

**Application of Machine Learning in Spatiotemporal Analysis of Traffic
Flow Characteristics: Classification, Prediction, and Representation**

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Hairuolong Zhang
August 2024**

Copyright © 2024 by Hairuolong Zhang
All rights reserved.

DEDICATION

To my parents, Shaonan Zhang, Hui Long, for their endless love and support.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Lee D. Han, for his unwavering support and guidance throughout my graduate studies. It has been a privilege to work under his mentorship, benefiting from his keen insight and wise judgment in my research endeavors. His ability to ignite my passion for research and inspire me to think critically and creatively has been invaluable. Dr. Han's dedication to aesthetics and his devotion to the field have profoundly influenced my academic journey, shaping me into the scholar I am today.

I am also deeply grateful to my committee members, Dr. Russell Zaretzki, Dr. Candace Brakewood and Dr. Hairong Qi, for their invaluable support and guidance. Their collective expertise and encouragement have been instrumental in my academic success. I would also like to thank my friend and mentor, Dr. Hao Zhou and Dr. Fengqi Li, for their support and advice on my research and career development.

My heartfelt thanks go to all my teammates: Dr. Pankaj Dahal, Dr. Zhihua Zhang, Dr. Nima Hoseinzadeh, Dr. Yangsong Gu, Dr. Yuandong Liu, Ruqing Huang, Dr. Diyi Liu, Jinmu Lv, and Mohammad Khojastehpour. It has been a wonderful experience collaborating with such a talented and supportive group of individuals. Together, we have explored various projects and research ideas, sharing our experiences and life lessons along the way. Working with all of you has been a rewarding journey.

Lastly, I want to express my deepest love and appreciation to my family. My father, Shaonan Zhang, has always loved me unconditionally and encouraged me to pursue a Ph.D. degree abroad. My mother, Hui Long, has instilled in me the values of responsibility, honesty and love, qualities that have been essential not only in my academic pursuits but also in life.

ABSTRACT

Traffic data, regardless of its collection method—detector-based, probe-based, or crowdsourced approaches—fundamentally constitutes spatiotemporal information. This intrinsic characteristic underscores the complexity and dynamic variability of traffic flow across space and time. This dissertation explores the innovative application of machine learning techniques to the spatiotemporal analysis of traffic flow characteristics, offering a novel perspective on managing and understanding traffic dynamics. The dissertation consists of three interrelated chapters that collectively aim to enhance the accuracy and reliability of traffic data analysis, focusing on both truck volume estimation and short-term traffic volume forecasting under various conditions, alongside a study on representing network traffic condition.

The dissertation begins with a chapter introducing a novel approach that combines quantile regression and Light Gradient Boosting Machine to improve the accuracy of truck volume estimates from single loop detectors. This method significantly enhances the precision of truck volume estimations, providing a solid foundation for freight management and planning from a microscopic perspective. The second chapter proposes and reviews a new training and testing framework in predicting short-term traffic volume. This investigation highlights the strengths and limitations of existing research, proposing improvements that could lead to more reliable traffic forecasts. The final chapter presents a case study on the application of matrix decomposition to infer and represent traffic conditions across a traffic network by focusing on a group of key detectors, which contributes to better traffic management and operations.

Overall, this dissertation contributes to the field of data science and traffic engineering by leveraging machine learning for the spatiotemporal analysis of traffic flow, offering new frameworks for spatiotemporal analysis including truck estimation, volume prediction, and network condition representation. Its findings have significant implications for improving traffic management and operation with advanced data-driven methods, particularly in the face of growing urbanization and the increasing complexity of transportation networks.

TABLE OF CONTENTS

INTRODUCTION	1
CHAPTER I IMPROVING TRUCK VOLUME ESTIMATION FROM SINGLE LOOP DETECTORS: A METHOD INTEGRATING QUANTILE REGRESSION WITH LIGHT GRADIENT BOOSTING MACHINE	3
Abstract	4
Introduction.....	4
Related work	7
Methodology	9
Truck Volume Estimation Framework	11
Light Gradient Boosting Machine (LightGBM)	14
Quantile LightGBM (QLightGBM).....	15
Model Performance Measurements	16
Data	16
Training: Dual Loop Detector.....	16
Test: Single loop detector and associated WIM station.....	21
Results and Discussion	21
Conclusions.....	29
References.....	33
CHAPTER II SHORT-TERM TRAFFIC VOLUME PREDICTION ON FREEWAYS UNDER INCIDENT AND NON-INCIDENT CONDITIONS: A PERSPECTIVE FROM TRAINING AND TESTING PHILOSOPHY	35
Abstract	36
Introduction.....	36
Related work	38
Research gaps.....	42
Methodology	44
Gate Recurrent Unit (GRU).....	44
Support Vector Regression (SVR).....	48
Historical Average (HA).....	48
Random Walk (RW)	49
Performance Metrics	49
Data	50
Training Framework	51
Testing Framework	52
Results and Discussion	53
Traffic Volume Prediction Performance.....	53
Comparative Analysis of different training and testing scenarios	56
Conclusions.....	62
References.....	65

CHAPTER III REPRESENTING AND INFERRING NETWORK TRAFFIC CONDITION: A METHOD INTEGRATING CLUSTERING INTO CUR DECOMPOSITION	68
Abstract	69
Introduction.....	70
Related Work	73
Materials and Methods.....	75
Data and study area	75
CUR Decomposition.....	77
CUR Integrated with Clustering	82
Results.....	85
Weighted Average Method	85
Clustering Approach 1	87
Clustering Approach 2	91
Discussion	99
Conclusions.....	102
References.....	103
CONCLUSION.....	106
VITA	109

LIST OF TABLES

Table 1 Descriptive statistics of dual loop detected individual vehicle effective length..	19
Table 2 Model performance comparison at hourly resolution.....	24
Table 3 Model performance comparison at daily resolution.	28
Table 4 Pros and cons of different models.....	30
Table 5 List of Acronyms in alphabetical order.	39
Table 6 Training data conditions.	54
Table 7 Testing data conditions.	54
Table 8 MAPE (%) performance on different scenarios for one-step ahead prediction; red bold numbers denote the best model for each scenario.	57
Table 9 Average model performance for one-step prediction.	59
Table 10 MAPE (%) performance on different scenarios for six-step ahead prediction; red bold numbers denote the best model for each scenario.	61
Table 11 Average model performance for six-step prediction.	61
Table 12 List of Acronyms and Symbols.	74
Table 13 Comparison of the performance of different methods without clustering.....	86

LIST OF FIGURES

Figure 1 Truck volume estimation framework.	10
Figure 2 Location of training and testing detectors.	17
Figure 3 Distribution of AEVL when volume = 1.	19
Figure 4 Speed distribution of seven training detectors.	20
Figure 5 FHWA Scheme-F vehicle classification.	22
Figure 6 Traffic volume report difference (WIM - Loop) distribution by time of day. ...	23
Figure 7 Estimation results for Loop 3412044 on I80-W: (a) error change over time for middle lane; (b) error change over time for outer lane; (c) boxplot of error by time of day for middle lane; (d) boxplot of error by time of day for outer lane.	26
Figure 8 Estimation results for Loop 3412041 on I80-E: (a) error change over time for middle lane; (b) error change over time for outer lane; (c) boxplot of error by time of day for middle lane; (d) boxplot of error by time of day for outer lane.	27
Figure 9 Structure of RNN.	45
Figure 10 Structure of a) LSTM and b) GRU.	47
Figure 11 Profile of training data, the blue shaded area denotes the incident impact duration.	55
Figure 12 Profile of testing data as well as GRU performance for one-step ahead prediction.	55
Figure 13 (a) Location of Nashville, TN, on US map; (b) RDS detector locations in Nashville, TN.	76
Figure 14 Input data matrix with training period in red rectangle and testing period in the green rectangle; columns in blue rectangles represent the selected group of detectors and blue shaded areas are data entries for testing period at selected detectors.	78
Figure 15 CUR decomposition.	79
Figure 16 Column selection and decomposition process for clustering approach 1.	84
Figure 17 Column selection and decomposition process for clustering approach 2.	84
Figure 18 Performance of the proposed method without clustering.	86
Figure 19 Performance heatmap of the proposed weighted method when CR = 16.	88
Figure 20 Clustering results: (a) elbow method showing total within-cluster sum of squares versus number of clusters; (b) clustering results showing two principal components of observations; (c) cluster centroids versus time index.	89
Figure 21 Comparison of three methods with and without clustering: (a) L2 norm-based importance; (b) SVD-based leverage importance; (c) weighted average of L2 norm and leverage.	90
Figure 22 CUR Performance of speed for different column selection methods.	92
Figure 23 CUR Performance of v/c ratio for different column selection methods.	93
Figure 24 Clustering result for speed when number of clusters is equal to four.	95
Figure 25 Clustering result for v/c ratio when number of clusters is equal to four.	96
Figure 26 Location of the selected detectors for speed (blue dot) and v/c ratio (black circle).	98

INTRODUCTION

Analysis of traffic flow characteristics is a critical aspect of urban transportation management, essential for ensuring efficient mobility, planning infrastructure, and mitigating congestion. As cities expand and urban traffic dynamics become increasingly complex, the ability to accurately monitor and predict traffic flow patterns has become more challenging and vital. The advent of advanced technologies such as GPS-enabled vehicles, smartphones, sensors, and various traffic detectors has led to an explosion of traffic data, presenting both opportunities and challenges in traffic flow analysis.

Traffic flow characteristics refer to the various attributes and behaviors of traffic as it moves through transportation networks. These characteristics include flow rate, speed, and density, which are essential metrics for understanding and managing traffic dynamics. Flow rate measures the number of vehicles passing a point per unit of time, while speed refers to the distance covered by a vehicle per unit of time. Density indicates the number of vehicles occupying a given length of the roadway. These characteristics are interrelated and influence each other, making their analysis crucial for traffic management, congestion mitigation, and transportation planning.

To effectively manage traffic flow and optimize urban transportation systems, it is essential to analyze these characteristics in both spatial and temporal contexts. Spatial analysis focuses on understanding how traffic conditions vary across different locations, such as urban centers, highways, and suburban areas. Temporal analysis, on the other hand, examines how these conditions change over time, influenced by factors such as daily commuting patterns, seasonal variations, and unexpected events like accidents or construction. The integration of spatial and temporal dimensions, known as spatiotemporal analysis, is therefore vital for a comprehensive understanding of traffic patterns and behaviors, enabling more accurate predictions and better-informed decision-making.

Machine learning techniques have become increasingly important in spatiotemporal analysis of traffic flow due to their ability to handle large datasets and uncover hidden patterns and relationships. These techniques, including neural networks,

support vector machines, and ensemble learning methods, can learn from historical data to predict future traffic conditions with high accuracy. By considering both spatial and temporal factors, machine learning models can provide more reliable predictions and insights, which are crucial for optimizing traffic flow and reducing congestion.

The fundamental aim of this research is to develop and validate novel machine learning approaches or frameworks to improve the accuracy and reliability of spatiotemporal analysis of traffic flow characteristics. Traditional traffic analysis methods often struggle with the complexity and variability inherent in traffic flow, especially under atypical conditions such as accidents, road closures, and special events. By leveraging machine learning, this dissertation seeks to address these limitations and enhance predictive capabilities.

The dissertation is organized into three interrelated chapters, each targeting a specific aspect of traffic flow analysis.

- The first chapter presents a framework that combines Quantile Regression with Light Gradient Boosting Machine (LightGBM) to improve truck volume estimation using single-loop detectors, addressing the limitations of traditional methods and enhancing accuracy for better traffic operations and freight management.
- The second chapter proposes a novel training and testing framework for traffic flow prediction models that incorporates incident data and identifies common evaluation pitfalls. The results demonstrate that including incident data significantly improves the accuracy and robustness of predictive models across various traffic conditions.
- The third chapter proposes a novel approach for representing and inferring network-wide traffic conditions using CUR decomposition integrated with clustering method. The results demonstrate that this framework provides a scalable and efficient solution for monitoring and analyzing traffic conditions across complex urban networks.

CHAPTER I
IMPROVING TRUCK VOLUME ESTIMATION FROM SINGLE
LOOP DETECTORS: A METHOD INTEGRATING QUANTILE
REGRESSION WITH LIGHT GRADIENT BOOSTING MACHINE

Abstract

Monitoring Long Truck (LT) volume plays an important role in traffic operation, transportation planning, freight transport, and roadway pavement design. This paper presents a novel framework based on Light Gradient Boosting Machine (LightGBM) to estimate LT volume using single-loop measurements. The motivation is to improve the accuracy of estimation for long trucks at enhanced resolution, as traditional methods have proven to be unreliable or unavailable under certain traffic conditions. A crucial step in this process is speed estimation, which the proposed framework handles using quantile regression to account for uncertainty. This approach works under both high temporal (30 seconds) and spatial (lane-by-lane) resolution. Data from seven selected dual loop detectors from July 11 to July 17, 2022, in California Freeway Performance Measures System (PEMS) were used as ground truth speed to train the quantile models. The temporal and spatial transferability of the proposed framework was validated by applying the algorithm to another two loop detectors from Nov 25 to Dec 14, 2022. The proposed approach generates close-to-ground truth estimates of LT traffic volume with a Mean Absolute Percentage Error (MAPE) of 5.5% and increased the accuracy by more than 40% compared to prevailing methods. The contributions of this paper include introducing quantile regression with LightGBM for handling uncertainty, offering high-resolution and improving LT volume estimation accuracy. The framework's results outperform previous models in terms of accuracy, generality, and computational cost, making it a valuable tool for transportation practitioners and policymakers to achieve reliable LT volume estimation.

Introduction

Traffic speed and vehicle classification data are crucial for a variety of transportation-related purposes, including traffic operations, transportation planning, and pavement maintenance (1; 2). Vehicles with longer lengths, such as buses, long trucks, and recreational vehicles, have slower acceleration, inferior braking, and larger turning radii (1). These types of vehicles have a greater impact on pavement wear and tear, especially

when fully loaded, and are associated with higher rates of fatal crashes compared to other vehicle types (1). Inductive loop detectors, particularly single-loop detectors, are the most commonly used traffic detectors on highways (2). However, single-loop detectors are unable to directly measure both traffic speed and vehicle classification (3). While dual-loop detectors can measure classified vehicle volumes and speeds, they are not widely available in existing roadway infrastructure due to high installation costs (2). In recent years, new types of detectors, such as the Remote Traffic Microwave Sensor (RTMS) and Radar Detection Sensor (RDS), are gradually deployed on prioritized roadways to enhance traffic monitoring, which can capture vehicle classification information. Nonetheless, they are less commonly used and not as extensively deployed as conventional single-loop detectors in existing roadway infrastructure systems. Since upgrading single-loop detectors to dual-loop or other advanced detectors can be cost-prohibitive, it is important to enhance the accuracy of single-loop traffic speed estimation and vehicle classification to meet the demands of active traffic management and advanced traveler information systems. Despite numerous algorithms developed for this purpose and documented in the literature, speed and long truck (LT) volume estimation techniques, including the g-factor method currently implemented in the PEMS system, the speed correlation-based algorithm proposed by Kwon (4), and the optimization-based algorithm recently published by Gu et. al (5), has not proven to be reliable in practice or not available under certain traffic conditions like traffic congestion (5). Hence, addressing this issue and improving truck traffic estimation is the motivation behind this study. The objective of this study is to develop a truck volume estimation framework that is easy to implement and applicable to various traffic conditions.

In line with prior related research (1; 4-6), this study defines "short vehicle" (SV) as any vehicle type other than long trucks (LT). Long Trucks (LT) are specifically defined according to the Federal Highway Administration (FHWA) Scheme F classification system as vehicles belonging to class 8 and larger. This classification includes a wide range of heavy vehicles, such as single-unit trucks with three or more axles, multi-trailer combinations, and other large commercial vehicles. These LT vehicles are typically characterized by their significant size and weight, distinguishing them from

smaller, lighter short vehicles (SV) which include passenger cars, motorcycles, and light-duty trucks.

While the confidence of data quality of loop detectors may differ, a reliable loop detector must provide accurate data for two traffic fundamentals: volume and occupancy per lane and per time interval. Building upon this premise, our paper introduces a novel and straightforward framework that utilizes common vehicle detector measurements (volume and occupancy) to specifically identify LT traffic counts on a lane-by-lane basis for each time interval.

Data from dual loops, single-loop, and weigh-in-motion (WIM) detectors were used in this study. Through the analysis of cases where each detector interval contains only one vehicle, which is the key factor that affects speed estimation and vehicle classification. Empirical data shows that vehicle length on highways follows a bimodal distribution, with the majority of vehicles being passenger cars that are typically around 15 feet long, while most trucks are trailer trucks about 65 to 80 feet long (5). This actual length distribution is further leveraged to improve the accuracy of effective vehicle length estimates for traffic speed estimation. However, the key challenge is to determine the most possible fleet combinations within a sampling period (i.e., 30 seconds). Feeding ground truth speed data collected from dual-loop detectors into a quantile regression model-based Light Gradient Boosting Machine (LightGBM) framework, we quantify the uncertainty of speed estimation given volume and occupancy, thus helping determine the most possible fleet combinations.

The remainder of this paper is organized as follows. The next section presents a literature review of related work. The methodology section presents the truck volume estimation framework as well as different methods used in the paper. Next, the data and research area are introduced. Then the results section presents the estimation results on two test sites and the comparison with two previously published models. The work is wrapped up in the final section, followed by a discussion of the advantages, possible future endeavors, and limitations of the work.

Related work

In 1965, Athol proposed the first method for estimating traffic speed using data collected from a single inductive loop detector (7). The term called “g-factor” was introduced which converts the units to their proper values and is related to mean vehicle length plus detector size. With the g-factor, the speed can be formulated as:

$$v = \frac{N}{g \times T \times \phi} \quad (1)$$

In this context, v denotes the space mean speed, while T denotes the duration of the observation period (commonly 20 or 30 seconds). ϕ is the percentage of time that vehicles occupy the loop sensor during each interval, and N represents the number of vehicles observed during each interval. The reciprocal of the g-factor is exactly the average effective vehicle length (AEVL) which is equal to the average vehicle length plus the width of the detection zone at a particular location (Equation 2).

$$\bar{l} = \frac{v \times T \times \phi}{N} \quad (2)$$

Based on the relationship in equation 2, an accurate speed estimate is crucial in obtaining the AEVL, which could be further used in determining the LT volume via the combination of SV and LT of the fleet during the 30-second interval. Then, truck volume estimation from single loop detectors is formulated as a problem of estimating speed and AEVL. Following this rationale, many previous studies have been conducted to enhance the accuracy of speed and AEVL estimation, which are reviewed in the following part.

In the past, numerous studies have been conducted to estimate classified vehicle volumes and speeds based on measurements from single-loop detectors. Sun and Ritchie studied individual vehicle speed estimation using single-loop inductive waveforms and proposed an algorithm using signal processing and statistical methods to extract speeds from inductive waveforms (8). A study by Wang et.al. obtained freeway traffic speed estimation with single-loop outputs and the speed estimations with corrected g-values are very close to the speeds observed by the speed trap (9). Lin et. al. investigated the discrepancy between the speed estimated with single loops and the speed estimated with double loops (10). Lu et. al. proposed a new approach using event data and mode

occupancy within a moving window of variable length to estimate the effective vehicle length, or g-factor (11). A Gaussian mixture model (GMM)-based approach is developed to estimate more accurate traffic speeds and classify vehicle volumes using single-loop outputs (12). Li et. al. developed a new speed estimation model which uses detailed vehicle signature data from single inductive loop sensors equipped with advanced detectors to provide accurate truck speed estimates (13).

Wang and Nihan developed an algorithm using single-loop measurements to estimate vehicle speed and classification (14). They divided vehicles into two categories based on length, with a threshold of 11.89 meters. Their method had three main steps: first, identifying the intervals during which long vehicles were present (i.e., the amount of time a long vehicle spent over a loop in 20 seconds); second, using the time intervals of short vehicles and assuming a constant vehicle length to estimate speed; and third, using the estimated speed to identify the volumes associated with the time intervals of long vehicles. Nevertheless, the fundamental assumption that there are at least two intervals containing only short vehicles is easily violated in real-world scenarios.

Another noticeable research was proposed by Kwon et al. (4), who proposed a simpler method based on the speed correlation between lanes instead of using a complex relationship described by the traffic fundamental stream model to estimate speed. This approach is currently adopted by the PEMS system. However, this approach is only suitable for freeway locations with truck-free lanes, which may not always hold true in practice. With the estimate AEVL, the speed of the truck-free lane can be determined using Equation 2. Then, by applying correlation coefficients observed from real-world data, the speeds of other lanes are indirectly estimated. However, this approach is influenced by the coefficients between lanes, and the accuracy of speed estimation heavily relies on the selection of these coefficients.

Besides the abovementioned algebraic methods, some recent studies applied supervised learning methods such as Artificial Neural Networks (ANN) (1) to build the relationship between volume, occupancy, time of day, and volume of different vehicle classes (SVs and LTs). As mentioned in their study, this method is capable of handling heavily congested traffic by retraining the ANN model with data collected during

congestion conditions (1). Due to this training requirement, the approach heavily depends on the use of dual-loop detectors and the quality of the training data, yet its performance in other locations remains uncertain due to potential spatial variations in traffic flow.

In 2022, Gu et al. proposed an innovative method (5), which formulates an optimization problem to minimize the difference between observed estimated occupancy where the number of SVs and LTs are decision variables. This chapter is an extension and comparison study inspired by Gu's study.

To summarize, previous studies have shown that the effective length of vehicles is crucial for accurately estimating speed and truck volume from single-loop detector data, assuming that occupancy and vehicle count measurements are accurate. Moreover, truck volume and speed estimation are two closely related problems. Three major research gaps were identified as follows. First, most truck estimation methods rely heavily on an accurate estimate of traffic speed, which is hard to infer from only volume and occupancy due to the complexity and nonlinearity of the traffic stream. Second, many existing approaches are unable to handle diverse traffic conditions effectively, especially during congestion. Finally, previous work only evaluated their model for limited locations and within only a short period. In some instances where satisfactory performance is achieved throughout the day, their applicability is restricted to specific times or locations (1; 4; 5; 15).

Methodology

This section introduces the truck estimation framework and all relevant technical methods used in the framework. **Figure 1** presents the proposed methodological framework of this paper.

First, all the pertinent data should be collected, archived, and preprocessed for further analysis. In step 0, data preprocessing and filtering were implemented to obtain high-quality training data. Then, in step 1, the single-vehicle case was analyzed to obtain the actual vehicle length distribution and percentage of truck occurrence when volume equals one. Step 2 combined quantile regression with the LightGBM model to estimate the speed range given volume, occupancy, and lane number.

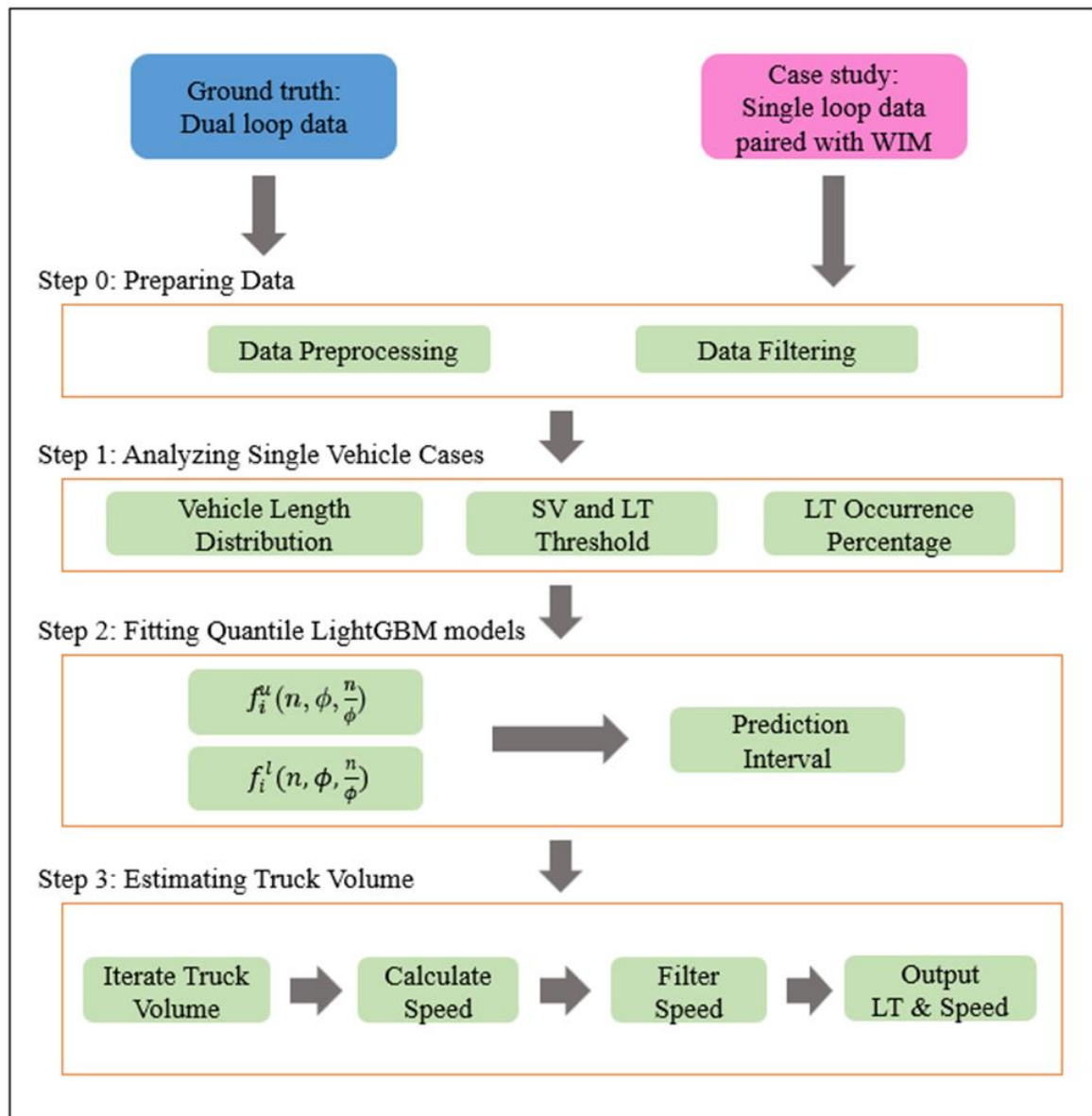


Figure 1 Truck volume estimation framework.

