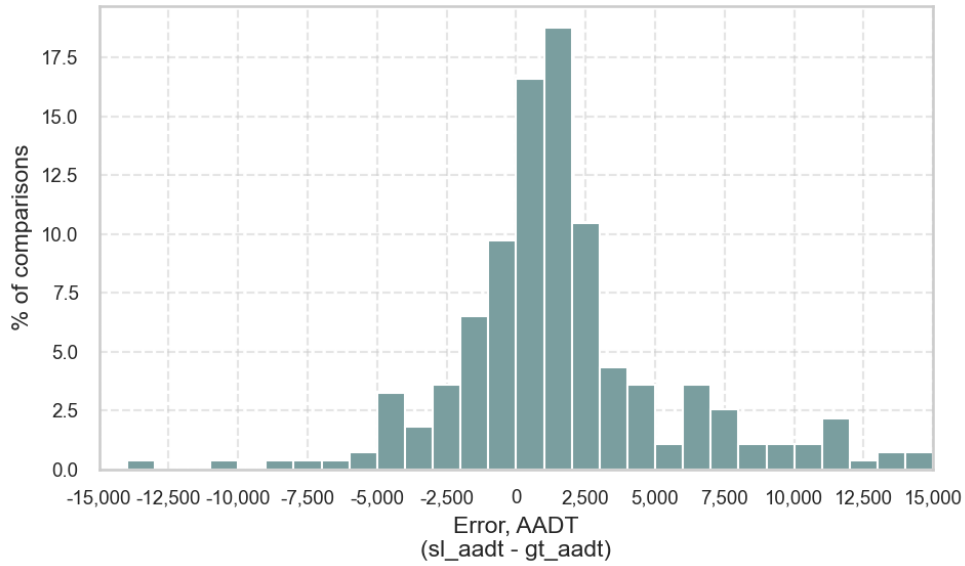

Figure 14. AADT Error for Urban Sites

The observed error for the StL AADT estimates is positively skewed.

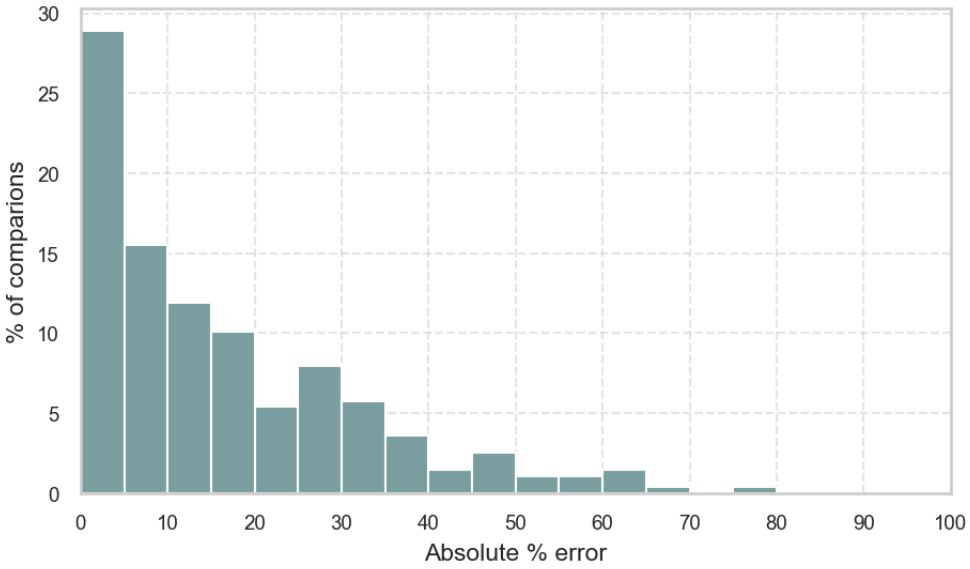

Figure 15. Absolute Percentage Error for Urban Sites

Half of all observations have an absolute percentage error less than 12.4%.

