Integrated Traffic Analysis and Visualization for Future Road Events

by

Taghreed Alghamdi

A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy of in **Computer Science**

Faculty of Science
University of Ontario Institute of Technology (Ontario Tech University)
Oshawa, Ontario, Canada
April 2023

Thesis Examination Information

Submitted by: **Taghreec Alghamdi**

Doctor of Philosophy in **Computer Science**

Thesis title: Integrated Traffic Analysis and Visualization for Future Road Events

An oral defense of this thesis took place on April 6, 2023 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Shahram Heydari
Research Supervisor	Dr. Khalid Elgazzar
Examining Committee Member	Dr. Jaroslaw Szlichta
Examining Committee Member	Dr. Jeremy Bradbury
External Examiner	Dr. Anwar Haque , University of Western Ontario
University Examiner	Dr. Richard Pazzi, Ontario Tech University

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

The existing traffic simulation methods are limited to specific synthetic scenarios. In addition, the natural structure of traffic and accident data requires modeling the dependent observations on multiple levels. Therefore, a system that utilizes hierarchical LMMs and GBM models are proposed which adaptively analyzes and predicts the traffic pattern based on hypothetical inputs. We developed a user-friendly interface to show the outcomes of the hybrid model. The proposed system encompasses three major components: (1) a road accident simulator and event profile to simulate an accident and predict its effects on traffic status; (2) a robust spatiotemporal traffic speed prediction model that integrates the impact of road accident with the prediction model to adaptively predict the future traffic status in response to this accident; (3) a traffic simulation tool to present the future traffic status. Our system provides satisfactory prediction results in terms of predicting with small errors, obtaining optimal hyperparameters, and less computational complexity.

The hierarchical structure of the spatial component in our approach effectively captures the correlation in traffic status across different spatial points on the same road. Furthermore, computing the traffic speed at different spatial levels and how it interacts with lagged prior traffic speed over the past four periods and a day prior up-scaled the system efficiency.

Evaluation is conducted to test the functionality, usability, and viability. Performance evaluation shows that the event profile model achieves small error rates with an MSE of 0.24 and an RMSE of 0.53 on the testing data, demonstrating satisfactory performance. For traffic status, the integrated model achieves high accuracy with low computational complexity. The boosted LMMs achieved high performance on the test data with an R^2 of 0.9190 and an R^2 of 0.9291 on the full-fitted dataset. The MAE and RMSE are 0.27 and 0.80, respectively, indicating that the fitness of our data was excellent.

Keywords— Random Effects Models; Spatiotemporal; Event Modelling; Traffic Prediction; Traffic Simulator

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

_____ Taghreed Alghamdi

Statement of Contribution

The work described in this thesis has been published as:

- Alghamdi, Taghreed, et al. "Forecasting traffic congestion using ARIMA modeling." 2019 15th international wireless communications mobile computing conference (IWCMC). IEEE, 2019.

- Alghamdi, Taghreed, Khalid Elgazzar, and Taysseer Sharaf. "Spatiotemporal Prediction Using Hierarchical Bayesian Modeling." 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA). IEEE, 2021

- Mostafi, Sifatul, Taghreed Alghamdi, and Khalid Elgazzar. "A Bayesian Linear Regression Approach to Predict Traffic Congestion." 2021 IEEE 7th World Forum on Internet of Things (WF-IoT). IEEE, 2021.

- Alghamdi, Taghreed, et al. "Improving Spatiotemporal Traffic Prediction in Adversary Weather Conditions Using Hierarchical Bayesian State Space Modeling." 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021.

- Alghamdi, Taghreed, Khalid Elgazzar, and Taysseer Sharaf. "Spatiotemporal traffic prediction using hierarchical bayesian modeling." Future Internet 13.9 (2021): 225.

- Alghamdi, Taghreed, et al. "A comparative study on traffic modeling techniques for predicting and simulating traffic behavior." Future Internet 14.10 (2022): 294.

Acknowledgements

I would like to express my deepest sincere gratitude to my advisor, Professor Khalid Elgazzar. His guidance, support, optimism, and encouragement have been invaluable throughout the entire time of my Ph.D. studies. I am grateful to the committee members of my dissertation defence, Dr.Shahram Heydari, Dr.Jeremy Bradbury, Dr.Jarek Szlichta, Dr.Richard Pazzi, and Dr.Anwar Haque.

I would like to thank my family for their endless love, encouragement, and sacrifices. My very sincere and special thanks and appreciation go to my father Abdulwahab and my mother Azzaha, whose continuous prayers, encouragement, and support helped me accomplish my goals. I'm also grateful to my siblings, who have always supported and encouraged me. Special thanks and appreciation are due to my fiancé Ahmed, who has stood beside me and tolerated me, and provided endless encouragement, support, and patience throughout my studies.

I would like to acknowledge everyone who played a role in my academic accomplishments, including my colleagues and friends who helped me with this project and shared all of their experiences with me, making it a memorable experience for me.

I thank Al-Baha University for giving me the opportunity to pursue my studies and for sponsoring me throughout these years.

List of Abbreviations

ITS	Intelligent Transportation System
GPS	Global Positioning System
API	Application Programming Interface
ANN	Artificial Neural Networks
GBDT	Gradient Boosted Decision Tree
LMMs	Linear Mixed Effect models
ST-Kriging	Spatiotemporal Kriging
$X_{Acc_{event}}$	Accident Duration for an event
$X_{accDuration}$	Accident duration
$X_{acctime}$	Accident occurrence time
Z_{acc}	The random effect parameter for accident description variable
MAE	Mean absolute Error
RMSE	Root Mean Squared Error
kph	Kilometers per hour
REML	Residual Maximum Likelihood criterion
NTS_{L1}	Normal Traffic Speed for a Location for the last 15-minute
NTS_{L2}	Normal Traffic Speed for a Location for the last 30-minute
NTS_{L3}	Normal Traffic Speed for a Location for the last 45-minute
NTS_{L4}	1 Normal Traffic Speed for a Location for the last hour
$NTS_{L_{day}}$	Normal Traffic Speed for a Location for the same time in the last 24 hours
RS_{L1}	The average of the NTS_{L1} values for all locations on a roadside
RS_{L2}	The average of the NTS_{L2} values for all locations on a roadside
RS_{L3}	The average of the NTS_{L3} values for all locations on a roadside
RS_{L4}	The average of the NTS_{L4} values for all locations on a roadside

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The transportation sector plays a significant role in the development of the economy, especially with the rapid evolution of intelligent transportation systems. This evolution significantly impacted the increase in the usage of transportation services such as buses, car-sharing, and other platform-based ride services (e.g., Uber, Lyft). As a result, traffic congestion and traffic accidents have increased and become a serious problem that is gradually growing worldwide. These two traffic problems have a direct cost to the environment, the economy, and the individual's well-being, such as fuel emissions, travel time reliability, and emotional stress. Although traffic congestion is considered a status and road accidents are considered an event, both often occur under similar circumstances, and both have a high impact on traffic speed. The association between traffic congestion and road accidents is complex and attracts attention from decision-makers and researchers, in particular with the advancement of the machine and deep learning algorithms.

A number of studies have focused on the impact of traffic congestion on accident frequency. Others have focused on the impact of accident occurrence on traffic congestion level [1] [2]. The findings show that the increase in accident occurrence is likely to cause

a high level of traffic congestion; however, the higher the level of traffic congestion, the lower the accident rates. The Maryland Department of Transportation [3] commissioned a survey to investigate the association between traffic congestion and accident rates. The study finds that the number of road accidents increases the level of traffic congestion. In 2021, INRIX Roadway Analytics [4] estimated that 50% of congested roads are a result of road accidents and often have a longer accident influence.

Furthermore, the Canadian Motor Vehicle Traffic Collision Statistics [5] have shown that road fatalities caused more than 1900 premature deaths and 9,000 serious injuries in Canada in 2018. According to the World Health Organization (WHO), the fatality rate and injuries increase annually due to road accidents, where 1.3 million deaths and approximately 50 million injuries are related to road accidents annually [6]. Besides the loss of lives, road accidents have an economic cost, and they can be direct or indirect. The direct economic cost of road accidents is measured in terms of property damage and the usage of public health services. Other indirect impacts of road accidents can be measured in terms of travel time, and fuel consumption [7]. For example, when an accident occurs, drivers attempt to reduce speed, increasing travel time and fuel consumption. According to Sheu et al.,[8] accidents are a major contributor to 60% of the delays on urban freeways.

Acknowledging the impact of road accidents on traffic status in terms of economic and social levels, numerous research studies have been conducted in order to improve various traffic modeling approaches that can simulate road accidents and predict traffic congestion. These traffic modeling techniques can be classified as spatiotemporal models, where they typically take into account factors that affect traffic estimation, such as the impact of a particular geographic area (spatial) within a specific time frame (temporal) [9]. The availability of these traffic prediction and simulation models plays an essential role in traffic engineering and in assessing the performance of road traffic facilities [10]. Transportation management systems can greatly enhance their services and facilitate real-time decision-making using these traffic models that simulate traffic status and

predict the characteristics of the traffic breakdown that may result in delays or accidents.

1.2 Traffic Modeling

Traffic modeling has been used quite extensively in recent years to analyze and predict traffic status at different levels of complexity, from congested urban settings to rural modeling at the macro and micro scale [11]. It has received a lot of attention due to the emergence of the Internet of Things (IoT) and the tremendous growth in the number of intelligent transportation services and traffic monitoring applications [12]. In traffic modeling, spatial and temporal relationships are heavily correlated and have a substantial impact on traffic status. Therefore, a number of research proposals have proposed spatiotemporal traffic models to analyze and evaluate the spatiotemporal relationships within the traffic data.

Spatiotemporal traffic modeling can be defined as a stochastic process that represents the behavior of traffic at a given location and time. Numerous research on traffic modeling has introduced various predictive models with different capabilities for examining specific time windows, such as short or long-time intervals [13]. Furthermore, other traffic models are proposed for analyzing and predicting traffic behavior in different road environments, including freeways, junctions, and intersections [14]. These proposed traffic models require different modeling techniques to account for the desired time period and the road network type.

1.3 Traffic Simulation

Traffic simulation systems simulate vehicle movements and analyze the traffic behavior [15]. These traffic simulators can be classified into three categories based on their level of representation, macroscopic, microscopic, and mesoscopic [16]. Macroscopic models formulate the relationships between traffic flow, traffic speed, and traffic density. Whereas

microscopic models capture the dynamics of traffic in more detail [17]. Therefore, microscopic models are suitable for simulating traffic in large network areas. Mesoscopic models provide combined features from both microscopic and macroscopic models [15].

All these three different types of traffic simulation models are used to simulate driving experiments and utilize their results in order to enhance the services of Intelligent Transportation Systems (ITSs). However, these traffic simulation systems are limited to specific scenarios where they simulate very specific events related to the interaction of vehicles [18]. For instance, it simulates the traffic flow behavior when vehicles travel from location A to location B. The simulation's output shows the road capacity as well as how traffic congestion breaks down or spreads out across different networks [19].

1.4 Traffic Prediction Models and Event Modeling

Assessing the impact of an event on the traffic status, such as road accidents or weather conditions, helps us to predict the consequences of these events on traffic, such as the level of congestion and the reduction of traffic speed. The nature of different events and the location where a specific event occurs make it challenging to propose a unified event-based predictive model. Planned events (e.g., special events, road construction) and unplanned events (e.g., accidents, weather) will have different patterns, such as the time of the event, the length of the event, and the risk associated with the event [20]. Planned events are often held at fixed periods of time; therefore, the impact of such events on traffic status is predictable. On the other hand, unplanned events or unexpected events such as extreme weather conditions and severe road accidents are difficult to predict due to their stochastic nature [21]. The dynamic patterns of these events make it difficult to determine the event duration, its impact on the traffic status, and when the traffic will resume normally. For example, the impact of weather events (e.g., rain, fog, and/or snow) on traffic status differs considerably from one spatial point to another spatial point.

It will have wider spatial coverage, unlike accident events that will have a smaller spatial coverage impact [22]. Predicting the traffic status under such events that severely affect traffic status is significantly important to enhance transportation management systems and support real-time decision-making.

1.5 Problem Definition

The significant advancements in the Intelligent Transportation System (ITS) have contributed to the increased development of traffic modeling. However, these traffic modeling techniques face several challenges for a number of reasons. First, the difficulties in determining the spatiotemporal dependencies between different geographic areas at different time frames [9]. These traffic models will require advanced techniques such as Bayesian Inference and Artificial Neural Networks (ANNs) that efficiently model large-scale spatiotemporal data [23]. Second, these advanced models require excessive time and memory to perform high computational complexity processes in order to provide accurate estimation [24] [25]. Third, the current traffic models lack the capability of predicting the traffic status under different future events that influence the traffic status [26]. Traffic status is heavily impacted by unexpected events such as road accidents or road closures. Simulating road accident events and predicting their impact in terms of the duration and the speed reduction can provide a glimpse into the traffic behavior if these events occur in the future [27]. On the other hand, traffic simulation techniques have been introduced to visualize the results of these traffic models graphically. However, the available traffic simulation tools are limited in their ability to dynamically visualize traffic status, especially when incorporating other factors such as road type, speed limit, and time of the day[28].

Recognizing these challenges, and the need for a system that computes spatiotemporal traffic data, simulates a scenario for a road accident event, and after that integrates the

effect of these events on prediction outcomes on a user-friendly interface, raises a number of interesting research questions which we address in the following section.

1.6 Research Questions and Objectives

This research contributes to the ongoing efforts to predict road accident impacts or at least reduce their impact by developing an integrated traffic analysis and visualization system for future road events. This system will be of high interest to both the public users and traffic planning authorities. The public can use the outcomes of these models to better plan their trips and make real-time decisions to the best of their interest. Traffic planning authorities (e.g., the Ministry of Transportation) can use the system to avoid congested roads due to accidents and improve route planning for drivers by predicting the traffic status under different future circumstances. This system will help the authorities more efficiently to guarantee better decision-making in designing road infrastructure. The proposed intelligent traffic system consists of different components, including accident event simulation, accident event impact prediction, accident impact integration, spatiotemporal traffic speed prediction, integration, and visualization. The first component involves simulating fabricated traffic scenarios, for example, an accident that blocks one lane at 5:00 p.m. on a highway or an accident that blocks two lanes at 12:15 p.m. on a bridge. We implement a hybrid Gradient Boosted Decision Tree (GBDT) with a hierarchical Linear Mixed Effect (LMMs) model to simulate different fabricated traffic scenarios and predict their impact at different spatial points. The second component of the system assesses the impact of these simulated events to predict the traffic speed in sequence over the accident duration using a boosted LMMs model. We leverage these advanced, reliable statistical models to integrate multiple data sources and ensure accurate estimation of traffic predictions. The third component is a traffic visualization tool to visualize the traffic prediction results in response to various events.