Finally, step 3 estimated the truck volume at the test sites to validate the performance of the proposed framework. The following section is a detailed description of all the steps and different methods used in this paper.

Truck Volume Estimation Framework

Step 0: Preparing data.

First, all the necessary data need to be archived and preprocessed, including:

- All dual loop detector data (volume, occupancy, speed) in the PeMS system from July 11, 2022, to July 17, 2022,
- WIM observed truck volume and associated single loop detector data (volume, occupancy) at two testing sites on I-80 from Nov 15, 2022, to Dec 31, 2022.

A series of preprocessing and data quality test procedures are then implemented to obtain high-quality training data:

- converted raw data into the proper format,
- removed duplicate records,
- matched WIM data with associated loop detector data,
- used detectors with 3 lanes for both training and test sites,
- checked the missing rate for each detector,
- filtered out dual loops with missing rates larger than 1%,
- calculated average effective vehicle length (AEVL) for all observations with the volume being one, and plot the distribution of AEVL for each detector, and
- selected detectors with obvious bimodal distribution for AEVL.

Step 1: Analyzing single vehicle case.

The objective of step 1 is to obtain the average effective vehicle length (AEVL) by investigating the observations only when the volume is equal to one and calculating EVL using the equation $EVL_i = v_i \times T \times \phi_i \times \frac{5280}{3600}$, where v_i is in mph, T is 30 seconds, ϕ_i is a percentage, and then EVL_i is expressed in feet for each sample i .

This individual vehicle inspection trick needs to be highlighted because it helps understand the derived actual effective vehicle length for the selected five dual loops, which to some extent represents the actual vehicle length of the fleet in the research area.

Then a K-means clustering model is applied to the derived EVL data to verify the

assumption that the fleet can be classified into SVs and LTs using a cutoff point l_{cutoff} determined by the clustering results. The center for the two clusters is the average effective vehicle length for a typical short vehicle l_{SV} and long truck l_{LT} , which will be used in the truck estimation model. The formula is written as:

$$\begin{aligned} l_{SV} &= \frac{\sum_{i=1}^M EVL_i}{M}, & \forall EVL_i \leq l_{cutoff} \\ l_{LT} &= \frac{\sum_{i=1}^N EVL_i}{N}, & \forall EVL_i > l_{cutoff} \end{aligned} \quad (3)$$

where M and N are the number of samples in SV and LT groups respectively.

With the inspection of individual vehicle scenarios, the truck occurrence likelihood p_{LT} when volume equals one was given by:

$$p_{LT}(LT|volume = 1) = \frac{\sum_{i=1}^N I_i}{N} \quad (4)$$

where N is the number of samples when the volume is equal to one, and I_i is the indicator:

$$I_i = \begin{cases} 1, & EVL_i > l_{cutoff} \\ 0, & otherwise \end{cases} \quad (5)$$

p_{LT} represented the percentage of truck occurrence when there was only one vehicle during the 30 seconds.

The output of this step is the average effective length for the short vehicle l_{SV} and for long truck l_{LT} , and the truck occurrence percentage under the single-vehicle condition p_{LT} .

Step 2: Fitting quantile LightGBM models.

The objective of the second step is to establish the relationship between the response variable speed, and the explanatory variable volume and occupancy. Instead of a point estimation of the mean of the speed for a given volume and occupancy, we adopted the quantile LightGBM model to generate prediction intervals. The fundamental relationship in traffic flow theory is given by:

$$q = k \cdot u \quad (6)$$

where q is the flow rate in vehicles per hour, k is the density in vehicles per mile, and u is the space mean speed in miles per hour. Inspired by this principle, we used high-

quality ground truth speed, volume, and occupancy measurements from the selected dual loops to fit one lower bound $f_i^l(n, \phi, \frac{n}{\phi})$ and one upper bound $f_i^u(n, \phi, \frac{n}{\phi})$ quantile model based on LightGBM framework for each lane respectively, where:

- l represents 1st percentile model,
- u represents 99th percentile model,
- $i = 1, 2, 3$ denotes the lane number from innermost to outmost,
- n is the volume within 30 seconds,
- ϕ is the occupancy within 30 seconds,
- $\frac{n}{\phi}$ is a surrogate for speed and it is included in the model for better performance.

The reason we considered fitting the model for each lane is that traffic behaves differently from lane to lane. Even with the same volume and occupancy, the innermost lane usually has a larger speed than the outer lanes. For consistency issues, it is necessary to keep the number of lanes the same, herein we considered 3 lanes for both training and test sites. The output of this step is three lower bound and three upper bound quantile models for lanes 1 (innermost lane), 2 (middle lane), and 3 (outmost lane).

Step 3: Estimating truck volume.

For each timestamp t at lane i , given total volume n and occupancy ϕ :

1. Iterate truck volume N_{LT} from 0 to n , calculate the associated speed $\hat{v}$ as:

$$\hat{v}_k = \frac{k \cdot l_{LT} + (n - k) \cdot l_{SV}}{T \cdot \phi} \times \frac{3600}{5280} \quad (7)$$

where $T = 30$ seconds, l_{LT} and l_{SV} are given by equation in step 2, $k = 0, 1, 2, \dots, n$.

This generated $(n + 1)$ solutions in the solution space.

2. For each $k = 0, 1, 2, \dots, n$, a corresponding probability was computed based on the assumption as follows:

Given volume n , assume that number of trucks per 30 seconds n_{LT} follows a binomial distribution $N_{LT} \sim B(n, p_{LT})$, so that the probability of truck volume being k is given by:

$$p_k = p(N_{LT} = k) = \binom{n}{k} p_{LT}^k \cdot (1 - p_{LT})^{n-k} \quad (8)$$

3. Generate 98% prediction interval (1st – 99th percentile) given i, n and ϕ :

$$v_l = \hat{f}_i^l \left(n, \phi, \frac{n}{\phi} \right) \quad (9)$$

$$v_u = \hat{f}_i^u \left(n, \phi, \frac{n}{\phi} \right) \quad (10)$$

$$PI_i \left(n, \phi, \frac{n}{\phi} \right) = [v_l, v_u] \quad (11)$$

4. Filter out $\hat{v}_k$ which falls outside the prediction interval PI_i , and denote the remaining solution space as $\mathcal{S}$.
5. If the number of remaining solutions is one, then the estimated truck volume $n_{LT} = k$ and the corresponding speed estimation is $\hat{v}_k$;
6. If the number of remaining solutions is larger than one, we propose two methods to determine the best solution. First, by using the large truck occurrence likelihood, the estimated truck volume is calculated as follows:

$$n_{LT} = \frac{\sum_{k \in \mathcal{S}} k \cdot p_k}{\sum_{k \in \mathcal{S}} p_k} \quad (12)$$

The second method is to take the solution which achieves minimum truck volume among the solution space:

$$n_{LT} = \min_{k \in \mathcal{S}} k \quad (13)$$

This method assumes that large trucks are less likely to occur than passenger cars. We adopt the second method in this paper.

For comparison, three well-established models, the integer programming model (denoted as Optimization) proposed by Gu et al. (16), the lane speed correlation model (denoted as LnCorrelation) developed by Kwon (8), the interval relationship model (denoted as Interval) published by Wang (9), are implemented on the same testing detector over the same period.

Light Gradient Boosting Machine (LightGBM)

LightGBM, announced by Microsoft Research Asia in 2017 (16), is an open-source and efficient decision tree-based framework for machine learning tasks such as sorting, classification, and regression. Compared to XGBoost (17), LightGBM offers

increased power with faster training speed and lower memory usage, resulting in reduced communication costs during parallel learning. Key features include the gradient-based one-side sampling (GOSS) decision tree algorithm, exclusive feature bundling (EFB), and a histogram and leaf-wise growth strategy with a depth limit. GOSS ensures a balanced relationship between sample quantity and decision tree accuracy. In the training process, down sampling focuses on samples with larger gradients for improved information gain, and EFB reduces feature dimensionality by combining mutually exclusive features. Overall, LightGBM is an improved variant of the gradient boosting decision tree (GBDT) model enhanced by adding GOSS and EFB to reduce the time complexity of processing data samples, but the underlying principle is still the gradient boosting tree algorithm which sequentially fits weak models (almost always simple trees) to the residuals from the previous iteration, taking the final model to be the sum of the individual models from each iteration.

Quantile LightGBM (QLightGBM)

Koenker and Bassett Jr. (18) introduced quantile regression as a method for estimating and conducting inference about conditional quantile functions. Unlike traditional least squares regression, quantile regression minimizes the mean absolute error, which assigns asymmetric weights to positive and negative errors, enabling the estimation of the entire conditional distribution (18). In essence, quantile regression replaces the mean squared error loss function with a defined loss function that predicts conditional quantiles instead of conditional means.

The quantile LightGBM combines quantile regression with LightGBM to generate prediction intervals (PIs) by fitting one lower bound $\hat{y}^l$ and the other upper bound $\hat{y}^u$ quantile models. Then the prediction interval of sample i is given by:

$$PI_i = [\hat{y}_i^l, \hat{y}_i^u] \quad (14)$$

The lower bound and upper bound were obtained by the estimates at τ^{th} percentile and $(100 - \tau)^{th}$ percentile. Hence, $(100 - 2\tau)\%$ of the time the true speed will fall into the output prediction intervals, which provides the uncertainty quantification

for the response variable given the specific input explanatory variables.

Model Performance Measurements

Two commonly used performance metrics, MAE and MAPE, are adopted to assess the model performance in estimating truck traffic.

- Mean Absolute Error (*MAE*): *MAE* measure shows the deviation of the estimated value from the actual observed value. A smaller *MAE* indicates a better model estimation. The formula can be written as:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (15)$$

- Mean Absolute Percentage Error (*MAPE*): This indicator captures the relative error between observed and estimated values as follows.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (16)$$

y_i , and $\hat{y}_i$ denote the actual observed and estimated LT hourly volume respectively, and n denotes the number of observations.

Data

Training: Dual Loop Detector

Volume, occupancy, and speed data collected at seven random selected dual loop detectors in the Interstate-80 Sacramento area from July 11, 2022, to July 17, 2022, were selected to be fed into quantile regression models to measure the uncertainty of traffic speed. **Figure 2** shows the location of the seven detectors used for training models (red, detectors at the same location in opposite directions might overlap on maps) and two detectors for testing model performance (green) on the map.

Data quality is important in determining the model's performance, so it is of utmost importance to assure good data quality before feeding data into models. The selection of the seven training detectors involved several quality checks mentioned in the Methodology section. Firstly, the entire dataset went through a missing rate calculation for each detector on I-80 and those with a missing rate of less than 1% (missing observations

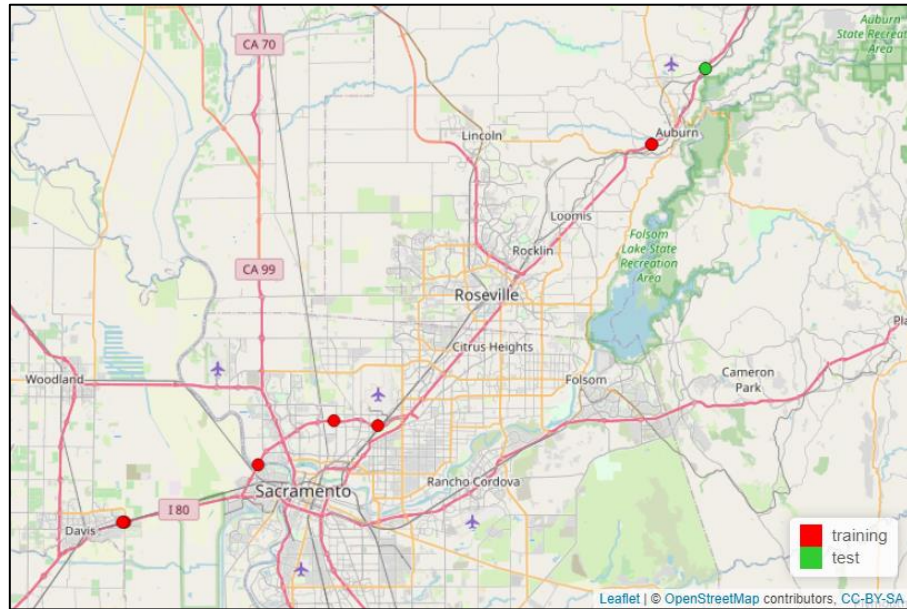


Figure 2 Location of training and testing detectors.

is less than $2/\frac{\text{minute}}{\text{lane}} \times 1440 \frac{\text{minutes}}{\text{day}} \times 7 \text{ days/lane} \times 3 \text{ lanes/detector} \times 1\% \approx 605/\text{detector}$) were passed into next check. To enhance training efficiency, detectors with a bimodal distribution for AEVL were selected when the volume equals one. This led to the inclusion of six detectors in the final training data set, of which three are westbound and four are eastbound.

The next step is to determine the typical vehicle length for SV and LT by checking the cases when volume equals one. Since the calculation results might be sensitive to outliers, we first filtered out observations below the 0.5th percentile (lower bound = 13 ft) and above the 99.5th percentile (upper bound = 90 ft). **Figure 3** illustrates the distribution of AEVL by looking at individual vehicle cases for the seven selected detectors over the one-week period. This individual vehicle effective length inspection along with a simple K-means clustering helps validate the assumption that the fleet could be divided into two clusters short vehicle (SV) and long truck (LT) with a typical length $l_{SV} = 23 \text{ ft}$ and $l_{LT} = 73 \text{ ft}$, respectively. In addition, the cutoff point to distinguish SV from LT is 48 feet. It should be noted that these numbers are “effective” length because it includes the longitudinal span of the detection zone. In this study, the length of the detection zone is the length of a typical loop, which is 8 feet. Subtracting the detector length from l_{SV} and l_{LT} , the actual vehicle length for SV and LT groups is 15 feet and 65 feet, which accords with the common knowledge and previous studies.

Table 1 shows the descriptive statistics of SV and LT groups measured by the selected seven dual loops setting volume equal to one. The average length of LTs is more than three times longer than that of SVs. Therefore, the average vehicle length for any 30-second interval is noticeably increased when long trucks (LTs) are present.

Figure 4 shows how traffic speed is distributed at each detector for each lane. Loop-314025 obviously experienced more congested states than the other six detectors did. To maintain the balance and generality of the training data, it is necessary to include detectors from both congested and uncongested conditions and measure the uncertainty of speed with respect to volume and occupancy during all kinds of traffic conditions.

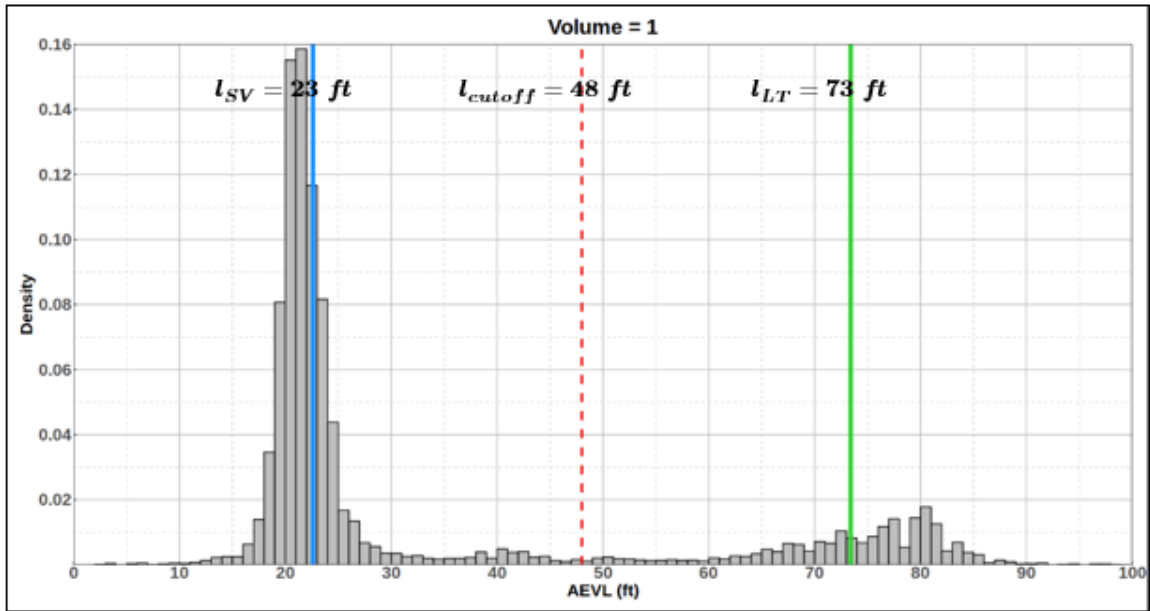


Figure 3 Distribution of AEVL when volume = 1.

Table 1 Descriptive statistics of dual loop detected individual vehicle effective length.

Class	Number of observations	Mean	Standard deviation	Minimum	Maximum
SV ($\bar{l} < 48 \text{ ft}$)	20173	22.58 ft	4.64 ft	13.02 ft	47.87 ft
LT ($\bar{l} \geq 48 \text{ ft}$)	5206	73.46 ft	8.65 ft	48.05 ft	89.76 ft

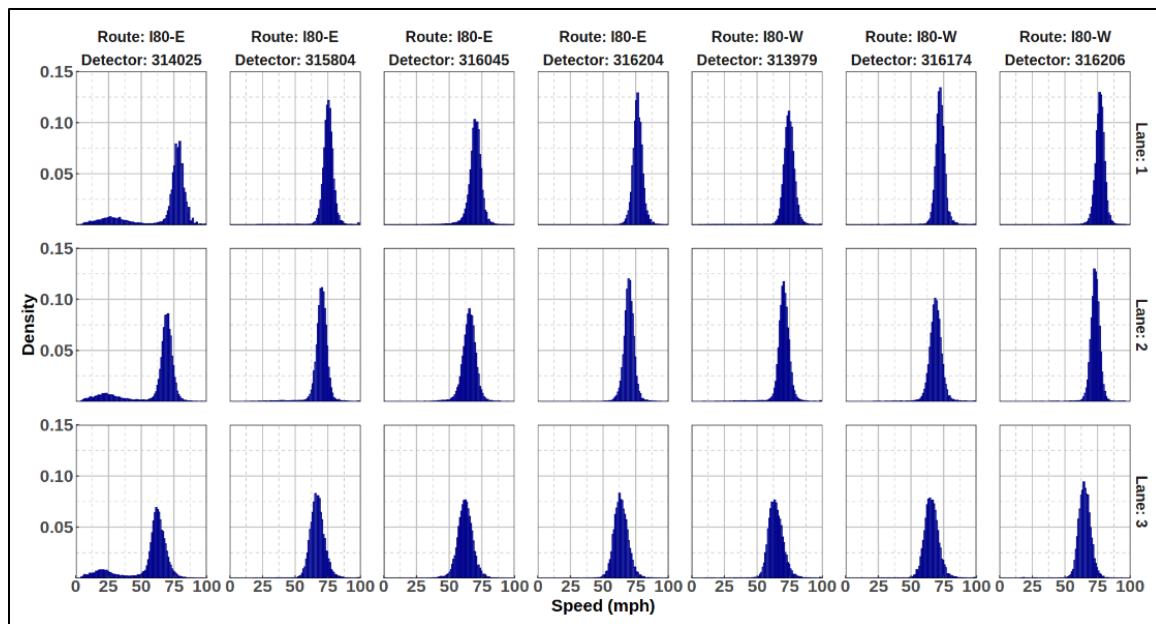


Figure 4 Speed distribution of seven training detectors.

In general, traffic speed increases from the outermost lane (lane 3) to the innermost lane (lane 1)

Test: Single loop detector and associated WIM station

To test the spatial transferability of the proposed framework, we selected another two different single loops paired with a corresponding WIM station which reported the ground truth FHWA scheme-F vehicle classification (19) (see **Figure 5**) results at hourly resolution. To categorize the vehicle into SV and LT as defined in our methodology, a vehicle class larger than and equal to 8 is regarded as LT and the remaining is SV since the length of class 8 begins to surpass 40 ft (20). For fair comparison purposes, we used the same detectors that Gu's work (5) used which is WIM-34630 on I-80. Besides, the author also implemented the well-known method proposed by Kwon (4) using lane-to-lane speed, and the interval method developed by Wang et al. (14). WIM-34630 covers both eastbound (3412041) and westbound (3412044) directions to monitor the passing by vehicles and their scheme F classifications.

With WIM and paired loop detector data, we validated that the traffic volume reported by WIM and loop detector can match each other. As shown in **Figure 6**, the WIM station and its paired loop detector both record traffic volume passing by, however, due to infrastructure and technology issues, they do not always match with each other. WIM-34630 tends to overcount volumes than loop does from 4 to 10 am while undercounting from 4 to 10 pm for all the lanes. The difference between the two observed values might also cause estimation errors because the proposed algorithm is taking loop measurements as input to estimate the truck volume which eventually needs to be compared with WIM observed truck volume.

Results and Discussion

The performance of the proposed algorithm for the two testing detectors over a 20-days period from November 25, 2022, to December 14, 2022, is presented in

Table 2. Gu's, Kwon's and Wang's model was selected as competitors.

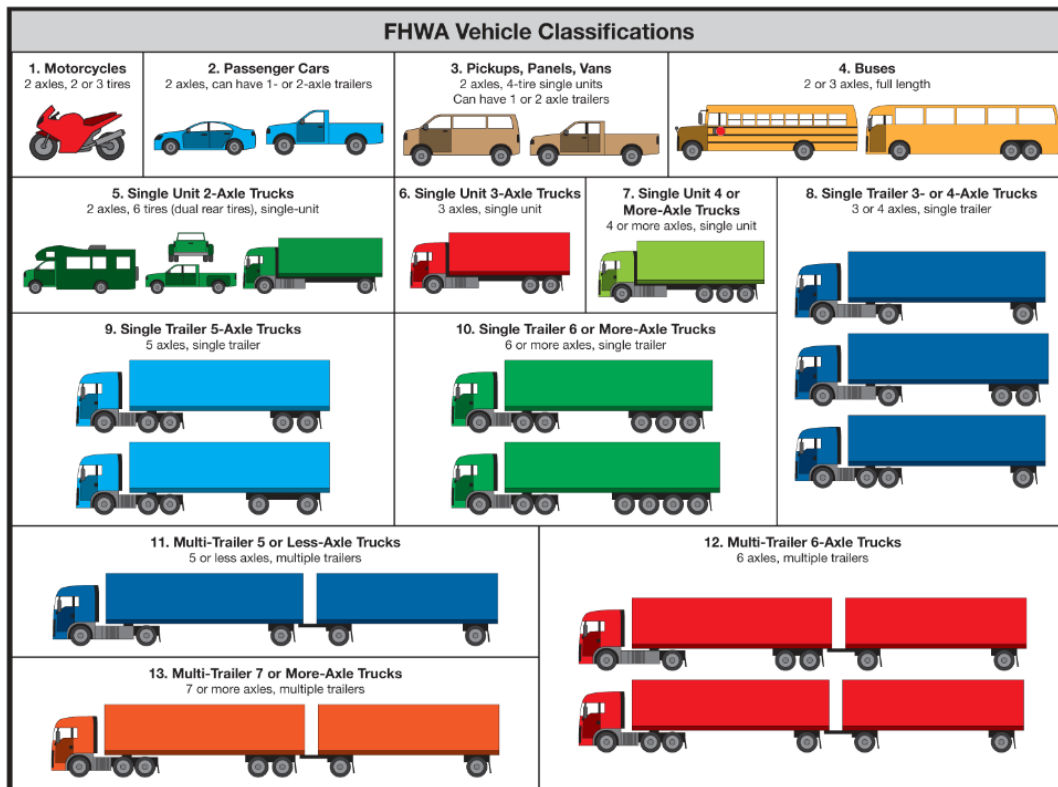


Figure 5 FHWA Scheme-F vehicle classification.