5.5 Low Volume Sites (500-4,999 AADT; n=122)

As noted previously, the validation data were divided into several volume ranges, based on direction from FHWA. The categorization was performed according to the ground truth AADT. This section focuses on results from the 'low volume' grouping, with AADT ranging from 500 to 4,999.[5]

Table 9. Summary of Model Validation Results for Low-Volume Sites

Validation Metric	Result
AADT error percentiles	2.5th: -769 16th: -303 50th: +243 84th :+1,365 97.5th: +6,117
95th percentile error range	-38% to +167%
Median percentage error (bias)	+12.6%
Median Absolute Percentage Error (MdAPE)	23.2%
Mean Absolute Percentage Error (MAPE)	32.3%
Normalized Root Mean Square Error	13%
Spearman's rho	0.88
Paired sample t-test	0.000
Wilcoxon matched-pairs signed-ranks test	0.000

Key Takeaways (Low Volume Sites)

- 68% of observations fall within an error range of -303 to +1,365 AADT
- 95% of observations fall within an error range of -769 to +6,117 AADT
- Compared to higher volume ranges, the StL estimates for low volume sites exhibit greater error.
- As with other groupings, the MdAPE for low volume sites is lower than the MAPE, indicating the mean error is influenced by outliers.
- The results suggest the StL and ground truth estimates for low volume sites are statistically different, regardless of normality assumptions.

[5] While some sites with volume lower than 500 AADT were included in the validation, these are not broken out separately. There were only 10 such sites, which is an insufficient sample size to draw meaningful conclusions.

Visualizations provided in Figure 16 through Figure 18 provide additional insight into the comparison of StL and ground truth estimates for urban sites.

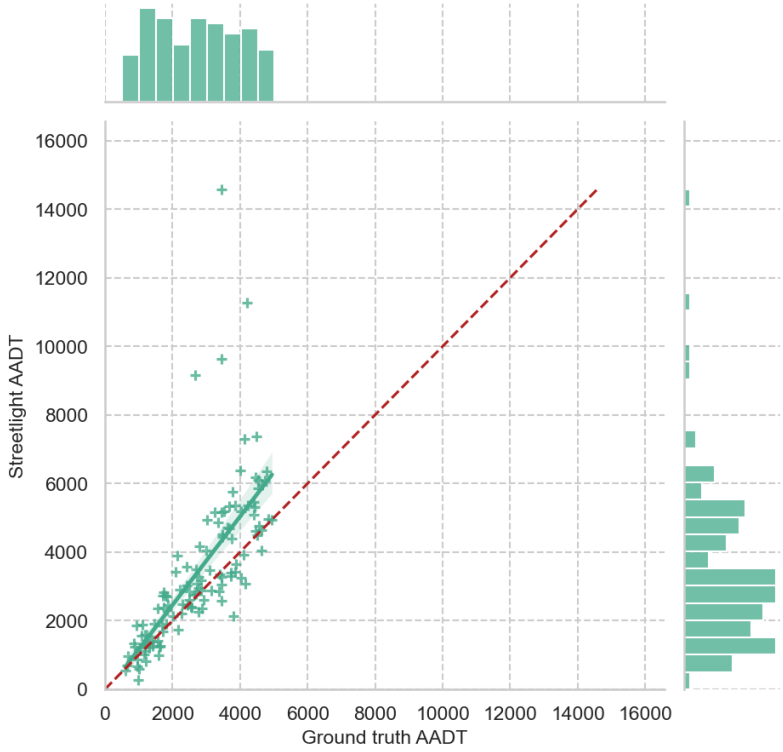

Figure 16. Scatterplot of Ground Truth and Streetlight AADT for Low Volume Sites

For low volume sites, the StL AADT estimates are less closely matched with ground truth AADT than for other comparison groups. StL estimates have a tendency toward overestimation, such that the regression line deviates considerably from the line of equality. This appears to be at least partly attributable to a small number of outliers with high error.