The major objective is to develop a hybrid spatiotemporal traffic speed prediction system that is capable of simulating a fabricated accident and predicting the time elapsed from the occurrence of the accident to the accident clearance time.

To carry out the research and achieve these goals, the following main research questions are formulated to be answered.

- Q1: Is it possible to build a robust spatiotemporal model capable of capturing the relationship between time and space with high computational efficiency and less computational complexity?
- Q2: How can we incorporate road accident impact into the spatiotemporal model and adaptively predict traffic status based on the predicted accident effect?
- Q3: How can we more effectively visualize the model results in a manner that is both user-friendly and interactive to present the insights obtained from this research?

These main research questions lead us to formulate further detailed questions to cover all aspects of this research in-depth, such as:

1. Which spatiotemporal traffic prediction models exist and give the best predictions of traffic within short-term intervals?
2. How to ensure traffic prediction accuracy without compromising model responsiveness?
3. What unexpected traffic events impact traffic status the most, and how do we analyze the patterns of these events on traffic behavior?
4. How to integrate these events into the predictive model, and how to generate an event profile for a future event?
5. What is the most appropriate simulation tool that can visualize the predictive model results and simulate the event's effect on the traffic status?

6. If the existing traffic simulation tools don't provide the flexibility to achieve the system requirements, what's the alternative?

7. What are the system framework limitations, and what are the performance measurement criteria to evaluate the system's performance overall?

The answers to these questions would entail extensive analysis and review of the state of the art of statistical and machine learning approaches. We evaluate the existing approaches that are suitable for implementing the three main components of our proposed system. In the literature review, we will examine studies on traffic prediction models, event models, and traffic simulation tools.

1.7 Thesis Contribution

The main contributions of this thesis are summarized as follows:

- We proposed a hybrid scheme that simulates an accident and predicts its impact on a given location at a specific time point.

- We predict the traffic speed with regard to the simulated accident to provide an overview of how the traffic responds to such an accident.

- We visualize the hybrid model output on a user-friendly interface and ensure its usability.

- We ensure the system's prediction accuracy, efficiency, and reduced complexity.

- We contribute to the research in the traffic domain area by developing a smart traffic system that integrates highly sophisticated methods for better prediction accuracy, efficiency, and lower complexity.

- We help with the traffic congestion problem by providing the ministry of transportation with insight into the traffic status when an accident occurs so they can efficiently manage the road infrastructure and improve route planning for drivers.
- We propose a system that provides solutions to guarantee better decision-making and road management under unexpected road circumstances.

1.8 Research Scope

This research mainly proposes a hybrid spatiotemporal traffic speed prediction system that uses spatiotemporal traffic data and integrates dynamic events to support real-time decision-making. Since traffic status is influenced by different events, we tackle the impact of these hypothetical events on traffic status by utilizing machine learning algorithms. In this research, the events of interest are road accident events and the severity of accidents. Events such as sports events, tourism events, festivals, and non-religious events can considerably impact traffic status. This is due to the difficulty of obtaining reliable data that can accurately quantify the impact of other events. Therefore, it is out of our scope, and we limit our research scope to road accident data.

1.9 Thesis Outline

- Chapter 1 provides an introduction to the research topic, the motivation behind this study, the problem statement, the research questions, and the objectives.
- Chapter 2 reviews the start of the art of traffic modeling and the most influential studies on traffic speed prediction, event modeling, and traffic simulation tools over the last ten years.
- Chapter 3 introduces the framework of the proposed system, describes the implementation details and provides deep insight into the proposed model.

- Chapter 4 explores and presents the road accident and traffic data used in this thesis.
- Chapter 5 presents the experimental results and the performance evaluation of the proposed system. It also describes the data acquisition methods, the experimental setup, and performance evaluation metrics.
- Chapter 6 summarizes the current work and presents the future plan and timeline.

Chapter 2

Background and Literature Review

This chapter is divided into two main sections that present a comprehensive literature review of the research related to traffic modeling studies and traffic modeling studies that incorporate event modeling. In the first section, we provide a review of the most common spatiotemporal prediction models and approaches that are used in smart cities, and intelligent transportation applications, in particular, traffic speed modeling. Furthermore, we summarize the available traffic simulation tools and define the required parameters and the limitations of each tool. In this section, we also provide a brief comparison of the time intervals used in spatiotemporal traffic prediction and demonstrate the challenges in long-term and short-term traffic modeling prediction. The second section looks into the existing literature on the impact of road events on traffic status and the event modeling techniques that have been proposed to observe how traffic behaves in response to these simulated events. Figure 2.1 highlights the structure of the literature review section where we review traffic prediction models, traffic prediction approach, traffic simulator tools and algorithms, and event-based traffic prediction models.

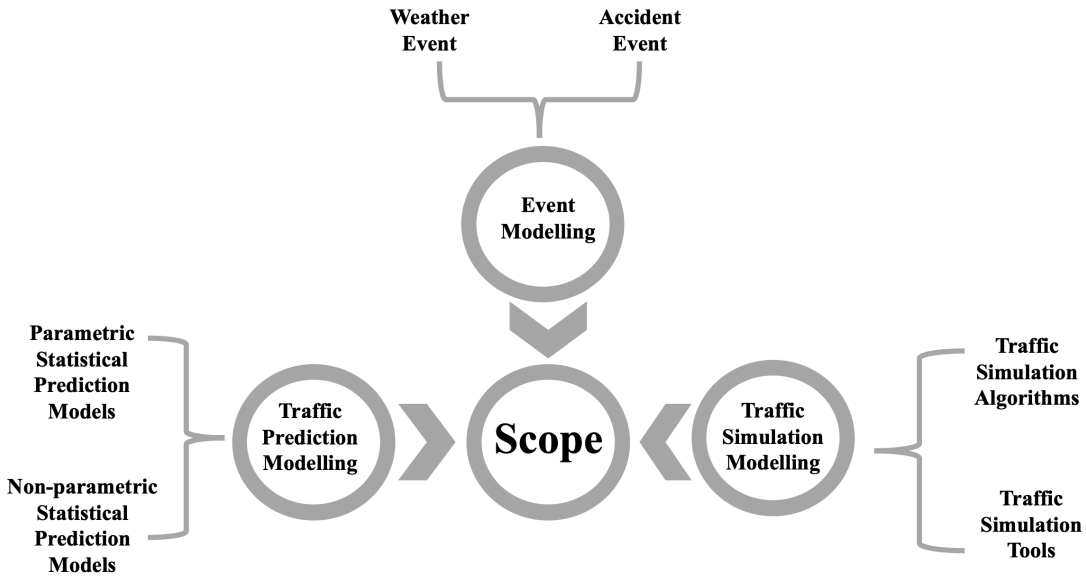

Figure 2.1: The structure of the literature review.

2.1 Traffic Data Prediction Approach

A number of traffic modeling studies have been proposed to predict traffic conditions according to the length of the prediction period, such as short-term prediction and long-term prediction. The characteristics of each prediction require a different modeling process to suit the temporal components and the targeted prediction period. Defining the length of the prediction period will help to decide the best technique to adopt for the traffic modeling study. Short-term and long-term predictions are used to define the time intervals in traffic flow prediction. Short-term prediction involves a short range of time periods such as seconds, minutes, hours, days, or weeks, while long-term prediction involves a long range of time periods such as several months or several years [29]. What follows will provide a description of each time interval in detail.

2.1.1 A Short-Term Traffic Data Prediction

Research on traffic prediction has been mostly restricted to short-term prediction. According to a definition provided by Vlahogianni et al. [30], short-term traffic predictions

are made to predict a period of time in the future that can range between a few seconds and a few days. The authors listed two reasons why short-term traffic prediction has become so dominant in the traffic prediction models field. The first reason is the availability of traffic data that represents a short period of time. The second reason is the availability of many traffic data analytical models that can be used to explore the data. However, traffic data that are collected every 10 s or less is meaningless and not useful for short-term traffic prediction [31]. A number of studies claimed that collecting traffic conditions every 15–30 min would be more effective for prediction results [32, 33]. A study by Song et al. [34] on short-term traffic speed prediction provided a comparison between four prediction methods under different data collected in time windows that ranged from 1 min up to 30 min. The study proposed a seasonal discrete grey model (SDGM) and compared the prediction accuracy with the seasonal autoregressive integrated moving average (SARIMA) model, artificial neural network (ANN) model, and support vector regression (SVR) model. The findings of this study show that the prediction accuracy increases when the targeted time window is more than 10 min, while the prediction of a time window that is less than 10 min suffers from instability. Additionally, the study shows that the SARIMA model's performance had the highest error indicator in the prediction results. A probable explanation regarding these results is that SARIMA cannot capture the variation characteristics of the traffic data in a small time window.

2.1.2 Long-Term Traffic Data Prediction

Regardless of the importance of long-term traffic prediction in enhancing the city roads infrastructure, most of the literature in traffic modeling is focused on short-term prediction. The time intervals of the long-term traffic prediction study the time window that ranges from several months to several years. Although studies claimed that long-term traffic prediction is not beneficial to obtain an accurate prediction, other studies that apply traffic modeling for long-term prediction highlighted the importance of long-

term prediction to improve traffic management systems [35]. Because of the large time window in the long-term prediction, seasonal patterns and cycle patterns will be detected in the traffic data; therefore, models that are able to identify these patterns are strongly recommended, such as the SARIMA model and seasonal autoregressive fractionally integrated moving average (SARFIMA) model [36]. Another study indicates that exponential smoothing models are powerful in capturing seasonality in the traffic data as well as handling trends and white noise satisfactorily [37].

2.1.3 Challenges in Traffic Data Prediction

There are a number of challenges in the field of traffic prediction modeling concerning time intervals. Employing the traffic prediction modeling for long-term time intervals faces several issues [38]. First, the uncertainties of the prediction increase due to the lack of data associated with short time intervals. Second, aggregating traffic data will lead to a high rate of errors in the prediction outcome. On the other hand, modeling short-term traffic prediction requires high computations and is highly sensitive to outliers. Therefore, using data analysis in traffic modeling plays an important role in reducing the drawbacks of predicting short-term or long-term traffic status, where it provides tools and functions that help in cleansing and transforming data into useful information before applying prediction models [39].

2.2 Spatiotemporal Traffic Prediction Models

Over the past few decades, traffic modeling has been associated with time series forecasting methods and spatial prediction methods. However, these two methods suffer from several major drawbacks that affect prediction accuracy. Time series forecasting methods only examine the time series of observations and build the forecasting on the time factor [40]. They are preferable when we only want to identify a directional movement

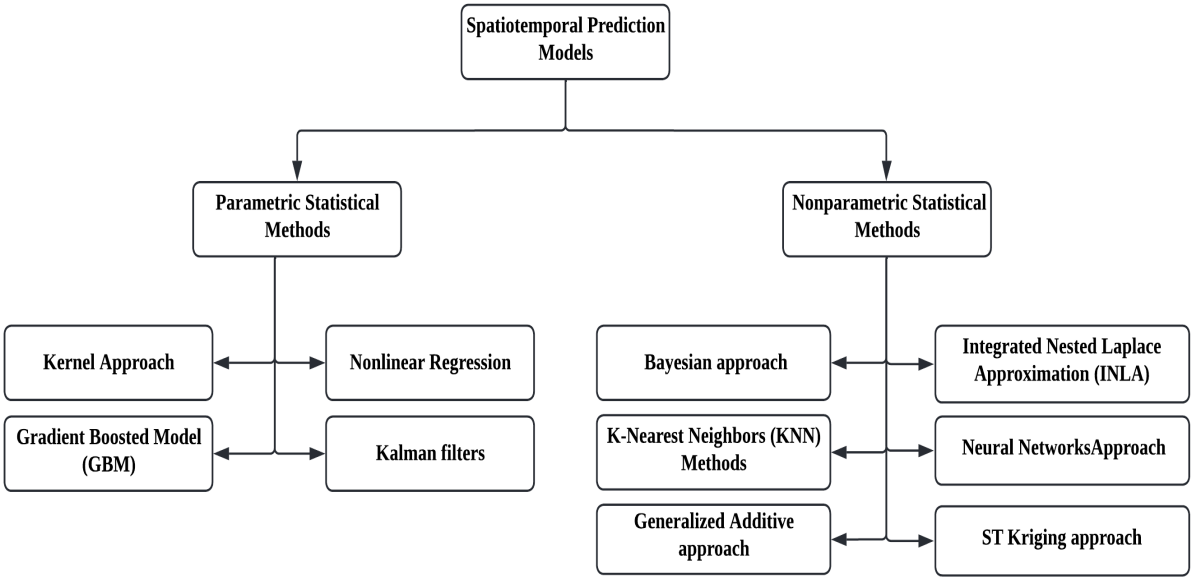

Figure 2.2: A schematic diagram of spatiotemporal models.

in a time series [41]. Spatial prediction methods only take into account the geographical space to build the prediction outcome. The lack of temporal factors in the spatial prediction decreases the prediction accuracy [42]. Combining the time and space factors in the prediction models can improve the prediction results [43]. Therefore, a variety of spatiotemporal methods have been proposed to analyze and predict traffic status to generate deep insight into large-scale traffic data.

Spatiotemporal traffic prediction techniques are statistical methods that model given traffic data to study the patterns of traffic flow and construct knowledge to predict the future traffic status [43]. Figure 2.2 shows the schematic diagram of the most common spatiotemporal models that are used in predicting different real-world applications, such as environment and healthcare. Among these spatiotemporal models are Bayesian inference, ST-Kriging, and artificial neural networks (ANNs). ANNs are considered one of the most widely used models in the traffic domain. The following section provides a comprehensive review of three state-of-the-art spatiotemporal methods and investigates the research gap in these existing models.