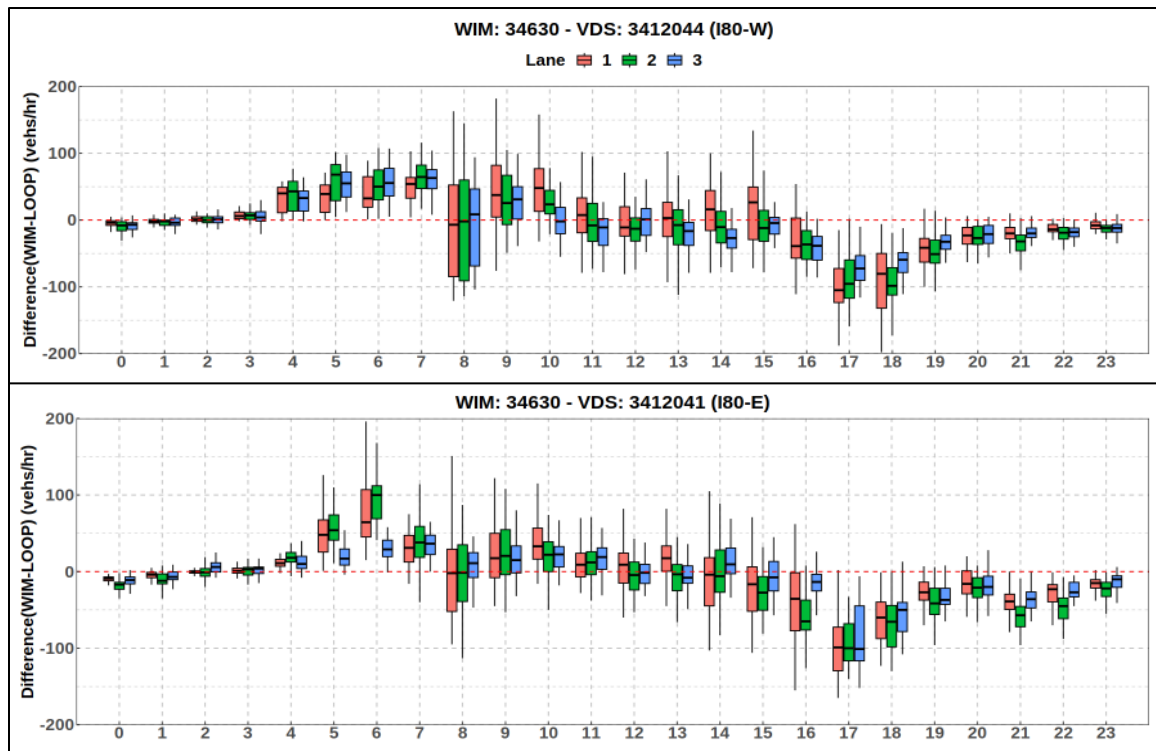


Figure 6 Traffic volume report difference (WIM - Loop) distribution by time of day.

Table 2 Model performance comparison at hourly resolution.

		Model	MAE (trucks/hour)				MAPE (%) (Truck > 5)			
			Inner	Middle	Outer	All	Inner	Middle	Outer	All
I-80E	3412041	QLightGBM	0.1(0.3)	4.2(5.9)²	8.1(10.9)	10.2(9.3)	-	25(24)	15(15)	18(27)
		Optimization	0.2(0.4)	4.9(6.7)	9.4(11.4)	11.5(9.7)	-	28(26)	20(21)	24(40)
		LnCorrelation	0.1(0.4)	7.8(7.3)	16.0(16.2)	23.6(18.8)	-	50(42)	29(24)	39(46)
		Interval	0.1(0.4)	13.5(11.6)	43.3(30.0)	56.7(39.4)	-	70(20)	66(20)	66(19)
I-80W	3412044	QLightGBM	0.2(0.6)	5.6(10.3)	10.4(15.1)	14.7(24.5)	-	28(25)	21(28)	24(44)
		Optimization	0.4(1.4)	6.1(10.7)	11.0(16.3)	16.1(23.5)	-	29(24)	21(26)	26(48)
		LnCorrelation	0.2(0.6)	7.5(10.3)	16.2(18.0)	23.3(24.9)	-	41(38)	29(33)	34(44)
		Interval	0.3(0.8)	15.5(18.6)	48.0(39.2)	63.5(55.6)	-	71(20)	71(16)	69(17)

Note:

- Number in brackets is the standard deviation, and bold font denotes the best performance among the four models.
- Truck volume is less than 5/hour on inner lane, so MAPE is not calculated

Lane 1 is the innermost lane with truck-free enforcement. The QLightGBM model and Optimization model by Gu do not presume a truck-free lane, instead, both models will detect truck volume automatically even though lane one can sometimes see several trucks per hour. In this study, we will mainly focus on the performance of the middle lane (lane 2) and outermost lane (lane 3), and MAPE performance will only be calculated when observed values are not missing and larger than zero due to the definition.

The proposed algorithm outperforms the other three models in terms of MAE for both stations. In addition, the overall MAE of the proposed method is about 12 trucks per hour on average and MAPE at hourly resolution is 21% on average.

Another advantage of the QLightGBM and Optimization method is that they can both automatically detect the very few truck cases on the inside lane instead of manually assuming the existence of truck-free lane, but the proposed method works better in light of a smaller MAE, MAPE and standard deviation.

Besides performance metrics, a graph showing the estimation error change over time and how error distributes by time of day is also of interest. **Figure 7** and **Figure 8** show the estimation results for Loop-3412044 (Westbound) and Loop-3412041 (Eastbound). The chart only displayed results from Nov 25 to Dec 14, 2022, when there were no missing observations from either Loop or WIM sources. The proposed model performs very well in the middle lane for both directions with an MAE of less than 5 trucks per hour, while generating bigger errors at the outer lane. This may be because of the high percentage of trucks resulting in congested traffic conditions.

In general, the proposed framework can estimate the truck volume with an MAE of less than 5 trucks per hour for the middle lane, and a MAE of less than 10 trucks per hour for the outer lane. Additionally, a day-by-day comparison of truck volume estimation is essential for understanding the temporal patterns of truck traffic and its implications on various aspects of transportation planning, management, and safety. Hence,

Table 3 is furnished to present the aggregated estimation error at the daily level.

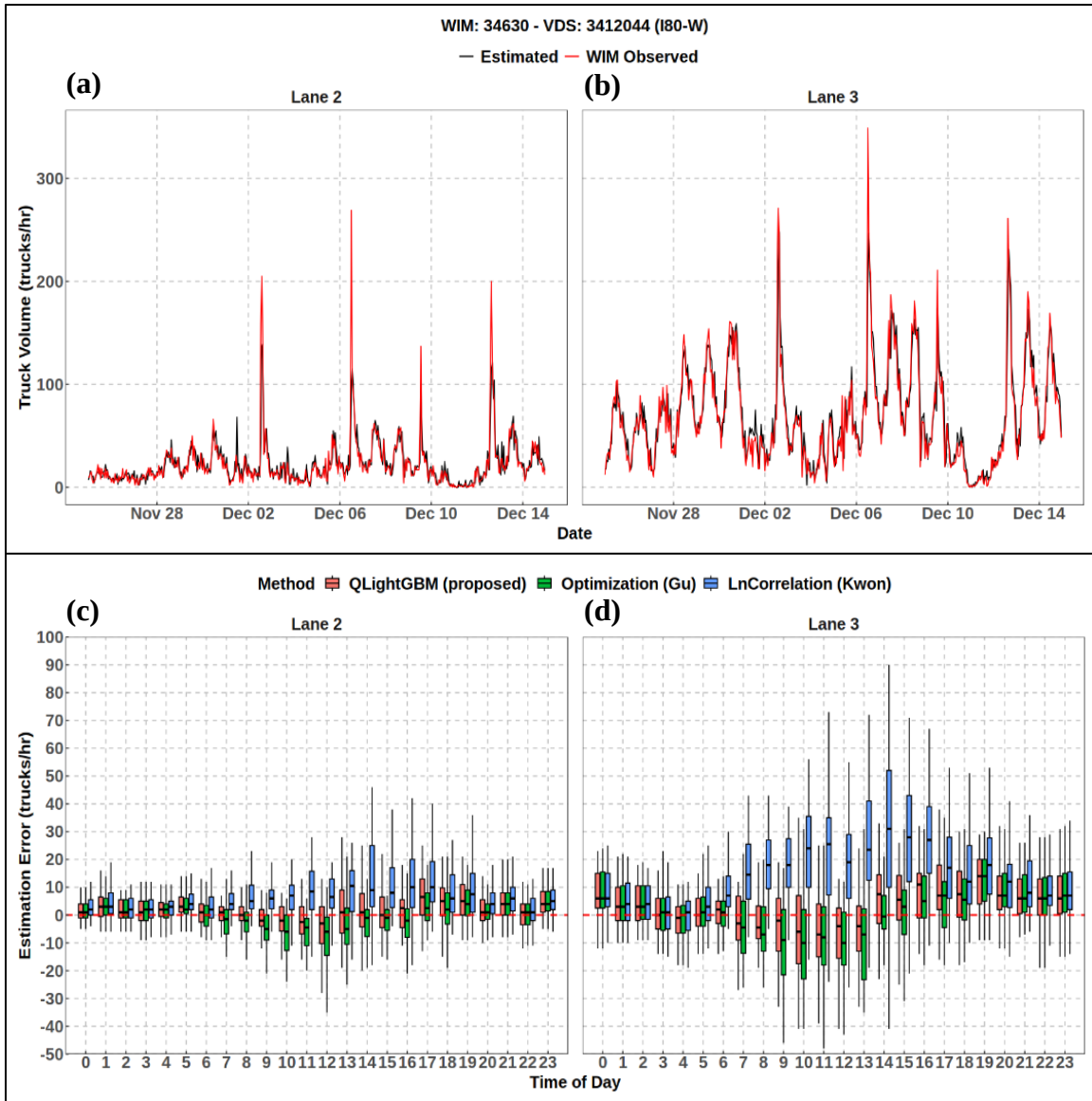


Figure 7 Estimation results for Loop 3412044 on I80-W: (a) error change over time for middle lane; (b) error change over time for outer lane; (c) boxplot of error by time of day for middle lane; (d) boxplot of error by time of day for outer lane.

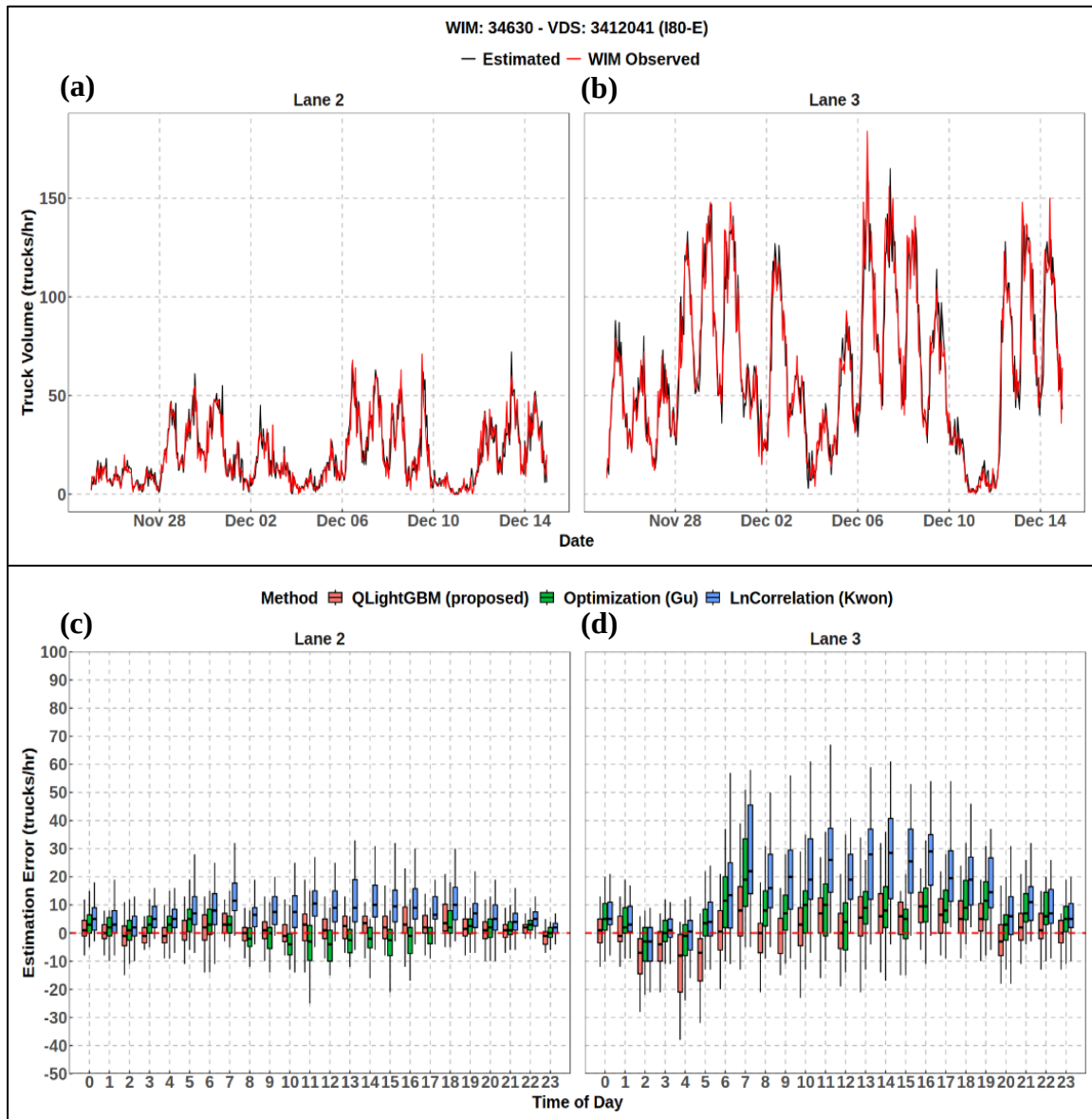


Figure 8 Estimation results for Loop 3412041 on I80-E: (a) error change over time for middle lane; (b) error change over time for outer lane; (c) boxplot of error by time of day for middle lane; (d) boxplot of error by time of day for outer lane.

Table 3 Model performance comparison at daily resolution.

		Model	MAE (trucks/hour)				MAPE (%)			
			Inner	Middle	Outer	All	Inner	Middle	Outer	All
I-80E	3412041	QLightGBM	1(2)	18 (12)	37(22)	45(32)	42(38)	6(4)	4(7)	4(5)
		Optimization	2(2)	29(18)	158(48)	155(57)	89(158)	10(11)	16(18)	13(14)
		LnCorrelation	3(3)	178(59)	360(137)	535(188)	100(0)	51(24)	27(13)	32(16)
		Interval	2(2)	323(187)	1035(529)	1360(713)	78(22)	70(5)	65(10)	66(7)
I-80W	3412044	QLightGBM	2(3)	47(50)	70(60)	94(96)	30(37)	9(6)	8(10)	7(9)
		Optimization	4(5)	54(42)	64(49)	117(66)	97(129)	11(10)	6(8)	8(9)
		LnCorrelation	6(6)	134(55)	327(126)	456(176)	100(0)	30(11)	22(8)	23(8)
		Interval	5(5)	368(193)	1148(549)	1522(731)	87(13)	70(3)	68(5)	69 (3)

Note:

- Daily aggregation smooths some inter-hour fluctuations,
- Bold font indicates the smallest value among the four methods,
- The numbers in parentheses represent the standard deviation.
- Kwon's lane correlation method assumes there is no truck on the inside lane so MAPE = 100% for nonzero cases

From the perspective of daily total truck traffic, the proposed method outperforms the other three models in terms of accuracy (smaller MAE and MAPE) and stability (smaller standard deviation). The proposed method achieves an MAE of about 70 trucks/day and MAPE of 5.5% on average at daily resolution. Besides, the proposed algorithm is computationally cheaper than other methods because once the supervised quantile models finished training, the established framework can be implemented in an online manner with just a few steps of simple calculation remaining. The following **Table 4** shows the pros and cons of the four models.

Conclusions

This paper presented a novel and straightforward framework for accurately estimating long truck (LT) traffic volume using single-loop measurements from traffic detectors. The motivation behind this study was the need to improve the accuracy of traffic speed estimation and vehicle classification, particularly for long trucks, as traditional methods have proven to be unreliable or not available under certain traffic conditions.

The proposed framework utilizes a Light Gradient Boosting Machine (LightGBM) and integrates quantile regression techniques to account for the uncertainty of speed estimation. Taking in one week of high-quality speed data from seven selected dual loops on the I-80, the trained quantile LightGBM framework provides a prediction interval based on volume and occupancy. The framework is then implemented at high temporal (30 seconds) and spatial (lane-by-lane) resolution, reducing the number of possible solutions and achieving accurate estimates. By capturing the relationship between speed, volume, and occupancy, the framework estimates the most possible fleet combination for the 30 seconds. The model generated close-to-ground truth estimates of long truck volume, with a Mean Absolute Error (MAE) of 12 trucks per hour and 70 trucks per day on average and the MAPE is 21% at hourly resolution and 5.5% at daily level on average.

The main contributions and findings of this paper are as follows:

Table 4 Pros and cons of different models.

Model	Pros	Cons
QLightGBM (Proposed)	- Low computational cost $O(n)$	- Requires high quality data for prior training
	- Easy to implement	
	- High accuracy	
	- High resolution	
Optimization (Gu)	- High resolution	- High computational cost $O(n^k)$
	- High accuracy	- Fail in congestion
LnCorrelation (Kwon)	- Easy to implement	- Assumption might not hold
		- Low resolution
Interval (Wang)	- N/A	- Assumption might not hold
		- Low resolution
		- Low accuracy

- To the best of our knowledge, this paper first adopts quantile regression based on LightGBM to quantify the uncertainty of speed estimation and truck volume estimation, allowing for more robust and accurate results.
- Estimation results are in a manner of high temporal (30 seconds) and spatial resolution (lane-by-lane), which enhances the precision and applicability of the framework by reducing the number of possible solutions per time interval (30 seconds) because traffic volume per 30 seconds reaches a peak value at around 25.
- The individual vehicle case analysis not only allows obtaining the actual vehicle length distribution but also facilitates the evaluation of the detector's measurement quality.
- The proposed method generates a close-to-ground truth estimate of LT volume with an MAE of 12 trucks per hour and 70 trucks per day on average. Truck estimates offer valuable insights for traffic operations, pavement design, and transportation planning, whereas speed estimates contribute to active traffic management and advanced traveler information systems.

The methodology and results were demonstrated on two test sites I-80 from the PEMS dataset, showing promising outcomes compared to three previously published models. The accurate and reliable LT traffic volume estimation makes this framework a valuable tool for transportation practitioners and policymakers, particularly those operating on tight budgets or lacking access to dual-loop detectors.

Looking ahead, this research paves the way for several promising avenues of further investigation. Firstly, evaluating the framework with higher-resolution data, such as 30-second data from video processing technology, could enhance the precision of truck volume estimation. Secondly, testing the model across different regions will help assess its spatial transferability, ensuring it can reliably adapt to diverse geographical and traffic conditions. Lastly, refining the LightGBM model by incorporating additional features may yield more accurate and robust estimates. Integrating real-time traffic data and leveraging advancements in big data analytics and machine learning could further improve the framework's performance and applicability in dynamic traffic scenarios.

In conclusion, the proposed framework represents a significant advancement in accurately estimating long truck traffic volume using single-loop traffic detectors. By addressing the limitations of existing methods and providing reliable results, it offers a valuable contribution to the transportation field. The findings presented in this study can support evidence-based decision-making for traffic management, infrastructure design, and sustainable transportation planning, contributing to safer and more efficient road networks in the future.

References

- [1] Zhang, G., Y. Wang, and H. Wei. Artificial neural network method for length-based vehicle classification using single-loop outputs. *Transportation research record*, Vol. 1945, No. 1, 2006, pp. 100-108.
- [2] Coifman, B., and S. Kim. Speed estimation and length based vehicle classification from freeway single-loop detectors. *Transportation Research Part C: Emerging Technologies*, Vol. 17, No. 4, 2009, pp. 349-364.
- [3] Dailey, D. J. A statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B: Methodological*, Vol. 33, No. 5, 1999, pp. 313-322.
- [4] Kwon, J., P. Varaiya, and A. Skabardonis. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. *Transportation research record*, Vol. 1856, No. 1, 2003, pp. 106-117.
- [5] Gu, Y., D. Liu, J. Stanford, and L. D. Han. Innovative Method for Estimating Large Truck Volume Using Aggregate Volume and Occupancy Data Incorporating Empirical Knowledge Into Linear Programming. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2676, No. 11, 2022, pp. 648-663.
- [6] Wang, Y., and N. L. Nihan. Dynamic estimation of freeway large-truck volumes based on single-loop measurements. In *Intelligent Transportation Systems*, No. 8, Taylor & Francis, 2004. pp. 133-141.
- [7] Athol, P. Interdependence of certain operational characteristics within a moving traffic stream; Highway Research Record No. 72. In, Highway Research Board Washington D. C, 1965.
- [8] Sun, C., and S. G. Ritchie. Individual vehicle speed estimation using single loop inductive waveforms. *Journal of Transportation Engineering*, Vol. 125, No. 6, 1999, pp. 531-538.
- [9] Wang, Y., and N. L. Nihan. Freeway traffic speed estimation with single-loop outputs. *Transportation research record*, Vol. 1727, No. 1, 2000, pp. 120-126.
- [10] Lin, W.-H., J. Dahlgren, and H. Huo. An enhancement to speed estimation using single loop detectors. In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, No. 1, IEEE, 2003. pp. 417-422.
- [11] Lu, X.-Y., P. Varaiya, R. Horowitz, Z. Guo, and J. Palen. Estimating traffic speed with single inductive loop event data. *Transportation research record*, Vol. 2308, No. 1, 2012, pp. 157-166.

- [12] Lao, Y., G. Zhang, J. Corey, and Y. Wang. Gaussian mixture model-based speed estimation and vehicle classification using single-loop measurements. *Journal of Intelligent Transportation Systems*, Vol. 16, No. 4, 2012, pp. 184-196.
- [13] Li, Y., A. Y. Tok, and S. G. Ritchie. Individual truck speed estimation from advanced single inductive loops. *Transportation research record*, Vol. 2673, No. 5, 2019, pp. 272-284.
- [14] Wang, Y., and N. L. Nihan. Can single-loop detectors do the work of dual-loop detectors? *Journal of Transportation Engineering*, Vol. 129, No. 2, 2003, pp. 169-176.
- [15] Ye, Z., Y. Zhang, and D. R. Middleton. Unscented Kalman filter method for speed estimation using single loop detector data. *Transportation research record*, Vol. 1968, No. 1, 2006, pp. 117-125.
- [16] Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30, 2017.
- [17] Chen, T., and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016. pp. 785-794.
- [18] Koenker, R., and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, 1978, pp. 33-50.
- [19] FHWA. *FHWA 13-Category Scheme for Vehicle Classifications*. http://onlinemanuals.txdot.gov/txdotmanuals/tri/classifying_vehicles.htm.
- [20] Minge, E. D., S. Peterson, H. Weinblatt, B. Coifman, and E. Hoekman. Loop-and length-based vehicle classification: Federal highway administration-pooled fund program [tpf-5 (192)]. In, Minnesota Department of Transportation, Research Services, 2012.

CHAPTER II
SHORT-TERM TRAFFIC VOLUME PREDICTION ON FREEWAYS
UNDER INCIDENT AND NON-INCIDENT CONDITIONS: A
PERSPECTIVE FROM TRAINING AND TESTING PHILOSOPHY

Abstract

Most research on short-term traffic flow prediction has primarily concentrated on designing new models and architectures, often overlooking the critical application of these models under atypical conditions. This paper aims to address this gap by reviewing and proposing a comprehensive training and testing framework for traffic flow prediction, emphasizing the importance of including incident data in both training and testing phases. Using traffic volume data from radar detectors in Nashville, TN, this study implements four widely used models: Gated Recurrent Unit (GRU) neural network, Support Vector Regression (SVR), Historical Average (HA), and Naïve Random Walking (RW). The focus is on evaluating the performance difference of these models with different training procedures in forecasting short-term traffic flow under with incident, without incident, and incident specific scenarios. Our findings demonstrate that including incident data in the training datasets significantly enhances the model's performance and reliability by reducing the mean absolute percentage error by 25% on average under incident specific scenario. This highlights the critical importance of maintaining the integrity of incident data, avoiding data manipulation prior to training, avoiding diluting model performance by including huge amounts of testing data. Notably, complex models do not always outperform simpler models such as random walk. This identifies a common evaluation trap in past research on traffic flow prediction, where new models are often not compared against basic models like Random Walk. In conclusion, this paper underscores the necessity of a robust framework that incorporates incident data to improve the accuracy and applicability of traffic flow prediction models, providing valuable insights for practitioners implementing intelligent transportation systems (ITS) in real-world scenarios.

Introduction

Traffic flow data, a specialized type of time series data, is significantly influenced by both temporal and spatial factors. Accurate prediction of traffic flow is crucial for Intelligent Transportation Systems (ITS), which have seen considerable advancements

since the 1990s due to the development of vehicle detection technologies. These technologies enable the efficient and accurate real-time acquisition of traffic data, allowing analysts to monitor and manage traffic systems effectively, thus improving operational efficiency and safety. Predicting traffic flow and speed, which are key components of ITS, helps control traffic flow and mitigate congestion (1).

Researchers are primarily interested in two major families of methods: parametric and non-parametric models. Among all parametric approaches, the Auto Regressive Integrated Moving Average (ARIMA) model is a prevailing framework and the benchmark for short term traffic flow prediction (1). While parametric methods perform well when traffic follows typical patterns, their accuracy significantly decreases under irregular and atypical conditions, such as unexpected events or long prediction intervals (2). This limitation has led researchers to focus on non-parametric methods, including support vector regression (SVR) and artificial neural networks (ANN). These methods, capable of capturing the non-linear and spatiotemporal characteristics in traffic data, have shown better performance than parametric approaches. However, they require extensive prior knowledge, data preprocessing, feature extraction, and parameter training to achieve satisfactory results.

Despite these advancements, the inclusion of incident data in both training and testing phases is often overlooked. Previous research has primarily focused on designing innovative and complex models to improve prediction accuracy, such as leveraging sophisticated machine learning algorithms and deep neural networks. However, these efforts often ignore the fundamental training and testing philosophy of short-term traffic prediction, which should account for real-world scenarios, including incidents and irregular traffic patterns.