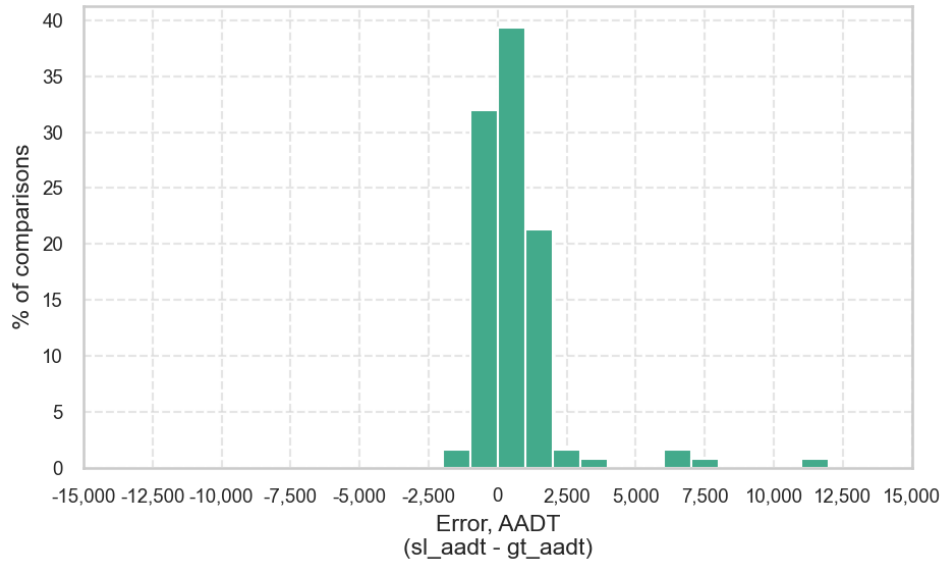

Figure 17. AADT Error for Low Volume Sites

The observed error for the StL AADT estimates is positively skewed with a few notable outliers.

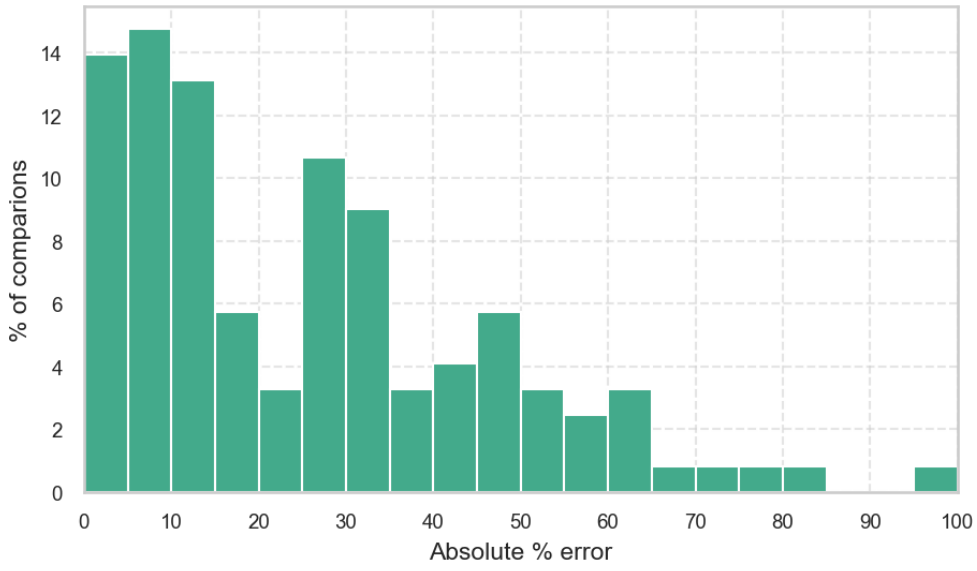

Figure 18. Absolute Percentage Error for Low Volume Sites

Compared to other groupings, absolute percentage errors for low volume sites tend to be less clustered at the lower end and instead are more evenly distributed up to around 50%. Half of all observations have an absolute percentage error less than 23.2%.