2.2.1 ST Kriging Approach

The Kriging method [44] was first developed by George Matheron in 1960 and is mainly used for spatial interpolation prediction. Kriging techniques are well known for optimal spatial prediction and Gaussian process regression. Kriging is a common statistical prediction method that is used by geologists. Subsequently, Kriging became widely used in numerous research works and studies, which made Kriging an essential tool in statistical studies of geographical data. Kriging was later generalized for spatiotemporal prediction and given the name ST-Kriging. The main idea in ST-Kriging is that spatial variability can be characterized by two major components. The first one deals with large-scale variation, exploring the data distribution and capturing trends and outliers. The second component deals with small-scale spatial variation to calculate the spatial autocorrelation and fitting semivariogram to obtain the prediction [45]. Spatial autocorrelation will take into account two functions—the distance and the degree of variation between known data points—when estimating values in unknown areas. Formally, the ST-Kriging equation can be derived from the following:

$$[Z(s,t) : s \in D_s \subset \mathbb{R}^d,\ t \in D_t \subset \mathbb{R}^d] \tag{2.1}$$

In Equation (3.6), the random variable Z is the value at location s at time t, and D_s is a vector of spatial coordinates (x_i, y_i), where $i = 1, 2, 3, ..., n$:

$$Z_{(s;t)} = \mu_{(s;t)} + \epsilon(s;t) \tag{2.2}$$

$$\mu_{(s;t)} = \chi\beta \tag{2.3}$$

In Equation (2.2), Z is a function of random variables at location s at time t, and μ is the conditional mean of large-scale variability. The second component that defines the spatial variability of the Kriging architecture is the small-scale variability that is represented as ϵ, or it can be defined as the noise that captures the large-scale variation.

The mean in Equation (2.3) refers to a function of the observed variables χ through the parameter β. A major step in fitting the ST-Kriging model is to estimate the space-time covariance model, which will be estimated by cross-validation (CV) methods. The covariance function shown in Equation (2.4) estimates the covariance of the observation of random variables at two spatial points. ST-Kriging is governed by a prior covariance matrix based on the data distribution:

$$Cov(X_i, Y_i) = covZ(X_i), Z(Y_i) \tag{2.4}$$

The ST-Kriging method yields the mean square error (MSE) of the variance (σ^2) and a number of linear predictors. It develops its prediction by selecting the minimum variance linear unbiased predictors. A study conducted by Brent and Kara [46] provided a comparison between Kriging methods and geographically weighted regression (GWR) to predict the annual average daily traffic (AADT) counts when the temporal component was excluded. The findings observed in this study have shown that the prediction accuracy of AADT that was provided by Kriging achieved more confidence than GWR. Kriging can control the spatial attributes at unsampled locations by calculating the distance using the spatial autocorrelation function. This function reduces the error in the AADT prediction, with their results indicating that the average absolute error was reduced by up to 63% and the mean square error was reduced by 50%. However, this study highlights a number of challenges when using Kriging for prediction. First, Kriging's prediction lies on a covariance matrix and an inverse covariance matrix, and with large-scale data, matrix inversion is difficult. Therefore, Kriging prediction is implemented on data with a limited size. Another challenge is optimizing the semivariogram estimation and selecting the optimal lag size and the number of lags.

Another research studies the problem of modeling the missing values in traffic data that is collected by road sensors [47]. One of the more significant aspects of this study is modeling traffic data that has a high ratio of missing values collected from 1000 road net-

works. To identify the most information-rich segments, the authors use a method called reduced measurement space [48]. The study indicates the ability of ST-Kriging methods to handle missing observations where they recommend modeling the road networks as one connected spatial component. This approach helps in reducing the impact of the missing observation on prediction accuracy. However, it does increase the computational overhead. In contrast, the prediction accuracy is reduced when each road network is considered separately. Therefore, the authors suggest using a distributed approach with a central control unit in future work.

Another study by Son et al. [49] applied ST-Kriging methods to handle road segments that take into consideration spatial characteristics and spatial homogeneity. Unlike other approaches, point-based Kriging considers the road segments as a single point and ignores these two factors, despite their importance in building more accurate traffic prediction. Their study proposes a segment-based regression Kriging (SRK) method to predict the traffic volume with a comparison between heavy vehicles, such as trucks, and light vehicles. There was a slight improvement in the prediction accuracy compared with point-based Kriging prediction. In the case of heavy vehicles, the prediction accuracy improved by 0.67%, whereas the uncertainty estimation showed significant results and improved by 53.63% compared with point-based Kriging. On the other hand, there was no increase in the prediction accuracy of the light vehicle, where the prediction accuracy results were less than the prediction accuracy in the point-based ST-Kriging approach.

Much of the usage of the ST-Kriging approach in traffic modeling research to date has been for improving the traffic system by modeling traffic conditions, such as by analyzing traffic congestion [50] or predicting traffic speed and travel time [51, 52].

2.2.2 Bayesian Inference Approach

There are several similarities between the ST-Kriging approach and the Bayesian approach in terms of employing the covariance matrix in estimating the minimum variance

and mean. However, the Bayesian approach yields a posterior and probability density function (PDF) of the conditional distribution, which defines the probability distribution of a random variable. In addition, the Bayesian approach does not depend on assumptions in the model settings, unlike ST-Kriging. It computes the prediction probability by sampling the data using the Markov chain Monte Carlo (MCMC) algorithm. Bayesian inference approaches use Bayes theory to produce statistical inference. To simplify the concept of Bayesian inference, three main terminologies need to be defined: prior, likelihood, and posterior. The prior refers to a prior probability of knowledge that is modeled by a probability distribution. This prior will be updated on a continuous basis as new data are acquired, as will the so-called likelihood probability. Incorporating the prior probability and the likelihood probability gives the posterior probability [53].

Equation (3.2) refers to the Bayes theorem that is used in the Bayesian inference process, where $P(\theta)$ is the initial prior probability distribution of the parameter from the current observation, also known as the initial hypothesis, and $P(Y|\theta)$ is the likelihood probability distribution of the observed data given a parameter value. The product of the likelihood and the prior gives $P(\theta|Y)$, which is the posterior probability of the parameter given the observed data [54]:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{2.5}$$

In the literature, $P(Y)$ tends to be used to refer to the marginal likelihood or the evidence, but Bayesian inference treats the evidence as a normalizing constant [55].

DAZIAN et al. [56] employed Bayesian methods in a study on analyzing road safety and modeling travel behavior. Additionally, a Gibbs sampler was used in MCMC computation, which is considered one of the common sampling methods used in Bayesian approaches. In their study, they modeled data into samples that were different in size, consisting of 30, 50, and 100 sites. Furthermore, they applied the experiment to data with missing observations. A comparison was introduced to evaluate two different Bayesian

approaches: the empirical Bayesian (EB) approach and the hierarchical Bayesian (HB) approach, which estimates the posterior within multiple levels. The results of this study show that in both approaches, modeling different sizes of samples is effective. However, the EB approach has a drawback in that other studies [57] have criticized the need for a repeated process, in which in the first run, the process uses the data to determine the model parameters, and in the second run, the process uses the data again to identify the posterior. In comparison, the HB approach can overcome this problem and provide a more flexible framework to determine the model hyperparameters and the posterior through its hierarchy. On the other hand, both the EB and the HB approaches handle missing observations and multidimensional attributes appropriately.

In 2019, Zheng and Sayed [58] proposed a study that applied to traffic safety, where they used the HB approach in predicting traffic accidents, particularly rear-end accidents that occur at intersections. The traffic data followed a generalized Pareto distribution, which is described as a probability distribution that is used to model the tails of another distribution. Additionally, a comparison was conducted between Bayesian hierarchical generalized Pareto distribution models (BHM-GPD) and the hierarchical generalized extreme value model (BHM-GEV), which models a distribution that has very rare or extreme behaviors [59]. The traffic data included a few traffic conflict (e.g., accident) observations that represented extreme events at a specific intersection. The results show that the BHM-GEV approach performs better when the traffic conflict observations are distributed over different intersections. However, the BHM-GEV approach may provide inefficient performance when there is a limited number of traffic conflict observations. A number of limitations are discussed in the study, where there are still some challenges in predicting traffic accidents at intersections, such as having short traffic observations at intersections, which are not preferable for modeling. However, the authors recommended collecting data over a longer period of time, with temporal dimensions such as days, weeks, and months. The limited number of traffic observations that are collected

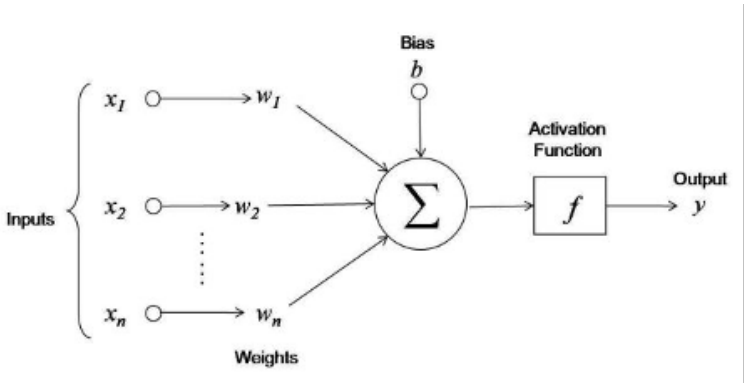

Figure 2.3: The Basic Components of ANN.

at intersections restricts other researchers from tackling this important topic in traffic modeling.

2.2.3 Artificial Neural Networks Approach

In the 1990s, artificial neural networks (ANNs) became a popular approach for binary and numerical data prediction. ANNs are data-driven machine learning algorithms that work similarly to smoothing algorithms in terms of learning the patterns from the data. It also works similarly to regression algorithms in that they are designed to capture a relationship between the input and output using cross-sectional data. In the literature, the ANN approach has mixed results regarding the performance of neural networks compared with other prediction methods, where neural networks work best with high-frequency data [60]. As we can see from Figure 2.3, a basic ANN model has an input layer, and an output layer [61]. All the layers in between the input and output layers are denoted as hidden layers. The neurons between different layers are connected via an edge associated with a certain weight. The ANN computes the values of these neurons in association with their weights and forwards the values to an activation function. An activation function maps the aggregated values from the input layer to the output layer.

One of the earliest studies on traffic modeling using ANNs was proposed by Ledoux in 1997 [39], where she designed a traffic modeling system based on ANNs. The system

has the capability to simulate the traffic flow for connected junctions and then model the traffic flow over a wide range of intersections. The study confirms the potential of using ANNs in traffic prediction modeling and recommends further investigation.

A predictive model based on the ANN approach was introduced by Li et al. [62] to predict traffic accidents and improve traffic safety. They discussed integrating back-propagation neural networks with genetic algorithms to identify potential jamming spots that were likely to cause traffic accidents. The model analyzes the traffic conditions and then produces samples of the possible road accident spots. Additionally, they applied the model to real-time data to predict traffic accident spots. Their conclusion was that integrating ANNs and genetic algorithms as a hybrid genetic algorithm backpropagation (GA-BP) model helped in optimizing the network. The computational overhead of this process produces the local minimum problem, which means that the ANN will continue training the data and updating the network's weights until it reaches the lowest point of the error function. The model has the ability to classify the static factors and the dynamic factors within the road traffic conditions to achieve a high prediction accuracy.

Çetine et al. [23] proposed a study to model historical traffic data using the ANN approach. The study focused on predicting the traffic flow at each main intersection in the city of Istanbul. The model predicted the traffic based on a specific scenario, such as during holidays and school hours. One of the model's features is informing the drivers of the traffic status for the next hour. This study proposed testing the feasibility of applying the ANN approach in the traffic modeling domain. The findings of this study show that ANNs successfully provide accurate predictions in different scenarios. However, the lack of long-term data might enhance the results, as was recommended by the authors.

2.2.4 Summary of Spatiotemporal Traffic Prediction Models

Having discussed the concepts of the previous models and how they were used in various traffic modeling studies, constructing a comparison to evaluate different aspects of each

model will help to decide which one is more suitable for our research. The comparison was restricted to evaluating the predictive accuracy, computational complexity, and evaluation criteria (see Table 5.3).

In terms of prediction accuracy, ST-Kriging prediction accuracy relies on the covari-

Table 2.1: Comparison between ST-Kriging, Bayesian inference, and ANNs.

	Bayesian Inferences	**ST-Kriging**	**ANNs**
Computational Complexity	NP-hard [63]	$O(N^2)$	$O(i \times o \times n + n \times o)$ or $O(n \times o \times (i+1))$ for training a single epoch. [64]
Performance Evaluation	Provides a posterior probability distribution with confidence interval.	Ensure linear unbiased predictors.	Epoch with the lowest sum of squared error.
Weaknesses	Very computationally intensive due to choosing the proper prior distribution.	• Missing value causes error in unmatched dimensions. • Can not handle large datasets. • Require normal distribution.	Require intensive data training, and this might lead to an overfitting problem.
Strengths	• Handle large and small data. • Handle missing values. • Prior knowledge about uncertain input is not required.	• Handle small data. • Computational efficiency.	• Handle big data and small data. • Accommodate missing values without a separate estimation step. • Computational efficiency due to the parallelity feature. • Prior knowledge about uncertain input is not required.
Overcoming the Limitation	Using uninformative prior to reduce the computational time, however, can affect the prediction accuracy negatively.	Remove observations that include missing values.	• Decrease the number of layers of the network. • Use iterative methods to stop the training process, such as gradient descent.

ance matrix to produce highly correlated data samples. Therefore, defining the correct correlation function in the correlation matrix is important for obtaining an accurate prediction. The Gaussian correlation function and Matérn correlation function are two of the most commonly used correlation functions in the correlation matrix in the ST-Kriging method. To identify the best correlation function, estimation tools are required to estimate the correlation function parameters, such as maximum likelihood estimation (MLE) and semivariogram estimation [65]. However, these tools suffer from a number of challenges. In the case of using semivariogram estimation, a plotted semivariogram will be given to determine the appropriate function parameter. Yet, the process of optimizing semivariogram estimation requires deep knowledge of the ST-Kriging approach.

