VTRC
Virginia Transportation
Research Council

We Bring Innovation to Transportation

Modeling Travel Time Reliability for Non-Interstate National Highway System Routes

https://vtrc.virginia.gov/media/vtrc/vtrc-pdf/vtrc-pdf/26-R23.pdf

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

CHIEN-LUN LAN, Ph.D.
Research Scientist

Final Report VTRC 26-R23

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-R23	2. Government Accession No.:	3. Recipient's Catalog No.:
4. Title and Subtitle: Modeling Travel Time Reliability for Non-Interstate National Highway System Routes		5. Report Date: November 2025 6. Performing Organization Code:
7. Author(s): Mo Zhao, Ph.D., P.E.; Chien-Lun Lan, Ph.D.		8. Performing Organization Report No.: VTRC 26-R23
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.: 123124
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:		

16. Abstract:

Under the Moving Ahead for Progress in the 21st Century Act, state departments of transportation are required to report travel time reliability and set performance targets for interstate and non-interstate National Highway System facilities. Although several Virginia Transportation Research Council studies have analyzed and predicted travel time reliability on interstates, a gap still exists in studies modeling travel time reliability on arterial roads. Using 4 years of INRIX and National Performance Management Research Data Set (NPMRDS) probe data, this study conducted a segment-level comparative analysis of travel time distributions and the Level of Travel Time Reliability (LOTTR). Results show that NPMRDS systematically reports higher travel times than INRIX, inflating reliability metrics. Consequently, NPMRDS classifies significantly more segments as unreliable, especially on short (less than 0.25-mile) segments. Applying sample size filters to exclude segments with low sample sizes does not substantially reduce the discrepancies between NPMRDS and INRIX travel times. Analysis reveals that a major contributing factor is the data density of NPMRDS. More than 80% of observations fall into the lowest category (density A), leading to greater volatility in the higher percentiles of travel times and inflated Level of Travel Time Reliability values.

In addition, this study developed planning-level models to predict the 50th, 80th, and 95th percentile travel times on arterial segments statewide. Linear mixed models, random forests, and Light Gradient Boosting Machine models were built using data from multiple sources, and model performances were compared. The results showed that linear mixed models perform reasonably well for segments with LOTTR below 1.5, but machine learning approaches offer superior accuracy across the full reliability spectrum. The study also found that the machine learning models trained with INRIX data underestimate the travel time percentiles calculated using NPMRDS data, with a mean absolute percentage error of more than 10 for the 50th percentile travel time.

The study recommends that the Virginia Department of Transportation further develop models trained with NPMRDS data to meet federal reporting requirements and improve accuracy. In addition, the Virginia Department of Transportation should consider improving consistency in travel time reliability analyses across divisions and programs where applicable. Researchers developed an implementation plan that includes stakeholder engagement, technical guidance development, and support for the next target-setting cycle of the system reliability performance measures set forth in 23 CFR Part 490 (May 30, 2025).

17. Key Words: Travel time reliability, National Highway Systems, NPMRDS, machine learning, performance measures, MAP-21 performance reporting	18. Distribution Statement: No restrictions. This document is available to the public through NTIS, Springfield, VA 22161.		
19. Security Classif. (of this report): Unclassified	20. Security Classif. (of this page): Unclassified	21. No. of Pages: 65	22. Price:

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

FINAL REPORT

MODELING TRAVEL TIME RELIABILITY FOR NON-INTERSTATE NATIONAL HIGHWAY SYSTEM ROUTES

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

Chien-Lun Lan, Ph.D.
Research Scientist

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

Charlottesville, Virginia

November 2025
VTRC 26-R23

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

Under the Moving Ahead for Progress in the 21st Century Act, state departments of transportation are required to report travel time reliability and set performance targets for interstate and non-interstate National Highway System facilities. Although several Virginia Transportation Research Council studies have analyzed and predicted travel time reliability on interstates, a gap still exists in studies modeling travel time reliability on arterial roads. Using 4 years of INRIX and National Performance Management Research Data Set (NPMRDS) probe data, this study conducted a segment-level comparative analysis of travel time distributions and the Level of Travel Time Reliability (LOTTR). Results show that NPMRDS systematically reports higher travel times than INRIX, inflating reliability metrics. Consequently, NPMRDS classifies significantly more segments as unreliable, especially on short (less than 0.25-mile) segments. Applying sample size filters to exclude segments with low sample sizes does not substantially reduce the discrepancies between NPMRDS and INRIX travel times. Analysis reveals that a major contributing factor is the data density of NPMRDS. More than 80% of observations fall into the lowest category (density A), leading to greater volatility in the higher percentiles of travel times and inflated Level of Travel Time Reliability values.

In addition, this study developed planning-level models to predict the 50th, 80th, and 95th percentile travel times on arterial segments statewide. Linear mixed models, random forests, and Light Gradient Boosting Machine models were built using data from multiple sources, and model performances were compared. The results showed that linear mixed models perform reasonably well for segments with LOTTR below 1.5, but machine learning approaches offer superior accuracy across the full reliability spectrum. The study also found that the machine learning models trained with INRIX data underestimate the travel time percentiles calculated using NPMRDS data, with a mean absolute percentage error of more than 10 for the 50th percentile travel time.

The study recommends that the Virginia Department of Transportation further develop models trained with NPMRDS data to meet federal reporting requirements and improve accuracy. In addition, the Virginia Department of Transportation should consider improving consistency in travel time reliability analyses across divisions and programs where applicable. Researchers developed an implementation plan that includes stakeholder engagement, technical guidance development, and support for the next target-setting cycle of the system reliability performance measures set forth in 23 CFR Part 490 (May 30, 2025).

FINAL REPORT

MODELING TRAVEL TIME RELIABILITY FOR NON-INTERSTATE NATIONAL HIGHWAY SYSTEM ROUTES

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

Chien-Lun Lan, Ph.D.
Research Scientist

INTRODUCTION

Under the Moving Ahead for Progress in the 21st Century Act (MAP-21), state departments of transportation must report and set targets for travel time reliability on both interstate and non-interstate segments of the National Highway System (NHS). The Virginia Department of Transportation (VDOT) and the Commonwealth Office of Intermodal Planning and Investment (OIPI) already rely on reliability metrics to communicate system performance and inform investment decisions. To help VDOT meet the federal performance reporting and target-setting requirement better, the Virginia Transportation Research Council (VTRC) conducted several studies to analyze and predict travel time reliability of interstate highways. A VTRC study by Zhang et al. (2021), *Methods to Analyze and Predict Interstate Travel Time Reliability* (VTRC 22-R2), and its implementation follow-up study, *Prediction of Interstate Travel Time Reliability: Phase II* (VTRC 23-R12) by Zhao and Appiah (2023), demonstrated that a generalized random forest (GRF) approach can predict interstate travel time reliability with reasonable accuracy with manageable data-processing effort. Both studies recommended that VTRC conduct additional research to explore the data sources and modeling techniques for analyzing and predicting travel time reliability on non-interstate NHS roads.

Numerous studies in the literature have developed models to predict travel time reliability, and these studies adopt a variety of reliability definitions and performance measures. The second Strategic Highway Research Program (SHRP2) provides valuable sources of travel time reliability-related information for state departments of transportation. The SHRP2 Project L03 developed linear regression models to predict the 10th, 50th, 80th, 90th, and 99th travel time indices for peak hours (Cambridge Systematics et al., 2013). The SHRP2 Project L07 modified the L03 model to predict the travel time index for all hours (Potts et al., 2014). The SHRP2 Project C11 developed regression models to estimate the percentiles of the travel time index for a road segment as a function of the mean travel time index (Economic Development Research Group et al., 2013). These SHRP2 methods are developed using field data and simulated data from a limited number of corridors. A study by Farzana (2018) evaluated three methods developed from the three SHRP2 projects to estimate the 50th, 80th, and 90th travel time index using data collected from one freeway corridor and one arterial corridor and found that these methods achieve reasonable prediction accuracy for both freeway and arterials. On the other hand, Hadi et al. (2014) found that the SHRP2 L07 method failed to produce reasonable estimates. Most models developed in the literature are for reliability estimation at the corridor level. However, for the NHS system performance target-setting required in 23 CFR Part 490 (May 30, 2025), the future travel time reliability at the segment level in the statewide system needs to be predicted. Currently at VDOT, travel time reliability targets are determined using a historical trend line method. This method is limited in its ability to predict reliability changes in the opposite direction of the fitted trend lines and account for the changes of factors influencing reliability. Furthermore, it cannot generate forecasts for new road segments. Therefore, a need exists to develop enhanced prediction models that can incorporate diverse factors potentially affecting travel time reliability.

VDOT Traffic Operations Division and OIPI often need to evaluate the reliability of a large number of road segments in different areas across the state. This study seeks to develop a methodology for estimating travel time reliability at a planning level across all statewide NHS arterial segments opposed to generating a more precise estimate for a specific corridor. Under the guidance of this study's technical review panel, the researchers focused on predicting several critical travel time percentiles rather than a singular reliability measure. The models to be developed must achieve a balance between predictive accuracy and ease of use for VDOT staff.

Understanding the factors influencing travel time reliability is essential for developing models to predict travel time reliability. Polus (1979) was one of the early attempts to model travel time reliability using arterial travel times from the Chicago area, where the arterial travel time, under some assumptions, follows a gamma distribution. Chen et al. (2003) demonstrated that percentile-based metrics reflect user-perceived service quality better than mean travel times alone. van Lint and van Zuylen (2005) also argued that the shape of the daily travel time distribution, for example, the width of 90th to 10th percentiles, is a better indicator of unreliability than either mean or variance. Dong and Mahmassani (2009) showed that the long right tail of the distribution is tightly coupled to breakdown probabilities when volumes hover near capacity. Additional background on the development and use of travel time reliability metrics can also be found in the literature reviews in VTRC reports 22-R2 and 23-R12 (Zhang et al., 2021; Zhao and Appiah, 2023).

From the literature, three primary challenges distinguish arterial reliability modeling from that of interstates:

1. *Traffic Signal Control*. Fu et al. (2001) showed that queue spillback can be the largest source of prediction error when cycle failure occurs. Yang (2006) utilized a state space approach and reached a similar conclusion that time-series models underestimated travel time when coordination between consecutive signals fails. Elefteriadou et al. (2008) also suggested considering intersection spacing, cycle length, and green ratios for arterial reliability models. However, other studies reported that signal timing factors can be less influential than geometric variables, such as left- and right-turn lanes and access density (Fils, 2012).
2. *Access Density and Land Use*. Fils (2012) found that driveway density, segment length, and the presence of two-way left-turn lanes explained a larger share of day-to-day travel time variability than posted speed limit or annual average daily traffic in corridor-level analyses. This finding highlights the disruptive role of frequent access points, particularly

in urban arterial environments. However, collecting such detailed geometric data statewide is resource intensive.

3. *Data Availability and Quality*. Unlike interstates, arterial networks often suffer from less complete and lower quality data. For example, third-party probe datasets, such as INRIX and the National Performance Management Research Data Set (NPMRDS), tend to have lower sample sizes and overage on arterials. A recent VTRC study comparing INRIX and NPMRDS datasets confirmed that data gaps and lower availability are more pronounced on arterial segments (Lan and Zhao, 2025).

Other factors from the literature are also reported to affect reliability. Zhao and Chien (2012) examined how adverse weather reduces traffic speed and leads to less travel time reliability. Kwon et al. (2011) decomposed travel times into additive contributions from incidents, weather, work zones, and special events, demonstrating no single dominating factor. Geometry also plays a role, in which Tu et al. (2006) found that shortening a weaving section below a critical threshold can significantly increase travel time variation.

Although VDOT and OIPI have conducted several studies on travel time reliability for interstate highways, relatively little is known about the factors influencing reliability on non-interstate Virginia highways. This gap is primarily due to the inherent variability of operating conditions, the broader and more complex influencing factors, and persistent data limitations. This study aims to address those challenges by evaluating reliability across VDOT-maintained non-interstate NHS arterials and developing predictive models.

PURPOSE AND SCOPE

The purpose of this project was to analyze and predict travel time reliability on VDOT-maintained, non-interstate NHS arterials. Specifically, the key objectives to achieve this purpose were to:

- Evaluate the completeness and quality of probe data sources on non-interstate NHS arterials.
- Assess the effect of data sources on the estimated travel time distributions and reliability measures.
- Investigate factors that affect travel time reliability on NHS arterials.
- Develop methods for estimating travel time reliability across all non-interstate NHS arterial road segments at a planning level.

This study focused on VDOT-maintained NHS arterials. The reliability metrics examined include the Level of Travel Time Reliability (LOTTR) and the 50th, 80th, and 95th percentile travel times.

METHODOLOGY

Researchers conducted the following major tasks to achieve the study objectives:

1. Data collection and preparation.
2. Comparative analysis of travel time distributions and reliability metrics.
3. Development of models for travel time reliability analysis and prediction.

Data Collection and Preparation

This study considered various factors that could affect travel time reliability on NHS arterials, including traffic characteristics, roadway geometry, crash and noncrash incidents, work zones, and adverse weather conditions. The study period was from 2021 to 2024. For comparative analysis of travel time distributions and reliability metrics, data in all 4 years were used. For developing prediction models, data from 2022 to 2024 were used to exclude the atypical traffic patterns associated with the COVID-19 pandemic.

The following sections describe the data sources and preparation procedures.

Data Sources

INRIX and NPMRDS Probe Datasets

This study used probe travel time data from both INRIX and NPMRDS. These datasets were downloaded through the RITIS Probe Data Analytics Suite (RITIS, 2025). This study used both datasets with the Traffic Message Channel (TMC) system, which defines the roadway network and organizes data spatially. The TMC system is a standardized digital referencing framework originally developed to deliver real-time traffic messages to navigation systems. In the context of transportation data analytics, TMCs define consistent and uniquely coded roadway segments, enabling systematic aggregation of speed, travel time, and volume data across space and time. Each TMC code is a nine-character alphanumeric identifier comprising a country code, location table reference, segment type and direction indicator, and a location identifier. The fourth character plays a dual role:

- A “+” or “–” denotes an external segment, typically representing a stretch of road between interchanges or intersections.
- A “P” or “N” denotes an internal segment, covering areas immediately after interchanges or within intersections.

These directional markers also imply travel direction: “+” or “P” typically indicate northbound or westbound, and “–” or “N” indicate southbound or eastbound. Understanding the distinction between internal and external segments is essential because traffic behavior often differs substantially around these locations. External segments tend to reflect broader traffic flow, whereas internal segments capture localized effects, such as merging near ramps and intersections.

The INRIX and NPMRDS data collected for this study include the following core fields:

- 15-minute average speeds.
- 15-minute traffic volume.
- Timestamps of travel time observations, recorded in local time zones.
- Segment metadata, including segment identifier (tmc_code), start and end coordinates (latitude and longitude), segment length, number of through lanes, and urban or rural designation.
- Data quality indicators, such as confidence scores (INRIX) and data density (NPMRDS), for which NPMRDS classifies data density as A, B, or C, depending on the number of reporting probe vehicles.

NPMRDS emphasizes field observed data and avoids imputation. All reported values are derived directly from observed probe vehicle trajectories (Schuman et al., 2024). NPMRDS uses a path processing algorithm that normalizes probe reporting frequencies and reduces bias from vehicles with slower speeds. Data density is classified into three categories—A: 1–4 vehicles; B: 5–9 vehicles; and C: 10 or more vehicles.

In contrast, INRIX uses a mix of real-time and historical data sources (RITIS, 2025). INRIX provides both current harmonic mean speed and historical average speed by the hour of day and the day of week. Each data record also includes a confidence score to reflect data quality:

- 30: High confidence, based on real-time data for that specific segment.
- 20: Medium confidence, based on real-time data across multiple segments or on a combination of expected and real-time data, or both.
- 10: Lower confidence, based primarily on historical data or road reference speeds.

For 15-minute data aggregation, the confidence score reflects an average of 15 underlying 1-minute confidence values. A 15-minute score of 30 means that all 15 1-minute intervals were based on real-time data (score 30). A score of 20 indicates that all 1-minute readings were derived from expected data (score 20). Scores between 20 and 30 reflect a blended dataset, combining both real-time and expected data.

VDOT-Maintained Data

The following data were collected from several VDOT databases:

- *Linear Reference System (LRS)*: LRS shapefiles were obtained to use as a backbone for matching the spatial information from TMC segments to the VDOT roadway system. The attributes used included the route name (RTE_NM), official state milepost, and traffic direction.
- *Annual Average Daily Traffic (AADT)*: At the time of this study, the 2024 volume data were not available in the INRIX and NPMRDS datasets. Historical AADT data were used to estimate 15-minute volumes in 2024 based on the growth rate from 2022 to 2023. The

information collected included the location (route names and start and end mileposts) and AADT from 2018 to 2023.

- *Crashes*: Crash attributes included the date and time of the crash, severity (fatal [K], serious injury [A], minor or possible injury [B], no apparent injury [C], property damage only [O]), and location information, including route names and mileposts.
- *Noncrash Incidents*: All traffic incidents during the study period were obtained, and the attributes used included the time of incident, time of incident cleared, incident type, number of lanes closed, and location (route name and milepost). The incident type indicates whether an incident was a crash (recorded as "accident" in the database) or other events.
- *Work Zones*: Work zone information included work zone start and end times, route names, start and end mileposts, the number of lanes affected, and the total number of lanes.
- *Weather Events*: Adverse weather events data included the level of impact (adverse, minor, medium, and severe), start and end locations represented by route name and milepost, and start and end times of the events.
- *Posted Speed Limit (PSL)*: PSL information included route name, road link start and end mileposts, and PSL on the link.
- *Traffic Signal*: VDOT-maintained information on traffic signals was collected. This information included the geographic coordinates and names of the major and minor roads at the intersections where signals are installed.

Data Quality and Availability

To analyze and enhance data reliability in this study, a filtered version of the INRIX dataset, referred to as INRIX-RT, was created. This version excludes any 15-minute intervals composed entirely of historical or expected data (i.e., intervals with a confidence score of 20 or less). Each retained interval in INRX-RT contains at least one real-time probe data point, ensuring a minimum level of real-time representation.

Data availability is crucial in estimating any measurements using probe datasets because it directly affects the accuracy and reliability of the analysis. Lan and Zhao (2025) discussed data quality and availability analysis on the INRIX and NPMRDS datasets in a separate VTRC report.

Data Preparation

Data from various sources were integrated through the data fusion process established in a previous VTRC study (Zhao and Appiah, 2023). Researchers determined the spatial information of the VDOT data by its position relative to LRS routes. Once the start and end points of the TMC segment are identified from LRS, all spatial data can be consolidated using the route name and milepost. However, the TMC segment locations cannot be directly identified from LRS. Aligning TMC segments with LRS is both challenging and time-consuming. This process employed ArcGIS spatial analysis tools, customized Python scripts, and manual validation. Although ArcGIS spatial analysis tools demonstrated efficacy for divided roadways, they exhibited reduced accuracy for routes where the LRS line features of both travel directions overlapped. For some TMCs, the start and end locations are on different routes, such as starting on a state route and ending on a local business route. Length-weighted average AADT and PSL were calculated for those segments. Customized scripts and manual verification were used to validate the results. A comprehensive manual verification process was conducted to ensure that the TMC segments were matched to the LRS line features of the intended travel direction, especially for the undivided two-lane highways, and the mileposts of the start and end points of the TMC segments were correctly identified. For the TMC segments that span multiple LRS routes, additional checks were performed to validate each portion's segmentation and attribution. Specifically, the length of the TMC segment associated with each route was carefully examined, and the corresponding start and end mileposts were cross-referenced to ensure consistency with route-specific data.

Travel time data from INRIX-RT, a filtered version of the INRIX dataset, were used to develop prediction models because of the project technical review panel's suggestion that INRIX is more widely used than NPMRDS at VDOT. The temporal information of VDOT-maintained data was synchronized with the 15-minute interval of the INRIX-RT timestamps. The study period included travel times during the a.m. peak (6 to 10 a.m.), p.m. peak (4 to 8 p.m.), mid-day (10 a.m. to 4 p.m.) on weekdays, and 6 a.m. to 8 p.m. on weekdays based on the MAP-21 performance measurement requirements in 23 CFR Part 490 (May 30, 2025).

Data Alignment Across Datasets and Years

To enable consistent comparison between the NPMRDS and INRIX datasets, TMC segments were aligned by matching their segment identifiers and start and end locations. Although the two datasets share similar TMC network definitions, they are not entirely identical, which necessitated careful cross-referencing to achieve consistent alignment. In addition, TMC definitions can vary across years, with segments being added or removed in both datasets. For the analysis of travel time distributions and reliability metrics, only TMCs that were present across all 4 study years were included, resulting in a standardized set of 6,029 TMCs covering approximately 6,763 directional miles of non-interstate NHS routes in Virginia. To build reliability prediction models, segments not presented in the entire study period were included to test the capacity of the models for predicting unseen conditions.

Comparative Analysis of Travel Time Distributions and Reliability Metrics

This task conducts a comparative analysis of travel time distributions and reliability metrics derived from three probe data sources: NPMRDS, INRIX, and INRIX-RT.

Travel time distributions were compared across datasets using percentile travel times (5th, 50th, 80th, 85th, and 95th percentiles) on matched TMC segments. In addition, the values of LOTTR, defined as the ratio of the 80th percentile to the 50th percentile travel time, calculated using the three datasets were also compared. Scatter plots and overlap analyses of unreliable segments (LOTTR $\geq$ 1.5) visualized and quantified differences across datasets. Detailed case studies examined how low sample size affects reliability measures.

The objective of this analysis is to develop a clear understanding of the characteristics of the three probe datasets by examining travel time distributions and reliability metrics calculated using them. By revealing how data quality and processing methodologies affect travel time distributions and performance measures, this analysis highlights the potential risks of drawing misleading conclusions when dataset differences are not carefully accounted for.

Develop Models for Travel Time Reliability Analysis and Prediction

The Zhang et al. (2021) study developed linear quantile mixed models and GRF models to estimate multiple travel time percentiles during peak hours for all interstates in Virginia. Between the two, the GRF quantile regression model outperformed the linear model. One key recommendation from that study was to explore the use of the GRF approach for predicting travel time reliability in nonpeak periods on interstates and to further develop prediction models for non-interstate NHS routes. Building on this recommendation, the implementation study by Zhao and Appiah (2023) expanded the GRF modeling framework to include all four MAP-21 time periods.

Following the procedure established in the implementation study (Zhao and Appiah, 2023), this study began by applying the GRF approach, which aggregates data at the hourly level for each of the four MAP-21 analysis periods on each day across all study years. It models travel time distributions on each road segment and can simultaneously predict the 1st through 100th percentiles. The GRF quantile regression for the interstate system with fewer than 2,000 TMC segments was manageable using a high-performance computing cluster. However, the arterial system, with more than 5,000 TMC segments, posed significant computational challenges. Preparing data and training models at this scale demanded substantial computing resources, often beyond the reach of most VDOT analysts. Even with access to high-performance computing clusters, the researchers found that the model training process is time-consuming. Consequently, this study investigated alternative methods for forecasting travel times at the 50th, 80th, and 95th percentiles.

Three widely used approaches for traffic prediction, linear mixed models (LMM), Light Gradient Boosting Machines (LightGBM), and random forests (RF) were examined to predict the 50th, 80th, and 90th percentile travel times. Unlike the GRF quantile regressor, which can predict multiple percentiles simultaneously, these three models require separate training for each percentile. However, the methods rely on less granular data and demand significantly fewer computational resources, making them more practical for implementation by VDOT analysts. The following sections describe each of these methods in more detail.

Linear Mixed Model

Linear Mixed Model (LMM) is a statistical model that incorporates both fixed effects (i.e., predictors that remain consistent across all observations, such as traffic volume) and random effects (i.e., predictors that vary across groups, such as individual TMC segments). This feature makes LMMs particularly suitable for hierarchical or grouped data structures. Because measurements for each TMC are often repeated over time, LMMs can capture within-group variability and improve estimation accuracy. Although LMMs are relatively easy to interpret and provide insight into the influence of each predictor, they are limited in their ability to model nonlinear relationships or complex interactions. This study used the Python statsmodels package to implement LMM. The five-fold validation using training data is employed to validate the trained models.

Random Forest

Random Forest (RF) is an ensemble learning model that constructs multiple decision trees and aggregates their predictions to improve accuracy and reduce variance. They are well suited for predictions involving nonlinear relationships and noisy data. RF can capture complex interactions between features, such as the combined effects of traffic volume, roadway geometry, and traffic incidents. However, they can become computationally intensive and less interpretable, especially when the number of trees is large. In this study, the Python RandomForestRegressor in the scikit-learn library was used to build the RF models. The hyperoperators tuned included the number of trees, maximum depth of trees, minimum samples to split, minimum samples per leaf, and the number of features for split. The tuning of hyperparameters was based on Bayesian optimization with the Python Optuna library.

Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is a machine learning technique based on gradient boosting decision trees. It is designed for speed and efficiency, making it capable of handling large-scale datasets with high-dimensional features. A previous study found LightGBM is particularly effective in modeling complex nonlinear relationships in traffic data, such as interactions between the time of day and historical traffic patterns (Cheng et al., 2022). LightGBM supports categorical variables and can manage missing data without preprocessing. Although its predictive performance is typically high, LightGBM models are less interpretable than linear models. Tuning hyperparameters is often more complicated than RF models, and the prediction accuracy is more sensitive to the selection of hyperparameters. LightGBM is known for its speed and efficiency for large datasets, but it can overfit or fail to generate solutions with small datasets. This study used the Python LightGBMRegressor to build prediction models. The hyperparameters tuned included the number of boosting trees, learning rate, maximum number of leaves in each tree, maximum depth of trees, minimum number of data needed in a leaf, subsample ratio of the training instance, subsample ratio of columns when constructing each tree, Lasso regularization term on weights, and Ridge regularization term on weights. The evaluation metric was the mean absolute error, and the Bayesian optimization of hyperparameters was implemented with the Python Optuna library.

SHapley Additive exPlanations

Because of the complexity of interpreting machine learning models, this study uses SHapley Additive exPlanations (SHAP) to help interpret their outputs. SHAP offers a method to clarify the results of any machine learning model by utilizing a game theory approach that assesses each participant’s contribution to the prediction result. SHAP assigns a Shapley value to each variable based on its marginal contributions across all possible combinations, providing a unique value to each variable to indicate its effect on the model’s output. The Python SHAP package was used to calculate feature importance.

Modeling Strategy and Evaluation

The VTRC study by Zhang et al. (2021) found that the development of linear quantile mixed models requires grouping TMC segments with similar travel time distributions to enhance predictive accuracy. That study further found a straightforward yet effective approach for interstate models, grouping TMCs by segment type (internal versus external) and reliability performance that is defined as reliable (LOTTR < 1.5) or unreliable (LOTTR ≥ 1.5). This study adopted the same grouping approach for LMM, classifying all TMC segments into internal and external segments based on the segment identifier (tmc_code). Within each segment type group, two subgroups were established according to the LOTTR values.

For machine learning models, the Zhang et al. (2021) study determined that prior grouping of segments was not necessary because of their ability to capture complex interactions inherently. To verify whether prior grouping improves model performance in the current context, the RF and LightGBM models in this study were developed both with and without such prior grouping. Initially, LMM, RF, and LightGBM models were constructed for the 50th, 80th, and 95th percentile travel times using data from peak traffic periods. The models demonstrating superior performance were then used to construct models for all other time periods.

Variables Selection and Model Evaluation

This study examined various factors that may influence travel time reliability on arterial roads. Table 1 lists the full set of variables explored. However, not all variables were retained in the final models. For work zones, the Zhang et al. (2021) study for interstate highways found the ratio of lanes affected was a good indicator compared with the count of work zones. This study considered the duration of the work zone in addition to the lane closure ratio using an intensity index. The work zone intensity is calculated by Equation 1:

$$Work\ zone\ intensity = \frac{Duration\ with\ work\ zone\ present}{Length\ of\ analysis\ time\ inteval} \times \frac{Number\ of\ lanes\ affected}{Total\ number\ of\ lanes} \times work\ zone\ length$$

(Equation 1)

Variable selection was primarily based on prediction accuracy derived from five-fold cross-validation of the training data, the significance of LMM coefficients, the variable importance in the RF and LightGBM models, and the Shapley values for the variables in the LightGBM models. Impurity-based variable importance was used for the RF models. A simplified method to demonstrate the RF importance score is illustrated in Equation 2:

$$Importance\ score\ (f) = \sum_{t \in trees} \sum_{s \in splits\ on\ f} \frac{n_s}{N_t} \cdot \Delta MSE_s \qquad \text{(Equation 2)}$$

Where:
n_s = the number of samples at the split.
N_t = the number of samples in the tree.
ΔMSE = the decrease in mean squared error (MSE) from the split.

Each time a feature is used to split a node, the algorithm measures the extent to which the MSE is reduced by the split. The values range between 0 and 1, and the importance scores sum to 1 for all variables.

Table 1. Variables Explored in This Study

Variable Category	Variable	Variable Name
Geometric features	Segment length (miles)	miles
	Number of through lanes (count)	throulanes_unidir
Traffic management	Traffic signal density (decimal)	signal_density
	Posted speed limit (mph)	PSL
Area type	Area type (urban = 0; rural = 1)	rural_des
Weather	Count of advisory and minor weather events (count)	count_1
	Count of medium impact weather events (count)	count_2
	Count of severe weather events (count)	count_3
	Count of total weather events (count)	Weather_count
Crashes	Frequency of fatal and severe injury crashes (count)	S
	Frequency of visible injury crashes (count)	B
	Frequency of nonvisible injury crashes (count)	C
	Frequency of property damage-only crashes (count)	O
	Count of total crashes (count)	Crash_count
Noncrash incident	Frequency of shoulder closure noncrash incidents (count)	shoulder_count
	Frequency of travel lane closure noncrash incidents (count)	lane_count
	Total number of noncrash incidents (count)	
Work zone	Work zone intensity	wz_intensity
	Lane closure ratio (lanes affected and total number of lanes)	Close_ratio
Traffic demand	Average 15-minute volume	Vol_15
	Percentage of heavy vehicles (decimal)	truck_pct

For LightGBM, a split-based importance score was utilized, indicating the contribution of each variable to the model's prediction by counting the number of times a feature was used to split the data across all trees. SHAP provides variable importance using Shapley values. The calculation of Shapley values considers both the main effect and the interaction effect. Main effect measures the individual effect of a variable on the model's output, ignoring interactions with other variables. This interaction effect captures how a variable's impact on the prediction depends on the value of another variable. Compared with the variable importance scores provided by RF and LightGBM, SHAP provides a better understanding of how each variable contributes to a specific prediction.

In addition to variable importance scores and Shapley values, the correlations among variables and the variance inflation factors were calculated for each model to identify multicollinearity among the variables.

The LMM, RF, and LightGBM models were created using different combinations of variables. The variable options for weather, crashes, noncrash incidents, and work zones were tested individually. For each combination of input variables, 12 models were created (3 target percentiles × 2 TMC segment types × 2 reliability subgroups) using peak-hour data (6–10 a.m. and 4–8 p.m. on weekdays). For LMM, to ensure the accuracy of predictions, variables with a significance level greater than 0.1 were excluded from the input data to mitigate potential overfitting concerns, and models were re-estimated using the subset of variables that were statistically significant ($p < 0.1$). The variable importance scores and Shapley values of the RF and LightGBM models were used to determine whether to include a variable.

The best performing models using peak-hour data were then selected to develop models for weekends and mid-days. Two methods were evaluated: a unified model applicable to all four time periods and four distinct models tailored for specific periods. The variable selection process used was analogous to that used for peak-hour data.

Both the training and testing errors were measured using the following four criteria, with the training error being assessed through five-fold cross-validation.

- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)
- Mean squared error (MSE)
- R-squared

RESULTS AND DISCUSSION

Comparative Analysis of Travel Time Distributions and Reliability Metrics

The comparative analysis consists of two main stages: (1) evaluation of travel time distributions and (2) assessment of travel time reliability measures.

In the first stage, the analysis examines the shape of travel time distributions by comparing the 5th, 50th, 80th, 85th, and 95th percentiles across matched TMC segments in the NPMRDS, INRIX, and INRIX-RT datasets. This comparison highlights areas of agreement and divergence among the datasets and evaluates how these differences shift when a minimum sample size filter is applied. To further illustrate the effect of data sparsity, two rural case studies are presented, demonstrating how limited sampling can either exaggerate or suppress extreme values.

The second stage explores how these percentile-based distributions translate into the LOTTR metric. By comparing LOTTR values across datasets, the analysis reveals the influence of dataset selection on reliability outcomes. Specifically, it shows how dataset selection affects the number of TMC segments classified as unreliable and the consistency of the identified unreliable list.

Analysis of Travel Time Distributions

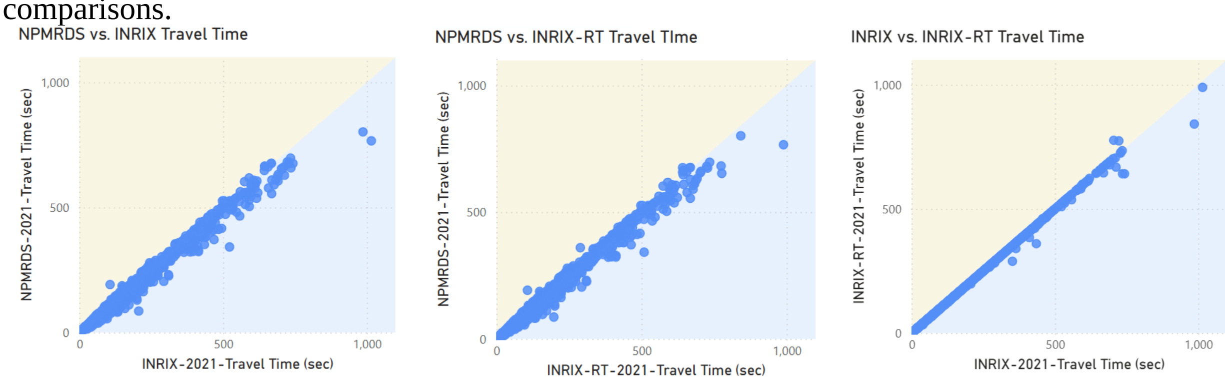
The first analysis focuses on identifying the underlying characteristics of travel time values across the three datasets by examining travel times at selected percentiles. This approach provides insight into the shape and behavior of the travel time distributions, particularly in the upper percentiles, which are critical to calculating travel time reliability metrics.

Specifically, travel time values at the 5th, 50th, 80th, 85th, and 95th percentiles are compared across the NPMRDS, INRIX, and INRIX-RT datasets. These percentiles are commonly used in the computation of reliability metrics, such as the Planning Time Index (PTI) and LOTTR, and differences in their values can significantly influence performance evaluations.

Because the data processing procedures applied by each probe data provider—including potential filtering, smoothing, or imputation—were not fully transparent, this analysis was intended to empirically reveal any artifacts or inconsistencies introduced during the transformation from raw probe data to TMC-level travel times. Particular attention was paid to the presence of extreme or irregular values and systematic differences in percentile estimates across datasets.

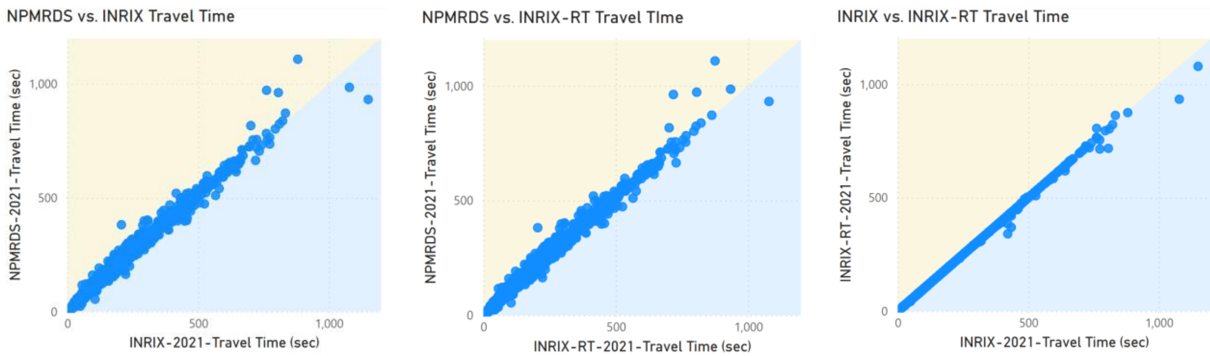
Comparison of Percentile-Based Travel Time Values

This section compares the travel time distribution derived from the NPMRDS, INRIX, and INRIX-RT datasets. The analysis focused on a.m. peak hours from 2021 and included 4,844 TMC segments present in all three datasets after excluding segments with no data or lengths shorter than 0.1 mile. Travel times at the 5th, 50th, 80th, 85th, and 95th percentiles were compared across datasets. Figures 1 through 5 provide scatter plots illustrating these comparisons.



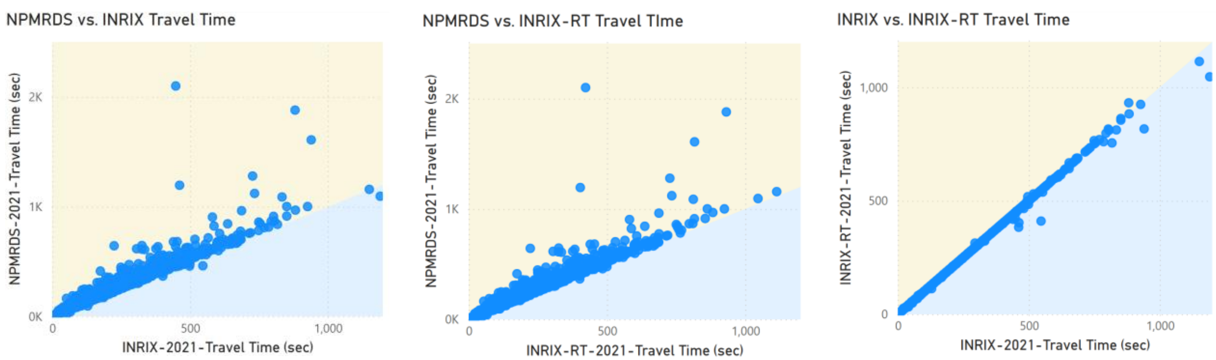
(a) (b) (c)

Figure 1. Comparison among 5th Percentile Travel Time Values across Traffic Message Channel Segments between: (a) NPMRDS and INRIX; (b) NPMRDS and INRIX-RT; (c) INRIX and INRIX-RT. NPMRDS = National Performance Management Research Data Set.



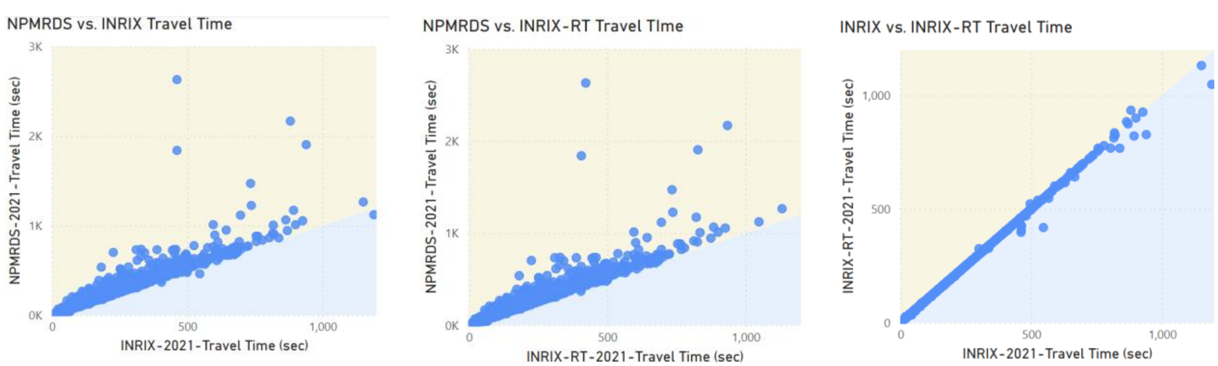
(a) (b) (c)

Figure 2. Comparison among 50th Percentile Travel Time Values across Traffic Message Channel Segments between: (a) NPMRDS and INRIX; (b) NPMRDS and INRIX-RT; (c) INRIX and INRIX-RT. NPMRDS = National Performance Management Research Data Set.



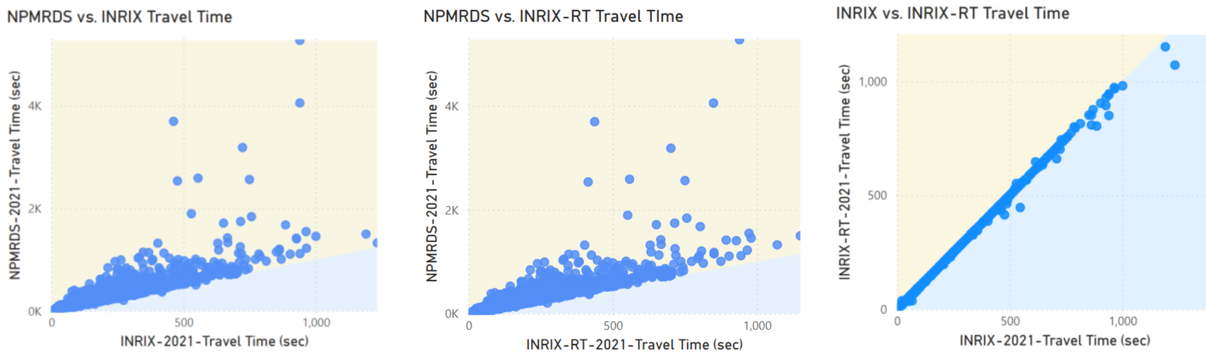
(a) (b) (c)

Figure 3. Comparison among 80th Percentile Travel Time Values across Traffic Message Channel Segments between: (a) NPMRDS and INRIX; (b) NPMRDS and INRIX-RT; (c) INRIX and INRIX-RT. NPMRDS = National Performance Management Research Data Set.



(a) (b) (c)

Figure 4. Comparison among 85th Percentile Travel Time Values across Traffic Message Channel Segments between: (a) NPMRDS and INRIX; (b) NPMRDS and INRIX-RT; (c) INRIX and INRIX-RT. NPMRDS = National Performance Management Research Data Set.



(a) (b) (c)

Figure 5. Comparison among 95th Percentile Travel Time Values across Traffic Message Channel Segments between: (a) NPMRDS and INRIX; (b) NPMRDS and INRIX-RT; (c) INRIX and INRIX-RT. NPMRDS = National Performance Management Research Data Set.

This analysis used all data from each database, without applying additional filters. Table 2 summarizes both the number and proportion of TMC segments for which NPMRDS reports higher travel time values than INRIX and INRIX-RT and the average percentage difference in reported travel times at each percentile.

Table 2. Comparison of Travel Time Percentiles between NPMRDS, INRIX, and INRIX-RT: Frequency and Average Differences (N = 4,844 Traffic Message Channels)

Percentile	**NPMRDS > INRIX, Count (%)**	**NPMRDS > INRIX-RT, Count (%)**	**Average Difference (NPMRDS – INRIX)**	**Average Difference (NPMRDS – INRIX-RT)**
5th	1,460 (30%)	1,462 (30%)	– 2.87%	– 2.71%
15th[a]	2,473 (51%)	2,478 (51%)	0.75%	0.87%
50th	4,020 (83%)	4,033 (83%)	8.13%	8.27%
80th	4,577 (94%)	4,580 (95%)	19.01%	19.16%
85th	4,630 (96%)	4,630 (96%)	23.14%	23.30%
95th	4,711 (97%)	4,711 (97%)	46.58%	46.76%

[a] Although the 15th percentile is not a standard metric for travel time reliability, it is included here to illustrate when one dataset reports travel times approximately 50% greater or less than the others. NPMRDS = National Performance Management Research Data Set.

At the 5th percentile, approximately 30% of segments show similar or slightly higher travel times in NPMRDS compared with INRIX and INRIX-RT, with modest average differences of -2.87% and -2.71%, respectively. This outcome suggests that under low-congestion or free-flow conditions, travel times are generally consistent across sources. This outcome is expected because the maximum operating speed governs the lower bound of travel time on a segment.

At the 50th percentile, the NPMRDS dataset starts to show systematically higher travel times compared with the other datasets. At the 50th percentile (median), NPMRDS shows higher travel times in approximately 83% of TMC segments, with an average difference exceeding 8%.

At higher percentiles, this divergence becomes more pronounced. At the 80th percentile, 94 to 95% of segments show higher travel times in NPMRDS, with an average difference of approximately 19%. This divergence increases to 96% at the 85th percentile and 97% at the 95th percentile, for which the average travel time difference peaks at 46.58% and 46.76%, respectively. In the upper percentiles, extreme values in NPMRDS are occasionally observed.

However, it remains unclear whether these data points represent true congestion spikes or anomalies in the dataset.

Analysis of Sample Size

The accuracy of travel time percentiles depends on adequate data per TMC segment. Segments with insufficient sample size may need to be excluded from the analysis to ensure the results are reliable. To understand the data availability, the cumulative distributions of 15-minute interval sample counts per TMC segment were examined for both NPMRDS and INRIX-RT datasets for the 2021 a.m. peak period (Figure 6).

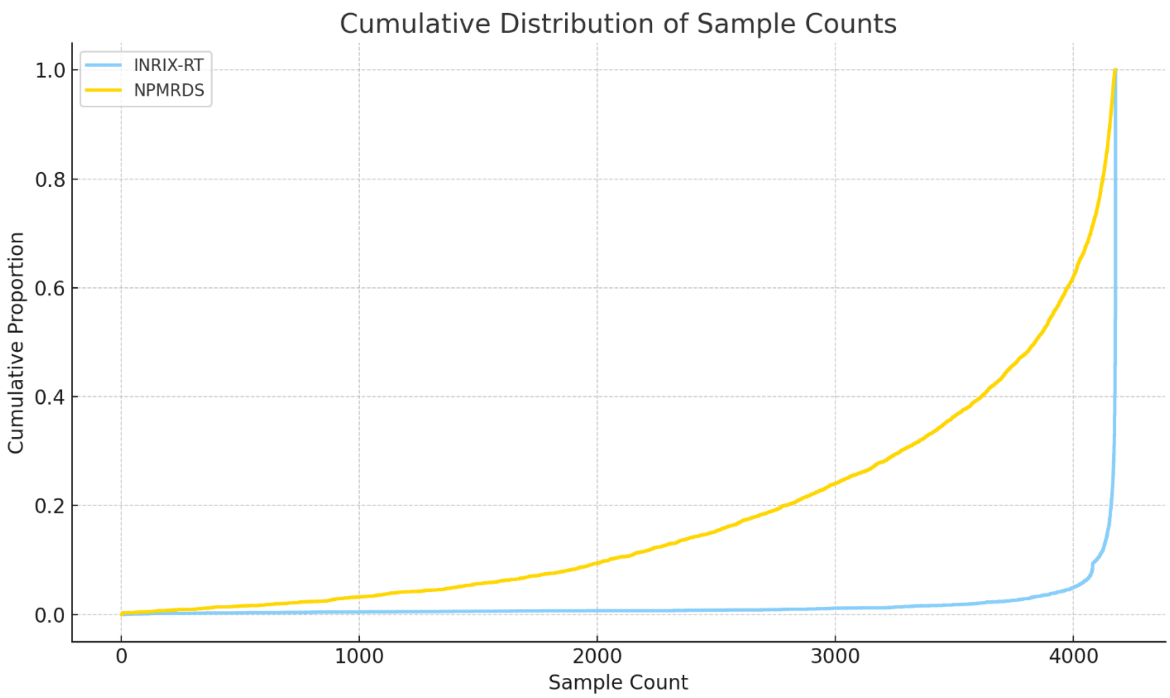


Figure 6. Cumulative Distribution of Sample Counts for NPMRDS and INRIX during a.m. Peak Periods in 2021. NPMRDS = National Performance Management Research Data Set.

Figure 6 illustrates notable differences in data density distributions between NPMRDS and INRIX-RT datasets. The INRIX-RT dataset generally provides much higher data density. To be more specific, more than 90% of TMC segments in INRIX-RT have sample counts exceeding 4,000 observations compared with less than 40% of segments in NPMRDS. To quantify the effect of potential minimum sample size filters on dataset coverage, Table 3 summarizes the number and proportion of additional TMC segments dropped under various sample size thresholds. These counts are calculated after already excluding segments shorter than 0.1 mile and segments with no data for the 2021 a.m. peak periods.

Table 3. Counts and Percentages of Traffic Message Channels Dropped at Different Sample Size Thresholds (a.m. Peak, 2021)

Sample Size Filter	Count of Traffic Message Channels Dropped			
	INRIX-RT		NPMRDS	
50	4	(0.08%)	16	(0.33%)
100	5	(0.10%)	23	(0.47%)
150	9	(0.18%)	32	(0.66%)
200	10	(0.20%)	37	(0.76%)
250	10	(0.20%)	44	(0.91%)
300	11	(0.22%)	45	(0.93%)
350	12	(0.24%)	54	(1.11%)
400	14	(0.28%)	66	(1.36%)
450	14	(0.28%)	67	(1.38%)
500	15	(0.30%)	75	(1.55%)
550	17	(0.33%)	79	(1.63%)
600	17	(0.33%)	85	(1.75%)
650	17	(0.33%)	95	(1.96%)
700	19	(0.37%)	100	(2.06%)
750	19	(0.37%)	108	(2.23%)
800	20	(0.39%)	115	(2.37%)
850	23	(0.45%)	118	(2.44%)
900	23	(0.45%)	137	(2.83%)
950	23	(0.45%)	146	(3.01%)
1000	23	(0.45%)	157	(3.24%)

NPMRDS = National Performance Management Research Data Set.