5.6 Medium Volume Sites (5,000-54,999 AADT; n=363)

The medium volume range covers roadways with AADT between 5,000 and 54,999. As such, there are a wide range of roadway types represented in this category, and this volume grouping accounts for the largest number of sites. Less aggregated results may be found at https://github.com/NREL/fhwa-streetlight-aadt-validation.

Table 10. Summary of Model Validation Results for Medium Volume Sites

Validation Metric	Result
AADT error percentiles	2.5th: -4,693 16th: -1,203 50th: +893 84th: +3,085 97.5th: +8,577
95th percentile error range	-27% to 46%
Median percentage error (bias)	+5.3%
Median Absolute Percentage Error (MdAPE)	10.9%
Mean Absolute Percentage Error (MAPE)	15.6%
Normalized Root Mean Square Error	10%
Spearman's rho	0.96
Paired sample t-test	0.000
Wilcoxon matched-pairs signed-ranks test	0.000

Key Takeaways (Medium Volume Sites)

- 68% of observations fall within an error range of -1,203 to +3,085 AADT
- 95% of observations fall within an error range of -4,693 to +8,577 AADT
- Consistent with all other groupings, medium volume sites are positively biased (+5.3% median error), albeit to a much lesser extent than low volume sites.
- Large positive outliers contribute to large MAPE relative to MdAPE.
- NRMSE of 8% for medium volume sites is the lowest among all reported groupings. The relatively large sample size may contribute to a lower NRMSE.
- The results suggest the StL and ground truth estimates are statistically different for medium volume sites, regardless of normality assumptions.

Visualizations provided in Figure 19 through Figure 21 provide additional insight into the comparison of StL and ground truth estimates for medium volume sites.

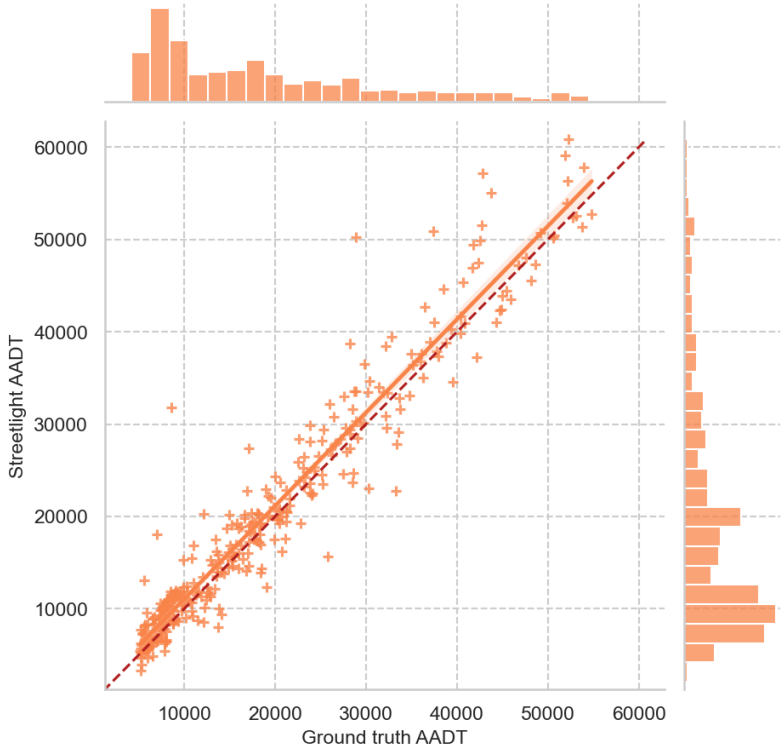

Figure 19. Scatterplot of Ground Truth and Streetlight AADT for Medium Volume Sites

The StL AADT estimates for medium volume sites are generally correlated with ground truth AADT. However, there is a slight tendency toward overestimation.