On the other hand, maximum likelihood estimation requires a large sample size to identify the correct function parameters. Additionally, the distance between the spatial points in each sample needs to be small [66]. These factors affect the prediction accuracy and need to be taken into consideration when applying the ST-Kriging approach in traffic prediction. Another point to consider is the computational cost of the model, wherein ST-Kriging, the computational complexity will be estimated based on the number of spatial data points N. When having a large number of spatial points, the covariance matrix becomes more complex, and thus detecting correlation in space and time becomes more complex as well [67]. In addition, ST-Kriging methods require high training times with a computational complexity of $O(N^3)$ [68]. This leads to the conclusion that the overhead cost of the ST-Kriging method is represented in the high complexity when computing traffic data that are large in size. In contrast, large traffic data produce samples that help to improve the prediction accuracy [68].

From the perspective of evaluating the model performance, ST-Kriging methods can be evaluated using cross-validation techniques and fundamental statistical parameters such as the variance of errors. Additionally, examining the model residuals helps assess the minimum variance of linear unbiased predictors [69]. Turning now to the data struc-

ture, ST-Kriging methods were implemented to model data with a Gaussian distribution. ST-Kriging does not perform the best when the value we want to predict indicates that there is a non-normal distribution, where the values either are higher or lower than the real values [70]. Cooper et al. [63] showed that probabilistic inference by using Bayesian belief networks is NP-hard. As a result, it is unlikely that a generalized algorithm will be designed in order to perform probabilistic inference efficiently in Bayesian belief networks over all possible classes. Therefore, for each of the special case, average case, and approximation algorithms, specific domain-centric Bayesian inference needs to be applied.

In Bayesian inferences, the prediction accuracy depends on reducing the uncertainty of the posterior distribution, where the Bayesian inference generates samples $\theta_1, \theta_2, \ldots, \theta_n$ from the posterior distribution. These generated samples will be updated using the Markov chain Monte Carlo (MCMC) algorithm until reaching the accurate posterior predictive distribution, which can be represented by the maximum likelihood [67]. Informative priors increase the accuracy of the Bayesian inference since they provide prior knowledge to help build the likelihood function. However, using informative priors requires more data to update the posterior since the posterior will be very much driven by the prior information. Computing more data can dominate the posterior distribution and cause an overfitting problem.

The computational complexity in a Bayesian inference manifests in the MCMC algorithm's intensive computation required to compute the maximum likelihood estimation. Furthermore, when modeling traffic data that have a short temporal component using Bayesian inference, the MCMC algorithm's computational cost increases dramatically due to the high dimension of the temporal component [68]. In addition, improper priors can maximize the variance in the posterior samples, and hence more computational time is needed to identify the proper prior in order to reduce the variance in each sample [71]. Overall, estimating priors is a computationally intensive process, and this is con-

sidered one of the drawbacks of the Bayesian inference approach. Despite this, Bayesian inference has the capability to handle large traffic data with missing values and assign priors to these missing values [69]. It can also model data that are small in size, such as one observation, and be able to compute the prior of one observation. This process can be performed iteratively in real time [70]. Another advantage of Bayesian inference is that it can handle multilevel models and compute its hyperparameters [58]. In terms of evaluating the model performance, it is recommended to use coefficient estimates and standard deviation errors to measure the uncertainty of the model performance.

When comparing the neural network approach to the previous approaches, specifying the proper network structure can affect the prediction accuracy, while optimizing the network structure can be achieved through experience [23]. Moreover, training ANNs can lead to an overfitting problem. Therefore, it is important to ensure that the validation accuracy is higher than the training accuracy [72].

Various ANN architectures, such as the multilayer perceptron (MLP) and fuzzy neural network (FNN), can be combined to predict the values of MPEG and JPEG video, Ethernet, and Internet traffic data one step ahead. The output of the individual ANN predictors is combined to enhance the prediction accuracy using an adaptive updating scheme that allows the predictors to be dynamic. Moreover, this type of combined model can capture the non-stationary traffic characteristics, as it considers prediction at different time scales so that the predicted values can be applied to the congestion control schemes. This approach outperformed the parametric autoregressive (AR) model, as the combination of ANN predictors enhanced the prediction accuracy [73]. The use of ANNs overcomes many failings related to traditional methods for the prediction of a congested freeway's traffic status, as most data prediction techniques highly depend on the accuracy of the stochastic processes governing the freeway [64]. The freeway modeling process is not mandatory for ANNs because the multilayer perceptron (MLP) type of ANN requires only an input training set along with appropriate outputs for prediction. As a result, this

ANN architecture can be applied generally since it is not dependent on the particular geometry of a freeway section.

Artificial neural networks are relatively insensitive to missing data for predicting traffic conditions and faulty data. In addition, ANNs can deal with nonlinear systems to handle highly dynamic traffic data. However, for traffic speed prediction problems, ANN models are time-consuming to train with high-dimensional data. Therefore, dimension reduction through proper feature selection would help to improve the modeling accuracy [74].

2.3 Traffic Simulation Models

In spite of the fact that traffic analytical models are helpful in providing insights into traffic status, traffic simulation systems play a significant role in representing and evaluating traffic behavior under a number of circumstances [17]. Traffic simulators are also considered a key enabler in the effective implementation of smart mobility services. Extensive simulation to evaluate and test the impact of such services will be essential prior to real-world testing. Hence, traffic analysis and modeling of *'what if'* scenarios assist policymakers and traffic planners with making informed decisions regarding infrastructure planning and investments. The ability of these traffic simulators to model various levels of traffic complexity and city-wide scales ranging from a single detailed intersection to a specific region will provide valuable insights into traffic modeling, and analysis [75]. This provides different levels of granularity among commercial and open-source traffic simulators which can vary extensively. Hence, these traffic simulators can be classified into three categories based on their level of representation, which are macroscopic, microscopic, and mesoscopic [16]. Macroscopic models formulate the relationships between traffic flow, traffic speed, and traffic density. These models adopt an abstracted level of traffic details, and the simulation occurs on a segment basis approach rather

than individual vehicle tracking [76]. The travel demand models associated with the macroscopic-based simulators have a prime focus on the traffic flow of vehicles and the vehicles' routing choices that are selected based on algorithms that optimize the vehicles' travel time. While microscopic models capture traffic, dynamic factors are processed in more detail [17]. Therefore, microscopic models are suitable when simulating traffic in large network areas. In these simulators, vehicles' movements are simulated according to car-following and lane-changing algorithms. Due to the high level of traffic details, these simulators are considered efficient in modeling and evaluating complex scenarios such as rush hour traffic congestion cases, complicated geometric traffic configurations, and many others [77]. Even with the aforementioned benefits offered by these simulators, microscopic models are considered time-consuming and expensive, and they suffer from calibration challenges [76]. The third model that represents some of the features of both microscopic and macroscopic models is the mesoscopic model [15]. All three of these different types of traffic simulation models are used to simulate the driving experiments and utilize their results in order to enhance the facilities and intelligent transportation systems (ITS). However, these traffic simulation systems are limited to specific scenarios, where they simulate the traffic status and interaction of vehicles under specific conditions [18].

To simulate the traffic status, simulation models primarily focus on the number of input and output parameters. The trip description is an input used to specify the destination and departure time. The second input is the network geometry layout, which describes the network's length, the number of lanes, etc. The third input is traffic flow, which indicates the number of vehicles on the network [78]. In terms of the simulation model output parameters, the outputs can be defined as the travel cost of the simulated scenario and the updated traffic flow value when the network layout has been changed. For instance, it simulates the traffic flow behavior when vehicles travel from location A to location B. The simulation's output will show the road capacity as well as how traf-

fic congestion breaks down or spreads across different networks [19]. Therefore, traffic simulation models may have various adjustable parameters that can detail underlying traffic behavior such as vehicles' routing choices, the selection of a shorter planned path, and driving behavior. Calibration, prediction, and validation of the inputs and parameters are considered data-demanding and require efficient computation tools [75]. These simulation models also use a number of algorithms, such as the car following algorithm, the lane changing algorithm, and the gap acceptance algorithm. These algorithms are used to view the traffic status dynamically when increasing the speed of the vehicle or driving within multiple lanes. We describe these different algorithms in the traffic simulation models to comprehend how these models work realistically. Figure 2.4 shows a typical illustration of the car following, lane change, and gap acceptance algorithms used in traffic simulation models.

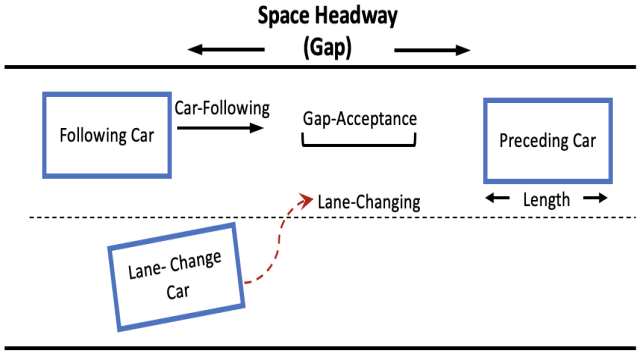

Figure 2.4: Illustration for the car-following, lane-change, and gap-acceptance algorithms.

Furthermore, these traffic simulation models are implemented in different traffic and transport planning software to show the traffic behavior in a graphical user interface, where the user can define the input parameters and view the output parameters [28]. According to Ejercito et al. [16], the seven most widely known traffic simulation tools are SUMO, MATSim, AIMSUN, CORSIM, Paramics, VISSIM, and TRANSIMS. We can consider traffic simulators to be dynamic visualization models that use statistical methods to examine traffic behavior and provide statistical reports for the simulated

scenario.

There are also some useful traffic simulators such as FreeSim [79], Traffsim [80], SUMMIT [81], and SifTraffic [82] that are designed for either microscopic or macroscopic traffic simulation. FreeSim traffic simulator is designed to conduct traffic simulations of freeways in real-time [79]. Traffsim simulator is widely known for modeling isolated traffic control strategies in different complex traffic environments [80]. In large traffic scenarios with massive and mixed traffic, the SUMMIT traffic simulator provides useful features and functionalities to simulate vehicle driving, especially in urban scenarios [81]. SifTraffic is a traffic simulation tool that provides practical implications for different types of traffic applications [82].

2.3.1 Traffic Simulation Algorithms

The car-following algorithm, the lane-changing algorithm, and the gap-acceptance algorithm are used in microscopic traffic simulation models. However, they can be implemented differently in terms of the vehicle's process of speed deceleration increasing, the gap size, and the accepted and rejected procedures for determining the safe distance between floating vehicles [83]:

- Car Following Models:

  A car-following algorithm is intended to describe how the simulated vehicles interact with the preceding vehicle in the same lane. For any car-following algorithm, the basic parameters used to define the speed–spacing relations are the capacity of a lane, the speed, and the average spacing between the preceding vehicle and the following vehicle [84]. Let n be the preceding vehicle, and $n+1$ be the following vehicle with a speed s and vehicle position x at time t. Therefore, the speed and position of the preceding vehicle are denoted by x_n^t and s_n^t, respectively. Similarly, the speed and position of the following vehicle are given by x_{n+1}^t and s_{n+1}^t,

respectively [85, 83]. The acceleration in speed is denoted by α at time t, and the difference in speed between the preceding and the following vehicle is denoted by $s\Delta$. Let $t+T\Delta$ be the time period when the vehicle accelerates, where $T\Delta$ is the time required for the driver to respond to a changing scenario. As a result, the safe distance between the preceding and the following vehicle is computed as $x_n^t - x_{n+1}^t$, which we refer to as the space headway $X\Delta_{safe}$. Let λ be the sensitivity coefficient parameter that is estimated by modeling the sensitivity of the relative distance between the following and preceding vehicles as well as the sensitivity of the relative speed for the subject vehicle [84, 86]. The notations used to describe the car-following algorithm are shown in Figure 2.5, and the basic equation of the car-following algorithm can be represented as follows:

$$\alpha_{(n+1)}(t + T\Delta) = \lambda s_{n+1}(t + T\Delta)\frac{T\Delta(t)}{X\Delta(t)} \tag{2.6}$$

Let λ be the sensitivity coefficient parameter that is estimated by modeling the sensitivity of the relative distance between the following and preceding vehicles, as well as the sensitivity of the relative speed for the subject vehicle.

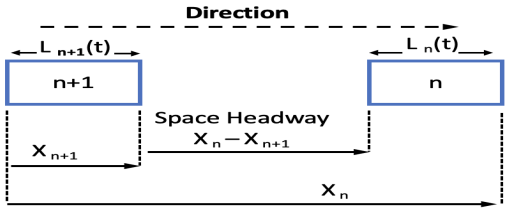

Figure 2.5: Basic Car-following Model.

In traffic simulators, car-following algorithms adopt exact replicas of the car-following maneuvers which are carried out by drivers or automated vehicles in real driving conditions. The essential concept of the car following algorithms is to control the longitudinal motion of vehicles [87]. In real-world settings, autonomous vehicles such as Google cars or Apple cars integrate the data-driven machine learning car

by following the model's approach. This approach extracts the patterns or associated rules of drivers' car following strategies and behaviors, in addition to capturing the relationships among variables that can have an impact on the car's following behaviors. This approach yields high accuracy in replicating drivers' car following behaviors for automated vehicles. Another car following model is the kinematics-based approach, which relies on kinematics processes such as the GM, intelligent driver, and safe distance approaches. These approaches adopt an explicit mathematical form, where most of the model parameters have physical meanings, and the model outputs can be controlled through refined adjustments of the model parameters [88].

- Lane-Changing Models:

  Lane-changing algorithms are used to simulate the impact of vehicles on adjacent lanes as they change lanes. These algorithms take into account the speed and position of the preceding vehicle as well as the time when this action takes place [89]. The concept of the lane-changing algorithm can be simply described as follows. When the vehicle intends to change lanes, the model assesses the existing headway space to determine whether changing lanes is achievable. If it is, then the process happens. If not, then the vehicle remains in the current lane [90]. A simple illustration of the lane-changing decision of a vehicle is depicted in Figure 2.6. The model must meet certain criteria such that for a given adjacent lane, both the space headway for the following and preceding cars must be more than the unsafe distance, which can be computed as follows:

$$d_{(safe)} = \frac{s_{n^2}^2 - s_{n+1}^2 + 3s_{n+1}b\lambda}{2b} \tag{2.7}$$

$$d = x_n - L_n - x_{n+1} \tag{2.8}$$

$$\text{Acceptable Headway Gap if} : d \geq Cd_{(safe)} \tag{2.9}$$

Let s_{n^2} and s_{n+1} denote the following and preceding vehicle speeds, respectively, and b be the vehicle's maximum deceleration. We refer to the actual vehicle following distance if the vehicle moved into the adjacent lane by d. Equation (2.9) is derived from Equations (2.7) and (2.8), which compute the smallest acceptable headway gap between each vehicle C and the minimum safe distance d_{safe} between the subject vehicle and the following vehicle [91, 83].