The analysis shows that the effect of dropping low-sample TMCs from the INRIX-RT dataset is minimal. To be more specific, less than 0.5% of TMCs would be dropped even at a filter of 1,000 samples. In contrast, NPMRDS would lose approximately 3% of segments under the same situation.

Given the observed distribution of sample counts and the trade-offs in data retention, a sample size threshold of 250 observations per TMC segment was selected for filtered analyses. In addition to minimizing the effect of dataset size (0.2% of INRIX-RT and 0.91% of NPMRDS TMCs dropped), the choice of 250 samples has a practical justification. Because the analysis focuses on a.m. peak periods, which are defined only for weekdays, the annual analysis window contains approximately 250 weekdays per year. Thus, a 250-sample threshold roughly corresponds to one observation per day. After applying this 250-sample filter to both INRIX-RT and NPMRDS, 4,786 matched TMC segments remained for the filtered comparison in the next section.

Noting that this sample size filtering applies only to INRIX-RT and NPMRDS datasets is important. The INRIX dataset, by design, fills all time stamps using a combination of real-time and historical or reference data. As a result, it contains no missing 15-minute intervals, and every TMC segment has full temporal coverage across the year. Thus, it is not necessary to apply a sample size filter to INRIX data. However, for INRIX-RT, which excludes intervals lacking real-time data, the sample count per segment can vary.

Analysis of Percentile Travel Time with Sample Size Filter

With the sample size threshold of 250 observations per TMC segment applied, a final sample of 4,786 matched TMC segments remained for comparative analyses. Travel times at the 5th, 50th, 80th, 85th, and 95th percentiles were compared across datasets. Figure 7 shows scatter plots illustrating these comparisons. Table 4 summarizes the number and proportion of TMC segments for which NPMRDS reports higher travel time values than INRIX-RT.

At the 5th percentile, approximately 30% of segments show similar or slightly higher travel times in NPMRDS compared with INRIX-RT. This proportion closely aligns with the unfiltered scenario (27%), indicating strong consistency across datasets at low percentiles, regardless of sample size filtering. The scatter plot in Figure 7a demonstrates tight clustering around the equality line, confirming minimal discrepancies.

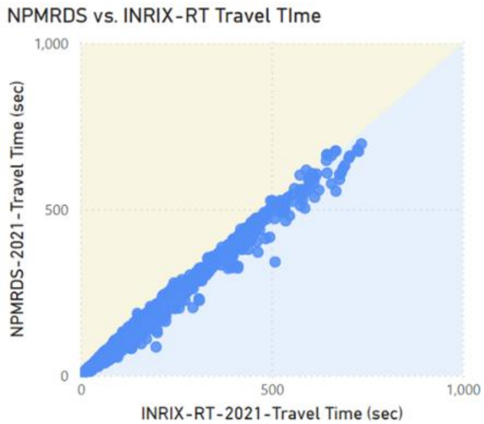
At the 50th percentile, 4,007 TMC segments (84%) report higher travel times in NPMRDS relative to INRIX-RT. This result closely aligns with the unfiltered comparison (83%), and all those TMCs are also present in the unfiltered set. This consistency suggests that applying the sample size filter does not change the general trend of NPMRDS reporting higher median travel times. On average, NPMRDS reports 6.4% higher median travel times than INRIX-RT in the filtered dataset compared with 6.3% in the unfiltered dataset. The average absolute difference between the two datasets is 7.7% in both cases. Although most TMCs (84%) follow the pattern, the TMC with the highest percentage difference, TMC 110-50086 in Newport News, shows the opposite. INRIX-RT reports a median travel time of 104.54 seconds compared with 56.37 seconds in NPMRDS, a 46.1% difference over a 0.5-mile segment.

A similar pattern is observed at the 80th percentile. In the filtered dataset, 4,540 TMCs (95%) report higher travel time in NPMRDS compared with INRIX-RT, similar to the unfiltered dataset (also 95%). On average, NPMRDS reports 13.7% higher 80th percentile travel times than INRIX-RT, with an average absolute difference of 14.3%. The TMC with the highest percentage difference at this percentile is TMC 110-07461, a 2.14-mile-long segment on US 58 in Carroll County, where NPMRDS reports 643.38 seconds compared with 222.75 seconds in INRIX-RT, which shows a 65.4% difference. Because both the 50th and the 80th percentile travel times are key inputs in calculating travel time reliability measures such as LOTTR, the persistent and sometimes large difference in travel time may lead to significant differences in the final reliability metrics.

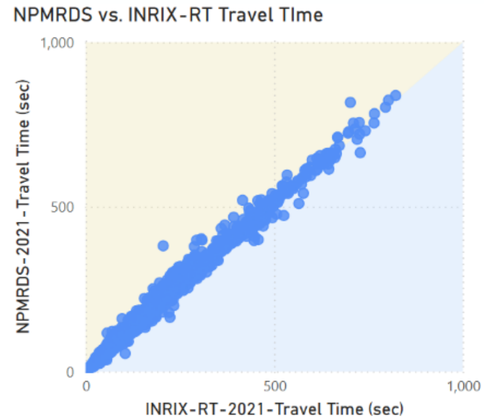
At the 95th percentile, the discrepancy widens further. At this percentile, NPMRDS travel times exceed INRIX-RT travel times by an average of 24.3%. The single largest gap of 79.6% is observed on TMC 110-08202 (US 211), where NPMRDS shows 384 seconds versus 78.5 seconds in INRIX-RT.

By comparing Figure 7 with Figures 3 and 4, one can observe that the sample size filtering substantially reduced variability. Despite the reduced variability, NPMRDS continued to report systematically higher travel times, possibly due to INRIX-RT's smoothing effects from the blending of historical data. Although the sample size filter eliminates many extreme outliers, substantial vertical dispersion remains. However, it remains uncertain whether these high travel time values reflect actual congestion events or are the result of data anomalies.

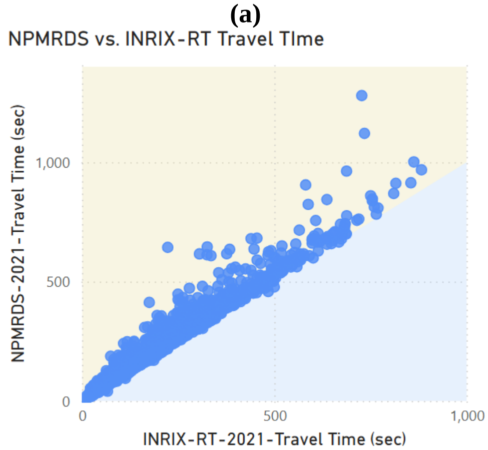
Comparing the filtered 4,786 TMC segments in INRIX-RT with INRIX revealed nearly identical travel time values across all percentiles. This finding clearly indicates that the differences initially observed between INRIX and INRIX-RT datasets predominantly result from segments with low sample counts. For segments with robust data coverage, INRIX inherently provides sufficient real-time data, making the INRIX-RT filtering largely unnecessary for well-sampled segments. In contrast, the gap between NPMRDS and the other two sources remains significantly large, indicating that the observed discrepancies are likely due to methodological differences in how travel times are estimated or aggregated rather than sample size limitations. To understand the source of this variability better, the following section explores how data density within NPMRDS may contribute to these differences.



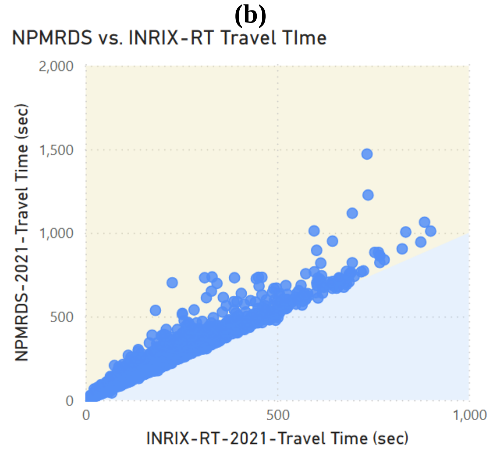
(a)



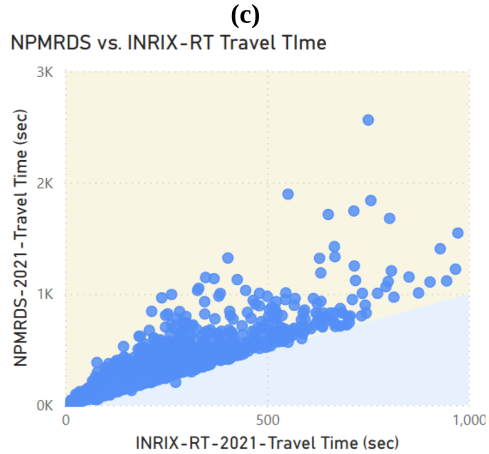
(b)



(c)



(d)



(e)

Figure 7. Comparison of Travel Time Values across Traffic Message Channel Segments between NPMRDS and INRIX-RT Datasets at: (a) 5th; (b) 50th; (c) 80th; (d) 85th; (e) 95th percentiles. NPMRDS = National Performance Management Research Data Set.

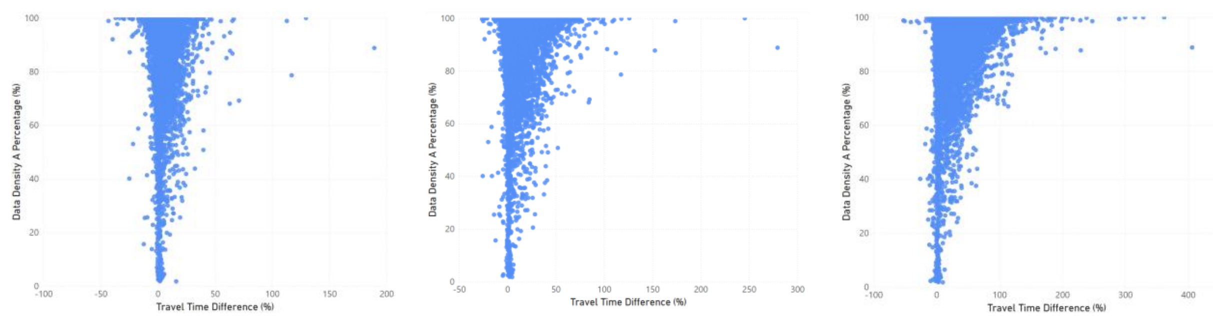
Table 4. TMC Counts and Percentages with Higher NPMRDS Travel Time Compared with INRIX-RT (N = 4,786 TMCs)

Percentile	NPMRDS > INRIX-RT
5th	1,456 (30%)
15th	2,470 (52%)
50th	4,007 (84%)
80th	4,540 (95%)
85th	4,586 (96%)
95th	4,665 (97%)

NPMRDS = National Performance Management Research Data Set. TMC = Traffic Message Channel.

Analysis of Data Density Effect on Travel Time Differences Between NPMRDS and INRIX

To further investigate the cause of the persistent differences between NPMRDS and INRIX travel times, this section examines whether those differences are associated with data quality, as indicated by the proportion of low-density (i.e., density A) observations in NPMRDS. Because density A reflects intervals with the fewest probe readings, a higher proportion of density A suggests lower data density. Figure 8 presents scatter plots showing the relationship between the percentage difference in travel time, calculated as the difference between NPMRDS and INRIX values divided by INRIX, against the proportion of density A observations. Separate plots are shown for the 50th, 80th, and 95th percentiles of travel time. Each data point in the plots represents a single TMC segment evaluated during a specific time period (e.g., a.m. peak, mid-day, p.m. peak, or weekend), enabling a direct comparison between datasets under aligned spatial and temporal conditions.



(a) (b) (c)

Figure 8. Comparison of NPMRDS-INRIX Travel Time Difference and Data Density A Proportion at: (a) 50th; (b) 80th; (c) 95th percentiles. NPMRDS = National Performance Management Research Data Set.

These plots reveal a consistent pattern across all percentiles. As the proportion of density A observations increases, the magnitude and variability of travel time differences between NPMRDS and INRIX also increase. At low-density A proportions, the differences between datasets are generally small and centered around zero, indicating good agreement. However, as data density declines, the differences become more volatile, with some TMCs exhibiting extreme deviations for which NPMRDS travel times are more than double those reported by INRIX. This funnel-shaped relationship is particularly pronounced at the 95th percentile, where reliability metrics are more sensitive to sparse data.

These findings are further contextualized by examining the distribution of data density classifications by time period in both 2021 and 2024. Table 5 summarizes the proportion of observations falling into density A, B, and C for each time period and year. In both years, most records fall into density A, underscoring widespread limitations in data quality. In 2024, density A accounted for 84.4% of records during the a.m. peak, 82.5% during mid-day, 80.7% during p.m. peak, and 85.8% on weekends. Similar patterns are seen in 2021, with slightly higher density A shares during some periods.

Table 5. Distribution of Data Density Classifications by Time Period in 2021 and 2024 (NPMRDS)

Time Period	Year	Density A (%)[a]	Density B (%)	Density C (%)
a.m. peak	2024	84.4	8.0	7.6
	2021	82.2	9.4	8.4
Mid-day	2024	82.5	9.1	8.5
	2021	80.9	10.6	8.4
p.m. peak	2024	80.7	9.2	10.2
	2021	84.5	8.5	7.0
Weekend	2024	85.8	7.6	6.6
	2021	90.9	5.6	3.5

NPMRDS = National Performance Management Research Data Set. [a] Percentages may not sum to exactly 100% due to rounding.

These distributions help explain the patterns observed in the NPMRDS-INRIX comparisons. The high prevalence of low-density intervals contributes to greater variability and larger travel time differences in NPMRDS, particularly at higher percentiles. Although some differences may be attributable to methodological variations between the datasets, the influence of data sparsity is evident and substantial.

Travel Time Cumulative Distribution Function for Low-Sample TMC Segments

To highlight how sparse observations can distort percentile travel times, two NHS segments were examined in this section.

- TMC 110+08589: US 58 in Grayson County (4.11 miles, AADT ≈ 600).
  2021 a.m. peak coverage: 48 NPMRDS records and 26 INRIX-RT records.
- TMC 110-08570: US 220 in Highland County (9.11 miles, AADT ≈ 500).
  2021 a.m. peak coverage: 51 NPMRDS records and 69 INRIX-RT records.
-

TMC 110 + 08589 (US 58). Figure 9 shows the cumulative distribution of travel time for TMC 110+08589 as derived from (a) NPMRDS, (b) INRIX, and (c) INRIX-RT datasets. In the NPMRDS dataset (Figure 9a), travel time increases smoothly from the 5th percentile (347 seconds) to the 50th percentile (456 seconds). The rate of increase accelerates between the 60th percentile (657 seconds) and the 80th percentile (1,195 seconds). Followed by a dramatic surge, the 85th percentile reaches 1,840 seconds, the 95th percentile reaches 2,811 seconds, and the maximum value reaches 4,931 seconds. For this TMC, 99.55% of 15-minute intervals are classified as data density A, meaning they are based on very few probe readings. This extremely low data density not only signals very limited sample counts but also introduces substantial uncertainty in the resulting travel time values. As shown in previous analyses, such conditions are strongly associated with inflated or unstable travel times, particularly at higher percentiles in which small sample fluctuations can significantly skew the results.

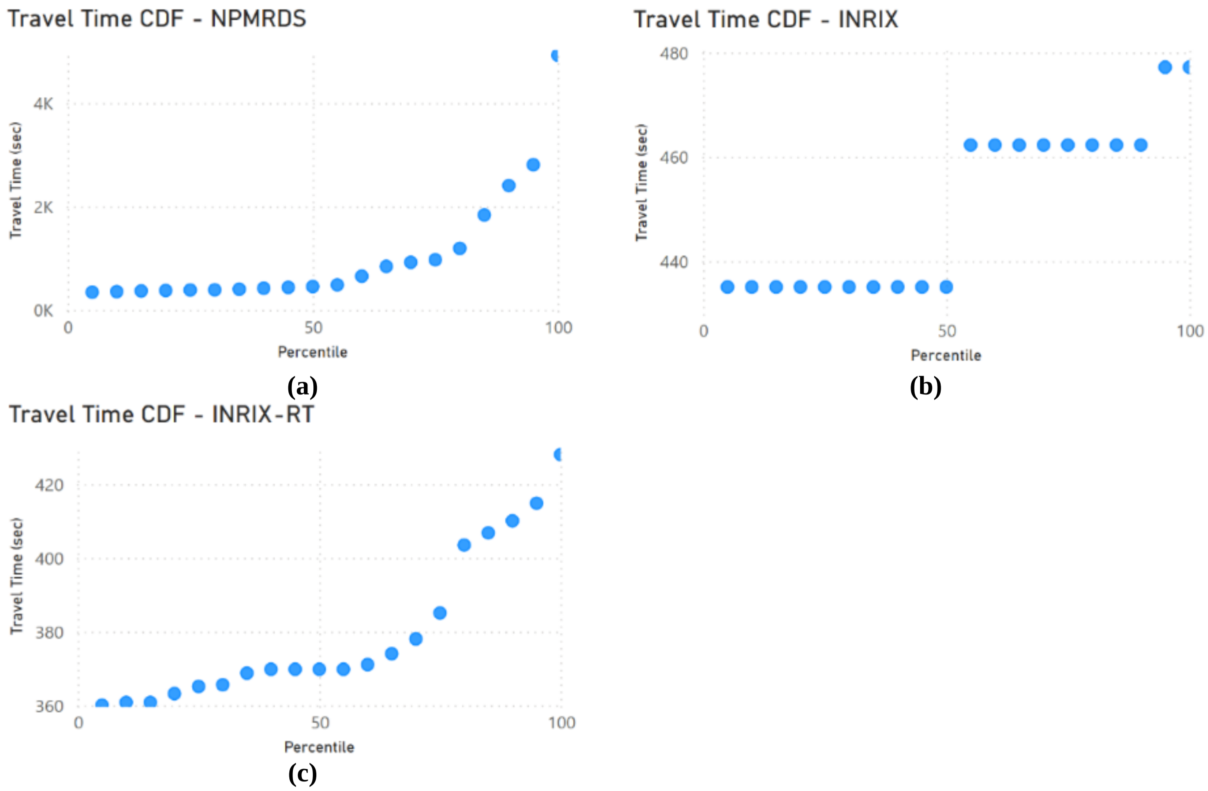


Figure 9. Comparison of Travel Time CDF Curves for Traffic Message Channels 110+08589 during 2021 a.m. Peak Hours Derived from: (a) NPMRDS; (b) INRIX; (c) INRIX-RT Datasets. CDF = cumulative distribution function; NPMRDS = National Performance Management Research Data Set.

In contrast, the INRIX distribution resembles a step function (Figure 9b). Travel time remains fixed at 435 seconds from the 5th through the 50th percentiles, then increases to 462 seconds from the 55th to the 90th percentiles. These plateaus reflect INRIX's practice of filling intervals with historical or reference values when no real-time probes data are available. Notably, the 5th percentile INRIX value (435 seconds) is higher than the maximum INRIX-RT value (428 seconds), indicating that all observations containing real-time data are effectively excluded from the unfiltered INRIX cumulative distribution.

The INRIX-RT curve is smoother than INRIX and spans from 360 to 428 seconds (Figure 9c). Although it captures some variability, it does not exhibit the extreme upper percentile values observed in NPMRDS. This result may stem from the aggregation of mixed historical and real-time data within each 15-minute interval or from inherent smoothing within INRIX's probe data processing methodology.

TMC 110-08570 (US 220). Figure 10 presents the cumulative distributions of travel time for TMC 110-08570, as derived from (a) NPMRDS, (b) INRIX, and (c) INRIX-RT datasets. In the NPMRDS dataset (Figure 10a), the curve rises gradually from the 5th percentile (571 seconds) to the 85th percentile (752 seconds). Travel times increase more steeply after the 90th percentile (872 seconds) and reach 1,876 seconds at the 95th percentile, with the maximum reaching 3,091 seconds. Similar to the US 58 case, a handful of very slow probe readings dominate the extreme upper tail when the overall sample size is small (51 observations).

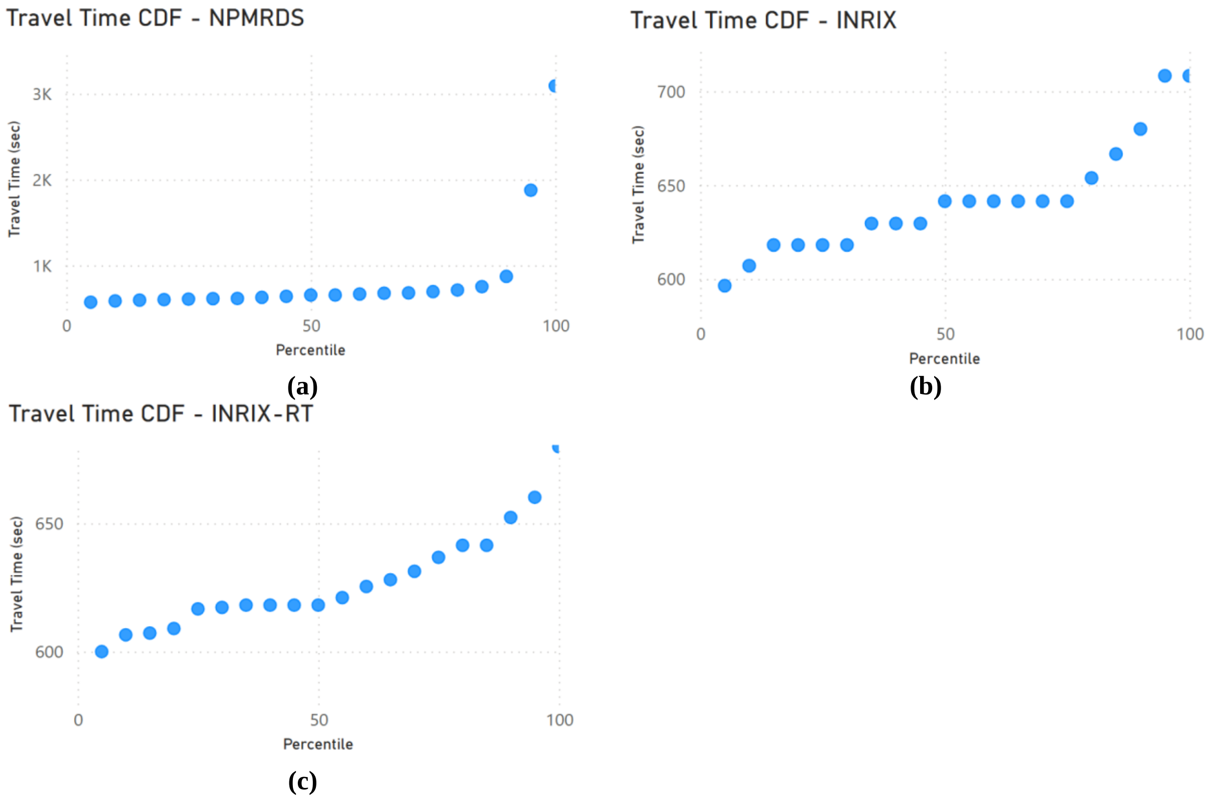


Figure 10. Comparison of Travel Time CDF Curves for Traffic Message Channel 110-08570 during 2021 a.m. Peak Hours Derived from: (a) NPMRDS; (b) INRIX; (c) INRIX-RT Datasets. CDF = cumulative distribution function; NPMRDS = National Performance Management Research Data Set.

In contrast, the INRIX distribution in Figure 10b again exhibits a step-like profile, but it is noticeably more gradual than the rigid three-plateau pattern seen on US 58. Travel time remains fixed at 618 seconds from the 15th through the 35th percentiles, increases to 630 seconds between the 35th and the 45th percentiles, and rises again to 641 seconds from the 50th through the 75th percentiles. This smoother, multistep shape occurs because the real-time readings on US 220 overlap the range of INRIX's historical and reference values. Where the ranges overlap, real-time intervals introduce slight variations into the otherwise uniform reference series, breaking long horizontal blocks into shorter steps and giving the cumulative distribution function a more continuous ascent. In contrast, on US 58, nearly all real-time readings fell below the 5th percentile, rendering them effectively invisible in the cumulative curve.

The INRIX-RT curve spans from 600 seconds at the 5th percentile to 680 seconds at the maximum (Figure 10c). The curve is the smoothest of the three, but it still fails to capture the extreme travel times reflected in the NPMRDS upper tail.

Overall, the US 220 case suggests that even modest real-time coverage can soften INRIX's cumulative distribution function plateaus when probe and reference ranges overlap, resulting in a more gradual distribution.