The omission of incident data leads to models that perform well under normal conditions but falter during unexpected disruptions, such as accidents, roadworks, or adverse weather events. This gap in the research can result in prediction models that lack robustness and fail to provide accurate forecasts when they are most needed. Therefore, addressing this issue by incorporating incident data into the training process is crucial for developing more reliable and resilient traffic prediction models.

This paper aims to address this gap by proposing a novel training and testing framework that emphasizes the inclusion of incident data. By using traffic volume data from radar detectors in Nashville, TN, and implementing models such as Gated Recurrent Unit (GRU), SVR, Historical Average (HA), and Random Walk (RW), this study evaluates the performance of these models under various scenarios. The findings highlight the critical importance of including incident data to enhance the generalizability and reliability of traffic prediction models. This approach ensures that the models are better equipped to handle the complexities and variabilities of real-world traffic conditions, leading to improved traffic management and congestion mitigation.

The remainder of this chapter is structured as follows: The next section offers a comprehensive literature review, highlighting key studies and identifying research gaps in traffic flow prediction. Following this, we describe the models implemented and the data used in our study, including the formulas for evaluation metrics. This section also introduces the proposed training and testing framework. The results and discussion section presents our main research findings, analyzing the models' performance and their implications. Finally, the conclusion discusses current research gaps and suggests possible future directions.

Related work

This section presents a thorough literature review of research in the last ten years about short-term traffic flow prediction. The review looked into details of prevalent and effective models utilized for traffic flow prediction including theoretical developments, different application scenarios, data sources and quality, performance evaluation and research gaps. **Table 5** lists all the acronyms and abbreviations used in this chapter, ordered alphabetically.

In 2014, Huang and Song et al. published the first paper that applies deep learning approach to traffic flow prediction research (3). They built a deep belief network (DBN) as the unsupervised feature extraction layer and a multitask regression output as the supervised prediction layer. Experiment results showed a 5% improvement over the state-of-the-art at that time. Niu and Zhu et al. proposed DeepSense, an innovative

Table 5 List of Acronyms in alphabetical order.

AE	Autoencoders
AI	Artificial Intelligence
ANN	Artificial Neural Network
APE	Absolute Percentage Error
ARIMA	Autoregressive integrated moving average
CNN	Convolution Neural Network
DBN	Deep Belief Network
DL	Deep Learning
FFNN	Feed Forward Neural Network
GAN	Generative Adversarial Network
GPS	Global Positioning System
GRU	Gated Recurrent Unit
HTM	Hierarchical Temporal Memory
ITS	Intelligent Transportation System
KF	Kalman Filtering
k-NN	k-nearest neighbor
LM	Levenberg-Marquardt algorithm
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Square Error
PCA	Principle Component Analysis
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machine
RW	Random Walk
RMSE	Root Mean Squared Error
RMSPE	Root Mean Squared Percentage Error
RNN	Recurrent Neural Network
SAE	Stacked Autoencoder
SGD	Stochastic Gradient Descent
SRCN	Spatiotemporal Recurrent Convolutional Network
SVM	Support Vector Machine
SVR	Support Vector Regression
TFP	Traffic Flow Prediction

spatiotemporal feature learning mechanism based on Restricted Boltzmann Machine (RBM) (4). This deep learning architecture helped reduce the dimension of the data. Then the low-dimension data was trained by a support vector machine (SVM) to achieve good prediction traffic flow. This model also achieved better accuracy with approximately 5% improvement compared to a normal neural network. Additionally, Lv and Duan et al. first applied stacked autoencoders (SAE) to represent traffic flow features in 2014 (5). The proposed deep-learning-based model considered the temporal and spatial correlations intrinsically and outperformed among different models. Zhao and Zhou designed a parallel computing method of DBN and applied it to traffic flow prediction (6). Experiments verified the effectiveness and time consumption economy of the proposed method.

In 2015, Ma and Yu et al. deployed RBM and Recurrent Neural Network (RNN) to forecast the evolution of traffic congestion using Global Positioning System (GPS) data from taxi (7). The prediction accuracy was as high as 88% and they also visualized the congestion evolution spatially and temporally. Leelavathi and Sahana exploited the spatial-temporal correlation within the simulation traffic flow data using SAE model in 2016 (8). Yang and Dillon et al. used spatial-temporal data from UK M-6 freeway to predict traffic flow based on SAE Levenberg Marquardt (LM) model (9). Results demonstrated the model with an optimal structure has superior performance. Duan and Lv et al. trained a SAE model and evaluated the performance at different time periods (10). Experiments indicated that parameters of SAE model varied with time periods.

Polson and Sokolov combined a linear model and a sequence of tanh layers to predict traffic during two special events in 2017 (11). Dawei Chen developed an optimized forecasting algorithm of radial basis function (RBF) neural network model (12). Simulation data proved the accuracy and effectiveness of the proposed model. Kunde and Hartenstein et al. (13) employed a Feed Forward Neural Network (FFNN) with stochastic gradient descent (SGD) algorithm on spatiotemporal VAMOS data. Lin and Dai et al. proposed a pattern sensitive network based on Wasserstein GAN with gradient penalty (WGAN-GP) (14). Experiments proved that the model excelled in both usual and unusual traffic states.

Some researchers tried to incorporate accident, weather, and social media data with traditional traffic flow data in order to improve the prediction accuracy because these atypical events have a great impact on traffic conditions. Soua and Koesdwiady et al. built a Deep Belief Network (DBN) and applied decision-level data fusion scheme to enhance the prediction accuracy (15). They also used DBN framework and adopted Dempster's rule to fuse information from both flow and event data (16). Results demonstrated the advantage of the designed framework over the state-of-the-art models. In 2018, Zhang and Kabuka proposed a gated recurrent unit (GRU) neural network (17) to predict traffic flow considering weather information (18). Jia and Wu et al. devised a rainfall-integrated DBN and LSTM model which can capture the traffic stream characteristics under various rainfall scenarios (19).

Long Short-Term Memory (LSTM) (20) is a type of RNN capable of leveraging long temporal sequences and dramatically minimizing the vanishing and exploding gradient problem encountered when training RNNs. LSTM was the most widely used deep learning architecture for traffic flow prediction among the selected articles because its inherent structure enables excellent capability of predicting time series with long and short temporal dependency. Ma and Tao et al. used LSTM neural network and achieved the best accuracy and stability compared to other prevailing models (7). Yu and Li et al. proposed an LSTM-based approach for extreme condition traffic forecasting (21). Yu and Wu proposed a spatiotemporal recurrent convolutional network (SRCN) integrating CNN and LSTM which outperformed other DL algorithms in both short-term and long-term traffic predictions (22). Chen and Lv et al. used a stacked-LSTM model with online spatial data to predict traffic flow (23). Liu and Zheng et al. proposed a Convolutional-LSTM and Bidirectional-LSTM module to extract the spatiotemporal and periodical features in traffic flow data (24). In 2018, Wu and Tan et al. used a hybrid model combining CNN and RNN to capture spatiotemporal information and successfully visualized the understanding process of CNN (25). Mackenzie, Roddick and Zito compared hierarchical temporal memory (HTM) with LSTM to show its potential as an effective tool for predicting short term traffic flow (26). Tian and Zhang proposed a novel approach based on LSTM to deal with missing data when predicting traffic flow (27).

Yang and Sun improved LSTM model with the attention mechanism to solve the issue of exceedingly long-term dependencies (28).

In 2020, Avila and Mezić (29) published a well-known article on Nature about forecasting traffic dynamics. The researchers present a novel, data-driven approach using Koopman mode decomposition (KMD) to analyze and forecast highway traffic dynamics. The study leverages data from the Federal Highway Administration and the California Department of Transportation to identify and reconstruct spatiotemporal traffic patterns without relying on traditional mathematical models or extensive parameter tuning. The researchers emphasize the model's capability to distinguish and forecast complex traffic patterns, such as traffic waves and localized clusters, across multiple lanes and varied conditions. However, a critical limitation noted in their methodology is the exclusion of incident data during the model training phase, which limits the model's generalizability and effectiveness in real-world applications where traffic disruptions are common. This gap highlights the necessity for future research to integrate incident data, enhancing model robustness and reliability in forecasting under both regular and irregular traffic conditions.

Research gaps

Based on the detailed and comprehensive review of traffic time series forecasting problem, this section will discuss the existing research gaps and possible future directions in this domain.

The commonly used performance metrics, RMSE, MAE and MAPE sometimes will experience a failure as discussed in the above section. Thus, how to define a universally applicable framework to evaluate the traffic flow prediction models remains an open issue and requires further statistical analysis. When a large number of testing data points are incorporated across various spatial and temporal dimensions, the average performance of traffic prediction models can become diluted. This dilution can obscure the model's true effectiveness under specific conditions, failing to provide a clear picture of its performance during critical traffic incidents. Addressing this gap is crucial for developing more robust and accurate models that can deliver reliable forecasts under

varied and challenging traffic scenarios.

The data collection process needs to be improved along with the development of big data technologies. The traditional loop detector-based approach is not stable enough thus causing problems of missing data, irregular sampling and so on. On the one hand, several articles have focused on traffic missing data imputation methods, such as kernel probabilistic principal component analysis (KPPCA) (30), LSTM (27), and so on. But the current imputation still remains at the first level which is to impute a group of missing points at a fixed location within a certain time period. Elevated to a route with multiple detectors and a complex network, imputation methods for inferring the missing data incorporating both time and space information is still an open topic and requires further studies.

On the other hand, it is advisable for traffic engineers to upgrade the vehicle counting technologies with more stable and convenient infrastructure. Traffic video data is a promising data source and computer vision-based vehicle tracking and counting techniques have been developed. However, the problem is the tradeoff between video resolution and data size. High accuracy of counting requires a relatively high video resolution thus making the data too big to be efficiently processed. With the advent of parallel and cloud computing technologies, non-structured data such as video and image has been gradually energized to provide fast and stable predictions. This remains an open challenge in the domain of traffic flow prediction.

Another problem lies in the occurrence of unusual and unexpected events such as traffic accidents, sudden rainfall, ball games in the neighborhood. Some works have considered some of these atypical events and the most widely used factor is the weather information. But few works have considered visibility caused by fog. Besides, the scheduled working zone effects, for example, lane closure, traffic guiding etc., are also rarely integrated in traffic flow analysis. Also, different types of event data may require different input formats and data fusion approaches.

Different impacts of data aggregation level and prediction interval require further research attention. Some parametric models are good at predicting the next one interval while non-parametric models are able to forecast the next multiple intervals. Moreover,

as the prediction interval increases, parametric models may fail to achieve satisfactory results whereas non-parametric models can capture long temporal dependency. It also needs to be investigated how different aggregation levels influence the prediction results with other factors being fixed.

The primary objective of this paper is to review and propose a novel training and testing framework for short term traffic flow prediction, shifting the focus from the design of complex new models to improving the practical application and evaluation of existing ones. This framework emphasizes the importance of incorporating incident data into both the training and testing phases to enhance the generalizability and reliability of traffic prediction models. By addressing the common pitfalls in current methodologies, such as the exclusion of incident data and the dilution of performance metrics, this paper aims to develop a more robust approach that can deliver accurate and reliable traffic forecasts under a wide range of real-world conditions.

Methodology

This section introduces the basic prediction models that were implemented in this study, including GRU, SVR, HA, and RW.

Gate Recurrent Unit (GRU)

GRU, which was implemented in this study, is a special variation of RNN, so we first introduce the structure of RNN which is depicted in **Figure 9 (31)**. Recurrent Neural Networks (RNN) are a class of artificial neural networks designed to recognize patterns in sequences of data, such as time series or natural language, by maintaining a hidden state that captures temporal dependencies. RNN works by updating a state s_t not only on current input x_t , but also upon the previous output at state (t-1). Equations are presented below to represent the operation of RNN (31), where:

$$s_t = f_s(w_i x_t + w_R s_{t-1} + b_s) \quad (1)$$

$$y_t = f_t(w_0 s_t + b_y) \quad (2)$$

w_i, w_R, w_0 represents weight matrix from input to hidden layer, recurrent weight matrix and weights associated with hidden to output layer.

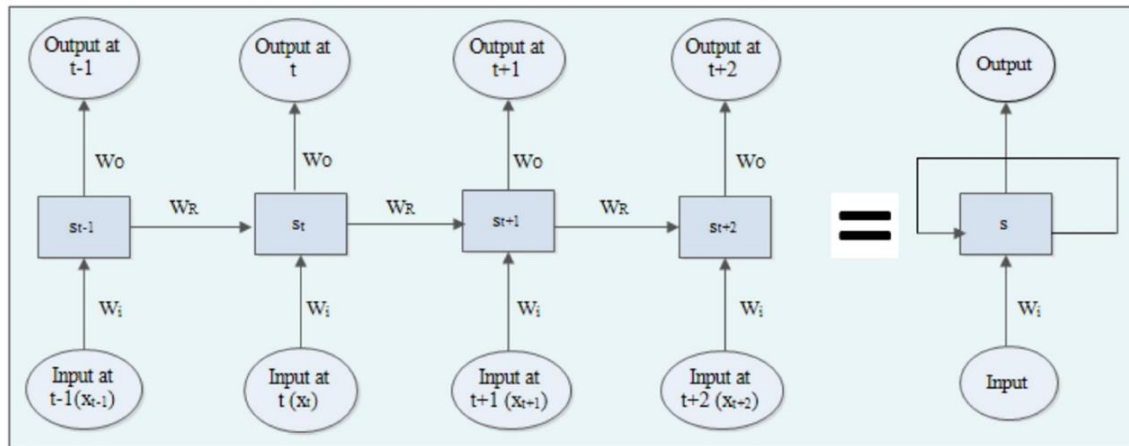


Figure 9 Structure of RNN.

In Conventional RNNs, the effect of past input on output declines exponentially as it cycles around the recurrent connection (32). Hence, RNNs can only deal with the problem of sequence handling for short-term dependencies, as for long-term dependencies, it encounters the problem of vanishing gradient. LSTM is a special subtype of RNN, and it was introduced by Hochreiter and Schmidhuber in 1997 (20). LSTM outperforms the traditional RNN due to its characteristics of selectively memorizing and forgetting patterns for a long time-duration. LSTM network has the capability of handling time series data for both long and short-term duration. GRU is a modified variation of LSTM, which is devised by Cho et al. in 2014 (17). Moreover, GRU performs similarly to LSTM but is computationally cheaper (33). The structures of LSTM and GRU are illustrated in **Figure 10** (31).

Equations (3), (4), (5), (6), (7) show how LSTM works.

$$i_t = \sigma(W_i x_t + U_i s_{t-1} + V_i C_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_f x_t + U_f s_{t-1} + V_f C_{t-1} + b_f) \quad (4)$$

$$o_t = \sigma(W_o x_t + U_o s_{t-1} + V_o C_{t-1} + b_o) \quad (5)$$

$$C_t = f * C_{t-1} + i_t * g(W_c x_t + W_c s_{t-1} + b_c) \quad (6)$$

$$s_t = o_t * h(C_t) \quad (7)$$

where i , f and o denote the output of input, forget and output gates respectively. C_t and s_t represent cell and memory block state vectors respectively. W, U, V are weight matrices and b represent bias.

The calculation principle for GRU is presented in the following equations (8), (9), (10) and (11).

$$z = \sigma(x_t U_z + s_{t-1} W_z) \quad (8)$$

$$r = \sigma(x_t U_r + s_{t-1} W_r) \quad (9)$$

$$h = \tanh(x_t U_h + (s_{t-1} * r) W_h) \quad (10)$$

$$s_t = (1 - z) * s_{t-1} + z * h \quad (11)$$

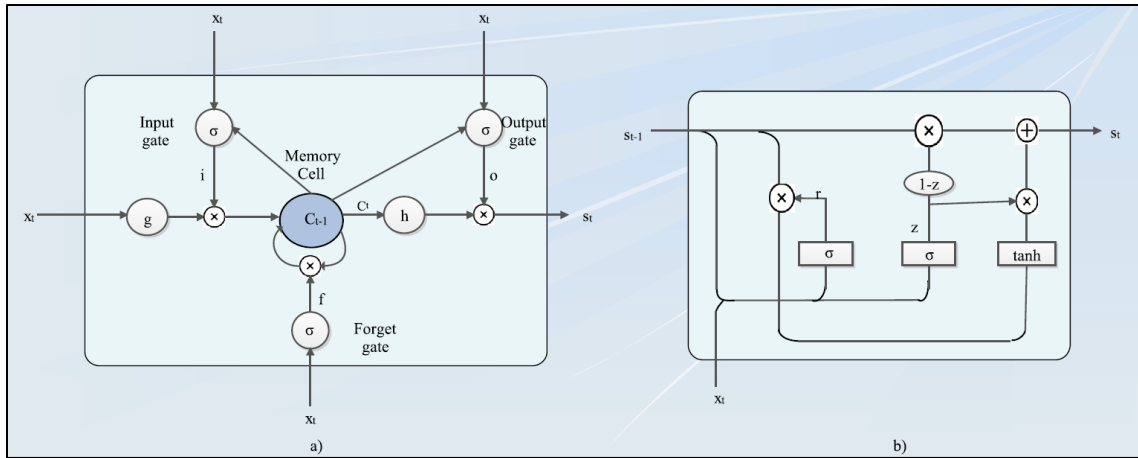


Figure 10 Structure of a) LSTM and b) GRU.

where, x_t, z, r, s_t represent input, update gate, reset gate and output vectors respectively, and W, U are weight matrices. σ is a sigmoid function that enables the output number between 0 and 1 where 1 signifies all information passes through and 0 represents no information is kept. A $\tanh$ function outputs numbers between -1 and 1.

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning method that uses a hyperplane to predict traffic flow by capturing non-linear relationships within the data. SVR works by minimizing the error within a certain threshold, making it robust against outliers. The SVR model can be expressed as:

$$\hat{y}_t = w \cdot \phi(x) + b \quad (12)$$

where $\hat{y}_t$ is the predicted traffic flow, w and b are the model parameters, x represents the input features (e.g., previous traffic observations), and $\phi(x)$ is a non-linear function that maps the input data into a higher-dimensional space. The parameters w and b are determined by solving the following optimization problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \max(0, |y_i - \hat{y}_i| - \epsilon) \quad (13)$$

where C is a regularization parameter, ϵ is a precision parameter, and N is the number of training samples. SVR is effective in handling complex, non-linear traffic patterns, providing more accurate predictions compared to simpler models like HA and RW.

Historical Average (HA)

The Historical Average (HA) model predicts future traffic flow based on the average of past observations over a specific period. This method assumes that traffic patterns are consistent over time and can be approximated by their historical means. The formula for HA can be expressed as:

$$\hat{y}_t = \frac{1}{N} \sum_{i=1}^N y_{t-i} \quad (14)$$

where $\hat{y}_t$ is the predicted traffic flow at time t , y_{t-i} represents the observed traffic flow at previous times, and N is the number of historical observations considered. Despite its simplicity, the HA model cannot accurately capture sudden changes in traffic flow due to

incidents or unusual events.

Random Walk (RW)

The Random Walk (RW) model predicts that future traffic flow will be the same as the most recent observed value, essentially assuming that traffic conditions will persist unchanged. The RW model is expressed by the formula:

$$\hat{y}_t = y_{t-1} + \epsilon \quad (15)$$

This model is easy to implement and often serves as a baseline for comparing more sophisticated prediction models. However, it can be inaccurate in dynamic traffic conditions, especially when traffic flow changes frequently.

Performance Metrics

There are many performance metrics for evaluating the prediction error or accuracy such mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and so on. $\hat{y}_i$ and y_i denote predicted value and observed value respectively. n represent the total number of predicted data points. Equations (16), (17), and (18) show the calculation of four mainly used measurements.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (16)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (17)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \quad (18)$$

These statistically designed metrics serve as a measure of effectiveness to train and assess different models' performance. The objective of these statistical metrics is to express the distance between actual observations and model predictions. Among all these measures, RMSE and MAPE are the most widely used (34). Having the same unit as the predicted variable, RMSE offers insightful assessment. It should be noted that a well-performed metric is able to evaluate the prediction results regardless of the absolute

magnitude. Sometimes, the effectiveness of metrics is greatly influenced by the magnitude of actual observations. For example, a 50% MAPE could be a good prediction if the actual volume is extremely low, say 4 vehicles per hour (vph); whereas a same MAPE for a saturated flow of 1900 vph is definitely a bad prediction.

Data

The datasets employed in this study are critical to understanding traffic flow variation and incident impact on roadways in Nashville, TN. This section briefly describes the datasets used in this study, as well as the proposed new training and testing framework for a more comprehensive evaluation. The two major datasets are traffic flow data and official incident reports which are described as follows:

- **RDS Traffic Volume Data:**
 - **Source:** Radar Detection System (RDS)
 - **Details:** This dataset contains traffic volume records from five detectors randomly selected from major interstates in Nashville, TN. These detectors provide continuous data on vehicle counts passing each point, offering insights into flow variations across different times of the day and during various conditions.
 - **Time Frame:** The dataset covers the period from July 10, 2022, to July 23, 2022, designated as the training period. This period was chosen to represent a mix of regular and irregular traffic conditions. The following week from July 24, 2022, to July 30, 2022 was used as the testing period.
- **Locate/IM Incident Data:**
 - **Source:** Locate/IM (Incident Management)
 - **Details:** The Locate/IM dataset includes official records of traffic incidents such as accidents, roadblocks, start time, and end time. This dataset helps in pinpointing specific time when incidents occurred, thus allowing the study to analyze how such events affect traffic patterns and to assess the prediction performance with and without the incident occurrence. The incident data is used to both filter out training data and

select test days.

- **Time Frame:** The duration of official incident logs covered the entire training and testing period.

The study explores three distinct training approaches to understand the impact of incident occurrences on model performance:

Training Framework

The study explores two different training approaches to understand the impact of incident occurrences on model performance:

- **Incident-Excluded Training Data:**
 - **Description:** The training data is filtered to exclude periods with recorded incidents. This approach aims to train the model on regular traffic patterns without disruptions, providing a clean dataset for learning typical traffic behaviors.
 - **Purpose:** To assess the model's ability to predict standard traffic volumes and its performance when not exposed to incident-related anomalies during training.
- **Incident-Included Training Data:**
 - **Description:** The training data includes all days within the training period, encompassing both regular days and those with incidents. This approach exposes the model to a broader range of traffic conditions, including the effects of incidents on traffic volumes.
 - **Purpose:** To evaluate the model's capacity to generalize from a diverse dataset and predict traffic volumes under varying typical and atypical conditions, including the impact of disruptions.
- **Incident-Specific Training Data:**
 - **Description:** The training data is specifically curated to include only periods with recorded incidents. This approach focuses on understanding the unique traffic patterns and anomalies caused by incidents, providing the model with targeted exposure to these situations.

- **Purpose:** To evaluate the model's ability to predict traffic volumes and patterns specifically during incident scenarios. This helps in assessing the model's robustness and accuracy in responding to disruptions, which is critical for applications in traffic management and incident response.

The comparison between these three approaches provides insights into how the presence of incident data during training influences the model's performance and its ability to handle both regular and disrupted traffic conditions.

Testing Framework

The model's performance is also tested under three distinct scenarios, mirroring the training data approaches. These scenarios are designed to provide a comprehensive assessment across various traffic conditions: incident-excluded, incident-included, and incident-specific.

- **Incident-Excluded Testing Data:**
 - **Description:** The testing data consists of periods without any recorded incidents. This scenario evaluates the model's performance in predicting traffic volumes during normal traffic conditions.
 - **Purpose:** To assess the model's accuracy in forecasting typical traffic patterns when no disruptions are present, ensuring the model can effectively handle standard scenarios.
- **Incident-Included Testing Data:**
 - **Description:** The testing data includes both regular days and days with recorded incidents. This scenario examines the model's ability to generalize and predict traffic volumes under a mix of typical and atypical conditions.
 - **Purpose:** To evaluate the model's robustness and flexibility in handling a diverse range of traffic situations, including those influenced by incidents.
- **Incident-Specific Testing Data:**
 - **Description:** The testing data is curated to only include periods with recorded incidents. This scenario focuses on the model's performance in

predicting traffic patterns during incidents.

- **Purpose:** To test the model's effectiveness in responding to disruptions and accurately forecasting traffic volumes during incident scenarios, which is crucial for applications in traffic management and emergency response planning.

This diverse selection of test days ensures a comprehensive evaluation of the model's performance under varying traffic conditions and its ability to adapt to both typical and atypical traffic scenarios. This analysis will provide valuable insights for optimizing traffic prediction models, particularly in the context of incident management and response. **Table 6** and **Table 7** shows the size of the data for the three randomly selected detectors under different training and testing scenarios. **Figure 11** shows the profile of training data for the selected detectors.

Results and Discussion

This section presents the results of the prediction model, focusing on the comparative analysis of the impact of three training conditions on three testing conditions respectively. The performance metrics across different testing scenarios for both one-step ahead and multi-step ahead predictions are discussed in detail.

Traffic Volume Prediction Performance

The graphs in **Figure 12** display the observed and one-step ahead predicted traffic volumes by GRU model for the three RDS detectors across different test scenarios. The blue lines represent observed traffic volumes, while the red lines show the predicted volumes from the model. The light blue shaded area highlights the time with a recorded incident occurrence. Overall, the model's predictions closely align with the observed data across three detectors and testing scenarios by visual inspection. The summary statistics will be presented in the next section.

Table 6 Training data conditions.

ID	RDS	# of incidents	Incident-included	Incident-excluded	Incident-specific
1	I24-58.5E	24	4032	3627	405
2	I65-91.2S	14	4032	3869	163
3	SR386-4.1E	6	4032	3970	62

Table 7 Testing data conditions.