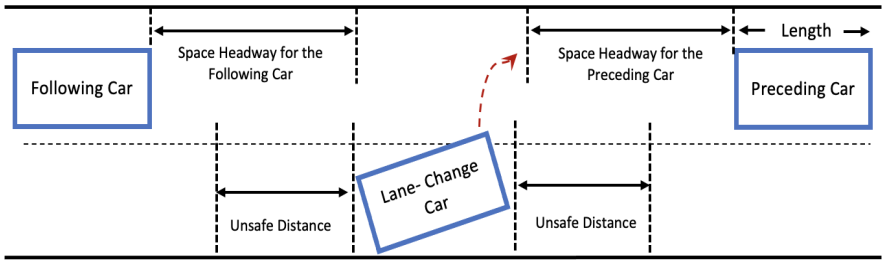

Figure 2.6: Basic Lane-change Model.

- Gap-Acceptance Models:

  Gap-acceptance models are mainly used to determine the traffic conditions in adjacent lanes prior to a vehicle accessing the available space. They are used to estimate the amount of space and time required to cross a junction, enter a roundabout, or change lanes [92]. These two factors are dependent on the traffic conditions, such as the road characteristics, the speed, and the lengths of the following and preceding vehicles as well as the passive vehicle. The minimum safe distance d_{safe} between the subject vehicle and the following vehicle, which is also known as the critical gap, is a significant parameter affecting gap acceptance behavior. An important assumption that has to be addressed in the gap acceptance models is the headway distribution in the circulating flow to measure the road capacity [93, 83]:

$$Y_{n(t)} = \begin{cases} 1, & \text{if } d_{n(t)} \geq d_{n_{(safe)}(t)}, \\ 0, & \text{if } d_{n(t)} < d_{n_{(safe)}(t)}, \end{cases} \tag{2.10}$$

where Y denotes the vehicle's decision of whether or not to overtake the adjacent lane at a given time t. The $d_{n(t)}$ is the available headway gap, and $d_{n_{(safe)}(t)}$ is the critical gap. The vehicle forces entry when the gap size is equally likely to be accepted ($Y_{n(t)} = 1$); otherwise, the vehicle rejects ($Y_{n(t)} = 0$) the observed gap and stays in the same lane [92].

2.3.2 Traffic Simulation Tools

In this section, we limit the focus to the most used simulation tools in the traffic modeling literature, where we explore the state of the art of these simulation tools and discuss their functionalities and characteristics. A number of criteria, such as the nature of the tool (e.g., free, open source, or commercial) and functional capabilities of the simulator, are addressed in this section [16].

- The Verkehr In Städten - SIMulationsmodell (VISSIM) is a commercial microscopic traffic simulation tool developed by Planning Transport Verkehr in Karlsruhe, Germany [94]. VISSIM is one of the common simulation tools used to simulate and evaluate traffic status and transportation control systems. It can simulate different elements, such as buses, trucks, pedestrians, and bicycles. VISSIM uses the component object model (COM) interface, which enables users to create and deploy a custom tool in VISSIM using C++, Visual Basic, or Python [17]. The latest versions of VISSIM incorporate additional autonomous vehicle-related features (communication and cooperation among vehicles) and detailed behavior specifications. The aforementioned features will utilize cooperation in lane changing, and advanced

merging algorithms for enhanced traffic network scaling. In this simulator, smaller headways have been chosen to model the cooperation among vehicles. Other add-on features are the new means of mobility that have also been introduced within the VISSIM simulator, which include cooperative autonomous vehicles (CAVs) and mobility as a service (MAAS) [95]. VISSIM is a microscopic traffic simulator for behavior-based multi-purpose traffic flow simulation [96].

- Advanced Interactive Microscopic Simulator for Urban and Non-Urban Networks (AIMSUN) is a new simulation tool that was developed by J. Barcelo and J.L. Ferrer in 2005 [97]. It is a commercial simulation tool that is capable of simulating real-world traffic situations in an urban network in order to build and validate traffic structures, public transportation networks, and new transportation infrastructure [17].

  AIMSUN is integrated with GETRAM, a simulation environment that includes different components: a traffic network editor (TEDI), a network database, a simulation module, and an application programming interface [98]. AIMSUN has developed AIMSUN LIVE, which integrates predictive-based systems that can provide real-time traffic prediction and management. In this aspect, AIMSUN LIVE can provide accurate real-time predictions of future traffic flow patterns that can be the outcome of a specific traffic management strategy. This is because AIMSUN LIVE leverages the combination of historical and real-time streaming data along with traffic congestion mitigation policies to provide accurate traffic forecasting. Subsequently, this can assist traffic control centers in utilizing the aforementioned traffic data to make real-time decisions about road network management [75].

- Multi-Agent Transport Simulation (MATSim) is another open-source simulation tool developed by the Polytechnic of Zurich that offers a range of tools for implementing very large simulation-based agents. In MATSIM, agents hold a list that

simulates the daily routine of traffic in a large area. MATSIM adopts activity-based methods that are used to model travel demand. Since MATSIM is an agent-based simulator, these agents hold a list of actionable plans and choices, which includes traditional traffic properties (e.g., travel routes and modes) and time schedules. In the MATSIM simulator, agents make their decisions according to the utilization of the integrated discrete choice models [99]. MATSIM mainly focuses on modeling individual vehicle behavior, which can be considered a drawback if we are interested in traffic behavior in general [100, 17].

- Simulation of Urban MObility (SUMO) is an open-source simulation tool that was developed in 2001 by the Institute of Transportation Systems at the German Aerospace Centre. It is capable of simulating traffic at the microscopic level and simulates moving vehicles and accidents [101]. In this simulator, the vehicle width is fixed, and it does not take into account the different types of vehicles such as buses, light rail, heavy rail, and trucks [94, 17]. SUMO is designed as an intermodal traffic-based simulator that includes public transportation, traffic road networks, and users such as pedestrians. SUMO simulators encompass a number of built-in features which include C2X communication among vehicles that are achieved through the integration of SUMO simulators with network simulators (such as OMNeT++ or ns-3), multi-modal traffic, and automated driving. Traffic management is also an additional add-on feature that can model vehicle detection loops and video detectors to manage and control traffic through traffic lights, monitoring vehicles' behaviors and adjusting traffic parameters such as vehicles' speed limits [18, 102].

- CORridor SIMulation (CORSIM) [103] is known as one of the most widely used microscopic traffic simulator software programs worldwide. CORSIM is used in thousands of applications as a standard traffic simulation tool. CORSIM is equipped

with reliable validation, continuous logic enhancement, solid verification, and calibration efforts. It can produce real-world traffic flow realistically and with high accuracy. All types of geometric conditions, including complicated traffic scenarios, can be handled virtually by CORSIM. Some of these conditions include the surfaces of streets that have different combinations of turning pockets and lanes, different types of on and off-ramps, and multi-lane freeway segments.

- Paramics addresses road networks with drivers and simulates the decisions, intentions, and subsequent actions of drivers when they move toward their destinations [104]. Depending on the characteristics of the basic network and the probability of encountering traffic congestion, drivers are considered to choose the possible route in the simulator. A set of decisions is prioritized by each driver throughout the network. These decisions include traffic speed and specific moments to change, cross, or merge into different traffic lanes. In the Paramics simulator, the network topology and travel demand drive the calibration. Flows of saturation and the proportion of lane usage are generated as outputs from the simulator to examine the road network's performance. However, these parameters cannot be provided as input for calibration assistance. Although Paramics does not prescribe the effect of a traffic model, it can simulate and model the cause of action. This way, the simulator preserves the predictive power of the simulation process in subsequent changes in the model and tests the change in the traffic road network.

- The TRansportation ANalysis and SIMulation System (TRANSIMS) creates an integrated regional transportation system environment by employing advanced computational and analytical techniques [105]. The simulation environment includes a regional population of individual travelers. TRANSIMS simulates the activities and individual interactions of travelers and their plans for the transportation system. It also simulates and determines the environmental impact of these activities.

TRANSIMS contains an interim operational capability (IOC) with numerous features, applicability, and readiness for each major module to complete different types of specific traffic case studies.

2.3.3 Summary of the Traffic Simulation Tools:

Several articles have focused on the comparison of urban road traffic simulators and provided comprehensive assessments of the existing simulation tools. Table 2.2 provides a comparison between these traffic simulators based on seven features along with their strengths and weaknesses points. We also list several key challenges in these traffic simulators that conflict with our research goal.

- A major drawback of the exciting simulation tools is the inability to implement or integrate advanced Bayesian-based models or algorithms. They use the objective optimization algorithm to simulate traffic behavior based on different traffic parameters such as route choice and vehicle movement.

- Another issue that has gained the attention of the traffic simulation community is the CPU and memory performance. Adding a number of parameters to represent different aspects of the traffic simulation model, such as traffic speed, the number of lanes, route length, and the width of the lane requires high usage of memory and the CPU, thus increasing the computation time.

- These traffic simulators embed sample events that we examine for their impact on the traffic status. These events are implemented as modules to represent limited events. These traffic simulators share similar events, such as traffic acceleration events, traffic deceleration events, and traffic red signal events.

Table 2.2: Different traffic simulation Tools and their main features and capabilities.

Features Simulator	**MATSim**	**AIMSUN**	**VISSIM**	**SUMO**	**CORSIM**	**Paramics**	**TRANSIMS**
Open Source	Yes	No	No	Yes	No	No	Yes
Visualization	2D	2D, 3D	2D, 3D	2D, 3D	2D, 3D	2D, 3D	2D, 3D
output	Text	Graphs	XML	XML	Text	Graphs	XML
Import Map	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Programming Language	C++, Java	Python, C++	C++, VB, Matlab, Python	C++, VB, Matlab, Python	Python, C++	C++, VB, Matlab, Python	C++, VB, Matlab, Python
Flexibility in infrastructure Development	Limited	Flexible	Flexible	Limited	Flexible	Flexible	Flexible
Coding	Easy	Difficult	Easy	Difficult	Difficult	Easy	Difficult

2.4 Traffic Event Modeling

Despite a significant amount of research that has been conducted on predicting traffic status, especially traffic congestion, incorporating future events such as weather conditions and traffic accidents to predict traffic congestion status remains largely unexplored. Most of the recent research has concentrated on developing methodological methods to analyze the association between weather events and traffic congestion status [106].

In a study investigating the impact of the weather on traffic flow characteristics, Aki et al. [107] found that rainfall reduces the average traffic speed by 8% to 12% and the traffic volume by 7% to 8%. According to Rakha et al. [108] rain and snow are the most common weather conditions that affect traffic flow, where they significantly reduce traffic speed and cause traffic congestion. Andrey et al. [109] provided a comparative review and discussed the relationship between weather conditions and travel risks. According to the study, rainfall has a 75% effect on traffic accidents on average among all weather data.

Overall, these studies indicate that weather characteristics have a huge impact on traffic status. Thus, the need for developing a predictive traffic system incorporating weather factors is significantly important to enhance transportation management systems. Lu and Zhou [110] proposed a short-term traffic prediction model that computes irregular and regular traffic flow patterns on freeways using a state-space model. They referred to irregular traffic flow patterns as those that occur when non-recurring conditions, such as severe weather conditions or accidents occur. For the non-recurring conditions, they adopt a polynomial trend model to predict irregular traffic flow patterns. However, the polynomial trend model has a significant methodological flaw in terms of the presumption of linearity, which may result in a poor approximation of the relationship between the traffic flow pattern and the non-recurring conditions [111, 112].

These proposed studies on modeling traffic data under different weather events mainly focus on analyzing and predicting the correlation between traffic status and weather conditions. On the other hand, research on road accident events mainly focuses on detecting the accident occurrence and its severity [113]. Numerous studies proposed utilizing machine learning and deep learning techniques such as support vector machines (SVMs), Convolutional Neural Network (CNN), Gradient Boosting Machines (GBM), long-term short memory (LSTM), and Random Forest (RF) that model video data obtained from Traffic Surveillance cameras to determine whether or not an accident has occurred or not [114, 115]. Ravindran et al. [116] developed a hierarchical support vector machine (SVM) integrated with a Histogram of gradients (HOG) to detect damaged vehicles from footage received from traffic surveillance cameras that indicates an accident. Another study by Arceda and Riveros [117] proposed a novel approach that integrates the Violent Flow (ViF) descriptor with a Support Vector Machine (SVM) to extract a number of features, such as the position and direction of the vehicle from the video frames to detect road accidents. However, these proposed models served as detection systems that detect the accident occurrence using the object detection concept, which examines sequences of

static images and uses trajectory or tracking algorithms. These models also contribute to traffic accident prediction research by modeling the relationship between the accident severity and a number of influential factors such as the location and the time of the accident [118, 119]. Miaomiao Yan and Yindong Shen [114] incorporated the temperature, humidity, day, and month as influential factors in their hybrid model that integrates random forest (RF) and Bayesian optimization (BO) to predict the severity of an accident in an urban road network. The proposed model predicts the severity level of the accident where 1 refers to a minor injury, 2 is a major injury, and 3 is a fatal injury. Overall, the BO-RF model provides a better classification ability compared to the traditional SVM, where the BO-RF achieved the highest F1-score with 57% compared to the traditional SVM model that archives 54%. Yet, the BO-RF has some drawbacks where it employs the partial dependence plot (PDP) to model the marginal effect between the response variable and the influential factors. This approach is subjected to the over-interpreting problem and an independence assumption problem, thereby affecting the prediction accuracy. Moreover, the approach has some limitations where it doesn't predict the impact of the accident severity on the traffic status or the duration of the accident based on its severity level. Predicting the accident duration is significantly important so to better manage traffic status when an accident occurs. Banishree Ghosh [120] compared a number of regression techniques for modeling the relationship between traffic factors and the duration of traffic accidents. The authors perform numerical analysis on accident data from Singapore and the Netherlands and group different accident types that have a similar impact in multiple clusters. Then, they train different predictive models such as on different clusters. The models used for comparison are Classification And Regression Tree (CART), Multi-Layer Perceptron (MLP), Support Vector Regression (SVR), and LSBoost models. Nevertheless, applying a data-driven approach to a small dataset may reduce the prediction accuracy [121]. Zinat and Mahdi [122] applied ensemble models to identify the severity of accident impact on traffic flow. The proposed classifier integrates

three models; Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Machines (GBM). This approach predicts the time delay as a result of the accident severity. A long delay indicates a high accident severity, and a short delay indicates a low accident severity. However, the delay time can result from other factors than the accident severity level, such as weather conditions, the number of road lanes, and the permissible speed limit. Considering the street type, the average speed limit at different locations should provide delay time precisely.