Analysis of Travel Time Reliability Measures

The previous section showed that percentile travel times can differ depending on the data source. The natural next question is how those differences propagate into the performance measures that agencies report. VDOT's *Traffic Operations and Safety Analysis Manual (TOSAM)* lists four VDOT-preferred reliability measures, including 95th percentile travel time index, 80th percentile travel time index, 50th percentile travel time index, and LOTTR (VDOT, 2020). All three travel time indices require a free-flow travel time denominator, yet the definition of free-flow speed can change travel time index values by double-digit percentages and alter the list of segments deemed unreliable (Lan and Zhao, 2025). Moreover, under the Federal Highway Administration's PM-3 performance reporting regulation, state departments of transportation must upload four time-period LOTTR values, together with their underlying 80th and 50th percentile travel times for every NHS reporting segment to the Highway Performance Monitoring System each year. This section compares the LOTTR values calculated using the three datasets.

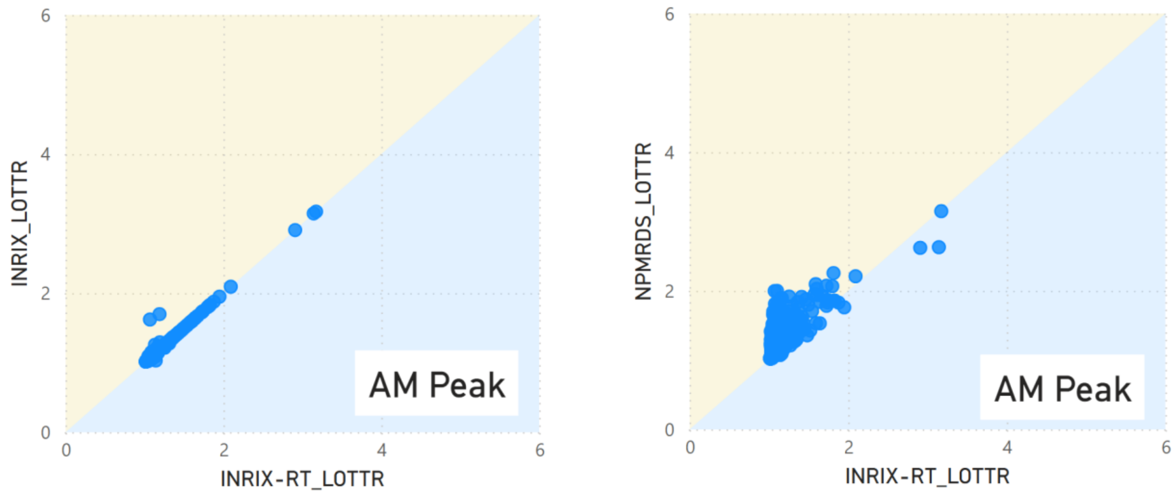
Level of Travel Time Reliability Scatter Plots

Figures 11 through 14 compare LOTTR values between datasets for the four reporting periods. Each figure contains two panels:

- Panel (a): INRIX-RT versus INRIX (effect of removing historical data backfill).
- Panel (b): INRIX-RT versus NPMRDS (effect of switching data source).

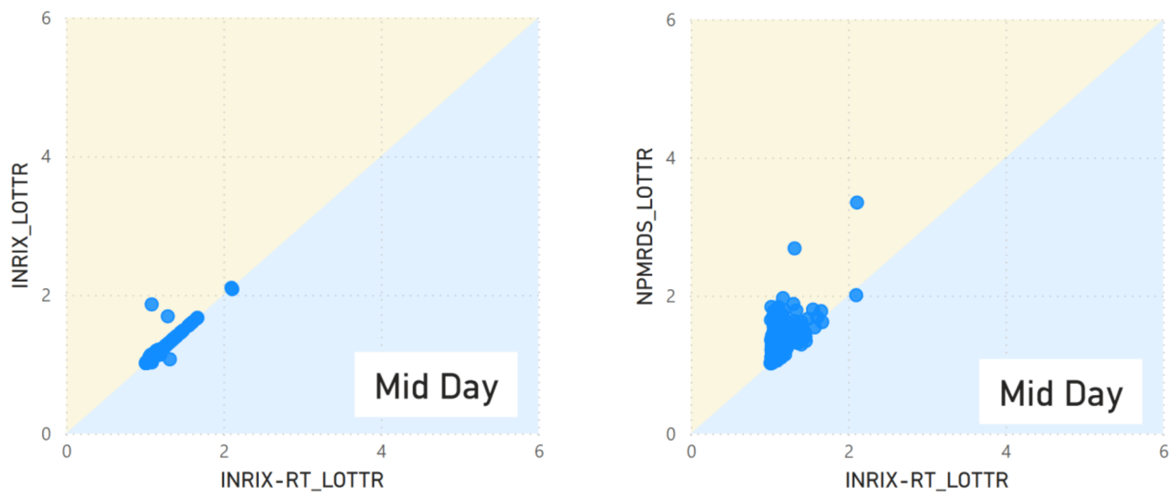
Figure 11a shows data points clustered closely around the equality line, indicating that the removal of pure historical values has a minimal effect on the results. Figure 11b shows data points well above the 45-degree line, highlighting a notable effect on the reliability metric. This observation aligns with the finding in previous sections that NPMRDS consistently reports higher travel times than INRIX, especially at upper percentiles, which contributes to the observed discrepancies in reliability measures. The NPMRDS flags 126 a.m. peak unreliable TMCs versus 27 for INRIX and 25 for INRIX-RT.

Figure 12 shows that mid-day traffic is generally better than that during a.m. peak periods. INRIX and INRIX-RT identify 9 and 7 unreliable TMCs, whereas NPMRDS flags 95 TMCs. Figure 13 shows that the p.m. peak period exhibits a wider spread in LOTTR values, with INRIX-RT identifying 62 unreliable TMCs, INRIX identifying 64, and NPMRDS identifying 192. Notably, all unreliable TMCs uniquely identified by NPMRDS but not by INRIX-RT have INRIX-RT sample counts exceeding 1,000, and only 4 TMCs have NPMRDS sample sizes less than 1,000 (N = 681 ~ 914). This observation suggests that tightening the sample size threshold further would not significantly change the unreliable TMC list. Therefore, the large discrepancy stems not from data density. As noted previously, the substantially higher count from NPMRDS is likely attributable to its tendency to report systematically higher travel times at upper percentiles compared with INRIX and INRIX-RT datasets.



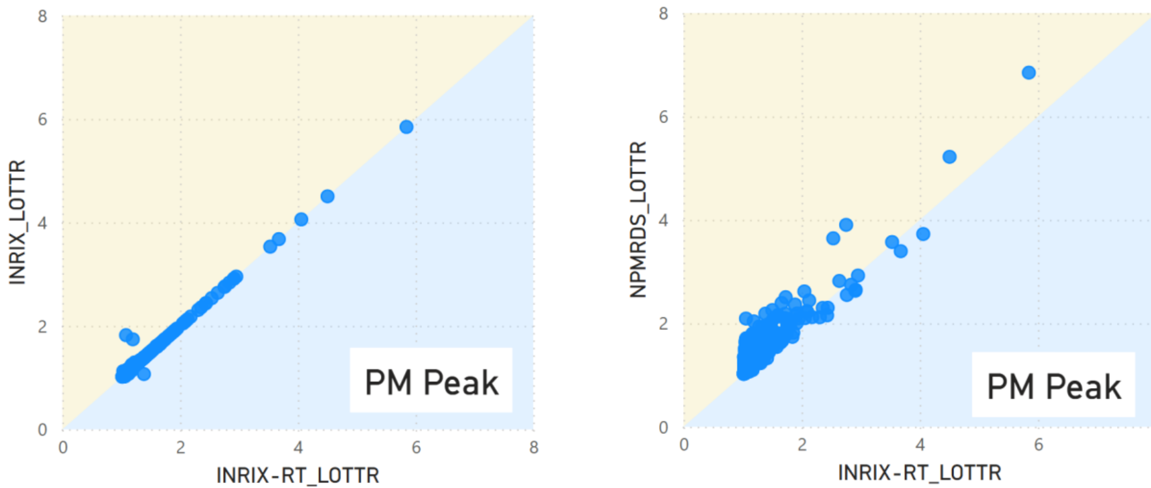
(a) (b)

Figure 11. a.m. Peak LOTTR Comparisons between: (a) INRIX-RT versus INRIX; (b) INRIX-RT versus NPMRDS Datasets. LOTTR = Level of Travel Time Reliability; NPMRDS = National Performance Management Research Data Set.



(a) (b)

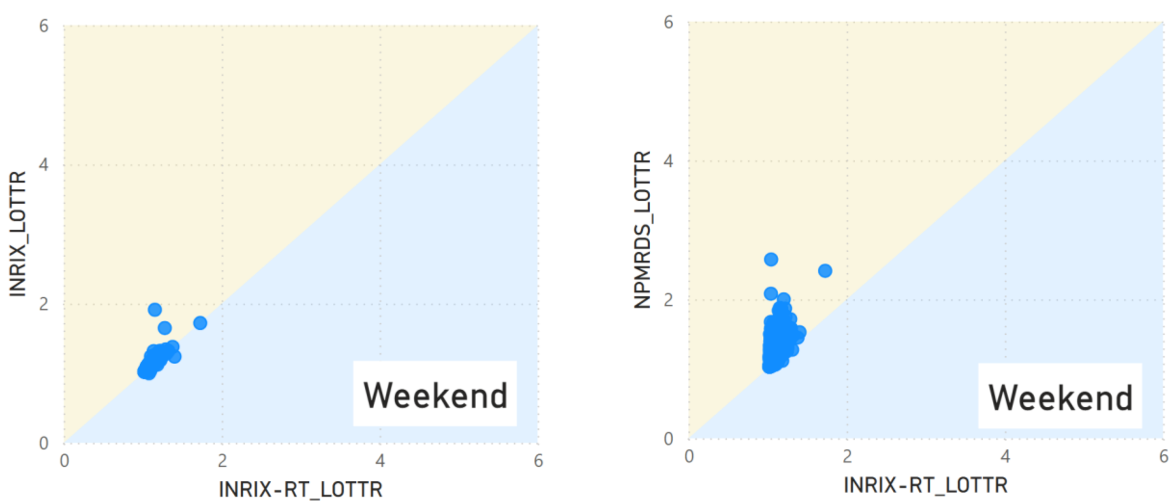
Figure 12. Mid-Day LOTTR Comparisons between: (a) INRIX-RT versus INRIX; (b) INRIX-RT versus NPMRDS Datasets. LOTTR = Level of Travel Time Reliability; NPMRDS = National Performance Management Research Data Set.



(a) (b)

Figure 13. p.m. Peak LOTTR Comparisons between: (a) INRIX-RT versus INRIX; (b) INRIX-RT versus NPMRDS Datasets. LOTTR = Level of Travel Time Reliability; NPMRDS = National Performance Management Research Data Set.

Figure 14 shows a significant discrepancy among datasets during the weekend period. INRIX-RT identifies only 1 unreliable TMC, INRIX identifies 3, and NPMRDS flags 100 TMCs. All those TMCs have sample sizes exceeding 1,000 in both INRIX-RT and NPMRDS, again confirming that data sparsity is not the most pronounced contributing factor, and the divergence is likely driven by NPMRDS's systematic tendency to report higher travel times at upper percentiles, which inflates LOTTR values and leads to substantially more TMCs being classified as unreliable.



(a) (b)

Figure 14. Weekend LOTTR Comparisons between: (a) INRIX-RT versus INRIX; (b) INRIX-RT versus NPMRDS Datasets. LOTTR = Level of Travel Time Reliability; NPMRDS = National Performance Management Research Data Set.

Observations and Discussions on Unreliable TMCs

Unreliable TMCs Identified from Different Datasets. Table 6 summarizes the overlap of unreliable TMCs across the three datasets during the a.m. and p.m. peak periods. The first column shows the number of TMCs all three datasets simultaneously identified as unreliable. The remaining columns report the total number of unreliable TMCs each dataset identified and the percentage of those segments that appear on the list of unreliable segments identified by all three datasets.

Table 6. Overlap of Unreliable TMCs Identified across Datasets (TMCs with LOTTR ≥ 1.5)

Time Period	# of Unreliable TMCs in All Three Datasets	# of Unreliable TMCs in INRIX	# of Unreliable TMCs in INRIX-RT	# of Unreliable TMCs in NPMRDS
a.m. peak	24	27 (89 %)[a]	25 (96 %)	126 (19 %)
p.m. peak	62	64 (97 %)	62 (100 %)	192 (32 %)

LOTTR = Level of Travel Time Reliability; NPMRDS = National Performance Management Research Data Set; TMC = Traffic Message Channel. [a] Percentages indicate the proportion of each dataset's unreliable TMCs that all three datasets also identified.

During the a.m. peak, 24 TMCs are identified as unreliable in all three datasets. These TMCs represent 89% of the unreliable TMCs identified by INRIX (24 out of 27) and 96% of those identified by INRIX-RT (24 out of 25). In NPMRDS, however, these 24 TMCs account for only 19% of its unreliable list (24 out of 126). The INRIX-RT TMC that was not included in the overlapped list is 110+05671 (LOTTR = 1.43), which overlaps with the INRIX list but falls slightly below the threshold by NPMRDS.

A similar pattern appears in the p.m. peak period, in which 62 TMCs are consistently identified as unreliable across all datasets. These TMCs represent 97% of INRIX's unreliable TMCs and 100% for INRIX-RT but only 32% of those NPMRDS identified.

These results confirm a high degree of alignment between INRIX and INRIX-RT, reinforcing previous findings that the removal of historical backfill has a minimal effect on the LOTTR reliability measure. In contrast, NPMRDS consistently identifies a broader set of unreliable TMCs, likely because of its systematically higher travel time at upper percentiles.

Higher Level of Travel Time Reliability Values in NPMRDS. The largest a.m. peak discrepancy occurs on TMC 110-07461, a 2.14-mile-long segment on US 58. NPMRDS yields a LOTTR of 1.998 from 1,186 observations, whereas INRIX and INRIX-RT produce values of 1.056 and 1.074 based on 2,125 INRIX-RT observations.

The second largest gap occurs on TMC 110+08505, a 0.14-mile-long segment on US 33. NPMRDS yields a LOTTR of 2.000 from 811 observations compared with 1.079 (INRIX) and 1.102 (INRIX-RT) from 1,111 INRIX-RT observations.

Because both TMCs have sample counts of more than 800 observations, tightening the sample size threshold would do little to narrow the discrepancy. However, after closer examination, segment length emerges as an impact factor. One can note that:

- Out of the 32 TMCs with the largest NPMRDS-INRIX-RT gaps, after excluding TMCs less than 0.1 mile, 16 are still only 0.1 to 0.25 mile long.
- When those less than 0.1-mile-long TMCs are restored, 99 of the 100 highest gap segments are less than 0.25 mile long. The only exception is TMC 110-07461, noted previously.

<u>*Disproportionate Presence of Short Segments Among Unreliable TMC.*</u> To investigate the relationship between segment length and identified unreliability further, Figure 15 compares length distributions for (1) all non-interstate NHS TMCs, (2) TMCs flagged as unreliable by NPMRDS, and (3) TMCs flagged as unreliable by INRIX-RT. Although 31% of all TMCs are longer than 1 mile, only 0.7% of those identified as unreliable by NPMRDS exceed this length. In contrast, 60% of TMCs flagged as unreliable by NPMRDS are shorter than 0.25 mile compared with 20% of all TMCs. This observation highlights that NPMRDS disproportionately identifies short segments as unreliable.

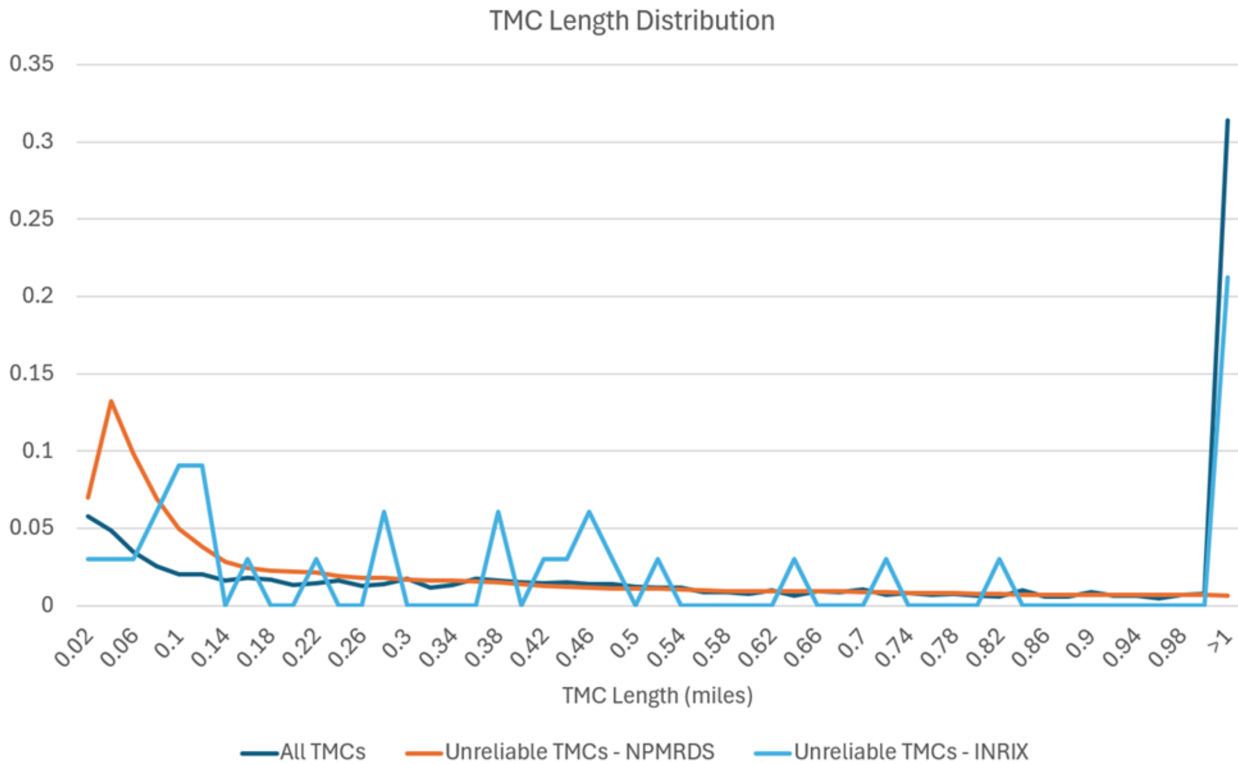


Figure 15. Length-Frequency Curves for All TMCs (Dark Blue) and for TMCs Labeled as Unreliable by NPMRDS (Orange) and INRIX-RT (Sky Blue). NPMRDS = National Performance Management Research Data Set; TMC = Traffic Message Channel.

The INRIX-RT curve also leans toward shorter segments, although the skew is much less pronounced. Because there are only 34 unreliable TMCs (LOTTR ≥ 1.5) in the INRIX-RT dataset, the shape of the sky-blue curve is more sensitive to the presence or absence of just one or two segments per bin, resulting in a less smooth profile compared with the other distributions.

Development of Models for Travel Time Reliability Prediction

Prediction models for the 50th, 80th, and 95th percentile travel times were developed using LMM, RF, and LightGBM. Models were trained separately for each target percentile. Data from 2022 and 2023 were used for training, and 2024 data were used for final testing. MAP-21-related travel time reliability analyses require data for an entire calendar year to calculate travel time percentiles. TMC segments shorter than 0.1 mile were not included in the modeling due to data quality issues presented in previous sections, which eliminated nearly 20% of the total number of all TMC segments and more than 45% of the total number of internal TMC segments. The total mileage of these short segments is 47 directional miles, which is 0.7% of the total mileage in the study network.

In addition, segments identified with data quality issues during the data preparation process, mostly internal segments, were excluded. For example, the 0.39-mile-long TMC segment 110P08603, highlighted in Figure 16, spans three LRS routes with different traffic and geometric characteristics: State Route 86 North with a PSL of 40 mph, State Route 413 East with a PSL of 35 mph, and State Route 413 East ramp with a PSL of 20 mph. This study used a length-weighted average method to estimate PSL and AADT on TMC segments matched to multiple LRS routes. For such short internal TMCs matched to multiple LRS routes, experiments using training datasets yielded poor prediction results across all modeling techniques. Therefore, such internal segments were excluded from the analysis.

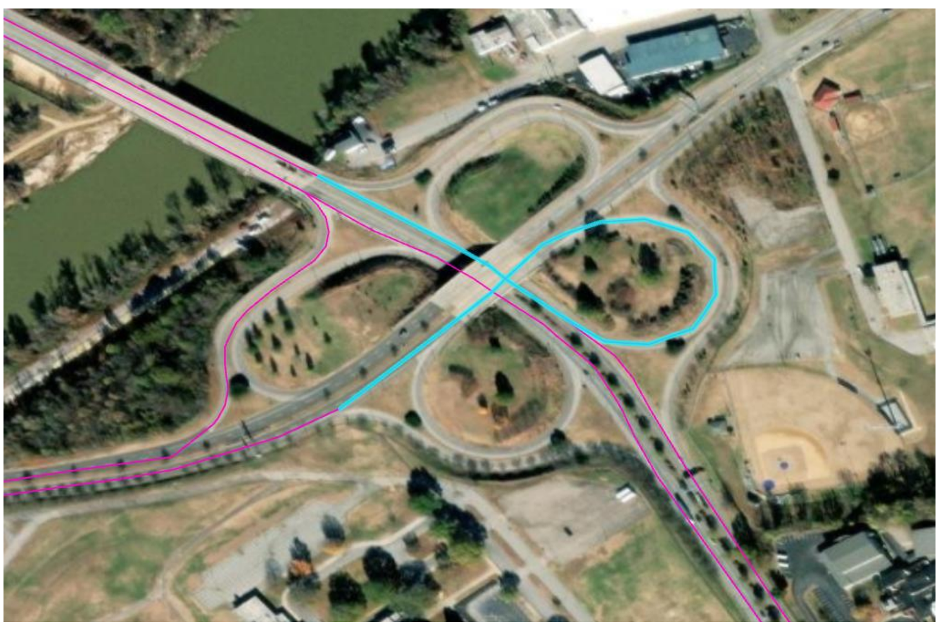


Figure 16. Example of Traffic Message Channel Segments with Data Quality Issues

Figures 17 through 19 show the distributions of the 50th, 80th, and 95th percentile travel times on all study segments. The number of internal segments was substantially less than that of external segments after data reduction, especially for unreliable internal segments. In addition, not all segments had data available across all four study periods. The distributions of the three target percentiles differed between reliable and unreliable segments within the same segment type. This variation highlights the need for grouping TMCs with similar travel time distributions as a prerequisite for the LMM model.