ID	RDS	# of incidents	Incident-included	Incident-excluded	Incident-specific
1	I24-58.5E	6	2016	1898	118
2	I65-91.2S	7	2016	1943	73
3	SR386-4.1E	1	2016	2007	9

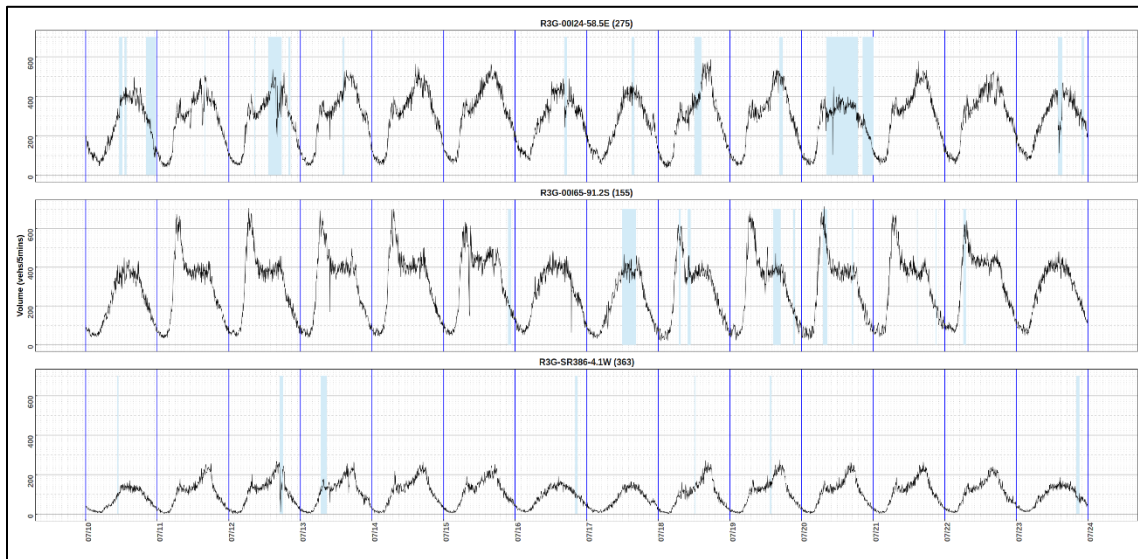


Figure 11 Profile of training data, the blue shaded area denotes the incident impact duration.

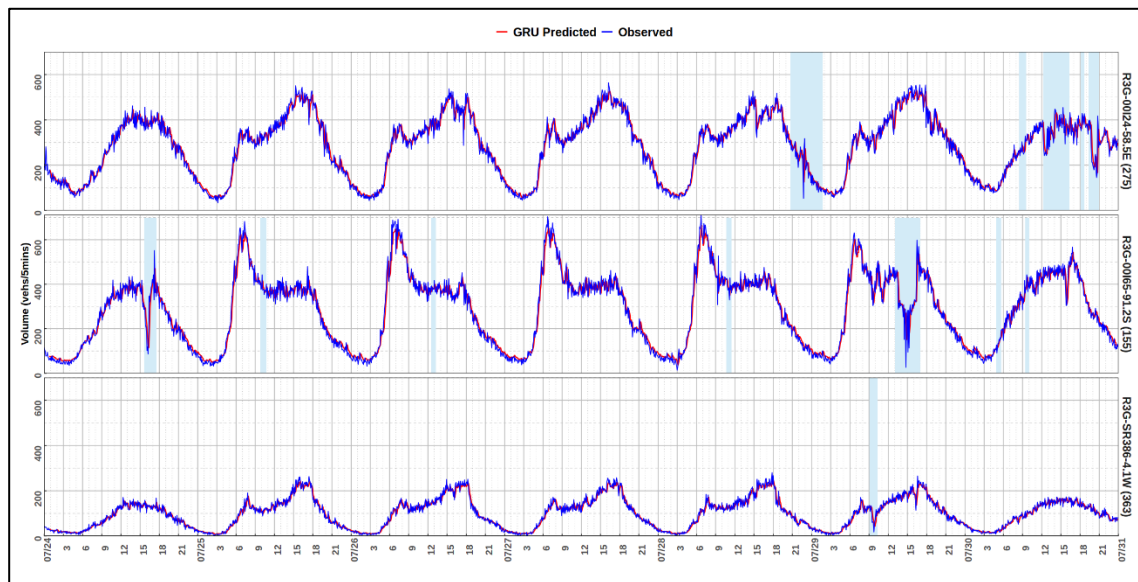


Figure 12 Profile of testing data as well as GRU performance for one-step ahead prediction.

Comparative Analysis of different training and testing scenarios

The second part of the analysis involves comparing the model's performance when trained and tested on three distinct scenarios. The performance metric used for comparison is the Mean Absolute Percentage Error (MAPE). The results for both one-step and 6-step ahead predictions will be presented in this section.

One-step ahead prediction

Table 8 presents a detailed comparison of the Mean Absolute Percentage Error (MAPE) for four different prediction models— GRU, RW, HA, and SVR—across three test scenarios. The table contrasts the performance of models trained on different conditions for detectors located on various roads in Nashville, TN.

For each test scenario, MAPE values are listed for individual detectors and averaged across all detectors to provide a comprehensive performance summary. The results indicate that models trained on non-filtered data consistently yield lower MAPE values across all test scenarios and detectors, demonstrating enhanced prediction accuracy. The findings indicate that the SVR model consistently outperforms other models, achieving the lowest average MAPE of 11.86% when trained with incident data and tested in the same condition. It also excels in incident-excluded scenarios with an average MAPE of 11.39%. The RW model shows competitive performance, particularly when trained without and tested with incident data, with an average MAPE of 13.24%. However, in incident-specific testing, GRU maintains a lower average MAPE of 26.35%, compared to RW's 26.47%. The HA model generally exhibits higher MAPE values, indicating lower prediction accuracy across all conditions. This underscores the value of including simpler models in traffic flow prediction studies to provide a reliable benchmark. The analysis also highlights the critical importance of including incident data in the training process for traffic flow prediction models. Incidents such as accidents, or sudden weather changes significantly change traffic patterns, creating unpredictable and often severe deviations from normal conditions. When incident data is incorporated into the training phase, models like GRU and SVR can learn and adapt to these disruptions,

Table 8 MAPE (%) performance on different scenarios for one-step ahead prediction; red bold numbers denote the best model for each scenario.

MAPE (%)			Testing											
			Incident-included				Incident-excluded				Incident Specific			
			GRU	SVR	RW	HA	GRU	SVR	RW	HA	GRU	SVR	RW	HA
Training	Incident-included	I24-58.5E	9.02	8.55	9.42	11.54	8.58	8.10	9.00	11.00	14.53	14.17	14.64	18.26
		I65-91.2S	11.74	10.41	11.56	14.39	11.06	9.64	10.72	13.34	22.94	23.08	25.51	31.59
		SR386-4.1E	17.81	16.63	18.72	20.15	17.63	16.43	18.57	19.78	41.59	42.16	39.25	68.11
		Average	12.85	11.86	13.24	15.36	12.42	11.39	12.76	14.71	26.35	26.47	26.47	39.32
	Incident-excluded	I24-58.5E	9.92	9.57	9.42	11.54	8.32	7.94	9.00	11.00	20.63	20.97	14.64	18.26
		I65-91.2S	13.62	12.49	11.56	14.39	10.51	9.44	10.72	13.34	33.95	33.93	25.51	31.59
		SR386-4.1E	21.19	18.46	18.72	20.15	16.92	15.61	18.57	19.78	60.72	61.14	39.25	68.11
		Average	14.91	13.51	13.24	15.36	11.92	11.00	12.76	14.71	38.44	38.68	26.47	39.32
	Incident Specific	I24-58.5E	29.13	27.86	9.42	11.54	30.63	28.50	9.00	11.00	10.46	10.49	14.64	18.26
		I65-91.2S	38.50	33.93	11.56	14.39	38.37	33.54	10.72	13.34	16.97	16.39	25.51	31.59
		SR386-4.1E	58.05	54.04	18.72	20.15	62.04	57.52	18.57	19.78	30.36	33.73	39.25	68.11
		Average	41.89	38.61	13.24	15.36	43.68	39.85	12.76	14.71	19.27	20.20	26.47	39.32

resulting in more accurate and reliable predictions during real-world incidents.

Without the inclusion of incident data in training, even sophisticated models may struggle to outperform basic models like Random Walk (RW) when tested in incident-specific situations. As shown in **Table 8**, the GRU model achieves a higher MAPE compared to RW when trained without incident data, indicating its performance can degrade when incidents are excluded from the training set. This underscores that the robustness and adaptability of advanced models depend heavily on their exposure to diverse traffic scenarios during training. These performance statistics are not intended to promote any specific model. Instead, they highlight the importance of maintaining Random Walk (RW) as the baseline model, as even sophisticated models like GRU can be outperformed by RW, particularly in incident-specific scenarios.

Notably, **Table 9** summarizes the average performance of traffic flow prediction models in terms of MAPE for one-step predictions under different testing scenarios. The table compares models trained and tested under three conditions respectively. It provides a comprehensive overview of the average model performance for one-step prediction, measured using the Mean Absolute Percentage Error (MAPE) across various training and testing scenarios. The findings highlight several key observations about the models' capabilities and the importance of data alignment in training and testing.

Firstly, the diagonal elements, representing scenarios where the training and testing datasets are aligned, consistently show the lowest MAPE values. This pattern underscores the significance of training models on data that closely mirrors the conditions they will encounter during testing. For example, when the model is trained on incident-included data and tested on incident-included data, the MAPE is 12.36%, which is notably lower than when tested on incident-excluded or incident-specific data. Similarly, the lowest MAPE for incident-excluded training (11.46%) occurs when testing on incident-excluded data, reinforcing the idea that alignment in data characteristics between training and testing phases enhances predictive accuracy.

Table 9 Average model performance for one-step prediction.

MAPE (%)		Testing		
		Incident-included	Incident-excluded	Incident-Specific
Training	Incident-included	12.36	11.91	26.41
	Incident-excluded	14.21	11.46	38.56
	Incident-Specific	40.25	41.77	19.73

Moreover, the table reveals that models trained on incident-specific data exhibit a significant increase in MAPE when tested on incident-included (40.25%) and incident-excluded (41.77%) data, compared to when tested on incident-specific data (19.73%). This disparity suggests that incident-specific training data might not provide sufficient generalization capability, highlighting the challenges in using specialized training datasets to predict more generalized traffic patterns. It also emphasizes the necessity of a diverse training dataset for models that are expected to perform well under a range of conditions, including both typical and atypical traffic situations.

The comparison also points out that models trained on incident-excluded data generally perform poorly when tested on incident-specific scenarios, with a MAPE of 38.56%. This result indicates that incident-excluded models may lack the robustness needed to handle the disruptions caused by traffic incidents. The inability to effectively predict traffic flows during such disruptions could significantly impact applications requiring real-time traffic management and incident response, where accurate predictions are crucial.

Six-step ahead prediction

Similarly, **Table 10** and **Table 11** showcases the MAPE for 6-step ahead traffic flow predictions, highlighting the best performance for each scenario in bold numbers. The findings are similar to those observed in the one-step model analysis. As seen in **Table 11**, the lowest Mean Absolute Percentage Error (MAPE) values occur along the diagonal elements, indicating that models perform best when trained and tested under consistent conditions. This reinforces the importance of including incident data in training datasets to improve model robustness across various scenarios.

In summary, findings from one-step and multi-step ahead prediction performance comparison highlight the importance of carefully selecting training datasets that reflect the conditions the model is expected to predict. While models can achieve lower error rates when trained and tested on similar data types, they may struggle with generalization if exposed to different scenarios during testing. This analysis underscores the need for balanced and comprehensive datasets that include a wide range of traffic conditions to

Table 10 MAPE (%) performance on different scenarios for six-step ahead prediction; red bold numbers denote the best model for each scenario.

MAPE (%)			Testing											
			Incident-included				Incident-excluded				Incident-specific			
			GRU	SVR	RW	HA	GRU	SVR	RW	HA	GRU	SVR	RW	HA
Training	Incident-included	I24-58.5E	12.90	11.12	11.13	13.53	12.31	10.55	10.57	12.94	20.32	18.21	18.14	21.02
		I65-91.2S	17.58	13.50	13.64	16.88	16.96	12.64	12.80	15.77	27.60	27.66	27.44	35.22
		SR386-4.1E	23.03	20.37	20.57	22.98	22.66	19.99	20.19	22.52	71.85	71.29	70.70	83.93
		Average	17.84	15.00	15.11	17.80	17.31	14.39	14.52	17.07	39.92	39.06	38.76	46.72
	Incident-excluded	I24-58.5E	15.35	13.12	11.13	13.53	12.19	10.13	10.57	12.94	26.41	22.77	18.14	21.02
		I65-91.2S	20.56	16.07	13.64	16.88	16.45	12.64	12.80	15.77	34.78	35.13	27.44	35.22
		SR386-4.1E	26.94	23.43	20.57	22.98	22.21	18.99	20.19	22.52	91.25	91.97	70.70	83.93
		Average	20.95	17.54	15.11	17.80	16.95	13.92	14.52	17.07	50.81	49.96	38.76	46.72
	Incident-specific	I24-58.5E	41.80	36.03	11.13	13.53	42.71	37.99	10.57	12.94	16.05	13.30	18.14	21.02
		I65-91.2S	58.00	43.61	13.64	16.88	59.54	42.97	12.80	15.77	21.81	20.75	27.44	35.22
		SR386-4.1E	76.00	66.21	20.57	22.98	78.86	69.55	20.19	22.52	51.73	49.90	70.70	83.93
		Average	58.60	48.61	15.11	17.80	60.37	50.17	14.52	17.07	29.86	27.98	38.76	46.72

Table 11 Average model performance for six-step prediction.

MAPE (%)		Testing		
		Incident-included	Incident-excluded	Incident-specific
Training	Incident-included	16.42	15.85	39.49
	Incident-excluded	19.25	15.43	50.38
	Incident-specific	53.61	55.27	28.92

ensure the model's robustness and adaptability across various real-world situations.

However, it's important to address the practical limitations we face in real-world applications of traffic prediction models. In real practice, we often lack the time and resources to verify the occurrence of incidents in real-time situations where quick predictions are crucial. This constraint makes it challenging to implement strategies that involve filtering out abnormal samples or meticulously curating training datasets based on incident occurrences.

Given these practical considerations, the most recommended approach for handling training data is to avoid over-manipulating or selectively filtering it. Instead, it is more pragmatic to use the available data in its entirety, including both regular and abnormal samples. This approach ensures that the models are exposed to a wide range of traffic conditions, which is critical for developing robust and adaptable models capable of handling unexpected disruptions. By training on comprehensive datasets, we can better prepare models for real-world scenarios, where the ability to generalize from diverse data is more valuable than perfect accuracy in narrowly defined conditions. This strategy not only saves time and resources but also aligns more closely with the realities of traffic management and prediction in dynamic environments.

Conclusions

This paper presents a comprehensive study on the importance of incorporating incident data into the training and testing phases of traffic flow prediction models. By proposing a novel training and testing framework and evaluating the performance of models such as GRU, SVR, Historical Average (HA), and Random Walk (RW) using traffic volume data from radar detectors in Nashville, TN, this research highlights the impact of including incident data on prediction accuracy and robustness. Through a detailed comparative analysis across various scenarios — with-incident, without-incident, and incident-specific—the study identifies critical findings that underscore the necessity of incorporating diverse traffic conditions in the training process.

The analysis revealed several key insights. Firstly, models trained on non-filtered data consistently outperformed those trained on filtered data across all test scenarios. This

improvement was most pronounced in the “with incident” and “incident specific” scenarios, where the non-filtered models demonstrated superior adaptability to traffic anomalies and disruptions. Specifically, MAPE values for “incident specific” predictions improved by an average of 25%, while “with incident” scenario predictions saw an average improvement of 10%. This suggests that incorporating diverse traffic conditions, including those influenced by incidents, provides the model with a broader learning base, enhancing its ability to generalize and handle atypical traffic patterns effectively.

Secondly, while the non-filtered training approach showed benefits for predicting traffic on regular weekdays, it is important to consider that performance metrics can sometimes be misleading. Including a large number of testing data points can dilute the perceived improvement. For instance, the inclusion of incident days in the training data contributed to a 10% average improvement in MAPE for typical days with incident, highlighting the value of exposing predictive models to diverse traffic conditions. However, in some cases, simpler models like Random Walk (RW) can outperform more sophisticated models like GRU and SVR in incident-specific scenarios when trained without incident data. This phenomenon underscores the importance of carefully interpreting performance metrics and ensuring that models are robustly tested across varied conditions to truly assess their predictive capabilities.

The study’s findings have significant implications for traffic management and predictive analytics. By demonstrating that models trained on more inclusive datasets can provide superior performance, the research advocates for the integration of diverse traffic conditions in model training. This approach not only enhances prediction accuracy but also prepares models to handle unexpected traffic scenarios, contributing to more effective traffic management solutions.

Future research could extend these findings by exploring the optimal balance between filtered and non-filtered data to further refine traffic prediction models. Additionally, incorporating other types of data, such as weather conditions and special events, could provide even greater context and improve model performance. Continuous updates and real-time data integration could also enhance predictive capabilities, making traffic management systems more responsive and adaptable.

In conclusion, the study underscores the critical importance of using comprehensive and representative training data in traffic volume prediction models. The findings consistently show that models trained with a mix of incident and non-incident data are more robust and accurate across diverse traffic conditions. This approach ensures that the models are better equipped to handle both typical and atypical scenarios, including unexpected disruptions caused by incidents.

The analysis revealed that the lowest Mean Absolute Percentage Error (MAPE) values are observed when the training and testing data conditions align, as seen in the diagonal elements of the performance tables. This indicates that maintaining consistency between training and testing datasets can significantly enhance model accuracy. Moreover, the study highlights that even simpler models, such as Random Walk (RW), can outperform more complex models in certain scenarios, particularly when dealing with incident-specific data. This finding emphasizes the importance of including baseline models like RW in evaluations to provide a benchmark for assessing the performance of more sophisticated approaches.

Furthermore, the inclusion of incident data in the training process proves to be crucial for developing models that are resilient and adaptable to real-world conditions. It suggests that rather than filtering out anomalies or focusing solely on regular traffic patterns, leveraging all available data—including incident days—provides a more realistic and comprehensive understanding of traffic dynamics. This holistic approach not only improves the predictive capabilities of the models but also supports better traffic flow management and planning in urban environments. Ultimately, the study advocates for a shift towards more inclusive training datasets, which can lead to more effective and reliable traffic prediction systems, enhancing overall urban mobility and safety.

References

- [1] Williams, B. M., and L. A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, Vol. 129, No. 6, 2003, pp. 664-672.
- [2] Zhao, Z., W. Chen, X. Wu, P. C. Chen, and J. Liu. LSTM network: a deep learning approach for short - term traffic forecast. *IET Intelligent Transport Systems*, Vol. 11, No. 2, 2017, pp. 68-75.
- [3] Huang, W., G. Song, H. Hong, and K. Xie. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 5, 2014, pp. 2191-2201.
- [4] Niu, X., Y. Zhu, and X. Zhang. DeepSense: A novel learning mechanism for traffic prediction with taxi GPS traces. In *2014 IEEE global communications conference*, IEEE, 2014. pp. 2745-2750.
- [5] Lv, Y., Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 2, 2014, pp. 865-873.
- [6] Zhao, L., Y. Zhou, H. Lu, and H. Fujita. Parallel computing method of deep belief networks and its application to traffic flow prediction. *Knowledge-Based Systems*, Vol. 163, 2019, pp. 972-987.
- [7] Ma, X., H. Yu, Y. Wang, and Y. Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, Vol. 10, No. 3, 2015, p. e0119044.
- [8] Leelavathi, M., and K. Sahana Devi. An architecture of deep learning method to predict traffic flow in big data. *Int. J. Res. Eng. Technol.*, Vol. 5, No. 4, 2016, pp. 461-468.
- [9] Yang, H.-F., T. S. Dillon, and Y.-P. P. Chen. Optimized structure of the traffic flow forecasting model with a deep learning approach. *IEEE transactions on neural networks and learning systems*, Vol. 28, No. 10, 2016, pp. 2371-2381.
- [10] Duan, Y., Y. Lv, and F.-Y. Wang. Performance evaluation of the deep learning approach for traffic flow prediction at different times. In *2016 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, IEEE, 2016. pp. 223-227.

- [11] Polson, N. G., and V. O. Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 79, 2017, pp. 1-17.
- [12] Chen, D. Research on traffic flow prediction in the big data environment based on the improved RBF neural network. *IEEE Transactions on Industrial Informatics*, Vol. 13, No. 4, 2017, pp. 2000-2008.
- [13] Kunde, F., A. Hartenstein, S. Pieper, and P. Sauer. Traffic Prediction using a Deep Learning Paradigm. In *EDBT/ICDT Workshops*, 2017. p. 4.
- [14] Lin, Y., X. Dai, L. Li, and F.-Y. Wang. Pattern sensitive prediction of traffic flow based on generative adversarial framework. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 6, 2018, pp. 2395-2400.
- [15] Koesdwiady, A., R. Soua, and F. Karray. Improving traffic flow prediction with weather information in connected cars: A deep learning approach. *IEEE Transactions on Vehicular Technology*, Vol. 65, No. 12, 2016, pp. 9508-9517.
- [16] Soua, R., A. Koesdwiady, and F. Karray. Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. In *2016 International joint conference on neural networks (IJCNN)*, IEEE, 2016. pp. 3195-3202.
- [17] Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [18] Zhang, D., and M. R. Kabuka. Combining weather condition data to predict traffic flow: a GRU - based deep learning approach. *IET Intelligent Transport Systems*, Vol. 12, No. 7, 2018, pp. 578-585.
- [19] Jia, Y., J. Wu, and M. Xu. Traffic flow prediction with rainfall impact using a deep learning method. *Journal of advanced transportation*, Vol. 2017, No. 1, 2017, p. 6575947.
- [20] Hochreiter, S., and J. Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, 1997, pp. 1735-1780.
- [21] Yu, R., Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu. Deep learning: A generic approach for extreme condition traffic forecasting. In *Proceedings of the 2017 SIAM international Conference on Data Mining*, SIAM, 2017. pp. 777-785.
- [22] Yu, H., Z. Wu, S. Wang, Y. Wang, and X. Ma. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors*, Vol. 17, No. 7, 2017, p. 1501.

- [23] Chen, Y.-y., Y. Lv, Z. Li, and F.-Y. Wang. Long short-term memory model for traffic congestion prediction with online open data. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2016. pp. 132-137.
- [24] Liu, Y., H. Zheng, X. Feng, and Z. Chen. Short-term traffic flow prediction with Conv-LSTM. In *2017 9th international conference on wireless communications and signal processing (WCSP)*, IEEE, 2017. pp. 1-6.
- [25] Wu, Y., H. Tan, L. Qin, B. Ran, and Z. Jiang. A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, Vol. 90, 2018, pp. 166-180.
- [26] Mackenzie, J., J. F. Roddick, and R. Zito. An evaluation of HTM and LSTM for short-term arterial traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 5, 2018, pp. 1847-1857.
- [27] Tian, Y., K. Zhang, J. Li, X. Lin, and B. Yang. LSTM-based traffic flow prediction with missing data. *Neurocomputing*, Vol. 318, 2018, pp. 297-305.
- [28] Yang, B., S. Sun, J. Li, X. Lin, and Y. Tian. Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing*, Vol. 332, 2019, pp. 320-327.
- [29] Avila, A. M., and I. Mezić. Data-driven analysis and forecasting of highway traffic dynamics. *Nature communications*, Vol. 11, No. 1, 2020, p. 2090.
- [30] Li, L., Y. Li, and Z. Li. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, Vol. 34, 2013, pp. 108-120.
- [31] Miglani, A., and N. Kumar. Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. *Vehicular Communications*, Vol. 20, 2019, p. 100184.
- [32] Ma, X., Z. Tao, Y. Wang, H. Yu, and Y. Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, Vol. 54, 2015, pp. 187-197.
- [33] Chung, J., C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [34] Van Lint, J., and C. Van Hinsbergen. Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues*, Vol. 22, No. 1, 2012, pp. 22-41.

CHAPTER III
REPRESENTING AND INFERRING NETWORK TRAFFIC
CONDITION: A METHOD INTEGRATING CLUSTERING INTO
CUR DECOMPOSITION

Abstract

Intelligent transportation systems (ITSs) usually require monitoring of massive road networks and gathering traffic data at a high spatial and temporal resolution. This leads to the accumulation of substantial data volumes, necessitating the development of more concise data representations. Approaches like principal component analysis (PCA), which operate within subspaces, can construct precise low-dimensional models. However, interpreting these models can be challenging, primarily because the principal components often encompass a multitude of links within the traffic network. To overcome this issue, this study presents a novel approach for representing and inferring network traffic conditions through weighted CUR matrix decomposition integrated with clustering analysis. The proposed approach selects a subset group of detectors from the original network to represent and infer traffic condition through a matrix decomposition method, allowing for more efficient management and analysis. The proposed method is evaluated using traffic detector data from the city of Nashville, TN. The results demonstrate that the approach is effective in representing and inferring network traffic conditions, consistently outperforming prevailing methods. Overall, this study contributes to the field of network traffic monitoring by proposing a novel approach for representing traffic networks and exploring the effects of incorporating clustering into CUR decomposition. The proposed approach can help traffic analysts and practitioners to more efficiently manage and analyze traffic conditions, ultimately leading to more effective transportation systems.