All this research has been proposed to predict the traffic status in the occurrence of an accident event. The key challenge in simultaneously modeling future events and predicting traffic status based on these events lies in obtaining an event profile and incorporating this profile into the prediction model. The event profile allows us to examine and track the event's patterns effectively. Thus, modeling future events is based on the correlation between the future event value and the event profile of an occurring event. Though these studies fall into the category of traffic event detection and prediction, they mostly propose models that are limited to modeling the traffic event occurrence along with a comprehensive analysis of the event's impact on traffic status. Furthermore, these studies propose approaches that are specific to a certain geographic area. The key challenge in modeling traffic events and predicting traffic status based on unplanned events simultaneously lies in obtaining an event profile and embedding this profile into the prediction model. Classifying a large number of events as a single observation in the model and tracking the event's patterns effectively at different locations and at different times is challenging due to the computational process to obtain an accurate prediction.

2.5 Summary

In this chapter, we conducted a thorough literature review of studies on traffic modelling, traffic simulators, and traffic modelling studies that incorporate event modelling.

In traffic modelling, we discovered that time series forecasting and spatial prediction methods lacking the temporal and spatial factors prediction tend to be less accurate. Therefore, we focused our evaluation on the spatiotemporal models that are widely used for analyzing and predicting traffic status. There are many methods for spatiotemporal analysis including Bayesian-based, ANN-based, or ST-Kriging-based approaches. These approaches are considered some of the most widely used models in the traffic domain. We provided a comparison evaluating various aspects of these methods in terms of their accuracy, computational complexity, and evaluation criteria. Based on the literature, these approaches suffer from several major drawbacks, such as requiring excessive time and computational resources to provide an accurate estimation. Furthermore, we defined the different traffic prediction approaches with respect to the prediction time intervals.

Both long-term and short-term time intervals were compared briefly to further illustrate the issues associated with both methods. One of the challenges with Short-term traffic prediction modelling is its sensitivity to outliers and requires high computational resources. On the other hand, aggregating traffic data in the long-term time interval modelling increases the prediction uncertainties, leading to a high rate of errors in the prediction outcome. We further highlighted how the current predictive traffic modeling approaches lack the capability of predicting the traffic status under unplanned future events that influence the traffic status. We found that most recent studies have centered on developing methodological techniques for analyzing the relationship between these events and traffic status. Furthermore, a comprehensive evaluation of the existing technologies for traffic simulation and their limitations was conducted. The evaluation given was based on seven features and their respective advantages and disadvantages. We identified a number of significant issues with these traffic simulators that were incompatible with our research objective. Recognizing the gaps in the literature with respect to spatiotemporal prediction models, event modelling, and the traffic simulator tool, we propose an integrated traffic analysis and visualization system for future road events in

Chapter 3. The system simulates accident occurrences and predicts traffic speed based on the simulated accident's influence on an interactive interface.

Chapter 3

Proposed Framework

This chapter describes the structure of the proposed approach and highlights the characteristics of each of the three components. We lay the theoretical foundations of our proposed novel system for the integrated traffic analysis and visualization system that leverages three components: the Gradient Decision Tree (GBDT) model, the boosted Linear Mixed-Effects Models(LMMs), and a visualization tool developed using the Shiny in R. These components are integrated to simulate accident occurrences and predict traffic speed based on the simulated accident's influence on an interactive interface.

3.1 Overview of the Intelligent Traffic Speed Prediction System

The integrated traffic analysis and visualization system for future road events consist of four main phases which are illustrated in Figure 3.1. We classify our components into the data collection component, the accident event and its impact component, the traffic status prediction component, and the interactive interface component. The data processing component includes event data, traffic status data, and user input data. The event profiling component comprises implementing hybrid Boosting Gradient Decision

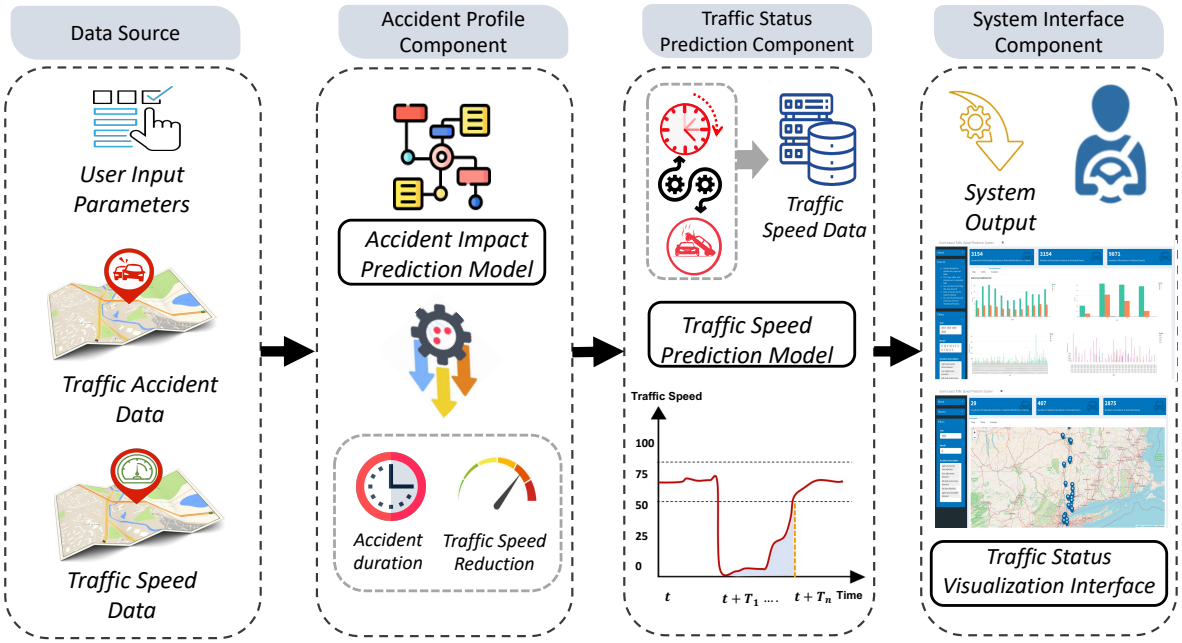

Figure 3.1: Methodology to Predict the Accident Impact on Traffic Speed.

Trees(GBDT) and boosted hierarchical Linear Mixed Effect models to simulate the accident impact at different spatial points. In this component, we predict the accident impact in terms of the duration of the event and the traffic speed reduction caused by the event. This event profile component computes the historical road event data, the historical traffic speed data, and the user input parameters. We obtain two parameters from the first component: the traffic speed reduction based on the event characteristics and the duration in a 15-minute interval. The next step in our modeling pipeline is to feed traffic speed and the event characterization received from the event profiling component to the traffic status prediction component. In this component, we utilize a boosted hierarchical Linear Mixed Effect (LMMs) model. This step integrates the predictions from the event profiling to predict the traffic status over the event duration on a 15-minute interval basis.

Figure 3.2 highlights in detail the system workflow that is used to simulate an accident event, predict its impact, predict the traffic speed over the accident duration, and finally visualize the output to the user.

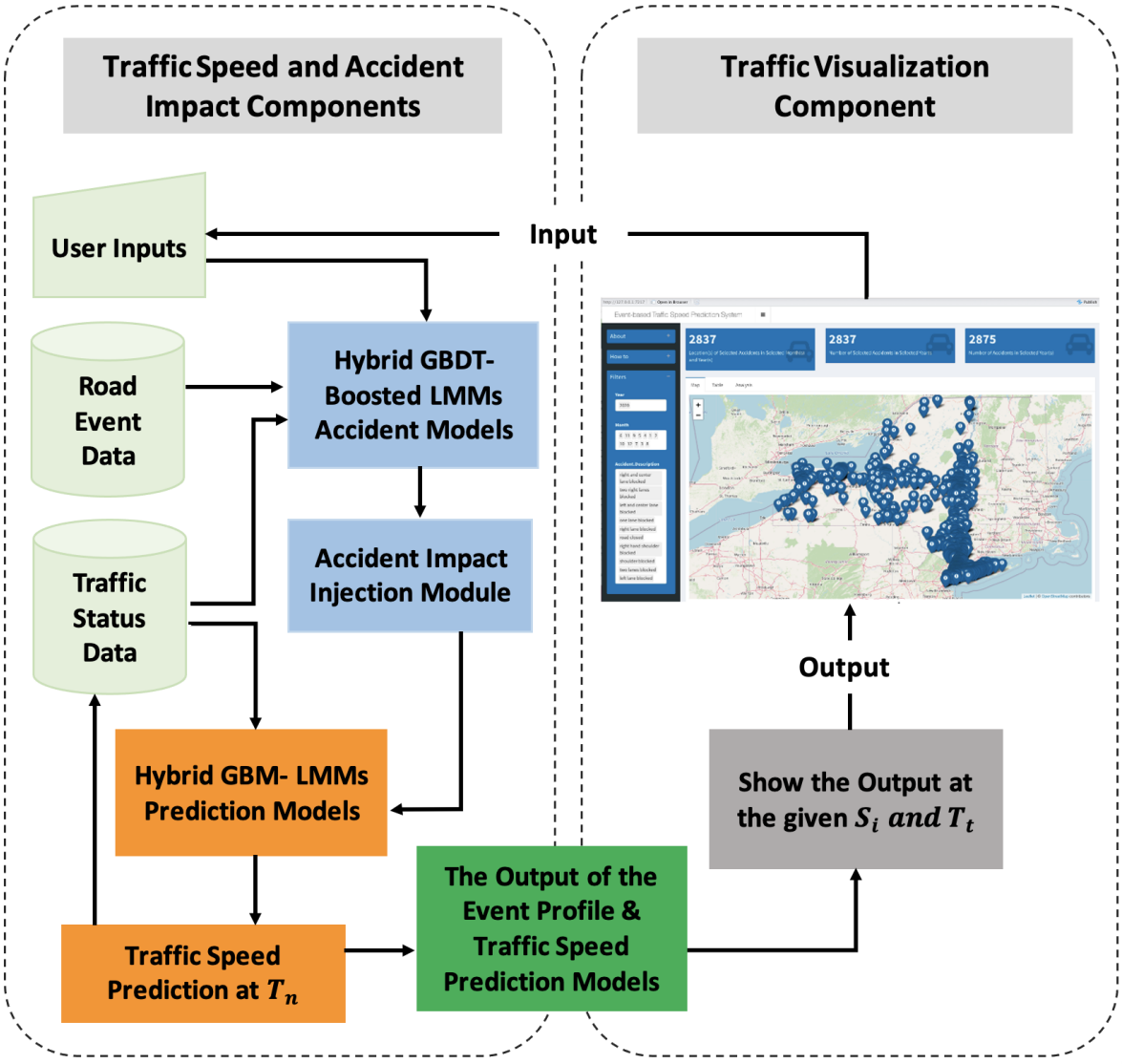

Figure 3.2: System workflow to Predict the Accident Impact on Traffic Speed.

The user input data contain the parameters that are only fed into our accident profile model. These parameters contain the selected spatial point from the map, the given time point, the type of accident, and the severity level of the accident. The event profile component takes input from three data sources; road accident data, traffic speed data, and user inputs. On the other hand, the traffic speed prediction component takes the prediction output of the event profile component as its input and computes it with the traffic speed data. Both the event profiling and traffic status prediction components run

in sequence and show the final output through an interface that represents our third component, which works as a support tool to visualize the output of the first and second components. The third component is developed using Shiny with R to offer an interactive interface to users to obtain a better understanding of future traffic status. Through the interface, the user will be able to create a specific event by selecting a location on the map and defining different parameters that formulate the event profile. Additionally, the user will be able to visualize the existing location and the historical road events data of these locations. Each of these three components is described and illustrated in more detail in the following sections.

3.1.1 Hybrid Boosting Gradient and Hierarchical Linear Mixed Effect Models

Our proposed system combines the GBM, LMMs, and boosted GBDT to perform the accident impact simulation and traffic speed prediction. The GBDT method is a particular case that follows the GBM algorithm in its implementation; however, it utilizes the decision stumps technique to estimate the targeted variable, which is the accident duration in our case. The GBDT obtains an optimal accident duration estimator $\widehat{F}(X_{accDuration})$ of an accident $X_{Acc_{event}}$ through a number of iterations when constructing decision stumps. Afterward, the GBM and the LMMs methods incorporate the $X_{accDuration}$ to predict the traffic speed y_i at the initial accident time $X_{acctime}$. The hybrid predictive system that combines the GBDT and the LMMs models outperforms these methods for accurately predicting the traffic speed over the $X_{accDuration}$. We fit the residuals r_i obtained from the LMMs with a GBM to reduce the residual error and improve the prediction accuracy. The GBM model excels at capturing predictor interactions and nonlinear effects missed by linear models, and as such, they can be extremely potent when used to boost LMMs.

Let G and F denote the LMMs and the GBM, respectively, in Equation 3.1 where both are integrated to estimate the traffic speed y_i. In the case of normally distributed

error, as in our model, the predictions from the two models can then be added together to produce new estimates for the observation y_i.

$$y_i \sim G(X_i) + F(X_i), \quad r_i \sim N(0, \sigma^2) \tag{3.1}$$

At the initial accident time $X_{acctime}$, a GBDT model predicts the accident duration $X_{accDuration}$ once. Similarly, the LMMs with the GBM model run once to predict the traffic speed for the first time point $X_{acctime}$. Afterward, the LMMs with GBM run iteratively at each 15-minute point using previous traffic speed values as autoregressive input variables and the explanatory variables obtained from the accident duration model on whether or not an accident is still occurring. These models process their predictors in a sequence where the GBM model runs after the LMMs model, incorporating the residuals from the LMMs model. We model the residual errors r_i from the LMMs model, which is used as the dependent variable for the GBM model to predict the error of the LMMs prediction over the accident duration. Figure 3.3 shows the three models' structure to predict the $X_{accDuration}$, and the traffic speed y_i.