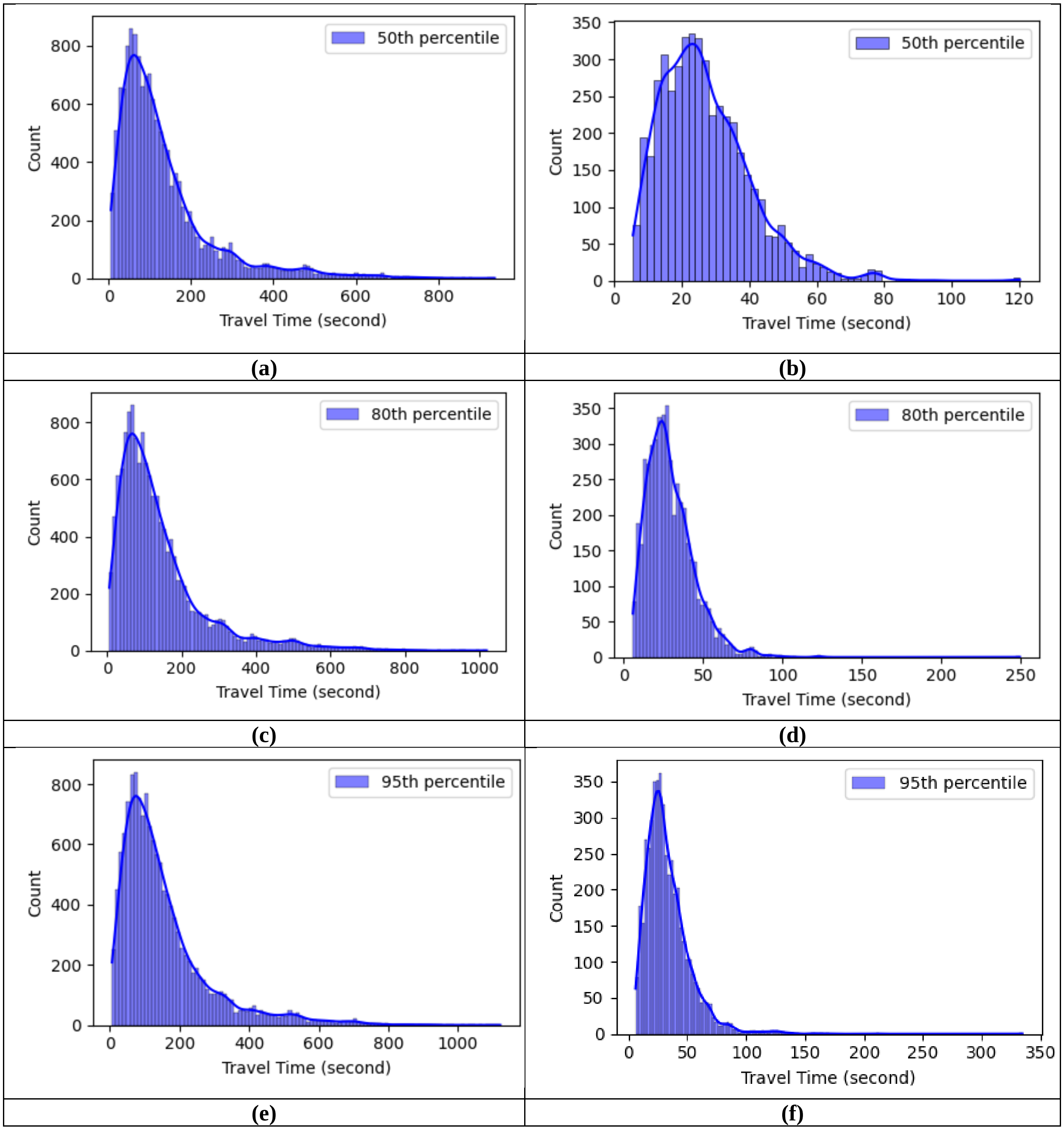
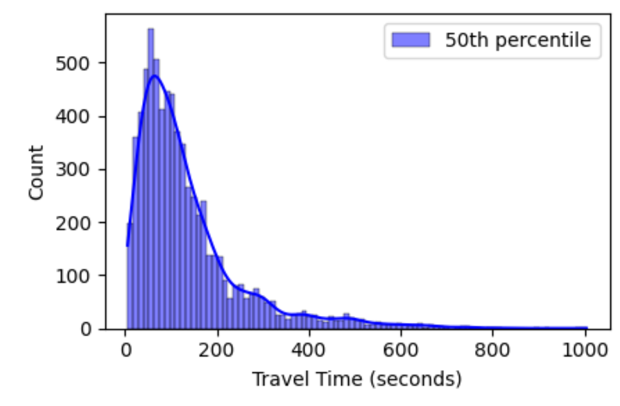
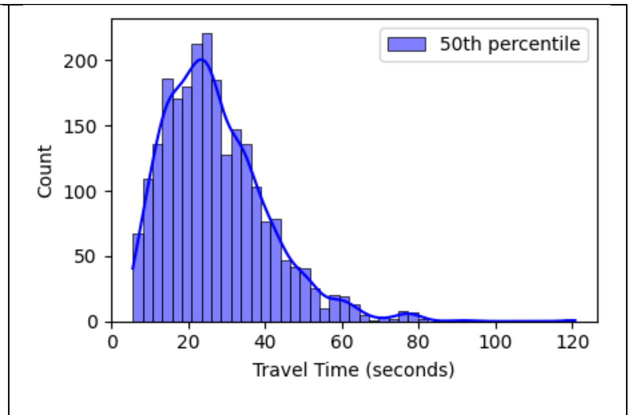


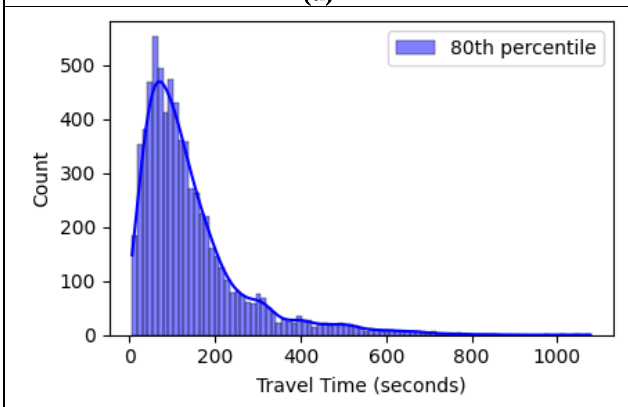
Figure 17. Distributions of the 50th, 80th, and 95th Travel Times in Peak Traffic Period: (a) 50th Percentile—External Segments; (b) 50th Percentile—Internal Segments; (c) 80th Percentile—External Segments; (d) 80th Percentile—Internal Segments; (e) 95th Percentile—External Segments; (f) 95th Percentile—Internal Segments



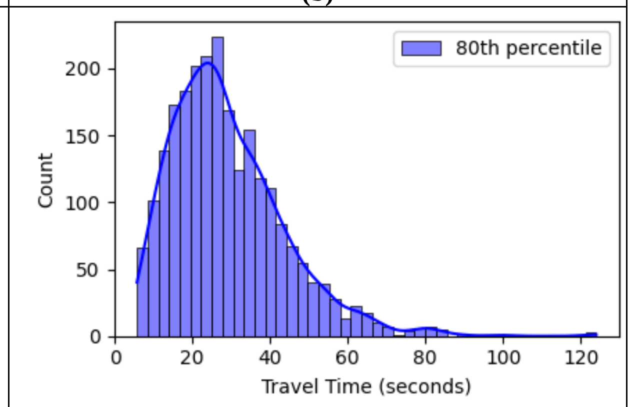
(a)



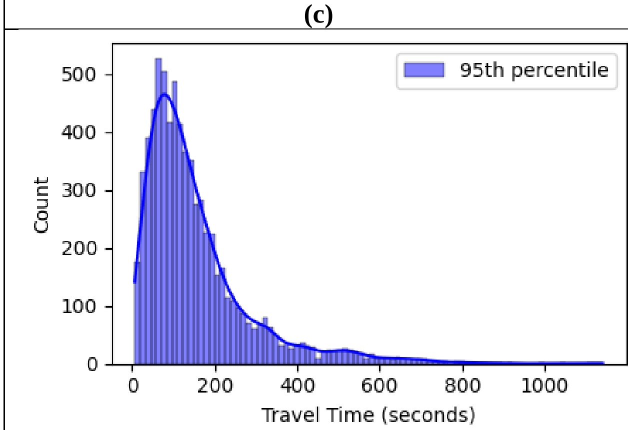
(b)



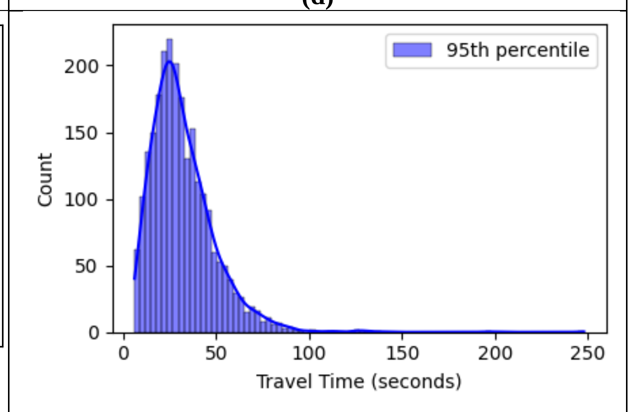
(c)



(d)



(e)



(f)

Figure 18. Distributions of the 50th, 80th, and 95th Travel Times in Mid-Day Period: (a) 50th Percentile—External Segments; (b) 50th Percentile—Internal Segments; (c) 80th Percentile—External Segments; (d) 80th Percentile—Internal Segments; (e) 95th Percentile—External Segments; (f) 95th Percentile—Internal Segments

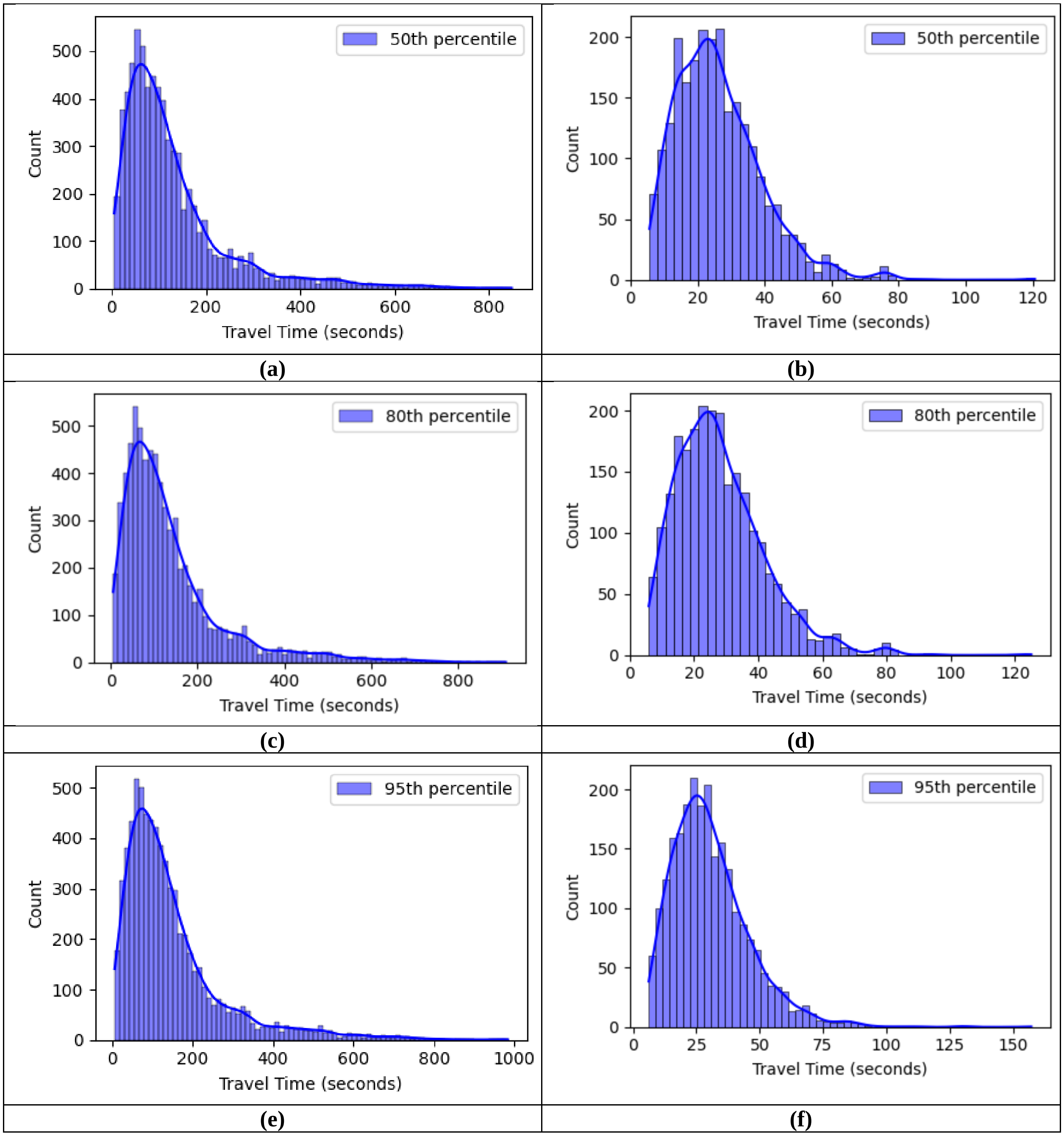


Figure 19. Distributions of the 50th, 80th, and 95th Travel Times in Weekend Period: (a) 50th Percentile—External Segments; (b) 50th Percentile—Internal Segments; (c) 80th Percentile—External Segments; (d) 80th Percentile—Internal Segments; (e) 95th Percentile—External Segments; (f) 95th Percentile—Internal Segments

Various combinations of the variables listed in Table 1 were tested as inputs. For the LMM, interaction terms were considered between volume and nonrecurrent events, including weather, incidents, and work zones. Based on the initial variable selection using the training datasets, the variables in Table 7 were used in the final models.

Table 7. Variables Used in Prediction Models

External Traffic Message Channel (TMC) Segments							
Variables	**Mean**	**Standard Deviation**	**Minimum**	**25th Percentile**	**50th Percentile**	**75th Percentile**	**Maximum**
Segment length (miles)	1.71	1.84	0.10	0.57	1.06	2.05	12.00
Posted speed limit (mph)	47.10	10.36	25.00	40.00	45.00	55.00	70.00
Number of through lanes	2.09	0.57	1.00	2.00	2.00	2.00	5.00
Traffic signal density	0.88	1.70	0.00	0.00	0.00	0.99	12.91
Average 15-minute volume (vehicles)	171.66	130.24	3.57	81.88	135.79	226.70	1443.37
Work zone intensity	3.85	23.07	0.00	0.00	0.00	0.81	995.18
Count of weather events	1.74	3.16	0.00	0.00	0.00	2.00	34.00
Count of crashes	1.96	2.91	0.00	0.00	1.00	3.00	55.00
Count of noncrash incidents	0.17	1.41	0.00	0.00	0.00	0.00	90.00
Internal Traffic Message Channel (TMC) Segments							
Variables	**Mean**	**Standard Deviation**	**Minimum**	**25th Percentile**	**50th Percentile**	**75th Percentile**	**Maximum**
Segment length (miles)	0.36	0.21	0.10	0.20	0.32	0.45	2.33
Posted speed limit (mph)	50.12	9.65	25.00	45.00	50.00	55.00	70.00
Number of through lanes	2.18	0.58	0.50	2.00	2.00	2.00	5.50
Traffic signal density	0.65	1.99	0.00	0.00	0.00	0.00	18.04
Average 15-minute volume (vehicles)	224.01	168.89	11.28	105.52	180.08	301.62	2415.16
Work zone intensity	5.07	51.08	0.00	0.00	0.00	0.74	2256.50
Count of weather events	1.45	2.78	0.00	0.00	0.00	2.00	20.00
Count of crashes	0.55	1.16	0.00	0.00	0.00	1.00	17.00
Count of noncrash incidents	0.24	2.39	0.00	0.00	0.00	0.00	69.00

Models for Peak Traffic Hours

Data from both the morning and afternoon peak periods were combined to create the LMM, RF, and LightGBM models for the 50th, 80th, and 95th percentile travel times for the four TMC subgroups. For LMM, the models were validated using five-fold cross validation using 2023 data and then tested using 2024 data.

Linear Mixed Models

The variables included in LMMs were selected based on prediction accuracy and evaluated using five-fold cross-validation on the training datasets. For each model combination of variables, correlation coefficients and variance inflation factors were calculated to assess multicollinearity among variables. Tables 8 and 9 present the estimated coefficients of the LMMs. Only variables that were statistically significant at the 0.1 significance level were retained in the final model. These coefficients are not standardized. Because of the limited sample size of unreliable segments, fewer variables were included in the models for unreliable segments compared with reliable segments of the same type.

Table 8. Estimated Linear Mixed Models Variable Coefficients for External Segments

Segments with Level of Travel Time Reliability (LOTTR) < 1.5						
Variables	50th Percentile		80th Percentile		95th Percentile	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	130.089	0.000	150.101	0.000	175.600	0.000
Urban/rural (rural =1)	– 14.601	0.000	– 16.616	0.000	– 19.579	0.000
Segment length (miles)	66.454	0.000	68.925	0.000	71.771	0.000
Posted speed limit (mph)	– 1.988	0.000	– 2.206	0.000	– 2.403	0.000
Number of through lanes	– 8.324	0.000	– 11.745	0.000	– 18.271	0.000
Traffic signal density	1.060	0.001	1.493	0.000	1.817	0.000
Average 15-minute volume (vehicles)	0.017	0.000	0.025	0.000	0.039	0.000
Work zone intensity	0.020	0.000	0.036	0.000	0.044	0.000
Count of weather events	0.108	0.000	– 0.233	0.000	– 0.338	0.000
Count of crashes	0.885	0.000	1.349	0.000	3.095	0.000
Count of noncrash incidents	0.201	0.009	0.299	0.009	0.765	0.000
Segments with Level of Travel Time Reliability (LOTTR) ≥ 1.5						
Variables	**50th Percentile**		**80th Percentile**		**95th Percentile**	
	Coefficient	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Intercept	126.123	0.000	187.973	0.000	237.251	0.000
Urban/rural (rural =1)	– 32.247	0.049	– 70.408	0.022	– 63.186	0.127
Segment length (miles)	89.795	0.000	161.773	0.000	242.194	0.000
Posted speed limit (mph)	– 1.996	0.000	– 2.711	0.000	– 3.567	0.000
Number of through lanes	– 14.623	0.028				
Traffic signal density	3.446	0.048				
Count of weather events					111.392	0.009

Table 9. Estimated Linear Mixed Model Variable Coefficients for Internal Segments

Segments with Level of Travel Time Reliability (LOTTR) < 1.5						
Variables	50th Percentile		80th Percentile		95th Percentile	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	34.730	0.000	40.745	0.000	49.554	0.000
Segment length (miles)	60.541	0.000	63.030	0.000	67.616	0.000
Posted speed limit (mph)	– 0.537	0.000	– 0.631	0.000	– 0.760	0.000
Number of through lanes	– 1.465	0.000	– 1.646	0.000	– 2.327	0.000
Traffic signal density	1.079	0.000	1.355	0.000	1.652	0.000
Average 15-minute volume (vehicles)	0.001	0.000	0.002	0.000	0.006	0.000
Segments with Level of Travel Time Reliability (LOTTR) ≥ 1.5						
Variables	**50th Percentile**		**80th Percentile**		**95th Percentile**	
	Coefficient	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Intercept	44.278	0.000	41.99	0.089	70.907	0.012
Segment length (miles)	68.889	0.000	186.973	0.000	249.936	0.000
Posted speed limit (mph)	– 0.792	0.000	– 0.823	0.009	– 1.807	0.009

Segment length and PSL are statistically significant across all models. According to the estimated coefficients, the 50th, 80th, and 95th percentile travel times increase with the length of the TMC segments. As expected, the 50th, 80th, and 95th percentiles are lower on segments with higher PSLs, assuming all other variables remain constant. On external segments and reliable internal segments, the 50th, 80th, and 95th percentile travel times are negatively associated with the number of through lanes. However, this relationship does not hold for the 80th and 95th percentiles on unreliable external segments. External segments in rural areas exhibit lower 50th, 80th, and 95th percentile travel times compared with urban counterparts under similar conditions. Traffic volume is negatively associated with the 50th, 80th, and 95th percentile travel times on reliable segments but does not emerge as a significant factor for unreliable segments. This outcome may be attributed to the small sample size of unreliable segments, which could increase the likelihood of the LMM incorrectly rejecting the null hypothesis. In addition, unreliable segments are generally shorter and may suffer from lower data quality issues. Traffic signal density is positively associated with the 50th, 80th, and 95th percentile travel times on reliable segments, but it is not a significant variable for unreliable segments except the 50th percentile on unreliable external segments. This result is likely due to the limited samples in this subgroup. On reliable external segments, nonrecurring events, such as work zones, adverse weather conditions, and incidents, can increase the 50th, 80th, and 95th percentile travel times when other conditions are held constant. Adverse weather is observed to significantly affect the 95th percentile on unreliable segments. Many studies in the literature have found that adverse weather can cause unreliable travel times and often lead to extremely high travel times. The nonrecurring event variables are not significant for internal segment models, possibly because of the smaller sample size and very short lengths of those segments.

The smaller sample size for internal segments, particularly for unreliable internal segments, poses challenges for variable selection in the LMMs because of limited evidence. For unreliable internal segments, the inclusion of too many variables in the LMM could lead to multicollinearity issues, potentially inflating the effect of significant variables such as PSL and segment length, and the random effects of LMM are less accurate.

Table 10 presents a summary of performance measures using both the testing data from 2024 and the training data from 2022 to 2023 through five-fold cross-validation. The prediction accuracy generally decreases as the target percentile increases. The models demonstrate reasonably better accuracy for reliable segments compared with the unreliable segments. This outcome was anticipated given that traffic conditions tend to be more complex on internal segments. Furthermore, LMMs may encounter underfitting issues when dealing with limited sample sizes.

Table 10. Linear Mixed Modes Performance Metrics

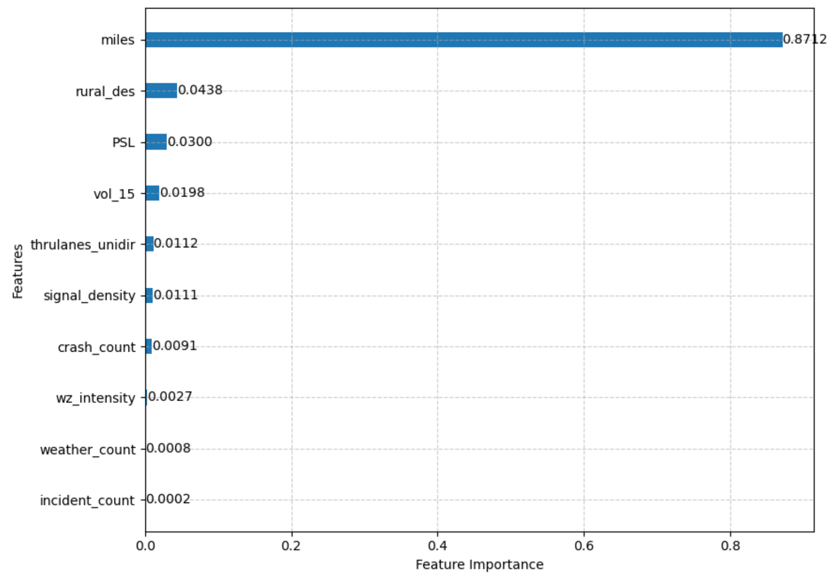
	2022–23 Cross Validation				Testing			
	MAE	**RMSE**	**R^2**	**MAPE**	**MAE**	**RMSE**	**R^2**	**MAPE**
External Segments with Level of Travel Time Reliability (LOTTR) <1.5								
50th	21.065	30.731	0.933	26.600	20.549	29.438	0.939	27.000
80th	23.736	34.695	0.922	27.680	23.204	32.734	0.930	28.824
95th	27.953	42.193	0.899	28.895	27.679	41.617	0.901	31.111
External Segments with Level of Travel Time Reliability (LOTTR) ≥ 1.5								
50th	25.235	33.368	0.926	21.154	15.743	19.233	0.948	19.577
80th	47.263	58.912	0.958	27.016	31.044	39.033	0.936	28.134
95th	63.992	81.028	0.914	31.291	53.822	66.902	0.926	32.134
Internal Segments with Level of Travel Time Reliability (LOTTR) < 1.5								
50th	4.278	6.021	0.796	17.828	4.592	6.362	0.781	18.570
80th	5.212	7.413	0.743	19.919	5.587	7.853	0.727	20.733
95th	7.311	11.290	0.614	24.010	5.033	7.462	0.755	17.780
Internal Segments with Level of Travel Time Reliability (LOTTR) ≥ 1.5								
50th	5.436	6.431	0.721	22.869	5.214	5.625	0.732	25.884
80th	10.973	14.457	0.698	23.762	17.405	18.697	0.646	26.881
95th	19.598	25.925	0.448	28.735	27.229	50.221	0.416	28.913

LOTTR = Level of Travel Time Reliability; MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean square error.

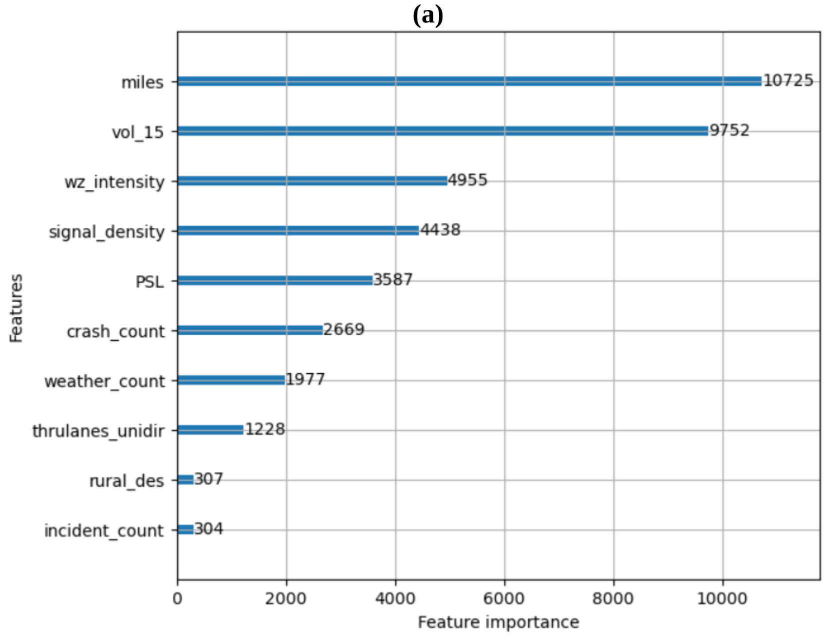
Random Forest and Light Gradient Boosting Machine Models

The variables in the RF and LightGBM models were selected based on prediction accuracy metrics, as indicated by the mean absolute error, RMSE, R^2, MAPE, and variable importance scores. Bayesian optimization was used to tune the hyperparameters for each combination of variables. Hundreds of models were tested with various variables combinations, and the models demonstrating the highest prediction accuracy were selected. To compare prediction performance, an identical set of variables was used to train both RF and LightGBM models.