Introduction

With the rapid development of intelligent transportation systems (ITSs) in the big data era, a huge amount of up-to-date data are being collected, archived, and analyzed from a variety of sources such as smart phones, probe vehicles, video cameras, and infrastructure-based detectors. Traditional detector loops and more advanced technologies, such as radar and microwave detectors, have been widely deployed in large metropolitan areas to monitor the traffic flow in a real-time manner. These systems deal with thousands of detector stations with high temporal resolution at 20 to 30 s per update, which poses great challenges for the efficiency and computational cost of analyzing ITS data. One of the main challenges is to assess the road network condition quickly and accurately for a given moment and efficiently infer and forecast meaningful spatial and temporal trends for the future, which can be useful for many ITS applications such as monitoring networks, planning traffic, and mitigating congestion (1; 2).

Previous research has aimed to model the road network by considering every road segment within it. However, this approach may not be practical for large traffic networks and real-time applications. Furthermore, missing data will significantly impair the effectiveness of this method. To address this issue, our focus is on creating low-dimensional network models that only require monitoring of a selected subset of road segments.

To develop accurate low-dimensional models for large and diverse road networks, prior research has primarily focused on techniques such as principal component analysis (PCA) (3-7). However, PCA models are difficult to interpret in terms of individual links in the network. Djukic et al. used PCA to analyze OD pair data in a small network (4; 5), while Asif et al. applied various subspace methods to compress traffic speed data (8). These studies demonstrated that subspace methods like PCA and DCT (discrete cosine transform) can effectively reduce the size of traffic data. However, they do not provide detailed insights into traffic patterns for specific roads and time periods. In addition, implementing PCA online still requires information from all detectors in the network because the principal component is basically a linear combination of all columns of the

original data, which greatly reduces the efficiency of data processing. This is the reason why PCA is mostly used offline for dimension reduction purposes.

In contrast, the CUR decomposition (9) considers selecting subsets from individual links and time instances, enabling us to directly extract underlying spatial and temporal patterns in large road networks. In 2013, Mitrovic et al. explored the use of the CUR matrix decomposition method for compressing and sensing traffic speed data. They compared this method to PCA and showed that the resulting low-dimensional models from CUR are much more easily interpretable. They also demonstrated how CUR can be used for compressed sensing of traffic data (2). In 2015, Mitrovic et al. employed column-based (CX) low-dimensional models based on their previous research to improve the scalability of compressed sensing and prediction. The researchers broke down the compressed prediction error into different parts and examined how they related to each other, and their numerical findings demonstrate that this approach considerably lowers computational expenses with little effect on prediction accuracy (10). The benefits of the CUR method are achieved with a trade-off of greater prediction errors as the compression ratio (the ratio of original number of columns to number of columns in a low-dimensional representation) increases. Nevertheless, there is significant room for improving the process of selecting columns based on statistical leverage (9) calculated from singular value decomposition (SVD) (11). Another issue is that the author only tested a limited range of compression ratios from 2 to 10 (2; 10), which is not sufficient to depict the performance of the proposed method. In addition to leverage-based column selection, Couras et al. proposed (12) an algorithm to perform the approximation of the tensors based on the CX decomposition for matrices. Han and Huang (13) proposed a road network compression method to improve the efficiency of data processing based on correlation analysis and CX decomposition, which is then integrated into a deep learning network to predict the short-term traffic flow.

Apart from matrix decomposition, transformation techniques such as discrete Fourier transform (DFT) and discrete wavelet transform (DWT), which are commonly used for compressing signals and images, can be adapted for spatiotemporal data. In the context of road traffic data compression, methods involving DCT and SVD are utilized,

leveraging Kronecker product and tensor decomposition (14). Additionally, by organizing the data in a multidimensional format, algorithms based on DWT are integrated to achieve dimensionality reduction (15). To harness the benefits provided by signal processing methods from other data types, the field of graph signal processing (GSP) (16) has been developed to conceptualize spatiotemporal data as a two-dimensional graph signal and establishes operations like linear filtering and linear prediction. Furthermore, Chindanur et al. (1) adopted graph Fourier transform (GFT) and achieved less than one percent reconstruction error (RE) on California's Interstate 605 (I-605) freeway data. Although the GFT-based method outperformed other methods, the problems are twofold: (1) it is difficult to interpret in terms of individual detectors, and (2) the model is too complex, possessing a mathematical form which is unlikely to be deployed and understood by practitioners in real world situations.

To summarize, previous studies have focused on the PCA approach, matrix decomposition methods, and signal transformation techniques to compress large datasets and represent the overall network condition using low-dimensional approximations. Three major research gaps were identified, as follows. First, the CUR decomposition method is widely used due to its strong interpretability, whereas the approach to selecting the optimal subset of columns remains to be improved. To this end, the clustering method is explored in this study to enhance the representing performance. Second, most previous studies failed to compare their method of selecting columns with a random sampling approach with equal probabilities. Finally, most studies only evaluated their algorithms in a limited range of compression ratios, but more interesting insights could be discovered if more cases, especially those under a high compression ratio, are tested.

To address these issues, this paper aims to utilize a column-based CUR matrix decomposition technique integrated with a clustering method to represent the original traffic network as a smaller subset of original columns with acceptable errors. The experiment was conducted using radar detector outputs (speed) aggregated at 5 min resolution from Nashville, TN. A group of representative detectors are selected using the proposed weighted average method to establish a relationship with the original network by analyzing the historical data and calculating the relationship matrix offline. In

addition, this study incorporates clustering analysis into CUR decomposition inspired by the methodology of computing stock indexes such as the S&P 500 and Nasdaq. Analogous to selecting stocks from different sectors to represent the whole market, a clustering analysis was first conducted, followed by the normal CUR decomposition within each cluster, and results from each cluster would be merged as the final outputs. In this regard, we also reviewed several papers applying clustering algorithms to tackle problems in the traffic domain. Nguyen et al. (17) applied clustering algorithms to obtain labels automatically from the data and presented the results of clustering analysis using both point-based and area-based features, highlighting the superiority of the area-based approach in producing meaningful clusters. Cheng et al. (18) proposed an improved fuzzy c-means clustering method to classify urban traffic states with real world traffic flow data. Chen et al. (19) utilized dynamic time warping (DTW) k-means clustering to classify lane-changing risk profiles into several categories. In the traffic domain, clustering methods are mostly used to categorize time series over different locations.

Related Work

This section presents a literature review of the research problem, mainly focusing on the dimension reduction of traffic network. **Table 12** shows all the abbreviations and symbols used in this article.

Prior research on low-dimensional representations for traffic networks has primarily focused on the use of PCA, as evidenced by previous studies (4-7). Djukic et al. used PCA to analyze OD pair data in a small network (4; 5), while Asif et al. applied various subspace methods to compress traffic speed data (8). These studies demonstrated that subspace methods like PCA and DCT can effectively reduce the size of traffic data. However, they do not provide detailed insights into traffic patterns for specific roads and time periods. In addition, implementing PCA online still requires information from all detectors in the network because the principal component is basically a linear combination of all columns of the original data. This is the reason why PCA is mostly used offline for dimension reduction purposes. In contrast, the CUR decomposition considers data from individual links and time instances, enabling us to directly extract

Table 12 List of Acronyms and Symbols.

Abbreviation	Definition
CR	Compression rate
CX	Column-based
DBSCAN	Density-based spatial clustering of applications with noise
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DTW	Dynamic time warping
DWT	Discrete wavelet transform
GFT	Graph Fourier transform
GSP	Graph signal processing
ITS	Intelligent transportation system
OD	Origin–destination
PCA	Principal component analysis
PRD	Percent-root-mean-square distortion
RDS	Radar detector system
RE	Reconstruction error
SVD	Singular value decomposition
TN	Tennessee
$a_{i,j}$	The i^{th} row and j^{th} column of matrix A
A	The original data matrix
$\hat{\mathbf{A}}$	The low-rank approximation of original matrix
$\mathbf{A}_k$	The best rank-k approximation of matrix A
C	The selected submatrix
$\mathbf{C}^+$	The Moore–Penrose pseudo-inverse of matrix C
p_j^{L2}	The L2 norm-based importance score for column j
p_j^{SVD}	The SVD leverage-based importance score for column j
p_j^{W}	The proposed weighted importance score for column j
$\mathbf{S}_i$	The i^{th} cluster
U	The unitary matrix with columns being left singular vectors of A
V	The unitary matrix with columns being right singular vectors of A
$\mathbf{v}_j^\xi$	The j^{th} coordinate of ξ^{th} right singular vector
X	The relationship matrix
x	The data point
Σ	The rectangular diagonal matrix with non-negative diagonal entries being the singular values of matrix A
μ_i	The mean (centroid) of points in cluster $\mathbf{S}_i$
$\ \cdot\ _{\text{F}}$	The Frobenius norm of a matrix

underlying spatial and temporal patterns in large road networks.

In 2013, Mitrovic et al. explore the use of the CUR matrix decomposition method for compressing and sensing traffic speed data. They compare this method to Principal Component Analysis (PCA) and show that the resulting low-dimensional models from CUR are more easily interpretable. They also demonstrate how CUR can be used for compressed sensing of traffic data. The paper includes numerical results analyzing the performance of the proposed CUR method for compression and compressive sensing of large-scale traffic data, using a specific traffic network as an example (2). The data they used was network traffic speed. By contrast, this paper utilized traffic flow data and explored ways of representing and indexing the network traffic from a more practical way.

In 2015, Mitrovic et al. employed column-based low-dimensional models (CX) to improve the scalability of compressed sensing and prediction. The researchers broke down the compressed prediction error into different parts and examined how they related to each other and their numerical findings demonstrate that this approach considerably lowers computational expenses with little effect on prediction accuracy (10).

Materials and Methods

This section mainly introduces the data, study area, and the proposed framework.

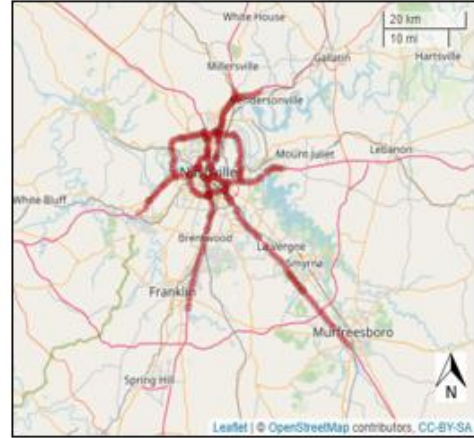
Data and study area

The radar detector system (RDS) in Nashville consists of 349 detectors covering 555 directional links. The study area and geographic locations of the detectors used in this study are displayed in **Figure 13**. **Figure 13(a)** displays where Nashville, TN, is in the US, and **Figure 13(b)** shows the location of all detectors in Nashville, TN.

The objective of this study is to represent the traffic condition of a massive network, and traffic speed is one of the most widely used and most intuitive characteristics to reflect the traffic condition at a specific location. Thus, traffic speed data were sampled from July 9, 2023, to July 29, 2023. The first two weeks of data were used as training data to select the most representative detectors and learn the relationship matrix, while the remaining one week was used to evaluate the performance of the



(a)



(b)

Figure 13 (a) Location of Nashville, TN, on US map; (b) RDS detector locations in Nashville, TN.

proposed framework on new data. **Figure 14** shows the input data matrix.

The raw data, updated each 30 seconds, was first aggregated to 5 min to reduce randomness and variability, which will greatly reduce the computational cost as well.

CUR Decomposition

This section will briefly describe the details of different CUR decomposition methods and present the proposed column-based CUR decomposition algorithm integrated with clustering method.

As shown in **Figure 15**, let $\mathbf{A} \in \mathbb{R}^{t \times n}$ denote the original data matrix with t time instances in row and n detector links in column. Firstly, we define a compression rate as the ratio of number of selected columns to the total number of columns in matrix $\mathbf{A}$ ($CR = c/n$). The objective of column-based CUR is to find a submatrix $\mathbf{C} \in \mathbb{R}^{t \times c}$ consisting of c columns of $\mathbf{A}$ to create a low-rank approximation $\hat{\mathbf{A}}$ as shown in Equation (1):

$$\mathbf{A} \approx \hat{\mathbf{A}} = \mathbf{C}\mathbf{X} = \mathbf{C}\mathbf{C}^+\mathbf{A},$$

where

- $\mathbf{C}^+$ is Moore–Penrose pseudo-inverse of matrix $\mathbf{C}$ (20).
- $\mathbf{X} = \mathbf{C}^+\mathbf{A}$ ($\mathbf{X} \in \mathbb{R}^{c \times n}$) is the relationship matrix which projects the selected columns back onto all the columns in original data space (21). For the given matrices $\mathbf{A}$ and $\mathbf{C}$, the relationship matrix is computed as the matrix product of $\mathbf{C}^+$ and $\mathbf{A}$.

The crucial problem with CUR method is how to select the most representative columns so that the relative error between approximation matrix and original matrix is as small as possible. The goal is to capture the essential information in the data while reducing dimensionality. To evaluate the approximation result, we calculated the percent-root-mean-square distortion (PRD), which is commonly used to assess reconstruction performance (22), provided by the following formula:

$$PRD (\%) = \frac{\|\mathbf{A} - \mathbf{C}\mathbf{X}\|_F}{\|\mathbf{A}\|_F},$$

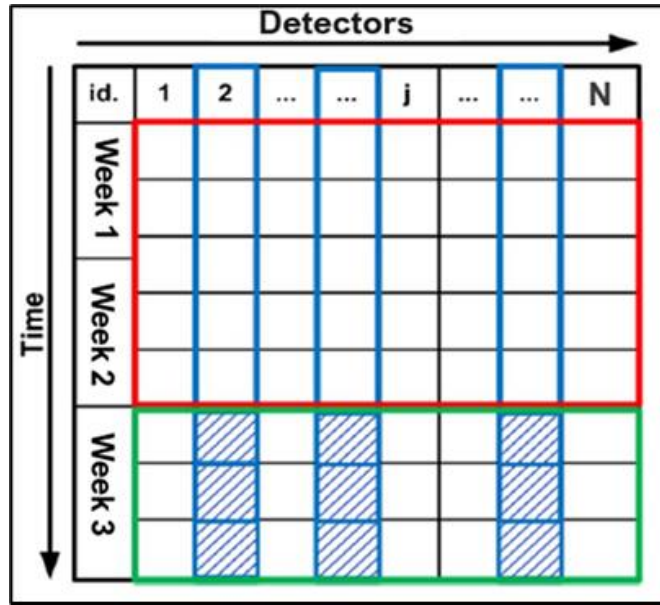


Figure 14 Input data matrix with training period in red rectangle and testing period in the green rectangle; columns in blue rectangles represent the selected group of detectors and blue shaded areas are data entries for testing period at selected detectors.

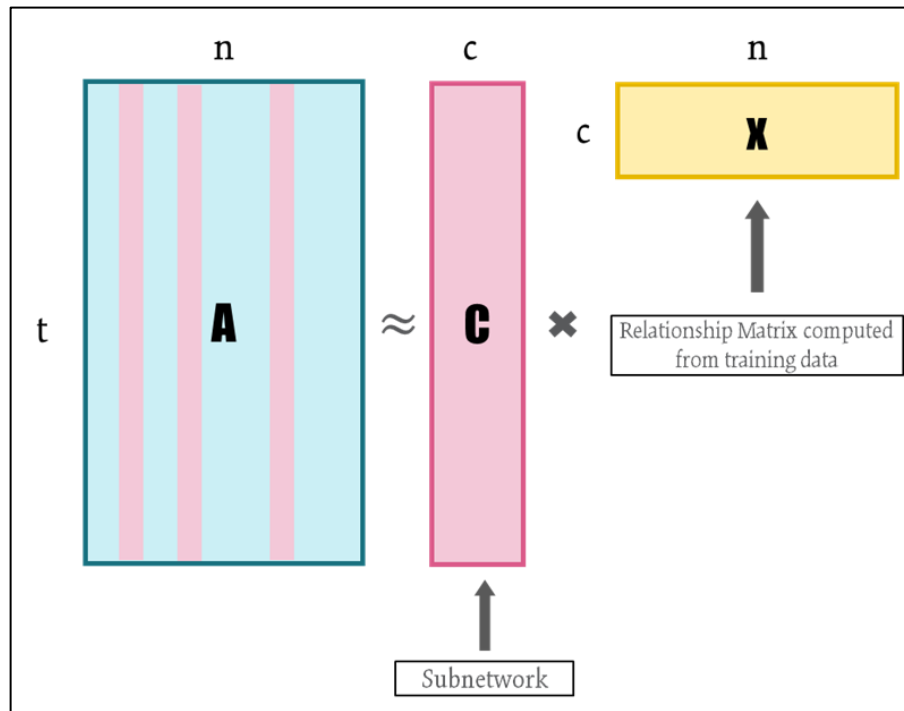


Figure 15 CUR decomposition.

where,

$$\|\mathbf{A}\|_F = \left(\sum_i \sum_j a_{ij}^2 \right)^{\frac{1}{2}}$$

And it represents the Frobenius norm of matrix $\mathbf{A}$. A lesser PRD represents a better compression result.

More specifically, we can break down the column-based CUR method into the following steps:

1. Determine the importance score of each column in the network matrix. Each column in the matrix $\mathbf{A}$ is assigned a probability score indicating its likelihood of being selected;
2. Create matrix $\mathbf{C}$ by selecting top c columns from matrix $\mathbf{A}$ in the descending order of probability scores calculated in the first step;
3. Calculate the relationship matrix $\mathbf{X} = \mathbf{C}^+ \mathbf{A}$ ($\mathbf{X} \in \mathbb{R}^{c \times n}$);
4. Assume traffic is stationary, infer the future traffic speed for the entire network using only speed measurements at the selected link, and calculate PRD for the testing data as the performance measurement.

Step 1 is clearly the most essential step in the CUR process. To identify the optimal set of columns for a specified number of columns, one would normally need to evaluate all possible $\binom{n}{c}$ combinations. Nevertheless, employing a brute-force approach entails a computational complexity of $O(n^c)$ (23). Given this computational burden, it is usually impractical to assess every potential selection of c columns. To tackle this challenge, two major randomized algorithms have been introduced (21; 24) to compute the importance score of each column and then sample columns based on the score:

- **L2 norm-based:** this method calculates the square of L2 norm for each column and divides it by the sum of squares of all entries of the matrix, expressed as the following equation:

$$p_j^{L2} = \frac{\sum_{i=1}^t A(i, j)^2}{\sum_{i=1}^t \sum_{j=1}^n A(i, j)^2},$$

where $j = 1, 2, \dots, n$. The concept behind this method is to select detector links

with high speed. L2 norm-based sampling offers benefits like rapid computation and a better understanding of the column magnitudes. The drawback lies in the neglect of detectors that record a significant portion of low-speed intervals, leading to substantial estimation errors during congestion periods.

- **Leverage-based:** this method utilizes statistical leverage (9), which measures the contribution of each column to the overall variance of the data, as the importance score for each column. The importance score can be conceptually understood as quantifying the “statistical leverage” or “impact” of a specific column on achieving the most accurate low-rank approximation of the data matrix. By prioritizing the selection of columns that has a disproportionately significant influence over the low-rank approximation, as opposed to L2 norm method that samples columns with higher Euclidean distances, we can ensure that CUR performs nearly as effectively as the best rank- k approximation $\mathbf{A}_k$ in capturing the predominant portion of the spectrum of $\mathbf{A}$ (9). Columns with high leverage scores are often considered important. The underlying concept is to consider detector links with large variations in speed, thus covering various traffic conditions and providing a better capture of the detectors with different traffic states. However, the limitations associated with this approach include the substantial computational cost. The detailed steps are described as follows. Firstly, singular value decomposition (SVD) is performed on the original data matrix $\mathbf{A}$ (11) to represent $\mathbf{A}$ as:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{t \times t}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are unitary matrices, and the columns of $\mathbf{U}$ and $\mathbf{V}$ are left and right singular vectors of $\mathbf{A}$, respectively. Matrix $\mathbf{\Sigma} \in \mathbb{R}^{t \times n}$ is a rectangular diagonal matrix with non-negative diagonal entries, known as the singular values of matrix $\mathbf{A}$. Then, the best rank- k approximation can be obtained by keeping top k columns of $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ so that the explained variance is at least 80%:

$$\bullet \quad \mathbf{A}_{t \times n} = \mathbf{U}_{t \times k} \mathbf{\Sigma}_{k \times k} \mathbf{V}_{k \times n}^T.$$

Then, the leverage score of column j can be calculated as:

$$p_j^{SVD} = \frac{1}{k} \sum_{\xi=1}^k \left(\mathbf{v}_j^{\xi} \right)^2,$$

where $\mathbf{v}_j^{\xi}$ is the j^{th} coordinate of ξ^{th} right singular vector. It is obvious that the sum of p_j is equal to one because matrix $\mathbf{V}$ is unitary.

Based on the two previous options, we propose a weighted method considering both L2 norm and statistical leverage, expressed as Equation (8):

$$p_j^W = w \cdot p_j^{L2} + (1 - w) \cdot p_j^{SVD},$$

where w is a weight parameter, and w is set to be 0.5 in this study to assign equal weights to column magnitude and column variation.

In addition to the three options, a random sampling method to select c columns with equal probability for each column is also implemented. This process was repeated five times, and average performance was recorded. Prior research failed to consider the comparison with the simplest random sampling method, thus making the performance less persuasive.

CUR Integrated with Clustering

The primary objective of this study is to explore the effects of incorporating clustering method into CUR sampling process. The reason for considering clustering method is to investigate whether homogenous subnetworks can enhance the overall performance of the network representation. The idea of integrating clustering analysis into CUR method is inspired by the methodology of devising a composite index for stock market such as Nasdaq-100. There are two main approaches of integrating clustering into CUR decomposition process. Approach 1, which was also adopted by previous studies, is described as follows.

Clustering Approach 1

Stocks from different sectors will go through a rigorous screening process and then be selected to represent the whole market using a weighted average index. Similarly, to select the most representative traffic detectors in a road network, the first step is to filter out columns with missing rate greater than 5% and to impute the remaining columns

using data from adjacent time intervals (linear interpolation). Then, we can classify the detectors into different clusters based on their feature values during the training period, followed by the normal CUR decomposition process within each individual cluster given a fixed compression rate (CR). Finally, the selected columns and corresponding relationship matrix will be merged as the final CUR outputs. The entire column selection process is displayed in **Figure 16**.

Mitrovic et al. has taken into account the clustering method to compress large-scale traffic data (10); however, they simply clustered the network according to road category instead of using a real clustering method based on the speed condition on the road. It is obvious that roads in the same category do not necessarily have similar traffic conditions.

Clustering Approach 2

This clustering approach aims to enhance column selection for improved performance of the final CUR process. Unlike the first approach, the second step involves dividing the data into clusters equal to the number of columns to be selected. The column closest to the center of each cluster is then chosen to represent the entire cluster, followed by the CUR decomposition. **Figure 17** shows the entire framework of the second clustering approach.

In this study, k-means clustering was adopted due to its simplicity, computational efficiency, and interpretability. A detailed description of k-means clustering is provided as follows.

In k-means, clusters are defined so that the total intra-cluster variation (known as total within-cluster variation) is minimized. Within-cluster variation is defined as:

$\sum ||x - \mu_i||^2$, where we only consider x of a given cluster $\mathcal{S}_i$, and μ_i is the mean (centroid) of points in cluster $\mathcal{S}_i$. Then, total within-cluster variation is to perform the previous calculation for each cluster and take the sum of all clusters.

There are five overall steps in naïve k-means method:

- Step 1: specify k (number of clusters).
- Step 2: randomly select k instances from the data as the initial cluster centroids.

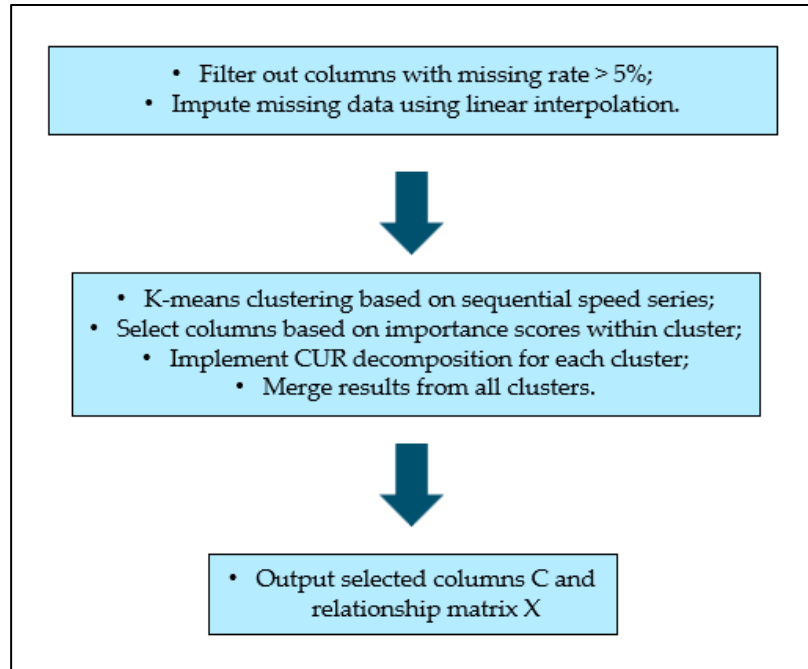


Figure 16 Column selection and decomposition process for clustering approach 1.