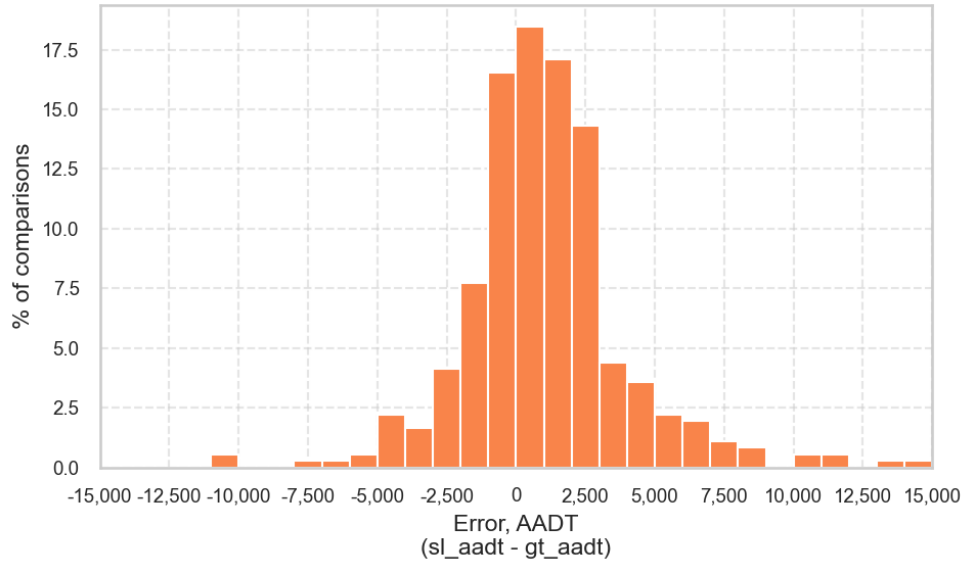

Figure 20. AADT Error for Medium Volume Sites

The observed error for the StL AADT estimates is positively skewed with a few notable outliers.

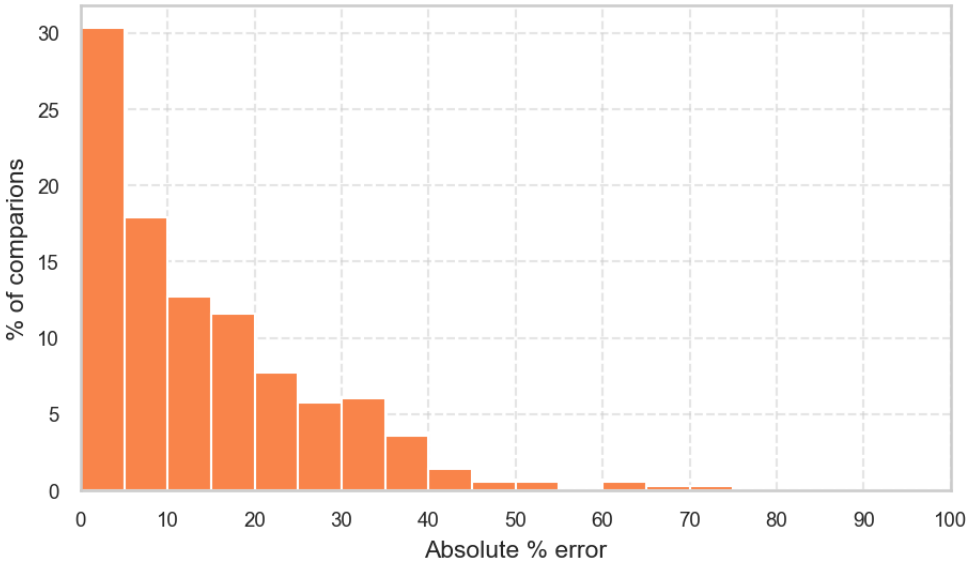

Figure 21. Absolute Percentage Error for Medium Volume Sites

Half of all observations have an absolute percentage error less than 5.3%.

5.7 High Volume Sites (55,000+ AADT; n=71)

High volume sites with AADT above 55,000 are evaluated in this section.