The final RF and LightGBM models included 10 variables. Figure 20 shows the variable importance scores for the 50th percentile models for the reliable external segment subgroups. In general, the variables identified as most influential are consistent across both RF and LightGBM models. The variable importance scores of both RF and LightGBM models represent how often a variable is used for splitting across all decision trees. The variables with the highest and lowest scores are the same, and most mid-ranked variables differed by no more than two ranks. Segment length is the most important variable across all models. This finding is also observed in interstate GRF models developed in the Zhang et al. (2021) study. PSL and signal density also have a high prediction power for both RF and LightGBM models, and the frequency of noncrash incidents had a limited contribution to predictions. The rural or urban designation has low importance for both models, despite being a highly ranked variable in the interstate GRF models from the Zhang et al. (2021) study.



(a)

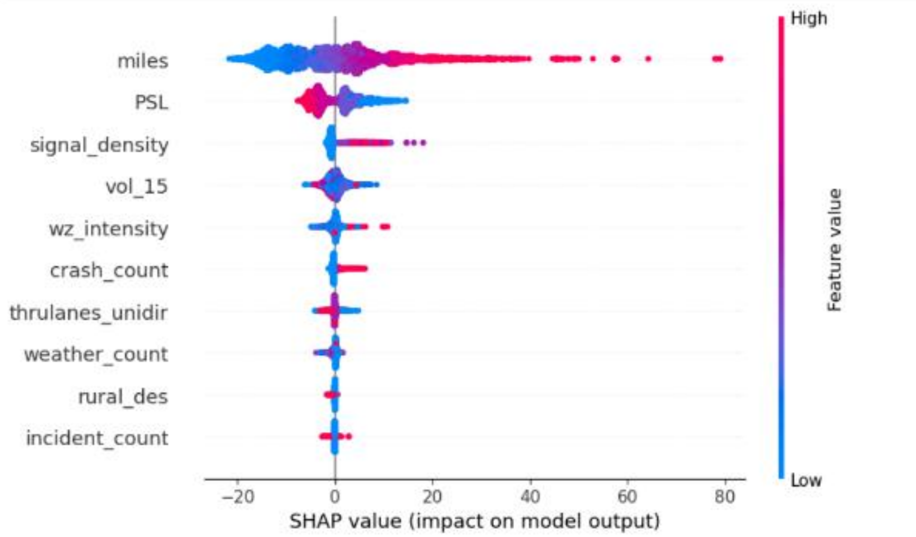


(b)

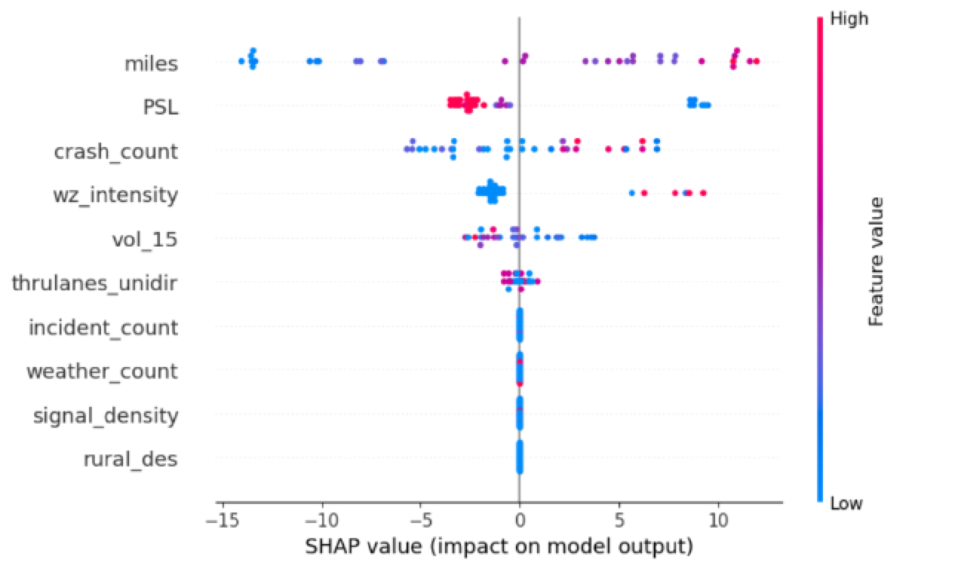
Figure 20. Variable Importance Scores of RF and LightGBM Models for 50th Percentile Travel Time: (a) RF Variable Importance Scores of Models for Reliable External Segments; (b) LightGBM Variable Importance Scores of Models for Reliable External Segments. LightGBM = light gradient boosting machine; PSL = posted speed limit; PSL = posted speed limit; RF = random forest.

The variable importance score represents the prediction power but not the actual effect of the variable on the target variable. It cannot be used to interpret the relationships with the target variables. The difference between the importance scores of RF and LightGBM models could be caused by the different perspectives they represent. The RF model may assign high importance scores to correlated features, whereas the LightGBM model is more sensitive to correlated variables because it uses a gradient boost approach. In addition, the RF importance score tends to be more global because it aggregates the effect of each variable on all decision trees in the forest. The LightGBM scores reflect a more local view, and they can change significantly during the training process. To obtain a more "global" view, the SHAP values were used to show the average contribution of a variable in LightGBM to the prediction outcome.

Figure 21 illustrates the SHAP values of the variables in the LightGBM model for the 50th percentile travel time. In Figure 21, variables with the highest importance are positioned at the top. Each point in Figure 21 corresponds to a row of the training data. Red and blue denote high and low values of a variable, respectively. The position of a point on the x-axis indicates whether the value of a variable caused a higher or lower prediction. A positive SHAP value indicates that the variable increases the prediction above the expected value, and a negative SHAP value suggests that the variable decreases the prediction below the expected value.



(a)



(b)

Figure 21. Light Gradient Boosting Machine SHAP Values for 50th Percentile Travel Time Models: (a) SHAP Values from Model for Reliable External Segments; (b) SHAP Values from Model for Unreliable External Segments. PSL = posted speed limit; SHAP = SHapley Additive exPlanations.

The variable importance, as indicated by the SHAP values in Figure 21, aligns more closely with the RF importance scores than the variable importance LightGBM generated. The segment length shows the highest overall impact on model predictions. The samples with very high positive SHAP values and long segment lengths, represented by the red dots on the far right of the x-axis, indicate that a small number of very long segments have very high 50th percentile travel times. The samples with large negative SHAP values and very short lengths (blue dots on the far left of the x-axis) tend to have low 50th percentile travel times. As Figure 21a shows, the few samples with high SHAP value about 80 indicate that segment length significantly influences the prediction of a limited number of samples of reliable external segments. Segment length also influences the prediction of a large number of samples to a moderate to low degree. Other variables, like PSL and average 15-minute volume (Figure 21a), affect a larger number of samples to a lesser extent.

PSL shows a generally higher negative impact. Segments with higher PSLs tend to have shorter travel times and vice versa. Positive associations were observed in certain samples, potentially representing congested segments, very short links on business roads, or TMC segments matched to multiple LRS links with varying PSLs. The signal density and crash count generally increased the prediction of the 50th percentile travel time. Although SHAP values help explain model behavior, they reflect correlation rather than causation. The high total importance of the segment length and PSL does not imply that these two variables affect the 50th percentile travel time directly.

Comparison of Model Performance

Tables 11 and 12 present the performance metrics for the RF and LightGBM models. Overall, the predictive performance of the two models is highly comparable. For reliable TMC segments, differences across most metrics were within 1%. Both models perform significantly better on reliable segments than on unreliable ones. For unreliable segments, the LightGBM model shows slightly higher accuracy at the 80th and 95th percentiles, although the improvements are generally within 10%. Similar to LMMs, the prediction accuracy declines as the target percentile increases.

The predictive performance of the commonly used trend line method was calculated, with results presented in Table 13, and compared with the RF and LightGBM model results in Tables 11 and 12. Both the RF and LightGBM models showed significant improvements in prediction accuracy over the traditional trend line method. The mean absolute error was reduced by approximately 50% for reliable segments and more than 10% for unreliable segments, highlighting the effectiveness of these machine learning models in capturing trends and making more accurate predictions.

Table 11. Prediction Performance of Random Forest and LightGBM Models for Reliable Segments

External Segments								
	RF				LightGBM			
	MAE	RMSE	R^2	MAPE	MAE	RMSE	R^2	MAPE
50th	2.101	3.496	0.935	7.653	2.260	3.623	0.931	8.368
80th	2.862	4.749	0.901	9.517	2.889	4.707	0.902	9.742
95th	4.566	9.608	0.751	12.166	4.696	9.012	0.749	15.785
Internal Segments								
	RF				LightGBM			
	MAE	RMSE	R^2	MAPE	MAE	RMSE	R^2	MAPE
50th	2.356	3.520	0.911	8.517	2.7279	4.1535	0.9070	10.416
80th	2.348	3.736	0.903	7.736	3.1436	4.8363	0.8925	10.909
95th	3.759	7.713	0.768	10.479	4.6961	9.0196	0.7592	13.558

RF= random forest, LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean square error.

Table 12. Prediction Performance of Random Forest and LightGBM Models for Unreliable Segments

External Segments								
	RF				LightGBM			
	MAE	RMSE	R^2	MAPE	MAE	RMSE	R^2	MAPE
50th	9.560	13.399	0.943	14.272	9.504	13.242	0.975	13.582
80th	21.312	27.592	0.591	27.289	18.401	24.556	0.659	21.451
95th	25.895	35.446	0.574	29.039	24.740	36.979	0.529	30.889
Internal Segments								
	RF				LightGBM			
	MAE	RMSE	R^2	MAPE	MAE	RMSE	R^2	MAPE
50th	3.291	5.258	0.863	15.848	3.249	5.444	0.857	12.576
80th	15.723	22.892	0.648	30.677	12.830	18.591	0.694	23.613
95th	25.270	49.840	0.394	25.501	20.397	39.150	0.425	22.869

LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error; RF= random forest, RMSE= root mean square error.

Table 13. Prediction Performance of Trend Lines

External Segments								
	Reliable Segments				Unreliable Segments			
	MAE	RMSE	R^2	MAPE	MAE	RMSE	R^2	MAPE
50th	4.175	7.235	0.997	13.444	11.730	17.591	0.959	14.727
80th	4.684	10.810	0.993	13.690	32.404	41678	0.911	29.552
95th	7.504	20.323	0.978	15.054	34.546	56.708	0.926	36.467
Internal Segments								
	Reliable Segments				**Unreliable Segments**			
	MAE	**RMSE**	**R^2**	**MAPE**	**MAE**	**RMSE**	**R^2**	**MAPE**
50th	4.951	5.775	0.823	17.749	5.061	6.332	0.868	16.400
80th	5.828	8.083	0.711	19.167	19.055	24.369	0.474	33.921
95th	7.953	15.860	0.319	22.211	32.680	48.836	0.106	35.094

MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean square error.

Figures 22 through 25 compare the prediction performance of the LMM, RF, and LightGBM models. The machine learning models outperformed LMM, with predictions close to the 45° lines in the scatter plots. Several consistent trends emerge across all models:

1. For the same target percentile, models for reliable segments (LOTTR < 1.5) demonstrate higher accuracy than those for unreliable segments. Similarly, models for external segments generally outperform those for internal segments.
2. Prediction accuracy decreases as the target percentile increases. The highest prediction accuracy is observed in the 50th percentile models for the reliable external segment subgroup (Figure 22). Among the three modeling techniques, RF models demonstrated a superior performance in handling outliers, which is also reflected in the performance metrics in Tables 11 and 12, for which the RF model outperformed both LightGBM and LMM.
3. For the 80th percentile models, LightGBM generally produced the lowest number of errors. However, RF outperformed it for the unreliable internal segment subgroup, as Figure 23 shows. LightGBM produced several large errors for this subgroup, likely due to noisy input data or the inclusion of unimportant variables, or both. RF showed the advantage of handling variables with a low importance score compared with LightGBM, primarily because of its use of variable bagging. This technique also enables RF to effectively average out noise in the data, whereas LightGBM may be prone to overfitting to noise if it is not adequately tuned.

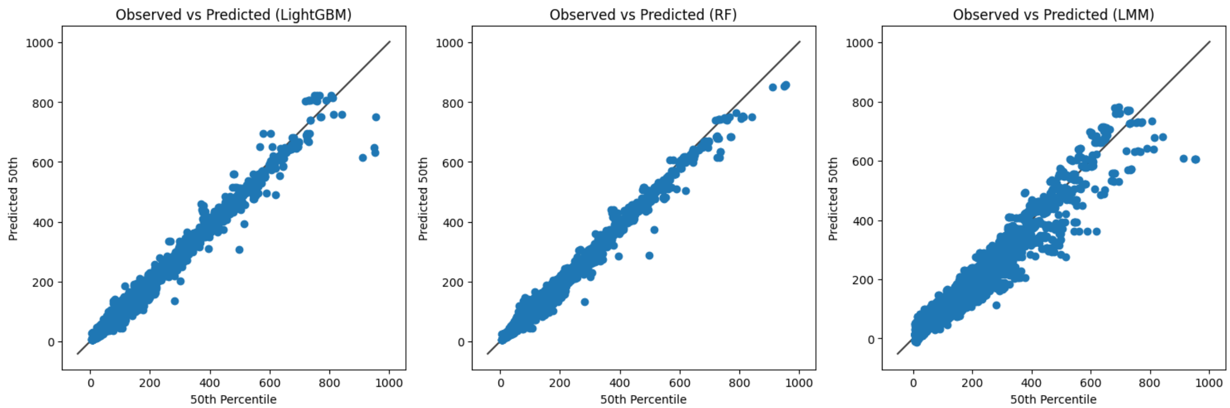


Figure 22. Observed versus Predicted 50th Percentile Travel Time Models for Reliable External Segments. LightGBM = light gradient boosting machine; LMM = linear mixed model; RF = random forest.

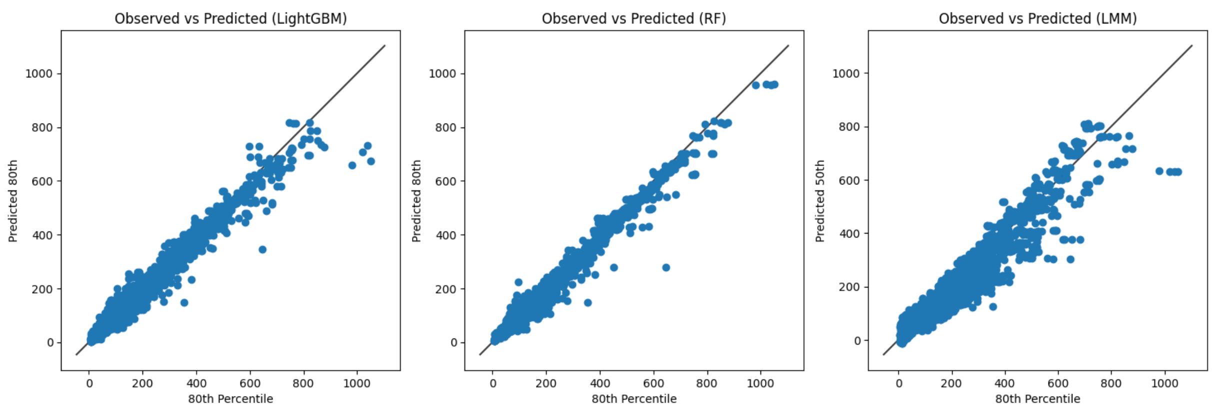


Figure 23. Observed versus Predicted 80th Percentile Travel Time Models for Reliable Internal Segments. LightGBM = light gradient boosting machine; LMM = linear mixed model; RF = random forest.

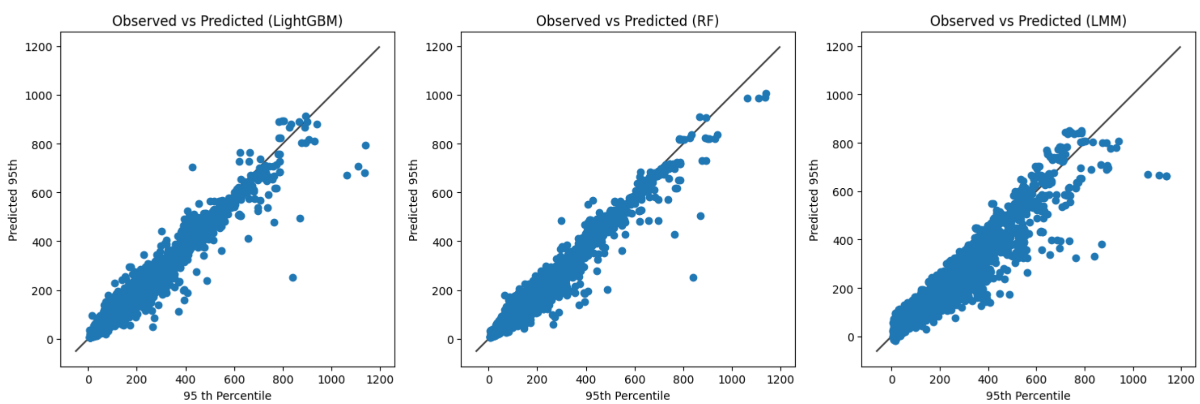


Figure 24. Observed versus Predicted 95th Percentile Travel Time Models for Reliable Internal Segments. LightGBM = light gradient boosting machine; LMM = linear mixed model; RF = random forest.

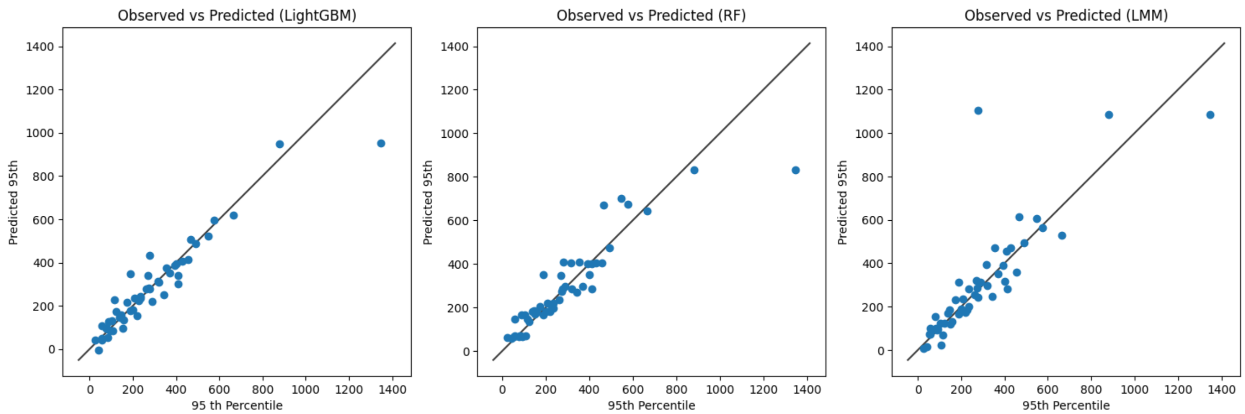


Figure 25. Observed versus Predicted 95th Percentile Travel Time Models for Unreliable External Segments. LightGBM = light gradient boosting machine; LMM = linear mixed model; RF = random forest.

Among all the models trained for the three target percentiles and four segment subgroups, the 95th percentile models for the internal unreliable segment subgroup showed the lowest prediction accuracy. As Figure 25 shows, LightGBM has better performance than LMM and RF methods. Among all peak-hour models, the LightGBM model generally produces higher accuracy than LMM and RF for the 80th and 90th percentile predictions. LMMs, although with comparatively lower accuracy, offer the advantage of interpretability through model coefficients and are more computationally efficient to train. In contrast, RF and LightGBM models can capture complex nonlinear relationships but are less interpretable than LMMs for the datasets in this study. In addition, tuning the LightGBM model is more challenging than tuning the RF model.

Given that both RF and LightGBM outperformed LLM and achieved similar levels of predictive accuracy, they were selected to develop models for all four time periods.

Model for All Time Periods

The variables used in the development of the peak-hour models were used to construct models that can predict the target percentile in any of the four time periods. Time periods were encoded as dummy variables within the models, with the a.m. peak as the baseline. Models were trained for each of the external and internal segment groups and for each target percentile. The previous VTRC studies by Zhang et al. (2021) and Zhao and Appiah, (2023) found that building models for internal and external segments separately enhances accuracy while reducing computational complexity. These two studies also recommend combining reliable and unreliable segments within each segment type because RF models can learn distinct patterns of those segments. Following these recommendations, this study developed RF and LightGBM models for each segment type group to predict travel times at the 50th, 80th, and 95th percentiles. Models for each percentile were trained separately.

Variable Selection

Input variables were selected using the same approach used for the peak-hour models. Figure 26 shows the variable importance scores of the RF model for the external segments.

Figure 27 illustrates the SHAP values of the LightGBM model. The top three most important variables are consistent across both models, with segment length demonstrating the most impact. The rural and urban designations and the count of adverse weather events showed low predictive power in both models. Traffic volume is a relatively strong predictor in the RF model. The SHAP scores of the LightGBM model also indicate a generally high positive relationship between volume and the predicted percentile, with volume significantly increasing the prediction for certain samples.

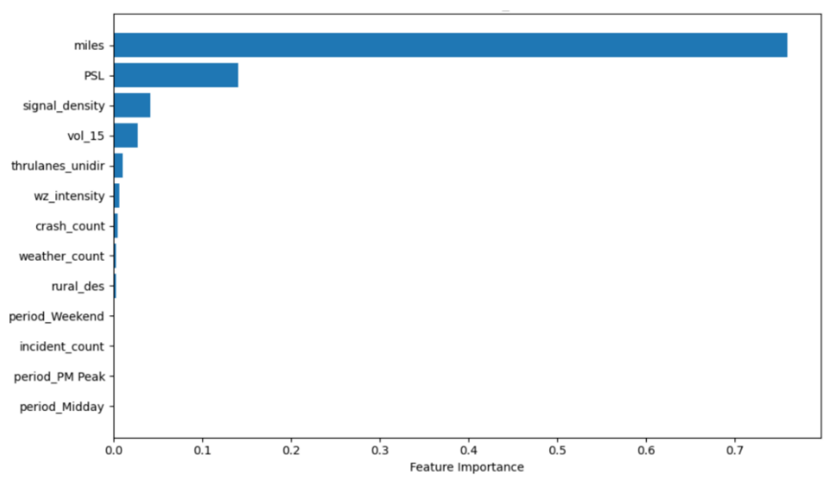


Figure 26. Random Forest Model Variable Importance Scores for External Segments in All-Period Model

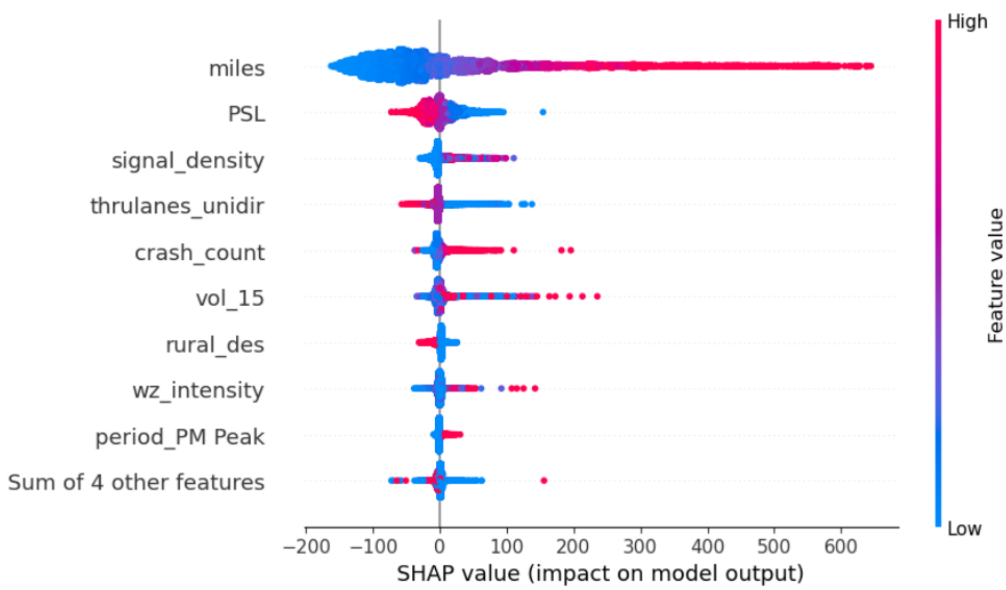


Figure 27. Light Gradient Boosting Machine Model SHAP Scores for External Segments in All-Period Model. PSL = posted speed limit; SHAP = SHapley Additive exPlanations.