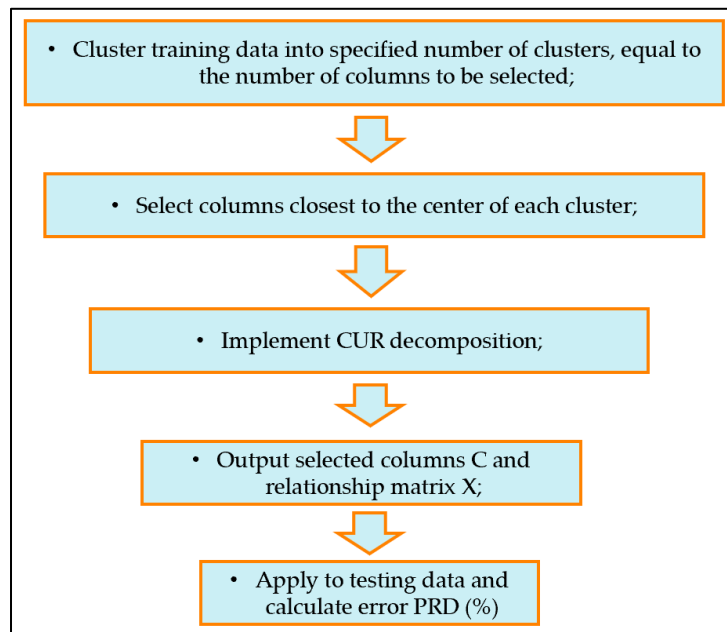


Figure 17 Column selection and decomposition process for clustering approach 2.

- Step 3: assign each instance to its closest centroid based on Euclidean distance.
- Step 4: for each of the k clusters, recompute the centroid by calculating the new mean of all the instances in the cluster.
- Step 5: repeat step 3 and 4 until centroids converge or the max number of iterations is reached.

Results

This section provides a detailed description of the results of all experiments. The first part of the study is representing the massive network using the CUR decomposition method based on the training data and inferring future conditions on the testing data. In addition, the second part is a new application scenario of indexing the network condition using the outputs from the CUR decomposition.

Weighted Average Method

Figure 18 shows the performance of the proposed weighted method compared with the L2 norm, SVD-based leverage, and random sampling methods. The red line in the graph indicates that the weighted method outperforms all other three options when the compression rate (CR) is lesser than 32 ($2^5 = 32$). The proposed weighted method resulted in lower PRD than other methods when the CR does not exceed 32, whereas the drawback of the weighted sampling method is relatively longer computational time. When the CR surpasses 32, the proposed weighted method is only better than the SVD-based leverage method, partly because the presetting default value of the weight (w) is 0.5, which provides too much weight for SVD-based leverage. Random sampling is even better than the L2 norm option when the CR is less than 64, which might be due to randomness and fluctuation. It is worth noting that SVD-based leverage and proposed weighted sampling perform best with a CR less than 8, whereas the error for the SVD-based leverage option increases dramatically as the CR becomes greater than 32.

Table 13 presents the summary statistics of the performance of the proposed method (weighted) compared to other three models (L2 norm, SVD-based leverage, and random sampling). It should be noted that random sampling is carried out to show the

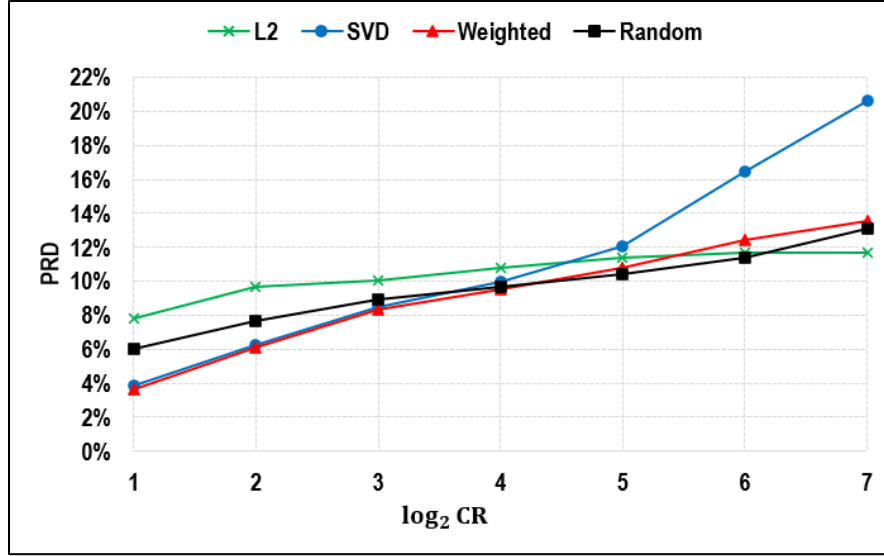


Figure 18 Performance of the proposed method without clustering.

Table 13 Comparison of the performance of different methods without clustering.

Compression Rate (CR)	PRD (%)				Computing Time (Milliseconds)			
	L2	SVD	Weighted	Random	L2	SVD	Weighted	Random
2	7.8%	3.8%	3.6%	5.8%	793.1	982.5	1075.5	645.9
4	9.7%	6.3%	6.1%	7.7%	692.0	870.9	964.9	562.6
8	10.1%	8.4%	8.3%	8.9%	705.0	899.3	939.8	534.3
16	10.8%	10.0%	9.5%	9.6%	648.2	860.0	973.0	512.7
32	11.4%	12.0%	10.8%	10.5%	691.7	915.1	939.8	512.0
64	11.7%	16.4%	12.5%	11.4%	631.2	835.1	940.4	567.0
128	11.7%	20.6%	13.5%	12.4%	624.3	882.4	905.8	522.4

minimal time needed to generate the decomposition results, since we assume that the time consumed by column selection could be ignored under a random sampling scenario.

However, random sampling is impractical because the performance fluctuates, and the selected columns are not fixed each time the sample is obtained. That is also why we repeated the process five times and used the average performance. Among the three methods, the L2 norm-based method is the most computation-friendly, and the proposed weighted method uses more computational resources than the other two methods. Overall, this is a trade-off between accuracy and computing cost.

Taking $CR = 16$ with the proposed weighted average method as an example, **Figure 19** depicts the absolute error of the original matrix and inferred matrix for the testing period (23 July 2023, to 29 July 2023). The horizontal axis represents the date and the vertical axis is the detectors. There are 31 links selected, and the $PRD = 9.5\%$. The absolute error is less than 10 mph for 94% of time and detectors. Some recurring patterns can be found, probably due to the recurring peak-hour congestion at some detector links.

Clustering Approach 1

Firstly, K-means clustering was implemented on the original data to classify detector links into different groups based on their speed measurements during the training period. **Figure 20(a)** shows the results of the elbow method which implied that two is more likely to be the optimal number of clusters, and **Figure 20(b)** displays the cluster results for all observations projected onto the dimension of two principal components. As can be seen, there are overlaps between the two clusters. **Figure 20(c)** displays the centroids of two clusters, respectively. The red line is the center of the first cluster, with speed around 70 mph, and the speed of the second cluster's centroids, shown in blue, fluctuates around 50 mph.

After clustering, the normal CUR process, with three different options, was conducted to explore the effects of incorporating clustering into CUR decomposition. **Figure 21** displays the comparison of the performance of the three options with and without clustering method 1. It was found that clustering almost failed to improve the performance for the SVD-based leverage and weighted average option, while the L2

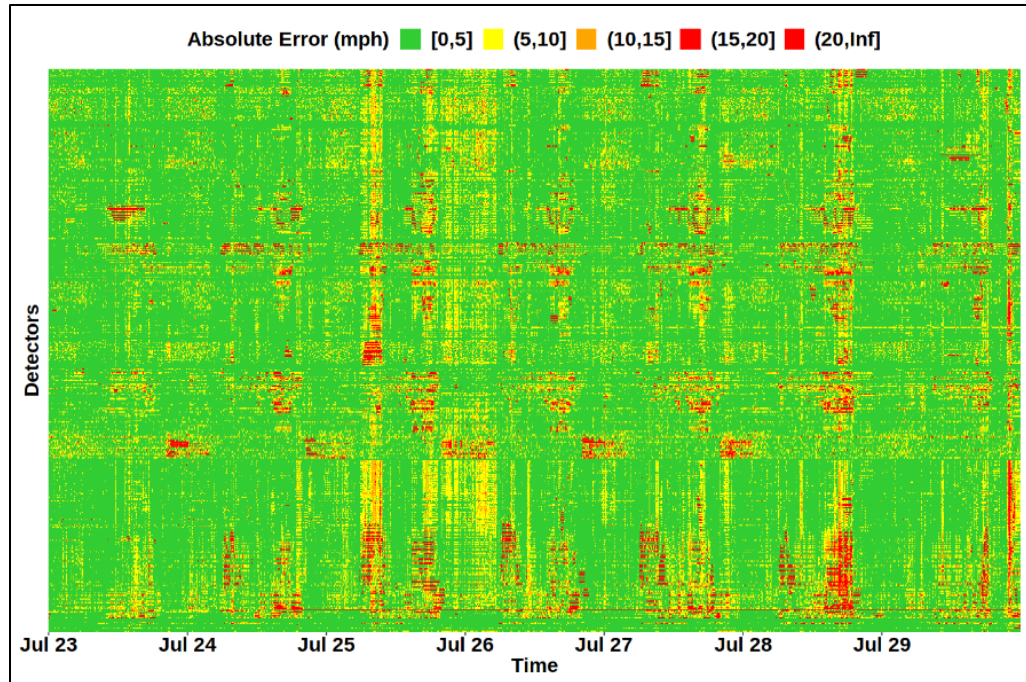
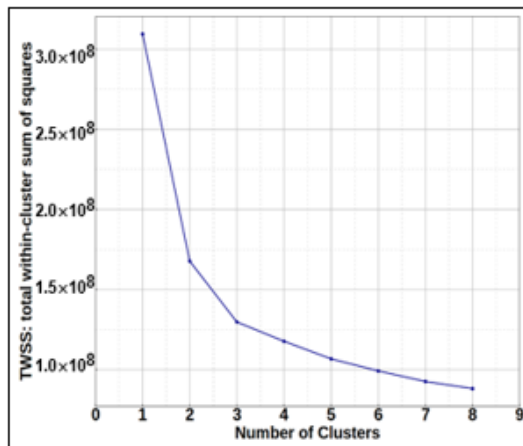
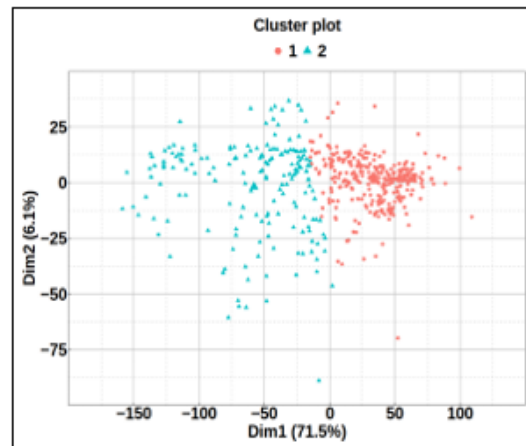


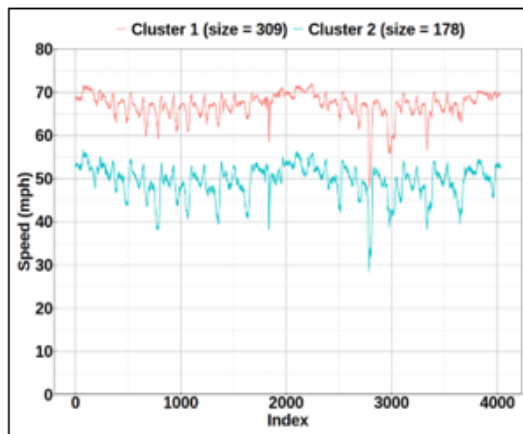
Figure 19 Performance heatmap of the proposed weighted method when CR = 16.



(a)

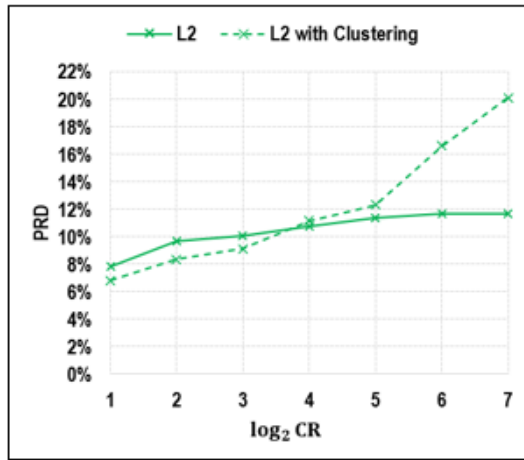


(b)

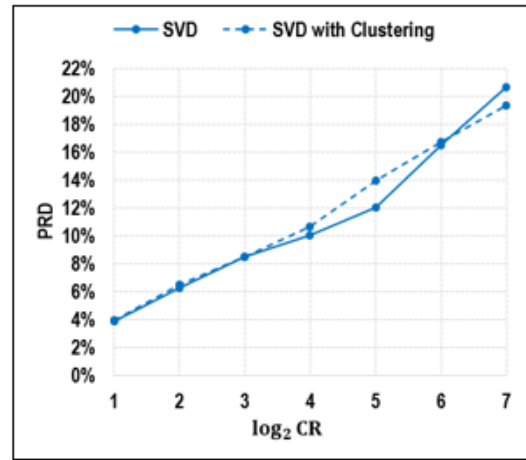


(c)

Figure 20 Clustering results: (a) elbow method showing total with-cluster sum of squares versus number of clusters; (b) clustering results showing two principal components of observations; (c) cluster centroids versus time index.



(a)



(b)



(c)

Figure 21 Comparison of three methods with and without clustering: (a) L2 norm-based importance; (b) SVD-based leverage importance; (c) weighted average of L2 norm and leverage.

norm-based option saw a minor enhancement when the CR was less than 16. This finding is somewhat intuitive and contradictory to Mitrovic's work (10), partly because Mitrovic only tested his method on CRs from 2 to 10. The reason why clustering failed to contribute will be discussed in the discussion section.

Clustering Approach 2

Clustering Approach 2 is an independent method for selecting representative columns, distinct from Approach 1, which combines L2 norms and SVD. This approach focuses solely on clustering data into a specified number of clusters, equal to the number of columns to be selected, and then choosing the column closest to the center of each cluster to represent the entire cluster. Analysis result illustrated in **Figure 22** and **Figure 23** has shown that the proposed clustering method consistently outperforms previous techniques, providing more accurate and reliable column selections. Furthermore, this approach has been applied to both speed and volume-to-capacity (v/c) ratio analyses, demonstrating its versatility and effectiveness in capturing key traffic flow characteristics across various metrics.

Notably, the Clustering method 2 exhibits the slowest growth rate in PRD among the methods. While the PRD for the L2, SVD, and Weighted methods increases significantly as the compression ratio increases, the Clustering method maintains a relatively lower and more gradual increase. This indicates that the Clustering method has better performance in maintaining lower PRD values, suggesting that it is more robust or efficient in preserving data accuracy under increasing compression compared to the other methods. This characteristic can be particularly beneficial in applications where maintaining data integrity is crucial despite aggressive compression.

These findings indicate that the new clustering approach is highly effective in selecting representative columns for the CUR decomposition, providing superior performance in both v/c ratio and speed metrics compared to other methods. Speed and v/c ratio are both critical metrics for analyzing traffic flow, but they serve different purposes and provide unique insights. Speed measures the rate at which vehicles travel along a roadway, reflecting the efficiency of traffic movement. In contrast, the v/c ratio is

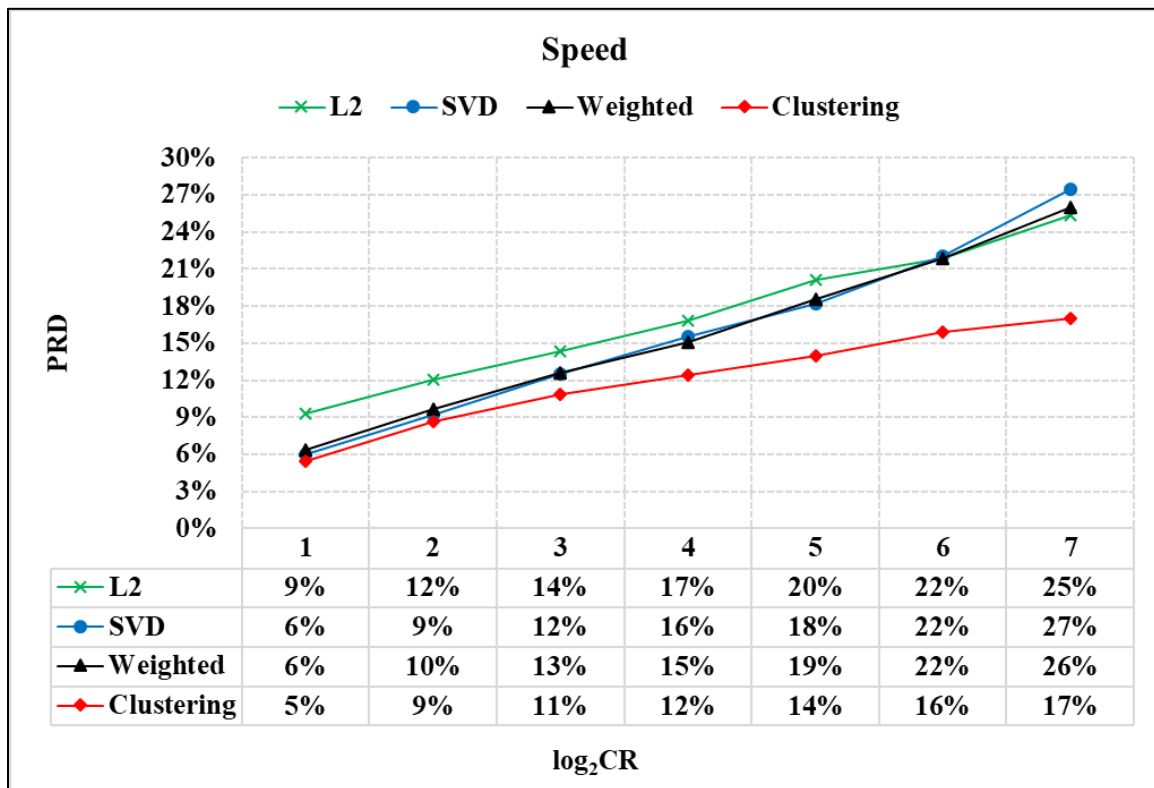


Figure 22 CUR Performance of speed for different column selection methods.

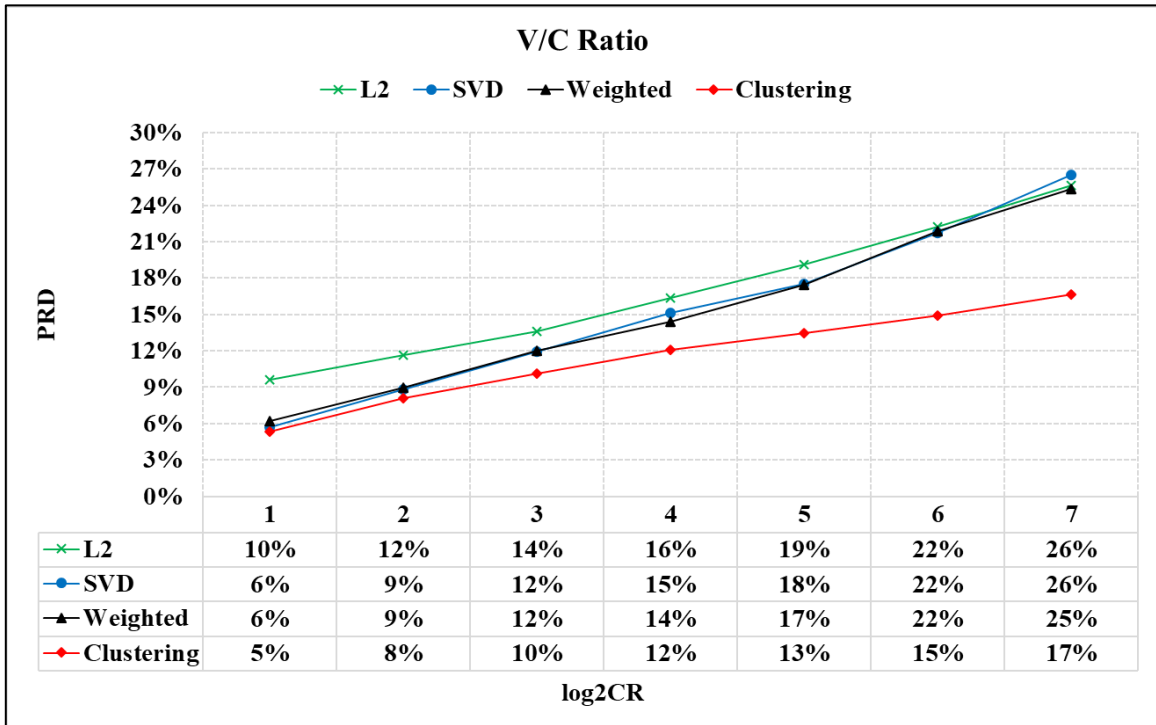


Figure 23 CUR Performance of v/c ratio for different column selection methods.

a measure of the roadway's demand relative to its capacity, indicating how congested a road segment is. While speed gives an immediate sense of how quickly traffic is moving, the v/c ratio offers a broader perspective on potential congestion and overall roadway performance.

Figure 24 and **Figure 25** consists of two maps displaying the clustering result when the number of selected clusters is equal to four. The left map illustrates traffic conditions based on speed, while the right map shows the Volume to Capacity (V/C) Ratio. Each color represents a different cluster, indicating varying traffic characteristics within each segment.

The first point to note is that the V/C ratio proves to be a more sensitive and effective variable for representing the overall network condition compared to speed. In contrast, speed can often remain relatively high during free-flow conditions, even when the network is experiencing varying levels of demand. This means that speed alone may not adequately reflect the underlying traffic congestion or road network stress, especially during off-peak hours when speeds can be uniformly high across the network. The clustering of V/C ratios provides a more nuanced understanding of congestion patterns, highlighting specific areas where road demand is consistently high relative to capacity, thus pinpointing critical congestion hotspots.

The concentration of V/C ratio clusters geographically suggests that certain sections of the network are more prone to congestion, offering a clearer picture of where traffic management efforts should be focused. This sensitivity to capacity utilization makes the V/C ratio a superior metric for evaluating network performance, planning infrastructure improvements, and implementing congestion mitigation strategies.

The maps also highlight distinct patterns in opposite traffic directions, which align with typical commuting behaviors. For instance, routes inbound to Nashville exhibit a pattern suggesting a morning peak in congestion and traffic volume. This is reflected in higher V/C ratios and lower speeds during morning hours as commuters travel towards the city center. Conversely, outbound routes from Nashville show a different pattern, with congestion peaking in the afternoon as people leave the city after work. This directional difference is typical of urban areas, where daily traffic flows are heavily

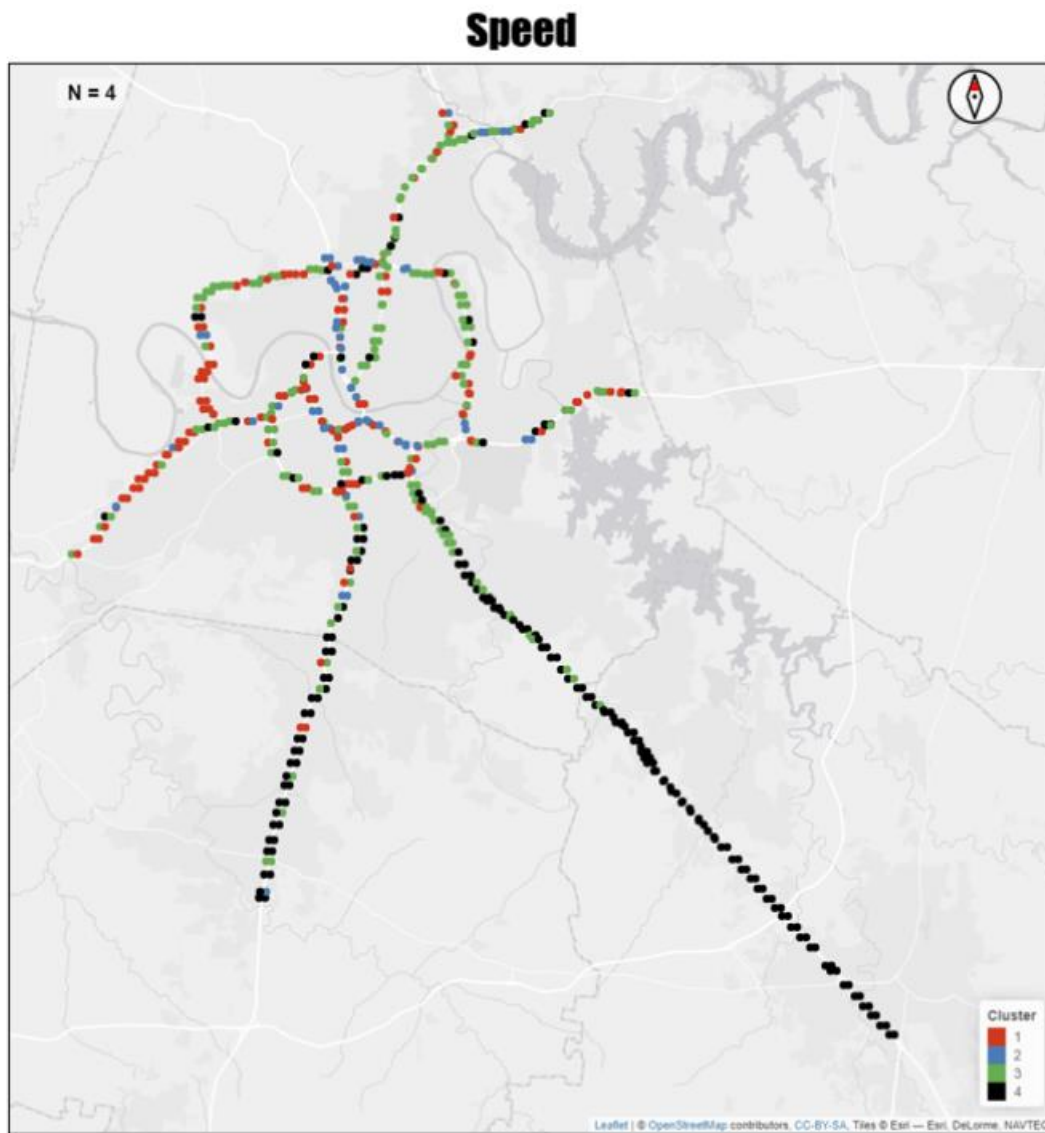


Figure 24 Clustering result for speed when number of clusters is equal to four.