Table 11. Summary of Model Validation Results for High Volume Sites

Validation Metric	Result
AADT error percentiles	2.5th: -14,129 16th: -5,256 50th: +1,055 84th: +12,116 97.5th: +26,506
95th percentile error range	-17% to +32%
Median percentage error (bias)	+1.1%
Median Absolute Percentage Error (MdAPE)	7.3%
Mean Absolute Percentage Error (MAPE)	10.1%
Normalized Root Mean Square Error	21%
Spearman's rho	0.90
Paired sample t-test	0.014
Wilcoxon matched-pairs signed-ranks test	0.063

Key Takeaways (High Volume Sites)

- 68% of observations fall within an error range of -5,256 to +12,116 AADT
- 95% of observations fall within an error range of -14,129 to +26,510 AADT
- The StL estimates for high volume sites exhibit the lowest median error among all groupings (1.1%).
- Although MAPE is higher than MdAPE due to some large positive outliers, the difference MAPE and MdAPE is lower than in other groupings.
- NRMSE for high volume sites is relatively large, indicating a higher degree of variance than other groupings. This may be attributable to a relatively low sample size for high volume sites.
- The results of the Wilcoxon matched-pairs signed-ranks test indicate the median ranks of the StL and ground truth estimates are not statistically different. The paired sample t-test reaches a different conclusion, but the error distribution suggests the Wilcoxon test is more appropriate.

Visualizations provided in Figure 22 through Figure 24 provide additional insight into the comparison of StL and ground truth estimates for urban sites.

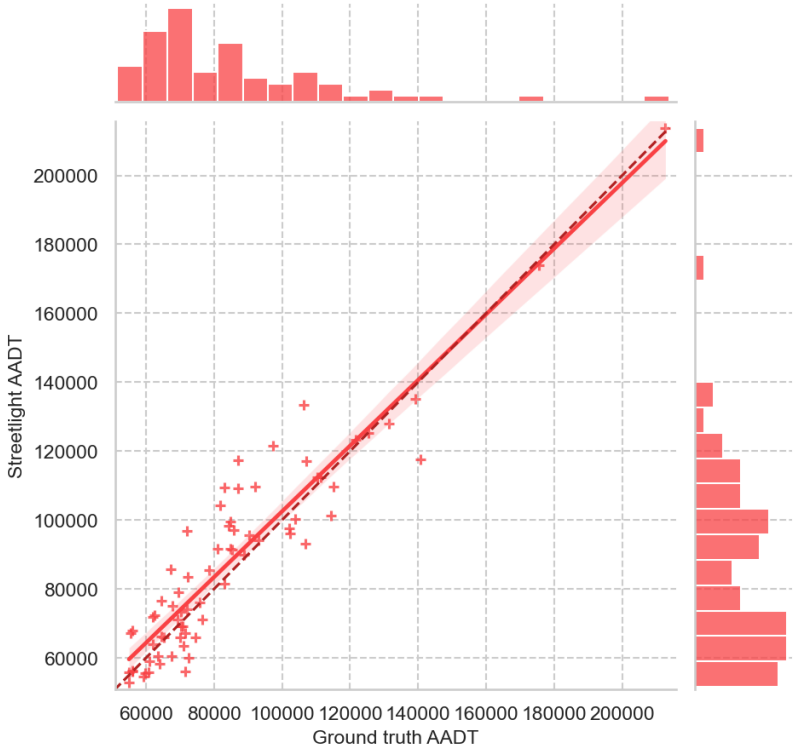

Figure 22. Scatterplot of Ground Truth and Streetlight AADT for High Volume Sites

The StL AADT estimates for high volume sites are generally correlated with ground truth AADT. There is no distinct pattern for over- or under-estimation.