3.1.2 Gradient Boosting Decision Trees

Gradient Boosting Decision Trees, also known as additive boosting, are ensemble learning algorithms commonly used in classification and regression problems. The GBDT uses a gradient descent algorithm which is described as a set of weak learner models, to eventually construct a single robust learner model. The GBDT follows an iterative and sequential approach wherein, in each iteration, we construct new decision stumps to classify a given dataset. These decision stumps are known as "weak learners" due to the high error rate. However, we train the new weak learner model using the errors that are obtained from the previous one to minimize the mean square error or the loss function. Thus, we will have a final learner model that achieves high accuracy and efficiency and

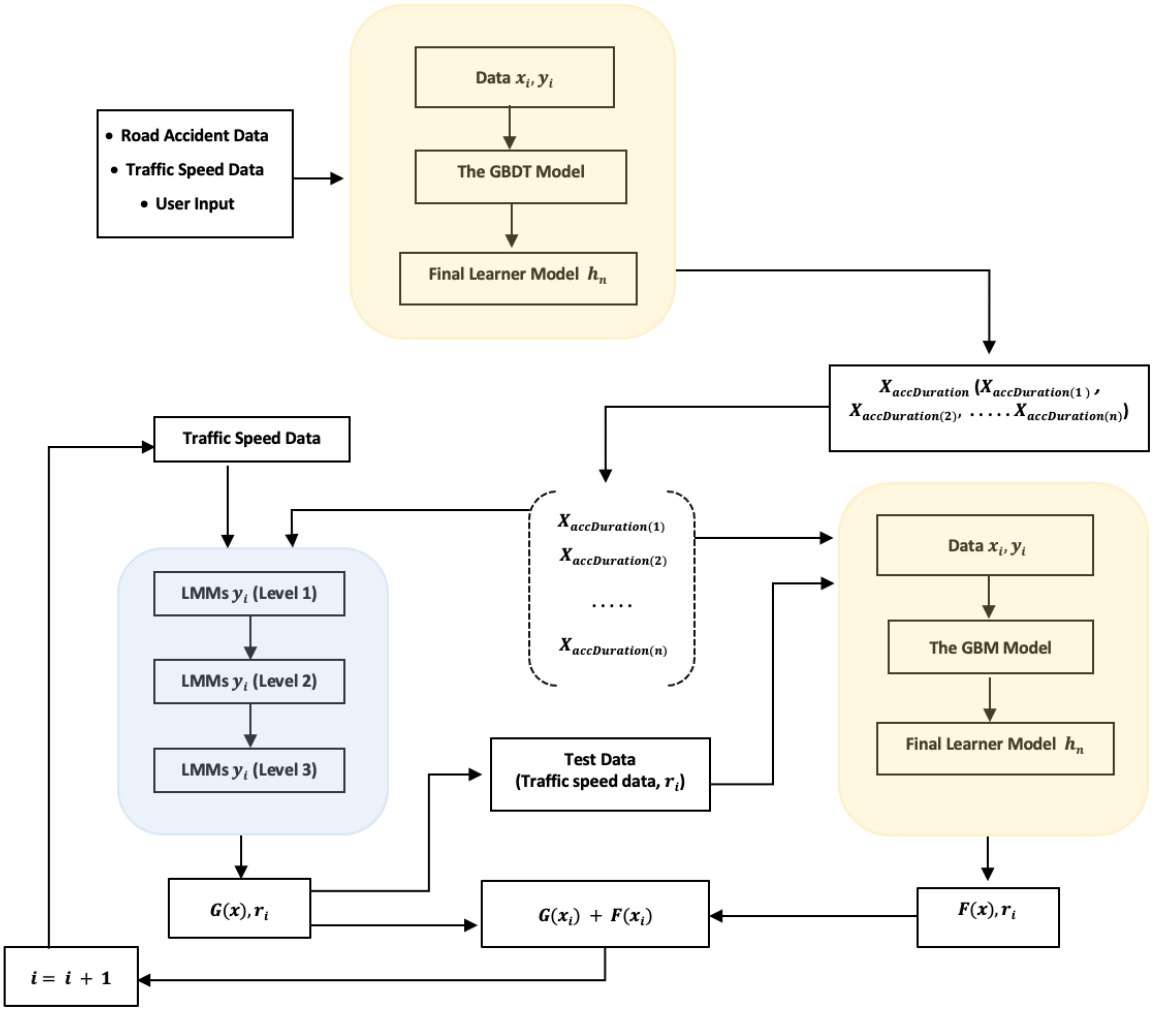

Figure 3.3: System Architecture to Predict the Accident Impact on Traffic Speed.

requires less computation time. Figure 3.4 shows the structure of the GBDT model and how each iterative stage is constructed, assuming that our dataset is (x_i, y_i).

3.1.2.1 Definition of GBDT

The GBDT is structured around three components: the loss function, the optimal learner model, and the weight of the selected data samples. These three components can be specified in Equation 3.2.

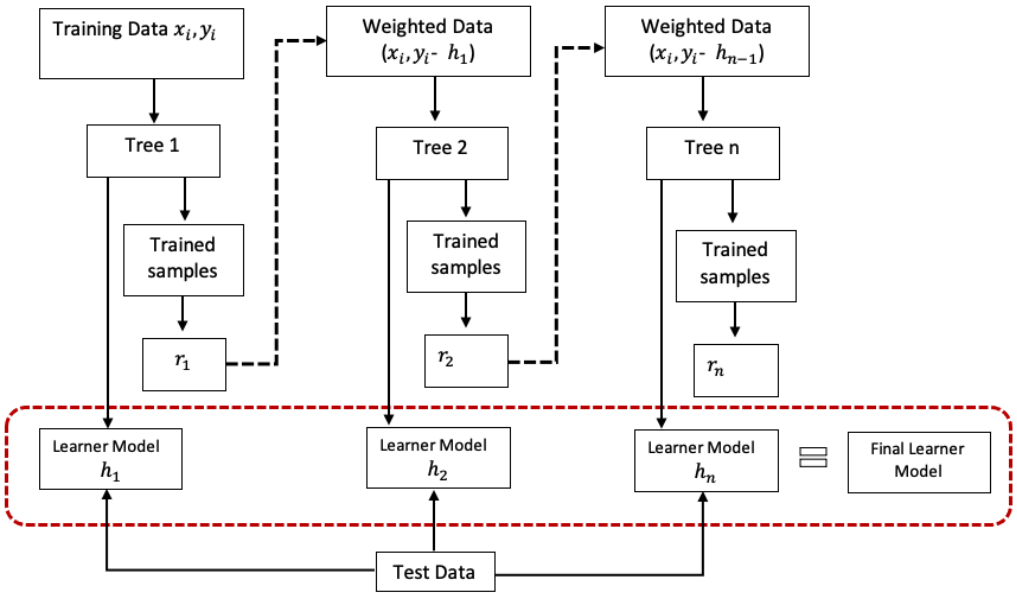

Figure 3.4: Gradient Boosting Decision Tree (GBDT)

$$\widehat{F}(x) = \arg\min_{F(x)} \mathcal{L}(F) \tag{3.2}$$

Let $\widehat{F}(x)$ be the optimal mapping function that can be obtained by minimizing the value of the loss function $\mathcal{L}(F)$. The optimal function is obtained after a number of iterations M through minimizing the value of the error/loss in the training set samples $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$, where x is the observed value, and y is the predicted value. Let $(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)$ be the training samples of the data set with n samples, and each sample has an initial weight w_i. The GBDT assigned the weight value to the new samples obtained from the previous iteration based on the loss function value. The minimum value of the loss function is the minimum assigned weight value. We define the optimized loss function $\mathcal{L}(F)$ in Equation 3.3 over the training set of N observations where $h(x)$ denotes the weak learner that models the training samples.

$$\mathcal{L}(F) = \sum_{i=1}^{N} L\left(y_i, h\left(x_i\right)\right) \tag{3.3}$$

In each boosting iteration M, a new decision stump is constructed by fitting the residual errors r_i obtained from the previous stumps to the updated weak learner model $h(x_i)$. After M iteration and achieving the minimum residual errors that can be calculated in Equation 3.4, the optimal mapping function $\widehat{F}(x)$ is obtained as shown in Equation 3.5 where f_m is the function increments that is obtained by $f_i(x) = -w_i \cdot r_i(x)$ and $i = 1, 2 \ldots M$. Algorithm 1 demonstrates the pseudocode of the GBDT model.

$$r_i = -\left[\frac{\partial L(y, F(x))}{\partial F(x)}\right]_{F(x)=f_{i-1}(x)} \tag{3.4}$$

$$F_M = \sum_{i=0}^{M} f_m \tag{3.5}$$

Algorithm 1 Gradient Boosting Decision Tree

begin

create the initial base learner model $\mathrm{L(F)}=\sum_{i=1}^{N} L\left(y_i, h\left(x_i\right)\right)$

for *iteration m= 1,2,3,...* **do**

Train $h_{(x_i)}$ from $D_{(x_i,y_i)}$

Compute $r_i = -\left[\frac{\partial L(y,F(x))}{\partial F(x)}\right]_{F(x)=f_{i-1}(x)}$

Fit h_{x_i} to the target r_i

Update the w_i

Compute $f_i(x) = -w_i \cdot r_i(x)$ and $i = 1, 2 \ldots M)$

Update the learner mode h_{x_i}

end

Output $\widehat{F}(x) = \arg\min_{F(x)} \mathcal{L}(F)$

end

3.1.3 Hierarchical Linear Mixed-Effects models (LMMs)

Linear Mixed-Effects Models (LMMs) are statistical multilevel models that model data with a complex hierarchical clustering structure. They are also referred to as hierarchical models, or conditional likelihood models, that estimate the residual error and the differences in variances at each level of the data hierarchy simultaneously [123]. The LMMs models are an extension of standard regression models that estimate predictor variables hierarchically by modeling the model's coefficients (fixed-effects) and random intercepts and slopes (random-effects) at multiple levels. In contrast to LMMs, traditional regression models assume that all observations have the same slope and intercept and only model the fixed-effects to estimate the predictor variables [124].

3.1.3.1 Definition of LMMs

The LMM shares the means and variances in hierarchically grouped data at every single model level. A detailed presentation of a given level i is provided in Equation 3.6

$$y_i = X_i\beta_i + Z_i\, u_i + \varepsilon_i \tag{3.6}$$

where y_i is the vector of the response variable measured for a given level i and $i = 1, ...N,$. We denote the number of observations in our data by n_i and $n_i \times 1$. The X_i and Z_i represent the design matrices of the fixed-effects and random-effects for the LMMs, respectively. Let $(n_i \times p)$ be the matrix dimensional of the fixed-effects coefficients vector β, and $(n_i \times q)$ be the matrix dimensional of the random-effects coefficients vector u_i specified in 3.7, and 3.8 where p is the number of fixed effect parameters, and q is the number of random effect parameters.

$$Z_i \equiv \begin{pmatrix} z_{i1}^{(1)} & z_{i1}^{(2)} & \dots & z_{i1}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{in_i}^{(1)} & z_{in_i}^{(2)} & \dots & z_{in_i}^{(q)} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_i^{(1)} & \mathbf{z}_i^{(2)} & \dots & \mathbf{z}_i^{(q)} \end{pmatrix}, \quad \mathbf{u}_i \equiv \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iq} \end{pmatrix}. \tag{3.7}$$

$$X_i \equiv \begin{pmatrix} x_{i1}^{(1)} & x_{i1}^{(2)} & \dots & x_{i1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i}^{(1)} & x_{in_i}^{(2)} & \dots & x_{in_i}^{(p)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i^{(1)} & \mathbf{x}_i^{(2)} & \dots & \mathbf{x}_i^{(p)} \end{pmatrix}, \quad \beta_i \equiv \begin{pmatrix} \beta_{i1} \\ \beta_{i1} \\ \vdots \\ \beta_{ip} \end{pmatrix}. \tag{3.8}$$

The ε_i represents the vector of the residual errors of level i, and we assume that the ε_i and u_i follow a normal distribution as described below:

$$\varepsilon_i \sim N(0, \sigma^2 I)$$

$$u_i \sim N(0, D)$$

, where $u_1, \dots, u_N, \varepsilon_1, \dots \varepsilon_N$ are mutually independent of one another for the same level i. The residual errors are distributed with zero mean and have the variance-covariance matrix σ^2. The I is the identity matrix denoted by $(n \times n)$ where the covariance matrix for u_i is denoted by D and has dimension $(q \times q)$.

3.1.3.2 Two-levels of LMMs

We can model the y_i at more than one level to evaluate the effects of higher levels on the model's coefficients (fixed-effects) and the model's slopes (random-effects) at the lowest level. These two levels of LMMs are indexed by i and j and are represented as follows:

$$y_{ij} = X_{ij}\beta_{ij} + Z_{1,ij}\, u_i + Z_{2,ij}\, u_{ij} + \varepsilon_{ij} \tag{3.9}$$

where y_i is the vector of the response variable measured for a given level i, and level j where $i = 1, ...N$, and $j = 1, ...M$. Let u_{ij} be a random-effects coefficient vector independent of the first-level i associated with the second-level j. The design matrices X_{ij} and Z_{ij} of the fixed-effects and random-effects are associated with the i level nested within the j level. The multilevel structure in LMMs succeeds in modeling multilevel data in order to obtain statistically efficient estimates of y_i.