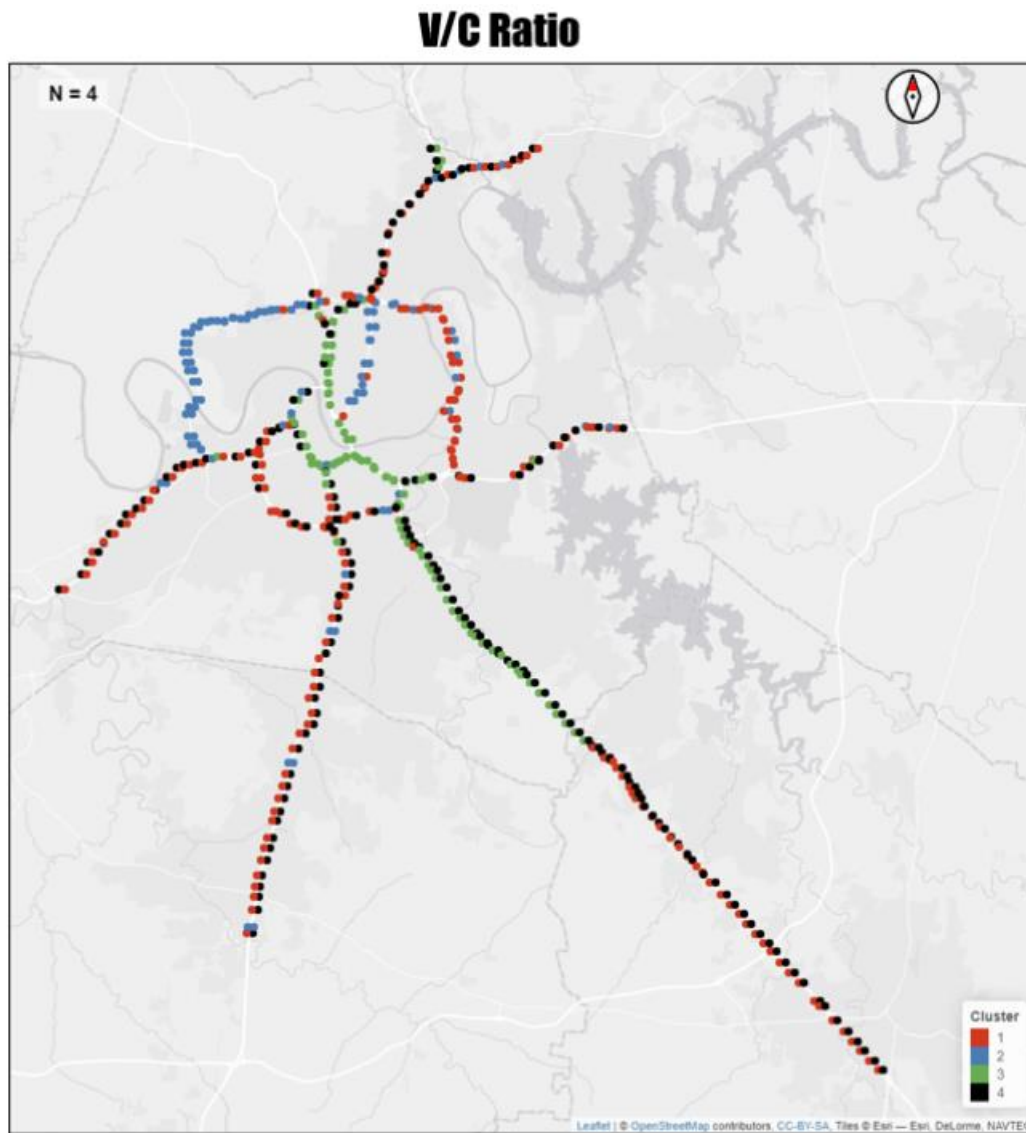


Figure 25 Clustering result for v/c ratio when number of clusters is equal to four.

influenced by work commute patterns.

Figure 26 shows a map illustrating the locations of the selected detectors for both speed and v/c ratio when the required number of columns is set to 10. The detectors, marked in blue dots for speed and black circles for v/c ratio, are strategically distributed across the Nashville area. This distribution ensures comprehensive coverage of the traffic network, capturing data from key points that are critical for accurate traffic flow analysis. The selected locations encompass major routes and intersections, providing a balanced representation of both urban and suburban traffic conditions. The speed detectors are predominantly concentrated along the I-24 corridor, indicating a focus on monitoring traffic speed variations on this major interstate. In contrast, the v/c ratio detectors exhibit a more dispersed coverage across the broader Nashville area, capturing data from various critical points throughout the network. This dispersed placement for v/c ratio monitoring helps to better understand congestion levels and traffic density across different parts of the city and its surrounding regions. Interestingly, there is one notable overlap where both speed and v/c ratio detectors are co-located. This overlap suggests a strategic point where both traffic speed and congestion levels are critically monitored, providing a comprehensive view of traffic dynamics at that location.

The more concentrated placement of speed detectors along I-24 reflects a targeted approach to monitor and manage traffic flow on a major corridor known for high traffic volumes. Meanwhile, the dispersed placement of v/c ratio detectors ensures a broader understanding of traffic conditions, allowing for effective congestion management across the entire network. This strategic distribution of detectors underscores the different objectives and needs for monitoring speed versus v/c ratio, enhancing the overall effectiveness of traffic management and planning efforts.

Conducting experiments on the v/c ratio is particularly important because it directly relates to the capacity and demand balance of a traffic network. High v/c ratios typically signal congested conditions, which can lead to delays and reduced traffic efficiency. By analyzing the v/c ratio, researchers can better understand the extent and impact of congestion, identify critical bottlenecks, and develop strategies for capacity enhancements. Additionally, speed on interstate highways often remains stable at free-

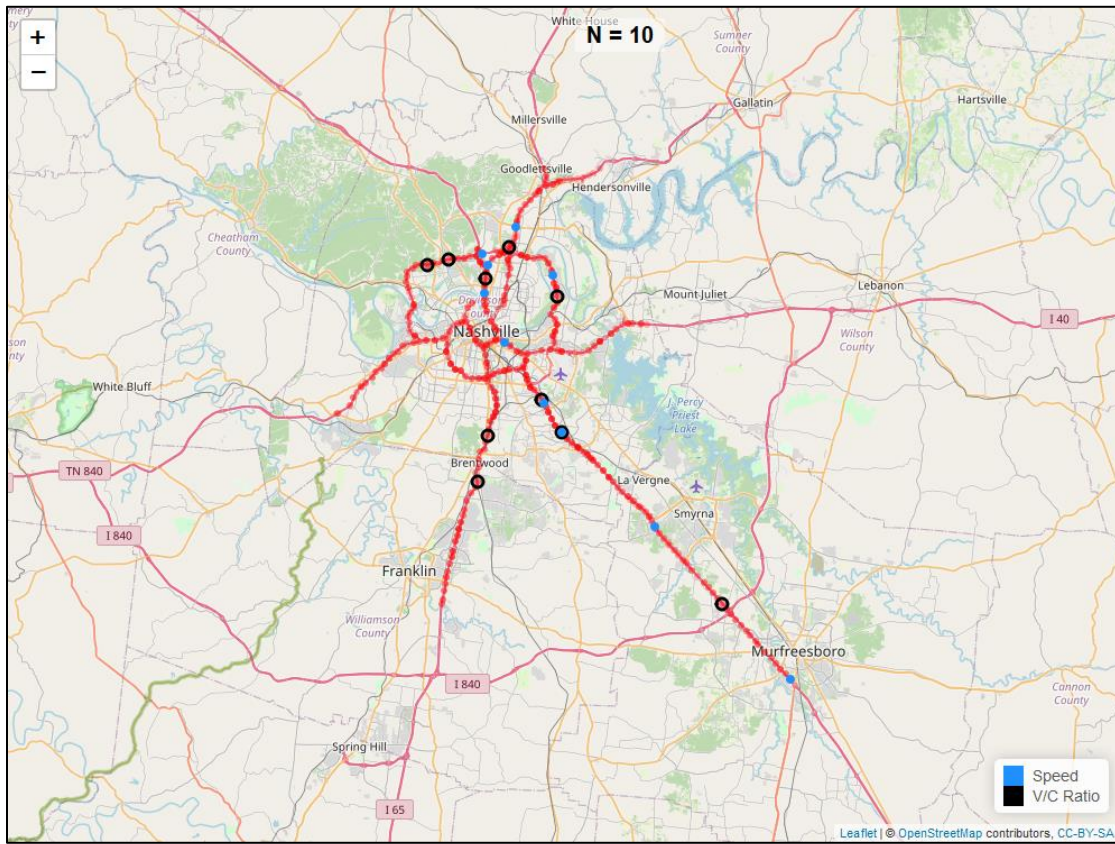


Figure 26 Location of the selected detectors for speed (blue dot) and v/c ratio (black circle).

flow conditions for most of the time, making it less sensitive to variations in traffic flow. In contrast, the v/c ratio captures both the variation of traffic flow and the congestion state, providing a more comprehensive and dynamic picture of traffic conditions. This comprehensive view of traffic conditions is essential for effective traffic management, infrastructure planning, and long-term transportation policy decisions.

Discussion

As shown in **Figure 18** and **Table 13**, the proposed weighted approach demonstrated lower PRD compared to alternative methods when the CR remained below 32. However, the downside of employing the weighted sampling method is the comparatively longer computational time. When the CR threshold surpasses 32, the proposed weighted method only outperforms the SVD-based leverage method, partially because the default weight value (w) is set at 0.5, assigning too much weight to the SVD-based leverage approach. A possible future direction could be tuning the hyperparameter w to achieve better performance over the L2 norm-based and SVD-based leverage methods on training data, followed by an evaluation of the method during the testing period. The PRD of the L2 norm-based method remains relatively stable (between 8% and 12%) as the CR increases from 2 to 128, while the SVD-based leverage method yields good performance with a CR lesser than 16, but its error rises very quickly as the CR becomes greater than 16.

It is worth noting that the performance of the random sampling method is even better than that of the other three methods when the CR changes between 16 and 64. Previous researchers either failed to consider testing their method within this range or forgot to compare with the performance of the random sampling method.

Figure 19 displays the heatmap of the absolute error between the original and inferred test data. A compression rate of 16 yields a PRD of 9.5%, and the absolute error is within 10 mph for 94% of the time and detectors. Additionally, the largest error always occurred recurrently during the peak hour at some specific detectors, which implies that the CUR decomposition method does not perform well during congestion periods. There also exist some non-recurring large errors, which might be due to the occurrence of some atypical incidents such as severe weather, special events, and scheduled road work.

Another interesting finding is that the first clustering approach almost has no room for improvement for SVD-based leverage and the proposed weighted average method. As for the L2 norm-based method, clustering helps reduce the error to as low as 1% to 2%. Integrated with the clustering method, all three methods even demonstrated an increase in error when the CR was greater than 16, except for the SVD method with CR = 128.

There might be various reasons accounting for this phenomenon. First, if the data does not exhibit clear clusters or if the clusters are not well defined, clustering algorithms may not yield meaningful or useful groupings. CUR decomposition relies on selecting representative columns and rows, and if the data does not naturally cluster, selecting clusters may not provide an advantage. Second, determining the optimal number of clusters (k) is challenging. Choosing an incorrect value for k can lead to poor cluster quality and, consequently, suboptimal CUR selection. Another reason could be the choice of the clustering algorithm. If the clustering algorithm is not well-suited to the data or the specific goals of CUR decomposition, it may not lead to improvements. Since we are clustering time series in the context of traffic speed, a dynamic time warping distance could be considered as the similarity metric to perform the clustering analysis. In addition, k -means clustering is simple and basic, so it might be beneficial to try other advanced clustering algorithms, such as hierarchical clustering or DBSCAN, to yield better clustering results and compare the CUR performance with different clustering methods. This study opens the door to various intriguing possibilities for further investigation of how to improve clustering results.

The second clustering approach presents a significant advancement in column selection for CUR decomposition by independently selecting representative columns, rather than combining techniques like L2 norms and SVD. Notably, the new clustering approach has demonstrated superior performance in both speed and v/c ratio metrics, as evidenced by lower PRD values across various scenarios. This method not only refines the accuracy of traffic state inference but also proves to be more effective in dealing with the complexities of traffic flow data, ensuring robust and reliable predictions. The enhanced performance of this approach underscores its potential for improving large-

scale traffic management and planning through more precise and efficient data representation.

In addition, the detectors used in this study are all from interstates without disruption by signal timing. Future work could be carried out to include different types of road networks to analyze their traffic speed patterns and investigate how clustering will impact the CUR decomposition performance on different types of road networks. As has been discussed in Mitrovic's work (10), clustering by different road categories improved the inferring performance for the entire network when the CR was between 2 and 10. Our future work could focus on analyzing the speed patterns on different road types and clustering based on speed profile and road type together.

Due to the data availability issues, this study only focused on a case study in Nashville, TN, and all detectors are located on urban interstates. To comprehensively evaluate the practical applicability, additional research could be carried out to investigate the generalizability of the proposed framework to other cities or regions and for other time periods. It is essential to extend this research to different locations and time frames to assess its spatial and temporal transferability.

In this study, external factors, such as adverse weather, special events and scheduled road work, were not included in the framework, which might impact the representing and inferring performance when there happened to be many non-recurring incidents during the training and testing period. One area for future work involves exploring the impact of external factors in greater depth to gain a more comprehensive understanding of its applications.

Another interesting avenue for future research is to explore the possibility of incorporating other traffic data, such as volume, density and vehicular types, into the methodology for a more comprehensive representation of the traffic condition at specific detectors, thus resulting in a better representation of the entire network. This involves methods like column-based decomposition for multidimensional tensors, proposed by Couras et al. in 2019.

Conclusions

In conclusion, this research addresses significant gaps in representing and inferring road network conditions by applying a column-based CUR matrix decomposition technique with a column selection method based on clustering. This novel approach efficiently represents the original traffic network with a reduced subset of columns while maintaining acceptable error levels. The proposed clustering approach not only offers enhanced interpretability and a more refined column selection process compared to previous methods but also demonstrates superior performance in both speed and volume-to-capacity (v/c) ratio metrics.

The main contribution of this research lies in the innovative application of a clustering-based column selection method integrated with CUR matrix decomposition. This approach significantly enhances the representation of traffic networks by efficiently reducing the dataset while preserving crucial information. The findings demonstrate that the proposed clustering approach outperforms traditional methods, particularly in terms of speed and volume-to-capacity (v/c) ratio metrics. The comprehensive evaluation revealed that the clustering method maintains lower error levels across various scenarios and compression ratios. Moreover, the practical application of this technique to radar detector data from Nashville, TN, showcases its utility in real-world settings, providing a scalable and effective solution for monitoring and analyzing traffic flows.

Overall, this research advances data compression and network representation, emphasizing the potential of combining CUR decomposition with clustering for various applications in big data analytics and network science. Future work will focus on assessing the spatial and temporal transferability of the proposed method and exploring frameworks that consider flow, density, and speed collectively to further enhance traffic prediction models and network analysis.

References

- [1] Chindanur, N. B., and P. Sure. Low-dimensional models for traffic data processing using graph Fourier transform. *Computing in Science & Engineering*, Vol. 20, No. 2, 2018, pp. 24-37.
- [2] Mitrovic, N., M. T. Asif, U. Rasheed, J. Dauwels, and P. Jaillet. CUR decomposition for compression and compressed sensing of large-scale traffic data. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, IEEE, 2013. pp. 1475-1480.
- [3] Wold, S., K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, Vol. 2, No. 1-3, 1987, pp. 37-52.
- [4] Djukic, T., G. Flötteröd, H. Van Lint, and S. Hoogendoorn. Efficient real time OD matrix estimation based on Principal Component Analysis. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, IEEE, 2012. pp. 115-121.
- [5] Djukic, T., J. Van Lint, and S. Hoogendoorn. Application of principal component analysis to predict dynamic origin–destination matrices. *Transportation research record*, Vol. 2283, No. 1, 2012, pp. 81-89.
- [6] Jin, X., Y. Zhang, and D. Yao. Simultaneously prediction of network traffic flow based on PCA-SVR. In *Advances in Neural Networks–ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, June 3-7, 2007, Proceedings, Part II 4*, Springer, 2007. pp. 1022-1031.
- [7] Li, Q., H. Jianming, and Z. Yi. A flow volumes data compression approach for traffic network based on principal component analysis. In *2007 IEEE intelligent transportation systems conference*, IEEE, 2007. pp. 125-130.
- [8] Asif, M. T., S. Kannan, J. Dauwels, and P. Jaillet. Data compression techniques for urban traffic data. In *2013 IEEE symposium on computational intelligence in vehicles and transportation systems (CIVTS)*, IEEE, 2013. pp. 44-49.
- [9] Mahoney, M. W., and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 3, 2009, pp. 697-702.
- [10] Mitrovic, N., M. T. Asif, J. Dauwels, and P. Jaillet. Low-Dimensional Models for Compressed Sensing and Prediction of Large-Scale Traffic Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 5, 2015, pp. 2949-2954.
- [11] Golub, G. H., and C. Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*, Springer, 1971. pp. 134-151.

- [12] Couras, M. F., P. H. de Pinho, G. Favier, J. P. da Costa, V. Zarzoso, and A. L. de Almeida. Multidimensional CX Decomposition of Tensors. In *2019 Workshop on Communication Networks and Power Systems (WCNPS)*, IEEE, 2019. pp. 1-4.
- [13] Han, L., and Y. S. Huang. Short - term traffic flow prediction of road network based on deep learning. *IET Intelligent Transport Systems*, Vol. 14, No. 6, 2020, pp. 495-503.
- [14] Feng, S., Y. Zhang, and L. Li. A comparison study for traffic flow data compression. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, IEEE, 2016. pp. 977-982.
- [15] Agarwal, S., E. E. Regentova, P. Kachroo, and H. Verma. Multidimensional compression of ITS data using wavelet-based compression techniques. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 7, 2016, pp. 1907-1917.
- [16] Sandryhaila, A., and J. M. Moura. Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure. *IEEE signal processing magazine*, Vol. 31, No. 5, 2014, pp. 80-90.
- [17] Nguyen, T. T., P. Krishnakumari, S. C. Calvert, H. L. Vu, and H. Van Lint. Feature extraction and clustering analysis of highway congestion. *Transportation Research Part C: Emerging Technologies*, Vol. 100, 2019, pp. 238-258.
- [18] Cheng, Z., W. Wang, J. Lu, and X. Xing. Classifying the traffic state of urban expressways: A machine-learning approach. *Transportation Research Part A: Policy and Practice*, Vol. 137, 2020, pp. 411-428.
- [19] Chen, T., X. Shi, and Y. D. Wong. A lane-changing risk profile analysis method based on time-series clustering. *Physica A: Statistical Mechanics and its Applications*, Vol. 565, 2021, p. 125567.
- [20] Golub, G., and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, Vol. 2, No. 2, 1965, pp. 205-224.
- [21] Drineas, P., M. W. Mahoney, and S. Muthukrishnan. Relative-Error CUR Matrix Decompositions. *SIAM Journal on Matrix Analysis and Applications*, Vol. 30, No. 2, 2008, pp. 844-881.
- [22] Dauwels, J., K. Srinivasan, R. M. Ramasubba, and A. Cichocki. Multi-channel EEG compression based on matrix and tensor decompositions. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2011. pp. 629-632.

- [23] Boutsidis, C., M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, SIAM, 2009. pp. 968-977.
- [24] Sun, J., Y. Xie, H. Zhang, and C. Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, SIAM, 2007. pp. 366-377.

CONCLUSION

This dissertation provides significant contributions to the application of machine learning in the spatiotemporal analysis of traffic flow characteristics. By focusing on truck volume estimation, short-term traffic volume prediction, and network traffic condition representation, the research offers new methodologies and insights that enhance both academic understanding and practical traffic management.

The first chapter of this dissertation is the development of a novel framework that integrates Quantile Regression with Light Gradient Boosting Machine (LightGBM) for accurate truck volume estimation using single-loop detectors. This approach effectively addresses the limitations of traditional methods, providing robust estimates of truck volumes across diverse traffic conditions. The framework's ability to produce high-resolution estimates, both temporally and spatially, with low mean absolute error (MAE), makes it particularly valuable for transportation planning and infrastructure design. It provides a cost-effective alternative to dual-loop detectors, enabling more reliable decision-making in freight management and roadway maintenance.

The second chapter explores the critical role of incorporating incident data into the training and testing phases of traffic flow prediction models, within the broader context of applying machine learning to spatiotemporal analysis of traffic flow characteristics. By proposing a novel training and testing framework, the study evaluates the performance of various predictive models, including Gated Recurrent Unit (GRU), Support Vector Regression (SVR), Historical Average (HA), and Random Walk (RW), using traffic volume data from radar detectors in Nashville, TN. The research highlights the significant impact that including incident data has on improving prediction accuracy and robustness across different traffic conditions. Through a detailed comparative analysis of scenarios with incidents, without incidents, and incident-specific conditions, the chapter identifies key insights that underscore the necessity of incorporating diverse traffic conditions in model training. These findings demonstrate the enhanced adaptability and reliability of traffic prediction models trained on comprehensive datasets, advocating for a more inclusive approach in the development of traffic

management solutions.

The third part of this dissertation is the representation and inference of traffic conditions across a large network using matrix decomposition techniques, specifically CUR decomposition integrated with clustering. This method provides an efficient way to represent network-wide traffic states using data from a subset of key detectors. The case study in Nashville, TN, demonstrates the effectiveness of this approach in capturing traffic dynamics and supporting large-scale traffic management. This technique offers a scalable solution for monitoring and analyzing traffic flows across complex urban networks, facilitating better traffic operations and planning.

The findings from these studies have several practical implications. The advanced modeling techniques developed for truck volume estimation and short-term traffic prediction can be directly applied to improve traffic management systems. By enhancing the accuracy of traffic forecasts and understanding traffic flow patterns, these models support proactive traffic control measures, reduce congestion, and improve travel times. Additionally, accurate truck volume estimates contribute to more effective infrastructure planning and maintenance. The ability to monitor and predict truck movements aids in designing roadways that can accommodate heavy vehicles, thereby extending pavement lifespan and reducing maintenance costs. Moreover, the demonstrated superiority of GRU models in handling irregular traffic conditions underscores their potential integration into ITS, which rely on accurate real-time data to manage traffic flow, respond to incidents, and provide timely information to travelers.

Future research should focus on integrating additional data sources such as weather conditions, social media, and real-time traffic camera feeds to further enhance the robustness and accuracy of traffic prediction models. Exploring the fusion of structured and unstructured data will be critical for developing more comprehensive traffic forecasting systems. Additionally, expanding the testing of these models to different regions and varying traffic conditions will help validate their generalizability and effectiveness, considering regional traffic behaviors and infrastructure differences. Developing real-time implementations of the proposed frameworks, particularly for truck volume estimation and short-term traffic prediction, will be essential for practical

deployment. This includes refining the models for faster processing and integrating them into live traffic management systems.

In conclusion, this dissertation advances the field of traffic engineering by applying machine learning techniques to critical challenges in traffic flow analysis. The developed methodologies for truck volume estimation, short-term traffic prediction, and network traffic condition representation provide significant improvements over traditional approaches. They offer more accurate, efficient, and scalable solutions for traffic management and planning. These contributions pave the way for more intelligent and responsive transportation systems, enhancing the ability to manage growing urban traffic demands and contributing to the development of safer and more efficient road networks.

VITA

Hairuiliong Zhang was born and raised in Yichang, China. In 2013, he began his academic journey at Shanghai Jiao Tong University, earning a bachelor's degree in Transportation with a focus on International Shipping. Seeking to expand his knowledge and expertise, Harry continued his studies in the United States, obtaining a Master of Science in Transportation Engineering from Northwestern University, Evanston, IL. His academic pursuits led him to the University of Tennessee, Knoxville, where he completed his Ph.D. in Data Science & Engineering.

During his Ph.D. program, Harry engaged in a variety of groundbreaking projects. As a Graduate Research Assistant, he led efforts in AI-enabled large-scale traffic sensor data mining, which involved data cleaning, aggregation, anomaly detection, and quality assurance. His work significantly improved the quality and reliability of traffic data analysis. He also spearheaded research on the U.S. freight network's resiliency, focusing on the impacts of infrastructure failures and the development of predictive analytics models. Additionally, Harry explored crowdsourced traffic data, implementing spatiotemporal visualizations and conducting comprehensive analyses to enhance traffic incident response and network reliability.

Harry's academic excellence is reflected in his perfect GPA of 4.0/4.0 at the University of Tennessee. His contributions to the field are highlighted through several publications in prestigious journals, including studies on mobility dynamics during COVID-19, the framework of representing traffic network conditions, and the modeling of spatiotemporal heterogeneity in traffic incidents. Harry has also been recognized with various awards for his research contributions, further solidifying his standing in the academic community.