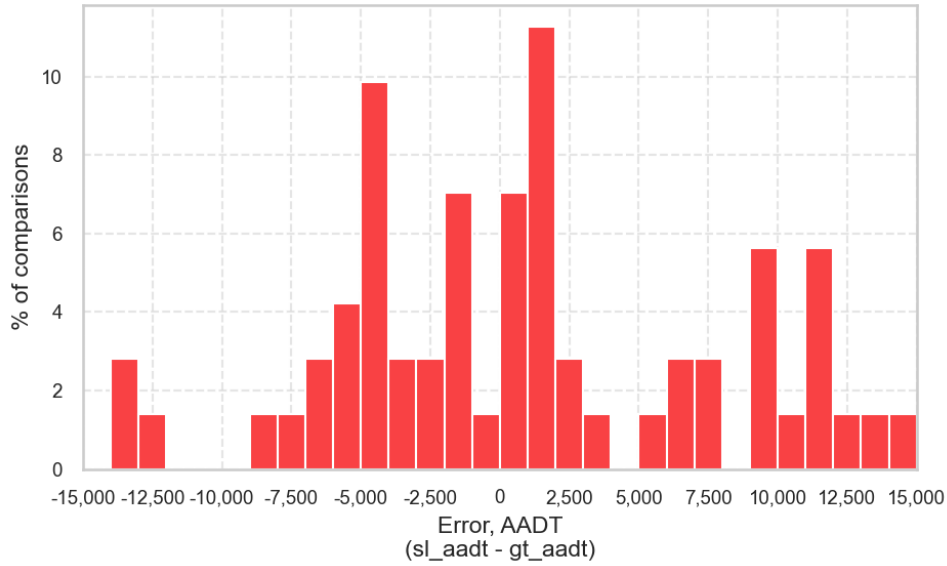

Figure 23. AADT Error for High Volume Sites

The observed error for the StL AADT estimates at high volume sites does not follow a predictable distribution.

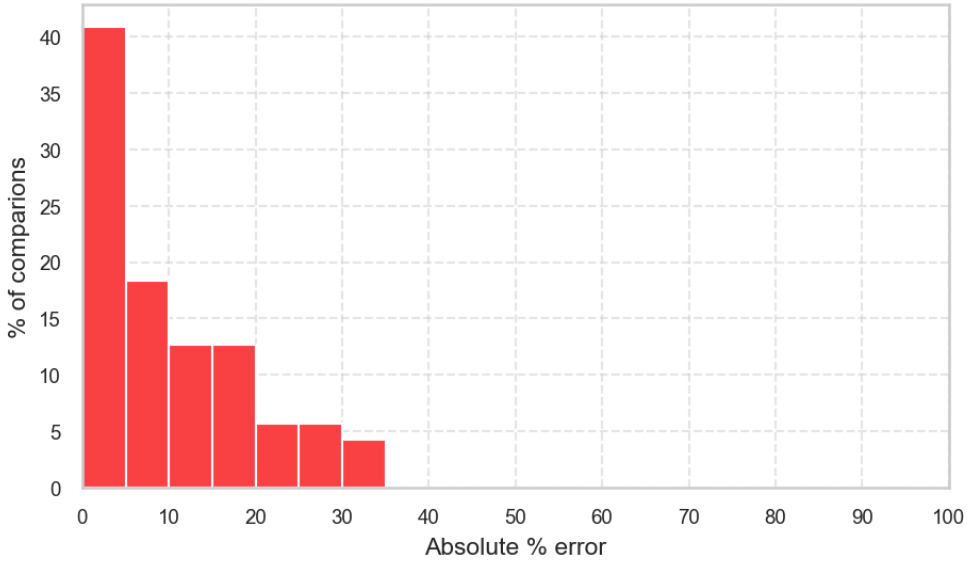

Figure 24. Absolute Percentage Error for High Volume Sites

Half of all observations have an absolute percentage error less than 7.3%. In general, the absolute percentage error for high volume sites is the lowest among all groupings.

6 Discussion and Conclusion

6.1 Interpretation of Findings

The results presented in this report demonstrate strong potential for the use of non-traditional approaches for estimating AADT. StL's estimates are well correlated with ground-truth data for all sites and across the different volume ranges evaluated. The overall observations are:

- Based on the statistical hypothesis test results, it is concluded that there are statistical differences between the AADTs obtained from the ground truth locations and StL's estimates. At a significance level of 0.05, the difference is unlikely due to random variation.
- StL results tend to overestimate relative to ground truth.
- Large positive outliers were observed across groupings, undermining the accuracy and reliability of the estimates.
- Errors at toll locations were generally higher and contribute to lower accuracy reported in various groupings. This may be due to complex geometries, differing vehicle occupancy rates, or other factors associated with toll locations.