Performance Comparison

Table 14 presents the performance metrics of RF and LightGBM all-period models. These metrics were calculated using data for all four time periods. Overall, the two modeling techniques demonstrated very similar levels of accuracy. The RF model slightly outperformed the LightGBM model for the 50th percentile, with a MAPE difference of less than 4%. The LightGBM model demonstrated marginally superior performance for the 80th and 95th percentiles. This result is consistent with the findings from the peak-hour models.

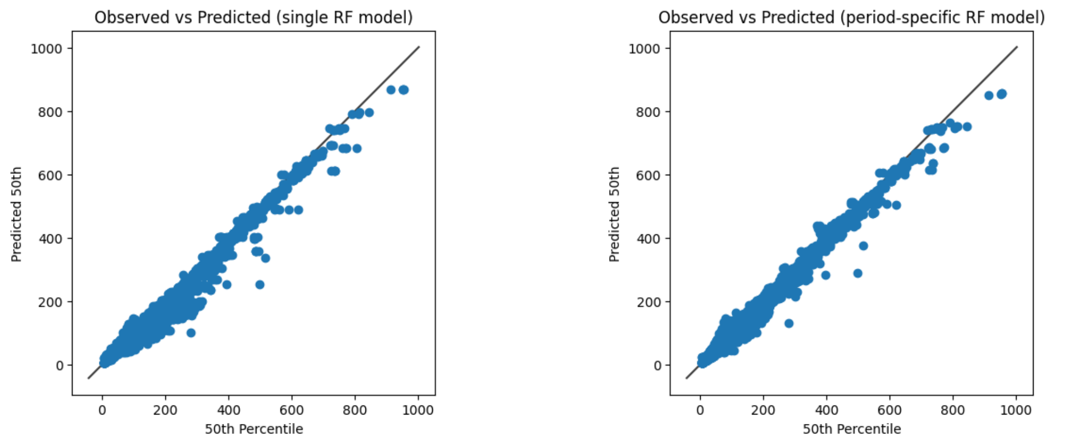
Table 14. Prediction Performance of RF and LightGBM All-Period Models

		MAE	RMSE	R^2	MAPE
50th	LightGBM	15.584	24.547	0.960	13.864
	RF	11.015	19.975	0.973	8.592
80th	LightGBM	13.384	16.393	0.836	10.831
	RF	14.006	24.056	0.965	10.260
95th	LightGBM	19.603	32.551	0.943	14.502
	RF	19.657	34.292	0.936	13.864

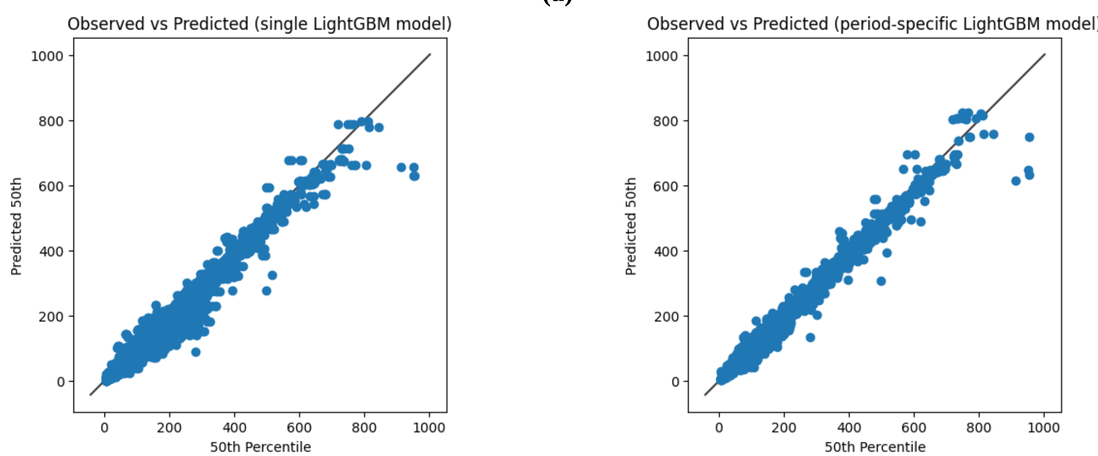
LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error; RF= random forest; RMSE= root mean square error.

For each target percentile, the predictions for the a.m. and p.m. peak hours from the all-period model were compared with the outputs from the peak-hour model. Figures 28 through 30 show the results for the 50th, 80th, and 95th percentile travel times on external segments. The plots on the left show the predictions generated by the all-period model in comparison with the observed peak-hour data, and those on the right depict the predictions from the peak-hour model against the observed data. For both RF and LightGBM, the peak-hour models consistently produced more accurate predictions than the all-period model, with the accuracy gap being largest for the 95th percentile models.

Both the RF and LightGBM models trained on all-period data produce less accurate predictions compared with the period-specific models. For example, comparing the all-period model performance metrics in Table 15 with the peak-hour model performance metrics in Tables 10 and 11, one can observe that MAPEs of the all-period model predictions are approximately 3% higher for RF models and 5% higher for LightGBM models when predicting the 50th, 80th, and 95th percentiles on reliable segments. For unreliable segments, both the RF and LightGBM all-period models show substantially increased errors, with the increase in MAPE values mostly ranging from 10% to 20%. When predicting the 95th percentile for the unreliable external segment subgroup, the all-period LightGBM model shows a significant increase in MAPE—from 30 to 60—compared with the peak-hour model. This outcome is likely because the dataset is not large enough to fully leverage the advantages of LightGBM. The study dataset includes a small number of unreliable segments, each with unique features. As a result, the LightGBM model tends to overfit the training data, capturing outliers rather than generalizable patterns. RF models rely on averaging multiple decision trees to reduce variance, and they appear to generalize better in this case.

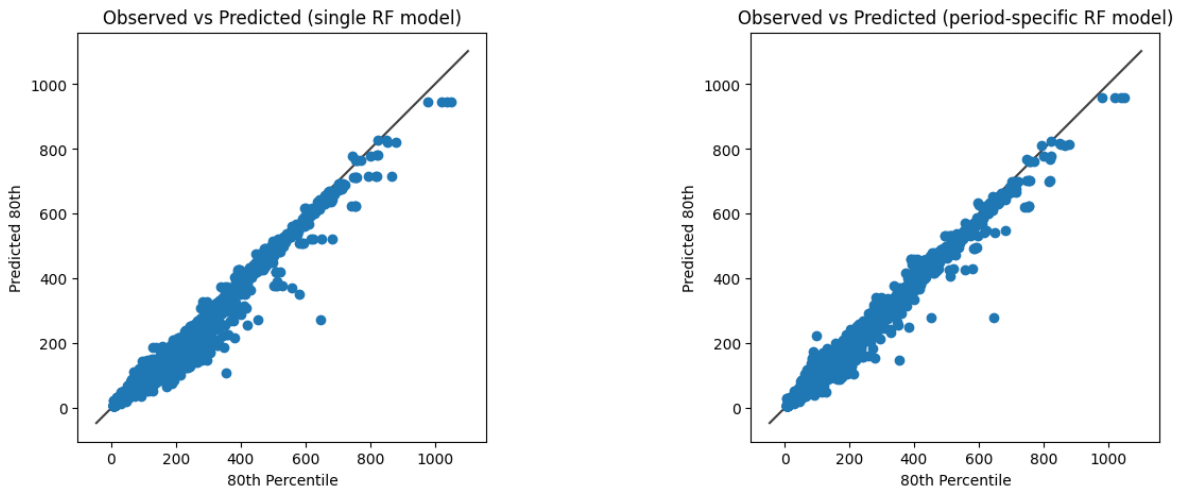


(a)

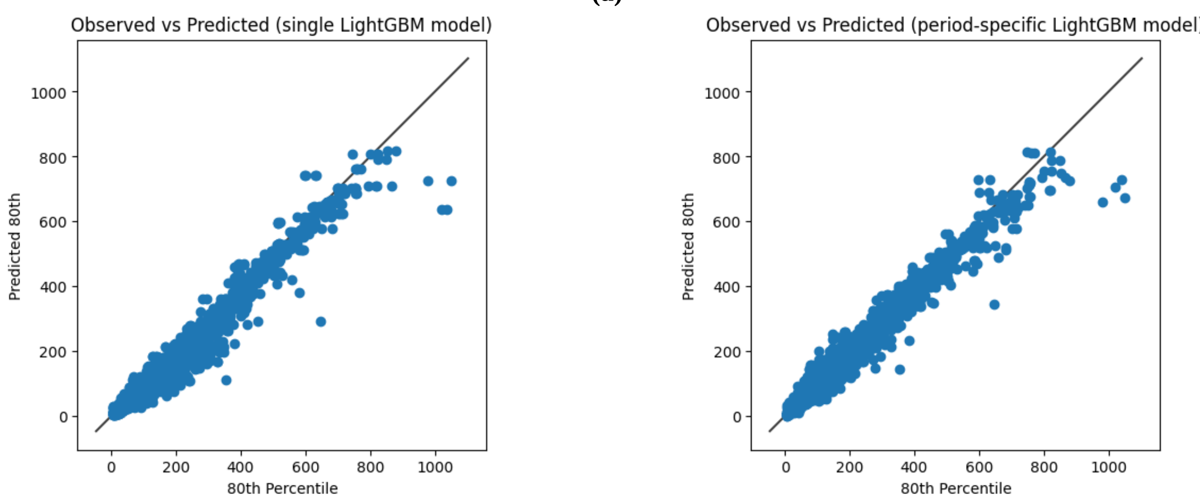


(b)

Figure 28. Single All-Period Models versus Period-Specific Models for 50th Percentile: (a) RF; (b) LightGBM. LightGBM = light gradient-boosting machine; RF = random forest.

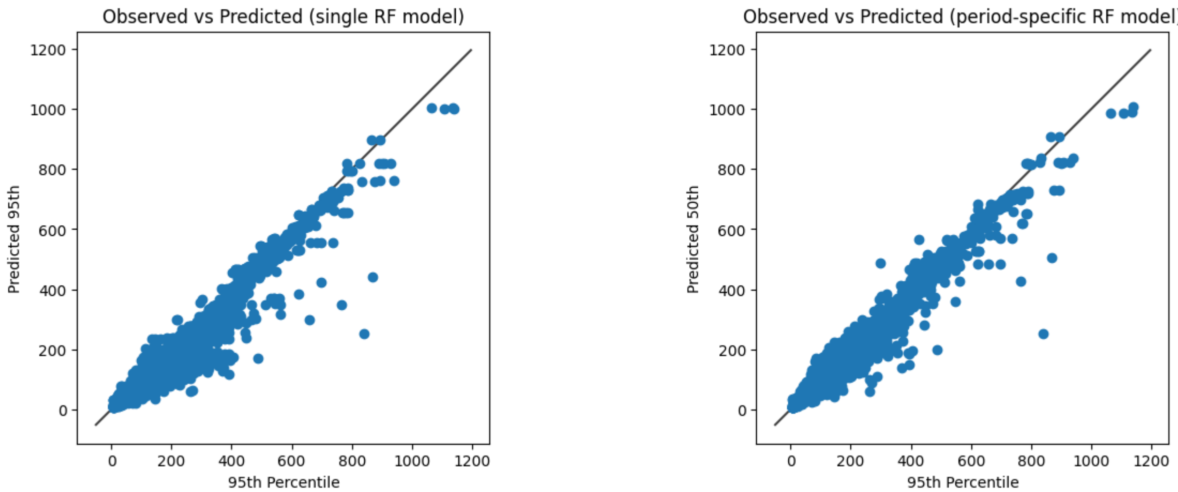


(a)

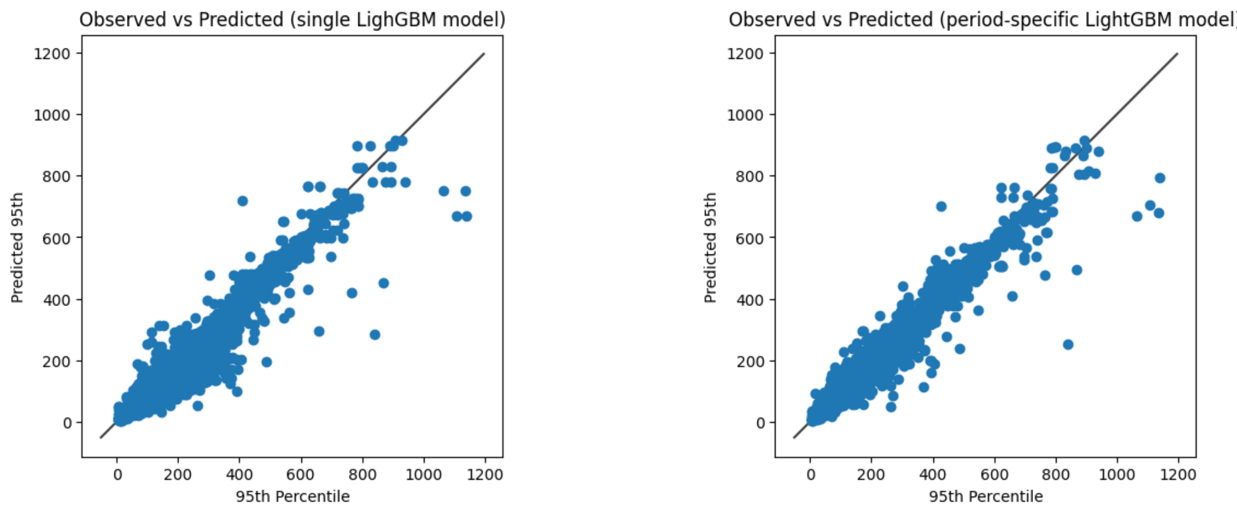


(b)

Figure 29. Single All-Period Models versus Period-Specific Models for 80th Percentile: (a) RF; (b) LightGBM. LightGBM = light gradient boosting machine; RF = random forest.



(a)



(b)

Figure 30. Single All-Period Models versus Period-Specific Models for 95th Percentile: (a) RF; (b) LightGBM. LightGBM = light gradient boosting machine; RF = random forest.

For peak hours only, the all-period RF model generally outperformed the all-period LightGBM model, except for the 80th percentile for the unreliable external segment subgroup.

The all-period model offers easier implementation compared with training separate models for each time period because it requires less effort in model tuning and management. However, its suitability depends on the purpose of the analysis and the accuracy requirements of the specific application. For quick exploratory analyses or high-level planning efforts, the all-period model may provide sufficient accuracy and operational efficiency.

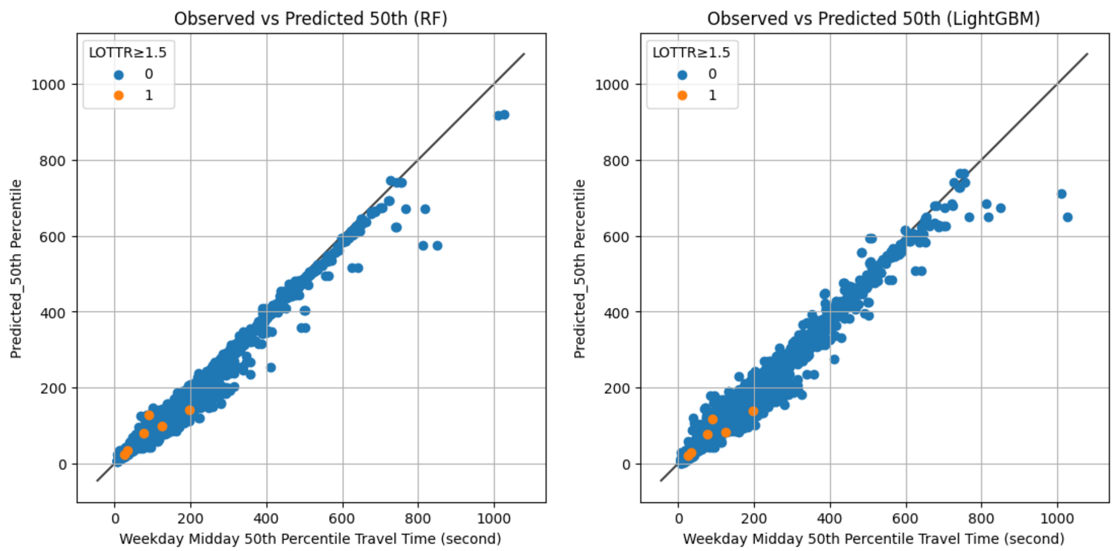
Table 15. Prediction Performance of Random Forest and LightGBM All-Period Models for Peak Hours

Reliable External Segments								
	RF				**LightGBM**			
	MAE	**RMSE**	**R^2**	**MAPE**	**MAE**	**RMSE**	**R^2**	**MAPE**
50th	11.950	21.085	0.970	9.175	16.847	26.203	0.954	14.760
80th	15.529	25.964	0.959	11.156	18.122	28.294	0.951	14.559
95th	22.080	37.380	0.925	15.109	22.671	36.989	0.926	16.260
Unreliable External Segments								
	RF				**LightGBM**			
	MAE	**RMSE**	**R^2**	**MAPE**	**MAE**	**RMSE**	**R^2**	**MAPE**
50th	9.123	15.250	0.971	10.989	13.411	17.098	0.963	19.839
80th	30.286	44.212	0.780	28.291	25.427	35.766	0.856	23.061
95th	52.454	71.904	0.549	47.140	78.090	113.789	0.129	60.712

LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error RF= random forest; RMSE= root mean square error.

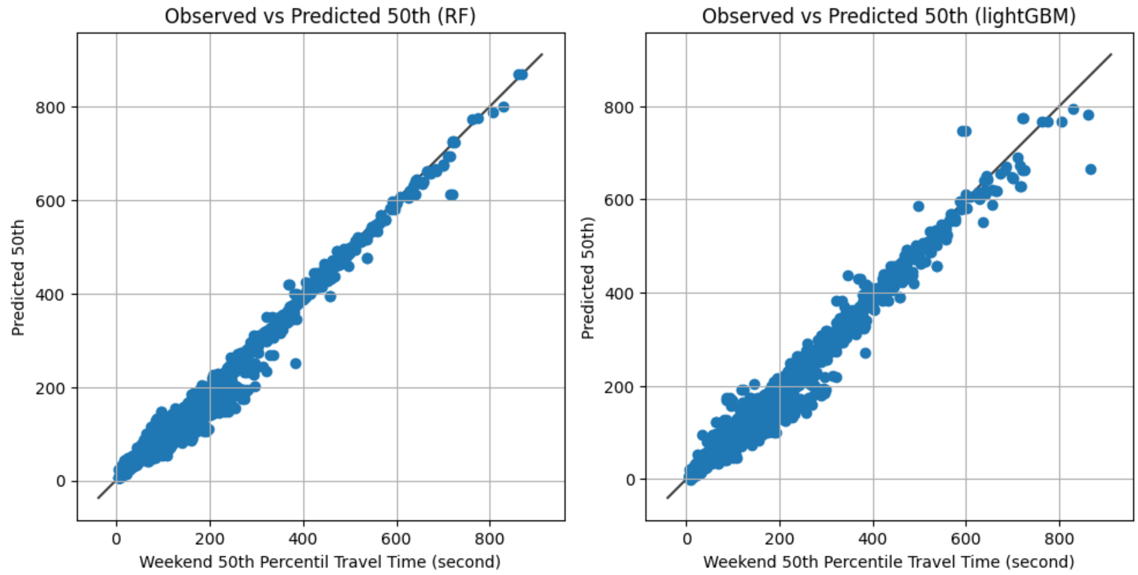
Predictions for Mid-Day and Weekend Periods

The 50th, 80th, and 95th percentile travel times in weekday mid-day and weekend periods were predicted using the all-period models. Figures 31 and 32 show the results of the 50th percentile for the RF and LightGBM models, respectively. Table 16 shows the performance metrics for the 50th percentiles. RF models outperformed the LightGBM models by about 5% in terms of MAPE. The errors for the 80th and 90th percentiles are larger, similar to the trends in Table 14 and Figures 29 and 30.



(a) (b)

Figure 31. Comparison of the Predictions of 50th Percentile Travel Times in Mid-Day Period; (a) RF; (b) LightGBM. LightGBM = light gradient-boosting machine; RF = random forest.



(a) **(b)**

Figure 32. Comparison of the Predictions of 50th Percentile Travel Times in Weekend Period; (a) RF; (b) LightGBM. LightGBM = light gradient boosting machine; RF = random forest.

Table 16. Prediction Performance of RF and LightGBM for Weekend

		MAE	**RMSE**	**R^2**	**MAPE**
Mid-day	LightGBM	16.722	26.431	0.954	14.511
	RF	11.968	21.625	0.969	8.835
Weekend	LightGBM	13.918	21.769	0.967	13.041
	RF	9.075	16.537	0.981	7.827

LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error; RF= random forest; RMSE= root mean square error.

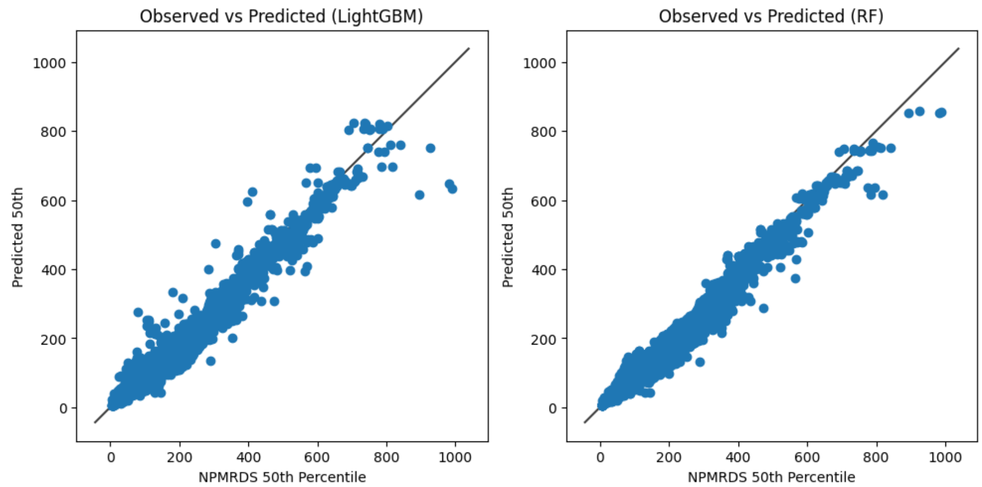
Prediction of NPMRDS Travel Time Percentiles

Although VDOT uses INRIX data for most mobility and reliability analyses, NPMRDS data are used for MAP-21 performance target setting. The peak-hour models trained with INRIX-RT data were used to predict the 50th, 80th, and 95th percentile travel times. Table 17 shows the performance metrics. Both RF and LightGBM produced reasonably good accuracy for the 50th percentile, with MAPE around 10%. The RF model achieved slightly better accuracy than LightGBM. MAPEs for both models are a little more than 25% for 95th percentile NPMRDS travel time. Figure 33 shows the observed and predicted NPMRDS travel time percentiles. The models trained with INRIX-RT generally underestimate the target percentiles generated using NPMRDS data, especially for the 80th and 95th percentiles. Figure 34 shows the distribution of prediction residuals for the 95th percentile. Both models generated some instances of large residuals. The LightGBM model appears to have more extreme values, suggesting that cases occurred in which the predictions deviated significantly from the observed values.

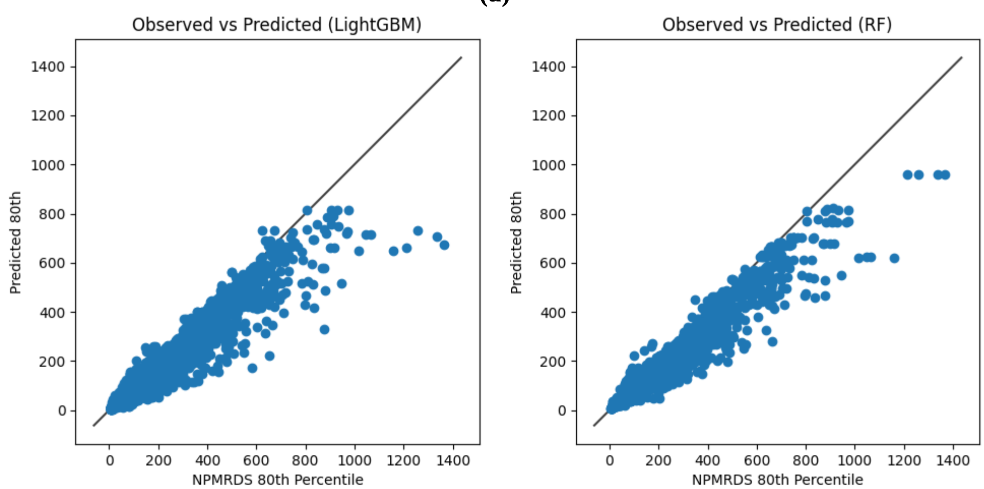
Table 17. Prediction Performance of RF and LightGBM Peak-Hour Models for the National Performance Management Research Data Set

		MAE	RMSE	R^2	MAPE
50th	LightGBM	13.767	22.443	0.967	11.041
	RF	12.818	21.281	0.971	9.853
80th	LightGBM	29.221	50.647	0.875	17.465
	RF	26.671	45.885	0.898	15.394
95th	LightGBM	69.620	123.540	0.629	27.372
	RF	68.243	121.333	0.642	26.247

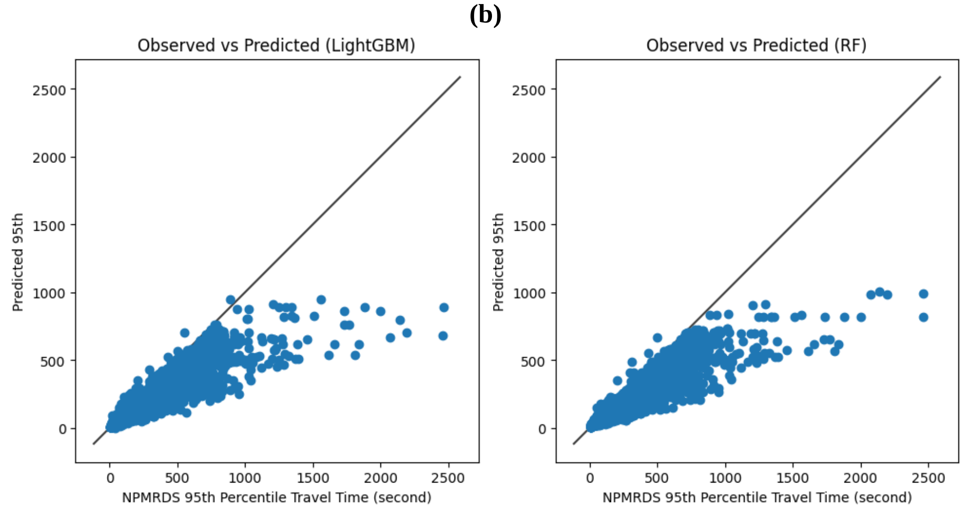
LightGBM = light gradient boosting machine; MAE = mean absolute error; MAPE = mean absolute percentage error; RF= random forest; RMSE= root mean square error.



(a)



(b)



(c)

Figure 33. Observed versus Predicted NPMRDS Travel Time Percentiles: (a) 50th; (b) 80th; (c) 95th. LightGBM = light gradient boosting machine; NPMRDS = National Performance Management Research Data Set; RF = random forest.

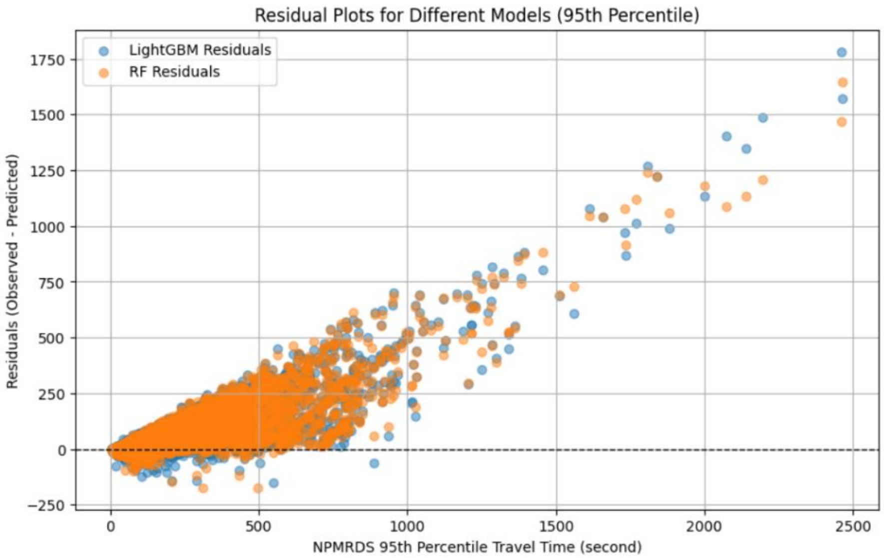


Figure 34. Residual of Predicted 95th Percentile NPMRDS Travel Times. LightGBM = light gradient boosting machine; NPMRDS = National Performance Management Research Data Set; RF = random forest.

Discussion

Analyses of Travel Time Distribution and Reliability Measure

NPMRDS Travel Times Are Systematically Higher than INRIX, Showing upward Bias that Affects Reliability Metrics

Across nearly all percentiles, particularly at the upper end, NPMRDS consistently reports longer travel time than either INRIX or INRIX-RT for the same roadway segments. This upward bias emerges early, at the 50th percentile. Approximately 80% of TMCs already show higher travel times in NPMRDS. The discrepancy becomes more pronounced at higher percentiles, with more than 95% of TMCs reporting longer times in NPMRDS. Although extreme upper-percentile values are occasionally observed in NPMRDS, it remains uncertain whether these values reflect true congestion spikes or anomalies in the dataset.

The systematically longer travel time at higher percentiles directly influences travel time reliability metrics. As the portion of higher reporting TMCs increases with the percentile, NPMRDS naturally yields higher LOTTR values. In all reporting periods, NPMRDS identifies two to three times more segments as unreliable than INRIX-based datasets. Notably, most TMCs flagged as unreliable by INRIX or INRIX-RT are also identified by NPMRDS, suggesting that NPMRDS captures a superset of unreliable segments. The additional TMCs uniquely identified by NPMRDS likely result from their higher travel time readings, reinforcing the effect of dataset bias on performance reporting.

Data Density and Distortion in Travel Time Distribution

Data density differed sharply between datasets. More than 90% of INRIX-RT TMCs exceeded 4,000 samples, whereas only about 30% of NPMRDS TMCs reached that level.

However, removing low-sample TMCs does not reduce the NPMRDS-INRIX travel time gap, showing that raw density is not the main culprit.

A more meaningful indicator of data quality is the distribution of data density classifications A, B, and C. In NPMRDS, density A dominates most time periods, accounting for more than 80% of observations across all time periods. This prevalence of low-density intervals suggests that a large share of travel time estimates are generated from sparse probe readings, which increases variability and reduces the stability of percentile calculations. These conditions are especially problematic for higher percentiles, for which even a few high readings can distort the shape of the travel time distribution.

On sparsely sampled links, INRIX backfills missing real-time data with historical values, smoothing variability and flattening its cumulative distribution curve, and sometimes causes step-function-like shapes. This practice may suppress short-term travel time spikes and hide some low real-time readings, producing narrower distributions and lower reliability metrics. NPMRDS, despite its lower overall sample size, retains these extremes, providing a fuller picture of volatile conditions. However, the downside is that when upper-percentile estimates are derived primarily from density A intervals, NPMRDS may also exaggerate unreliability, particularly if a single high reading disproportionately influences the 80th or 95th percentiles.

Thus, the divergent treatment of sparse or missing data, more so than data presence, explains the different behavior of NPMRDS and INRIX in travel time estimation and reliability metrics. The methodological choices in how datasets handle low-density intervals directly shape the observed distribution and performance measures.

NPMRDS Flags More Unreliable Segments Beyond INRIX

INRIX and INRIX-RT identify very similar sets of unreliable TMCs. NPMRDS captures this shared core, yet with higher LOTTR values and with many more unreliable segments. These discrepancies cannot be fully explained by sparse sampling alone because many of the largest LOTTR gaps occur on segments with more than 800 observations in both NPMRDS and INRIX-RT. However, closer inspection reveals that data density, not just raw sample count, plays a critical role. A large proportion of these segments, despite having sufficient total samples, are dominated by density A intervals, meaning that most 15-minute periods contain fewer than five probe readings. These low-density conditions make percentile travel time estimates in NPMRDS more volatile, often inflating upper percentiles and, in turn, LOTTR values.

Segment length also emerges as an influential factor. One-half of the 32 largest LOTTR discrepancies occur on segments between 0.1 and 0.25 miles long, and when sub-0.1-mile segments are included, 99 of the top 100 largest gap TMCs are under 0.25 mile. When both NPMRDS and INRIX-based datasets identify a TMC as unreliable, it can be reasonably considered a true reliability concern. However, NPMRDS often flags a much larger set of segments, particularly short ones, as unreliable because of its upward bias in travel times. Practitioners should use caution when interpreting these results because the discrepancies may reflect artifacts of data processing rather than meaningful reliability issues.

Predictive Models

This section presents the development of prediction models for the 50th, 80th, and 95th percentile travel times using LMM, RF, and LightGBM algorithms. The results indicate that both RF and LightGBM models significantly outperform the LMM approach when predicting travel times for reliable segments. For unreliable segments with a limited sample size, machine learning models still exhibit better performance than LMM, but the improvement is not as pronounced as for the reliable segments.

For the peak-hour models, the accuracies of RF and LightGBM models were very similar, and the discrepancy in MAPE was usually within 5%. RF generally performs slightly better for the 50th percentile, whereas LightGBM shows a slight advantage for the 95th percentile. For the all-period model, RF appears to be more robust in handling outliers and therefore performs slightly better than LightGBM. LightGBM performs best with large datasets greater than 10,000 rows but tends to overfit when applied to smaller datasets.

The period-specific model provides better accuracy than the all-period model when predicting the target percentiles for that particular time period, especially for unreliable segments. Compared with the peak-hour models, predictions generated by the all-period models show a higher number of errors, with MAPE increasing by approximately 3 to 5% for reliable segments and by 10 to 20% for unreliable segments. For both period-specific and all-period models, the accuracy of 50th percentile predictions is higher than that of 80th and 90th percentile predictions. In addition, prediction results for reliable segments are significantly better than those for unreliable segments.

The models trained with INRIX-RT data underestimate the target percentiles calculated using NPMRDS data. To adopt the developed models for MAP-21 performance target setting, additional research is needed to train models on NPMRDS data.

Although the travel time reliability prediction models for NHS arterials developed in this study show promise, further research is needed to apply them to MAP-21 reliability performance target setting. This research includes but is not limited to conducting case studies to evaluate the models' capacity to capture future reliability changes resulting from traffic improvement projects, adapting the models to enhance the accuracy of predicting NPMRDS travel time percentiles, improving accuracy for internal segments, and investigating the feasibility of employing the developed method to model the INRIX XD (eXtreme Definition) network.

CONCLUSIONS

- *NPMRDS travel times are systematically higher than INRIX, introducing an upward bias that affects reliability metrics.* At the 50th percentile, approximately 80% of TMCs show higher values in NPMRDS, which increases to more than 95% at the upper percentiles. Although some extreme values are observed at upper percentiles, whether these values reflect true congestion or are anomalies is unclear. This consistent overestimation leads to higher LOTTR ratios, with NPMRDS identifying two to three times more unreliable TMCs than

INRIX-based datasets. NPMRDS also captures the most unreliable TMCs flagged by INRIX, which adds additional segments, likely because of its inflated travel time estimates.

- *Applying sample size filters does not substantially reduce discrepancies between NPMRDS and INRIX travel time.* Although INRIX-RT generally contains far more observations than NPMRDS, removing low-sample TMCs does little to close the gap between the two datasets. The persistent differences are driven more by how each dataset handles limited data than by sample size alone. INRIX fills gaps by incorporating historical data, which smooths variability and produces more stable travel time distributions. This approach can suppress short-term spikes and flatten the upper tail of the distribution, leading to lower reliability metrics. In contrast, NPMRDS retains all values, including extreme travel times that may reflect true congestion or random noise under sparse probe readings. This results in greater variability in upper-percentile travel times and can lead to overstated unreliability, especially on segments with low data density. These differences in datasets explain why discrepancies persist even after applying sample size thresholds, particularly in reliability measures like LOTTR that depend heavily on the upper end of the travel time distribution.

- *NPMRDS identifies more unreliable segments than INRIX*. INRIX and INRIX-RT consistently identify the same core set of unreliable TMCs, but NPMRDS flags many more, often with much higher LOTTR values. These differences are not simply due to sample size but are observed to be linked to segment length and how each dataset handles sparse data. TMCs identified as unreliable by both NPMRDS and INRIX are likely valid concerns, but those identified only by NPMRDS, especially on short or low data density segments, should be interpreted with caution.

- *The RF and LightGBM models outperform LMMs in predicting the 50th, 80th, and 95th percentile travel times on non-interstate NHS arterials.* The performance improvement is significant for reliable segments with LOTTR less than 1.5. For unreliable segments with substantially smaller sample sizes, the RF and LightGBM models still show better performance than LMMs, but the improvement is less pronounced. The prediction accuracies of RF and LightGBM models are very similar, with a discrepancy within 5% in terms of MAPE in most cases.

- *Period-specific models are preferred over all-period models when higher prediction accuracy is desired*. For reliable segments, the prediction errors of the all-period models are approximately 3 to 5% higher than those of the peak-hour models. For unreliable segments, the differences are about 10 to 20%. The all-period models are easy to implement and thus may be used for quick exploratory or high-level planning analyses.

- *The RF and LightGBM models trained with INRIX-RT data underestimate the 50th, 80th, and 95th percentile travel times calculated using NPMRDS data.* MAPEs are around 10% for the 50th percentile, 15% for the 80th percentile, and in excess of 25 % for the 95th percentile.

RECOMMENDATIONS

1. *VDOT Traffic Operations Division (TOD) should apply the findings from this study to enhance travel time reliability analysis across planning and performance management practices.* TOD should collaborate with relevant stakeholders, including the Transportation and Mobility Planning Division (TMPD) and OIPI, to identify key areas where these findings can be applied and encourage and promote their use. Where applicable, the findings may be incorporated into internal guidance, tools, and procedures to encourage consistent reliability performance measurement approaches throughout the agencies.

2. *VTRC should conduct additional research to expand the application of the reliability prediction models developed in this study for MAP-21 performance target setting.* This future work should include developing and testing models using the NPMRDS dataset to align with federal reporting requirements. In addition, the research should further examine the models' capability to quantify changes in travel time reliability resulting from traffic improvement projects. Finally, VTRC should evaluate the feasibility of adapting the modeling framework to use with INRIX XD segments, identifying required modifications in data processing or model structures to ensure accurate and robust predictions.

IMPLEMENTATION AND BENEFITS

Researchers and the technical review panel (listed in the Acknowledgments) for the project collaborate to craft a plan to implement the study recommendations and to 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 6 months of the report's publication, TOD and VTRC will organize a stakeholder meeting to support implementing this study's findings. The Project Champion will invite the stakeholders, who will be identified in consultation with the technical review panel and will include, at a minimum, representatives from TOD, TMPD, and OIPI. VTRC will facilitate the meeting; present the key findings related to data sources, travel time distributions, travel time reliability measures, and prediction models; seek stakeholder input on implementation; and summarize the discussions. The objective of the meeting will be to identify potential areas for improvement in current practices and discuss how the research findings could be broadly applied across TOD, TMPD, OIPI, and others to enhance consistency in travel time reliability performance management while addressing each group's analytical needs and priorities. Based on the meeting outcomes, VTRC, under the guidance of the technical review panel, will develop a technical memorandum or guidance document and distribute it to the stakeholders to support the advancement of current practices.

Regarding Recommendation #2, within 3 months of the report's publication, VTRC will initiate a technical assistance project to support VDOT and OIPI during the next federally *mandated performance target-setting cycle in 2026.* The project will apply the reliability prediction framework developed in this study and fine-tune the models for NPMRDS to assist with setting appropriate targets for non-interstate NHS routes. Preliminary results will be prepared for VDOT and OIPI internal review by March 2026 to ensure the insights will inform the target-setting process in a timely manner.

Regarding Recommendation #2, within 6 months of the report's publication, VTRC will develop a research needs statement to address the additional research areas identified in this study beyond MAP-21 target setting. These needs include further evaluating the model's ability to capture reliability changes resulting from traffic improvement projects and exploring the feasibility of adapting the modeling framework for INRIX XD segments. The research needs statement will be submitted to the Data, Performance Measurement, Simulation, Connected and Automated Vehicle Research Advisory Subcommittee of the Traffic Operations and Safety Research Advisory Committee for consideration.

Benefits

The implementation of Recommendation 1 will enhance VDOT's ability to accurately analyze travel time reliability and interpret the results, and thus support more informed, consistent, and defendable decision-making across multiple planning and operation contexts. This study provides new insights into data quality, travel time distributions, and reliability metrics. The developed prediction models significantly improved travel time reliability prediction compared with the commonly used trend line approach, reducing prediction errors by 10 to 52% across different segments. By incorporating the study's findings into procedures, tools, and internal guidance for planning and performance measurement activities, VDOT will be better equipped to develop meaningful, actionable business insights and more accurately identify roads with reliability issues, allowing for targeted investment where most needed. Potential application areas may include:

- *Investment justification*. The reliability predictions could be used to support the justification of traffic improvement projects by quantifying the changes in reliability before and after the improvement.
- *Infrastructure investment*. By more effectively identifying areas with unreliability travel times, VDOT can better prioritize infrastructure or operational improvements in areas where reliability is most problematic.
- *Support for project selection through Smart Scale*. Travel time reliability is a critical input in the Smart Scale project selection process, particularly under the Economic Development factor area and as a Congestion and Accessibility measure. For example, peak period travel time reliability Buffer Time Index is used to calculate the Economic Development Travel Time Reliability measure for proposed projects. Travel Time Index, a measure of congestion closely associated with travel time reliability metrics, is used to scale peak-hour delay and accessibility throughput measures to peak period scores. The findings in this study can enhance the accuracy of these inputs by providing more reliable percentile-based travel time estimates on non-interstate NHS arterials. Improving the precision of reliability data used in the Smart Scale process can lead to better informed

scoring and prioritization, ultimately contributing to more effective transportation investment decisions.

- *Traffic management and operations*. Effectively monitoring and predicting travel time reliability allows VDOT to identify areas prone to congestion better, enabling the deployment of real-time traffic management strategies to alleviate congestion.
- *Operational strategies evaluation*. Accurate travel time reliability predictions could help evaluate the effectiveness of different operational strategies, such as evaluating work zone congestion management strategies and estimating the effect of weather incidents to prepare for disruptions better. Robust travel time reliability predictions could also support the assessment of the effectiveness of safety improvements in reducing crashes.
- *Traveler information*. Providing the public with information on expected travel time reliability can encourage safer driving behaviors because drivers may adjust their expectations or plans when traveling through areas with significant variability.

Implementing Recommendation 1 will also improve VDOT's ability to communicate reliability performance to stakeholders and the public with greater confidence and consistency. In addition, because INRIX and NPMRDS data are widely used nationwide, the findings on the comparative analysis of these two datasets could also be shared with VDOT partners and peers to encourage greater transparency in NPMRDS processing and quality improvements, benefiting not only VDOT but also other state departments of transportation and metropolitan planning organizations.

The implementation of Recommendation 2 will provide VDOT and OIPI with a robust set of machine learning models to predict travel time reliability on non-interstate NHS arterials. These models will help VDOT and OIPI set more realistic and achievable reliability targets to meet federal performance management requirements and enhance VDOT's performance measurement and operational planning practices. This increased level of precision not only provides a more reliable basis for setting performance targets but also strengthens VDOT's overall reliability analysis capabilities. By making more accurate predictions, VDOT can optimize its strategies and operations, leading to better resource allocation and more informed decision-making.

ACKNOWLEDGMENTS

The authors are grateful to the following technical review panel members for their guidance and support: Mena Lockwood, TOD Project Champion; Sanhita Lahiri, Traffic Operations Division; Jungwook Jun, TMPD; Justice Appiah, VTRC; and Margit Ray of OIPI of the Secretary of Transportation, Commonwealth of Virginia.

REFERENCES

Cambridge Systematics, Texas A&M Transportation Institute, University of Washington, Dowling Associates, Street Smarts, Levison, H., and Rakah, H. *Analytical Procedures for*

Determining the Impacts of Reliability Mitigation Strategies. SHRP2 Report S2-L03-RR-1. Transportation Research Board, Washington, DC, 2013.

Chen, C., Skabardonis, A., and Varaiya, P. Travel Time Reliability as a Measure of Service. *Transportation Research Record*, No. 1855, 2003, pp. 119–140.

Cheng, W., Li, J.L., Xiao, H.C., and Ji, L. Combination Predicting Model of Traffic Congestion Index in Weekdays Based on LightGBM-GRU. *Scientific Reports*, Vol. 12, 2022. https://doi.org/10.1038/s41598-022-06975-1.

Dong, J., and Mahmassani, H.S. Flow Breakdown and Travel Time Reliability. *Transportation Research Record*, Vol. 2124, No. 1, 2009, pp. 203–212.

Economic Development Research Group, Cambridge Systematics, ICF International, Texas A&M Transportation Institute, and Weris. *Development of Tools for Assessing Wider Economic Benefits of Transportation*. SHRP2 Report S2-C11-RW-1. Transportation Research Board, Washington, DC, 2013.

Elefteriadou, L., Xu, H., and Xie, Li. *Travel Time Reliability Models*. FDOT VD-545 RPWO #75. Florida Department of Transportion, Tallahassee, FL, 2008.

Farzana, F.H. *Estimation and Prediction of Mobility and Reliability Measures Using Different Modeling Techniques*. Florida International University, Miami, FL, 2018.

Fils, P.B. *Modeling Travel Time and Reliability on Urban Arterials for Recurrent Conditions*. Ph.D. dissertation. University of South Florida, Tampa, FL 2012.

Fu, L., Hellinga, B., and Zhu, Y. An Adaptive Model for Real-Time Estimation of Overflow Queues on Congested Arterials. *2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585)*, Oakland, CA, 2001, pp. 219–226. doi: 10.1109/ITSC.2001.948659.

Hadi, M., Xiao, Y., Qomn, S.F., and Cabrera, E. *Performance Measurements of Transportation Systems Based on Fine-Grained Data Collected by AVI and AVL Systems*. National Center for Transportation Systems Productivity and Management, Atlanta, GA, 2014.

Kwon, J., Barkley, T., Hranac, R., Petty, K., and Compin, N. Decomposition of Travel Time into Various Sources: Incidents, Weather, Work Zones, Special Events, and Base Capacity. *Transportation Research Record*, Vol. 2229, No. 1, 2011, pp. 28–33.

Lan, C-L, and Zhao, M. *Selection of an Appropriate Reference Speed for the Calculation of Highway Performance Measures*. 25-R21. Virginia Transportation Research Council, Charlottesville, VA, 2025.

Polus, A. A Study of Travel Time and Reliability on Arterial Routes. *Transportation*, Vol. 8, 1979, pp. 141–151.

Potts, I.B., Harwood, D.W., Hutton, J.M., Fees, C.A., Bauer, K.M., Lucas, L.M., Kinzel, C.S., and Frazier, R.J. *Identification and Evaluation of the Cost-Effectiveness of Highway Design Features to Reduce Nonrecurrent Congestion*. SHRP2 Report S2-L07-RR-1. Transportation Research Board, Washington, DC, 2014.

RITIS. Probe Data Analytics Suite. https://pda.ritis.org/analytics/. Accessed September 10, 2025.

Schuman, R., Jones, T., and Turner, S. National Performance Management Research Data Set (NPMRDS), Descriptive Metadata Document, Revision 2.1. 2024. https://files.ritis.org/public/pda/help/NPMRDS.pdf.

Tu, H., van Lint, H., and Van Zuylen, H. Travel Time Variability Versus Freeway Characteristics. *2006 IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, 2006, pp. 383–388. doi: 10.1109/ITSC.2006.1706771.

van Lint, J.W.C., and van Zuylen, J.H. Monitoring and Predicting Freeway Travel Time Reliability: Using Width and Skew of Day-to-Day Travel Time Distributions. *Transportation Research Record*, Vol. 1917, No. 1, 2005, pp. 54–62.

Virginia Department of Transportation. *Traffic Operations and Safety Analysis Manual—Version 2.0*. Richmond, VA, February 2020. https://www.vdot.virginia.gov/media/vdotvirginiagov/doing-business/technical-guidance-and-support/technical-guidance-documents/traffic-operations/traffic-operations-and-safety-analysis-manual-tosam.pdf.

Yang, J-S. *A Nonlinear State Space Approach to Arterial Travel Time Prediction*. MN/RC-2006-05. Minnesota Department of Transportation, St. Paul, MN, 2006.

Zhang, X., Zhao, M., Appiah, J., and Fontaine, M.D. *Methods to Analyze and Predict Interstate Travel Time Reliability*. VTRC 22-R2. Virginia Transportation Research Council, Charlottesville, VA, 2021.

Zhao, L., and Chien, S. Analysis of Weather Impact on Travel Speed and Travel Time Reliability. In *Proceedings of CICTP 2012: Multimodal Transportation Systems – Convenient, Safe, Cost-Effective, Efficient*. American Society of Civil Engenieers, Beijing, China, 2012, pp. 1145–1155.

Zhao, M., and Appiah, J. *Prediction of Interstate Travel Time Reliability: Phase II*. VTRC 23-R12. Virginia Transportation Research Council, Charlottesville, VA, 2023.