From a model calibration and validation standpoint, the more calibration data points a model uses, the more "accurate" the model tends to perform. However, as the number of calibration points increases over certain thresholds, the return on improving model "accuracy" diminishes. Given the number of such calibration data points (permanent continuous traffic monitoring stations) available, it is suspected that StL's model may benefit from more calibration or training data points. Similarly, this validation might have benefited from a larger number of ground truth sites.

6.2 Potential Sources of Error

There are a few potential sources of error that may go a long way in explaining the discrepancies observed between StL and ground truth estimates. From the standpoint of the validation dataset, it must be noted that while the ground truth estimates are thought to be reliable, they are not expected to be completely error free. Equipment failure or calibration errors are an unavoidable aspect of traffic data collection. While substantial effort was made to ensure the reliability of ground truth data sources and remove unreliable data, the possibility of error in the ground truth data cannot be entirely ruled out. This type of error could also impact StL's training dataset.

Another potential error source related to the validation dataset relates to the geospatial matching of counter locations with OSM segments. As discussed above, extensive quality control was conducted to ensure the best possible match between counters and segments, and the validation team erred on removing locations that seemed even slightly suspicious. However, the process of checking for alignment between counter locations and OSM segments still leaves some room for error. This is particularly true where counters are located at or near a ramp, and the OSM segmentation might not correspond exactly with the lanes counted.

Outside of the validation dataset, it seems likely that complex geometry might contribute to error on the part of StL's estimates. For example, toll locations often count vehicles adjacent to other

through lanes, which could make it difficult to obtain an accurate tally from passive data sources. Without a greater understanding of StL's algorithms for assigning vehicles to OSM segments, it's unclear whether and to what extent this might explain errors observed at toll locations.

6.3 What is "good enough"?

The validation team, along with the Pooled Fund's Technical Advisory Committee, participated in extensive discussions revolving around the question of "what is good enough?". This research effort does not make a determination as to what is an acceptable level of error, as that depends on the application. Conventional hypothesis-based statistical tests are one approach, but it isn't clear that the use of such tests for comparing model predictions with ground truth data collection is reasonable. This perspective is reinforced by the likelihood of some amount of error in ground truth data, as noted previously.

Other more commonly used model accuracy metrics, such as those provided in this report, offer another approach. These metrics do not offer a clear pass/fail threshold, but taken together, they provide helpful evidence to inform decisions around whether to use non-traditional AADT estimates to supplement or replace aspects of existing traffic monitoring programs. The NREL validation team believes the answer to the question of "what is good enough?" should be considered on a case-by-case basis, depending on the needs of the agency seeking AADT estimates and whether the risk of using data with the identified error is acceptable.

6.4 Recommendations for Future Research

This report and the associated Pooled Fund project represent an important step toward integrating non-traditional AADT estimation into traditional traffic monitoring programs. As the field of non-traditional AADT estimation is rapidly evolving, it is likely that the accuracy and precision of AADT estimates will improve over time and additional work in this area will uncover new insights about the utility of passive data collection. There are many areas of opportunity for future research along these lines, including:

- Understanding and controlling for the influence of complex geometries or other roadway configurations on AADT estimation,
- Identifying additional factors that may influence probe penetration and associated accuracy of AADT, such as socio-demographics or varying levels of engagement with smartphones during trips.
- Identifying and improving the accuracy of passive data collection for other traffic measures, such as hourly or daily estimates, of for other modes of travel.
- Determining the appropriate number of ground truth traffic monitoring stations to support robust model calibration and validation for varying applications.

U.S. Department of Transportation
Federal Highway Administration

U.S. Department of Transportation
Federal Highway Administration
Office of Highway Policy Information
1200 New Jersey Ave., SE
Washington, D.C. 20590
https://www.fhwa.dot.gov/policyinformation
September 2021
FHWA-PL-21-033