3.1.3.3 Boosted LMMs

Boosting the LMMs with the GBM model optimizes its prediction by reducing the prediction errors of the LMMs output. This process generates a sequence of intercepts and coefficient values and identifies the optimal value that maximizes prediction accuracy. Let β_{in} and u_{in} denote a sequence of the fixed-effects coefficient vector and the random-effects coefficient vector, respectively. The boosted LMMs given level i is provided in Equation 3.10

$$\begin{aligned} y_i &= X_i\beta_{i1} + X_i\beta_{i2} + \cdots + X_i\beta_{in} + Z_i\, u_{i1} + Z_i\, u_{i2} + \cdots + Z_i\, u_{in} + \varepsilon_i, \\ y_i &= X_i(\beta_{i1}, \ldots, \beta_{in}) + Z_i\, (u_{i1}, \ldots, u_{in}) + \varepsilon_i, \\ y_i &= X_i\beta_{in} + Z_i\, u_{in} + \varepsilon_i \end{aligned} \tag{3.10}$$

with the GBM model, we identify the optimal β_i and u_i by minimizing the sum of squared residuals as a loss function for both vectors obtained in Equations 3.11

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta}(y_i - X\beta)^t(y_i - X\beta), \\ \hat{u} &= \arg\min_{u}(y_i - Z_i u)^t(y_i - Z_i u) \end{aligned} \tag{3.11}$$

The boosting process handles the loss function value, unlike the GBDT model, which handles the predictor's parameters through incremental adjustment in each constructed decision stump. In the boosted LMMs, we fit the hierarchy LMMs instead of the decision stump and the ensemble classifier with an error probability in the boosting process. The boosting here optimizes the prediction execution time and the prediction accuracy. Its insensitivity to outliers makes the GBM a robust technique for improving prediction accuracy. The stage-wise sampling approach that the GBM algorithm uses to compute residual errors from each sample prior to fitting the new predictor allows it to optimize the prediction accuracy during the boosting process. Furthermore, in comparison to other boosting algorithms, such as the Adaboost boosting algorithm, the GBM boosting algorithm is more resistant to the effects of outliers since it utilizes various loss functions. The Adaboost boosting algorithm, on the other hand, uses the exponential loss function to optimize prediction accuracy, which makes the algorithm more sensitive to the influence of outliers.

3.1.4 Shiny Application Interface Design

Shiny is an R programming language web application framework that makes it simple to create interactive web applications directly from R. The application enables users to visualize data quickly and in a customizable manner. The layout of these Shiny applications is intended to be uncluttered so that users can quickly and easily comprehend how to interact with each app's individual components, as shown in figure 3.5.

Shiny applications consist of two key components; a server that runs the spatiotemporal prediction model that is developed in R code and a user interface (UI) that runs through a user's web browser. The UI contains a layout that can place input fields and output, such as visualizations. The input fields accept inputs from the user and then send these input values back to the R server. The R server passes the parameters to the model component for modeling and then sends back the model's output to the UI to be

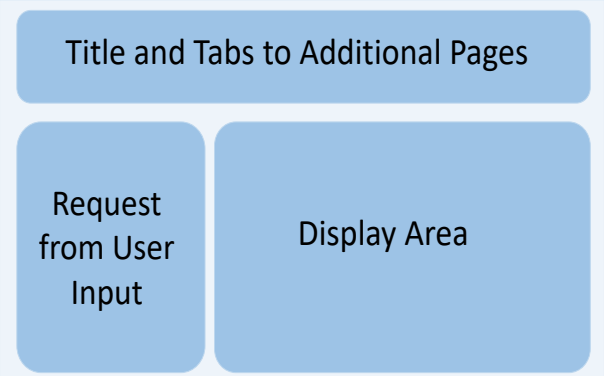

Figure 3.5: Shiny Application Layouts.

displayed (See figure 3.6). Shiny applications always need a server running R to work when we deploy the traffic visualization tool. The Shiny traffic visualization tool can be deployed on a local server or on Shinyapps.io.

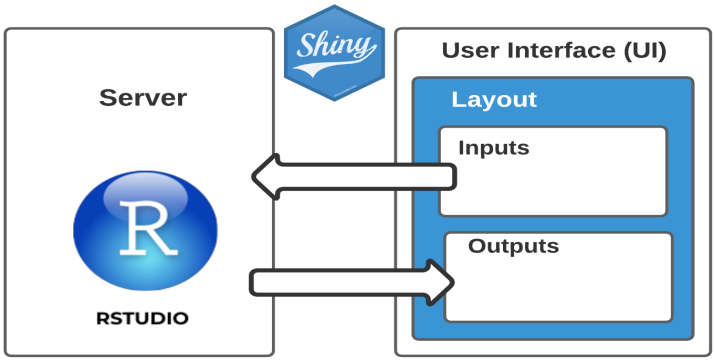

Figure 3.6: Shiny Application workflow.

3.2 System Implementation

In this section, we describe the implementation details of the event-based traffic speed analysis and prediction system that incorporates both spatial and temporal aspects to analyze the accident impact and predict traffic speed values. We discuss how to integrate the simulated accident event into the spatiotemporal traffic speed model and how to visualize the model outcomes on a user-friendly interface using Shiny with R.

3.2.1 Accident Impact Simulator Model for Duration Prediction

We model the impact of the time of the day on accident duration by dividing the day into two periods that represent the daytime and the evening time. We capture the period of the day using the Sine and Cosine functions, which give us a cyclic temporal component. Let $X_{t_{sin}}$ and Cosine $X_{t_{cos}}$ denote the time with a period of a day that is included to account for daily sinusoidally. The time period indexed by T where $cos = [cos_0, \ldots, cos_{T-1}]'$ and $sin = [sin_0, \ldots, sin_{T-1}]'$. We also included the cyclic temporal component to capture the impact of time of day or time of year on our dependent variable y_i. Additionally, a weekday variable X_{wkdy} is used to account for contrasts between weekdays and weekends. Modeling irregular time patterns in road accident data is challenging due to the differing time intervals when an accident occurs; however, defining these parameters enables the GBDT model to capture irregular time series. Incorporating a time series model here, such as AR, MA, or ARMA, won't be beneficial as each accident is only a single observation in the model.

3.2.2 Accident Impact Simulator Model for Traffic Speed Prediction

After predicting the accident duration, the boosted hierarchical LMMs incorporate the accident duration as a dependent variable to predict the traffic speed for the simulated accident. The boosted hierarchical LMMs model is structured on three levels. We pass the user input representing accident parameters to the LMM's model. The parameters specify the selected spatial point, the time point, the type of accident, and the severity level of the accident. In addition to the previously given parameters, we also compute the accident duration predicted by the GBDT model. The LMMs model predicts the traffic speed value yi at the time of the accident's occurrence. In the first level i of the

model, we predict the value y of the traffic speed, which is nestled within clusters in the second level j, which are nestled within superclusters in the third level k. The hierarchy structure of the accident data allows us to formulate the structure of the three levels of the LMMs as follows:

$$y_{ijk} = X_{ijk}\beta_{ijk} + Z_{1,ij}\, u_i + Z_{2,ij}\, u_{ij} + Z_{3,ijk}\, u_{ijk} + \varepsilon_{ijk} \tag{3.12}$$

where y_{ijk} is the vector of the traffic speed when an accident event $X_{Acc_{event}}$ occurs at a given location s and a certain time t. The traffic speed is measured on three levels: level i, level j, and level k. Let $i = 1, ...N$, defines the traffic speed value y_i when $X_{Acc_{event}}$ occurs at given spatial point $s_{(lat,long)}$, and $j = 1, ...M$, defines y_i of the $X_{Acc_{event}}$ at the same road segments s_{lane} where in $k = 1, ...K$, we define y_i when $X_{Acc_{event}}$ occurs at a larger scale level that captures the effect on neighboring road segments $s_{Municipality}$. The model hierarchy is implemented based on geographical levels of typical accident data to ensure accurate prediction for unseen observation at the $s_{(lat,long)}$ level. The hierarchical structure of the spatial component in our approach can be represented as shown in Figure 3.7.

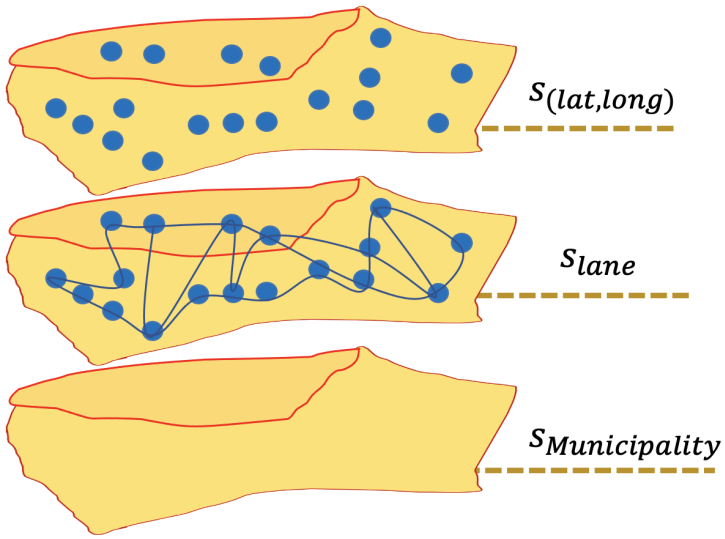

Figure 3.7: The hierarchical structure of the spatial component of the LMMs.

Additionally, we define y_i when the accident occurs $X_{Acc_{event}}$ according to the traffic flow stream direction where we incorporate the left side of the road and the right side of the road into our model. Afterward, we model the traffic speed data following the same approach as the road accident data; however, the spatial point-level $s_{(lat,long)}$ in the traffic speed data has interactions with lagged prior traffic speed over the past four periods $y_{i_{(s_{(lat,long)},t_T)}} = (y_{i_{(s_{(lat,long)},t_1)}}, y_{i_{(s_{(lat,long)},t_2)}}, y_{i_{(s_{(lat,long)},t_3)}}, y_{i_{(s_{(lat,long)},t_4)}})$. Let t_T denote the time where we want to predict the accident impact at, and t_4 t_3 t_2 and t_1 represent a full hour before the selected time t_T. We also capture the mean of the traffic speed at the same lane s_{lane} for all spatial points $s_{(lat,long)}$ in the past four periods and a day prior using 15-minute increments, which gives us 96 periods as shown in Equation 3.13, and 3.14.

$$y_{i_{(s_{lane},t_T)}} = y_{i_{(s_{lane},t_1)}} + y_{i_{(s_{lane},t_2)}} + y_{i_{(s_{lane},t_3)}} + y_{i_{(s_{lane},t_4)}} \tag{3.13}$$

$$y_i = \frac{D}{u_i} \tag{3.14}$$

Other fixed-effects in the model include two linear variables used to summarize accidents when they occur; time under accident $X_{acctime}$ and accident severity level X_{accsev} using a scale from 1 to 4. The scale represents the level of severity where 1 refers to a minor injury, 2 is a moderate injury, 3 is a major, and 4 is a fatal injury.

The random-effects in the model are the accident description Z_{acc}, municipality $Z_{Municipality}$, and location Z_s which has interactions with the past four-speed observations $Z_{(s,t-1)}$, $Z_{(s,t-2)}$, $Z_{(s,t-3)}$, $Z_{(s,t-4)}$. As we mentioned earlier, Z is used to represent the variables used to calculate random-effects, as X is for fixed-effects. The Z_s would be a dummy variable for whether or not an observation was at a particular location.

3.2.3 Sequence Traffic Speed Prediction Model

A second boosted LMMs model fits the accident duration and the initial predicted traffic speed value to predict the next traffic speed based on 15-minute time intervals. Let $y_{i_{(s_{(lat,long)},t_{+15})}}$) denote the value of traffic speed at the location $s_{(lat,long)}$ starting from the second 15 minutes of the accident occurrence to the end of the accident period. In each sequence prediction, the boosted LMMs model the previously predicted y_i to predict the next traffic speed observation y_{i+1}. In this iterative prediction process, we shift the window size for the $X_{accDuration}$ at 15-minute intervals and adjust our fixed-effects and random-effects parameters accordingly. This process is run for each of the ten simulated accident duration periods, and speeds are averaged for each time point to obtain the final predictions.

3.2.3.1 Predicting Accident impact and traffic speed at unseen locations

Our proposed system can simulate an accident at an unseen spatial point $s_{(lat,lon)_{new}}$ and predict the traffic speed based on the accident impact while taking into account the spatial point characteristics. We use the Geolocation API address from Google Maps to capture the latitude and longitude coordinates of the selected location and then convert them to its physical address. Before training the accident profile component, a function will detect the s_{lane} and $s_{Municipality}$ from the physical address of the new observation and then embed this new observation into our road accident data. We employ the e Euclidean distance function to compute the distance between the new location and existing locations that share the same characteristics. Let $s_{(lat,lon)_{new}}$ denote the spatial point of the simulated accident, and $s_{(lat,lon)_{exc}}$ be a vector of existent spatial points in our data that share the same characteristics. The Euclidean distance is defined as

$$d = \left[\left(s_{(lat)_{new}} - s_{(lat)_{exc}}\right)^p + \left(s_{(lon)_{new}} - s_{(lon)_{exc}}\right)^p\right] \tag{3.15}$$

, where d is the distance between a new spatial point and the existent spatial points, and p is the cluster of the spatial points with the same characteristics. For better accuracy, we limit the modeled observations to incorporate locations that have the same street type, whether it's an interstate highway, highway, boulevard, bridge, etc. The historical data for the same street category will be applied at a random location, and data from other locations on the same street in the same Municipality will be used to predict future traffic speed. It is assumed that no traffic signals, intersections, or other factors exist at the selected location. The prediction for the new observation follows the same steps as any other observation in our dataset. Due to the lack of historical data for the unseen location, we limit the random-effects and mixed-effects vectors to observations that have similar location characteristics.

3.2.4 Shiny Design Interface

The interface allows the user to create an accident event by selecting a location on the map and defining different parameters that formulate the accident profile, such as the location, the time under accident $X_{acctime}$, the severity level, and the type of accident. The user will be able to visualize the existing location and the historical road accident data of these locations. Also, the user can filter the locations based on the street category, where we categorized the streets into ten categories: Avenues, Boulevards, Bridges, Drives, Streets, Roads, County Highways, Interstate Highways, State Highways, and U.S. Highways. The accident type uses a natural language description of the accident, and we limit the type of the accident to 12 types of accident. Table 3.1 shows the parameters that the user will define on the interface.

3.2.4.1 Shiny Design Interface Feature

- User Inputs: this feature allows the user to choose from a list of inputs that describe the event and the other variables associated with the selected event. A user can

Attribute	Description
Accidents Description	•Left and center lanes blocked •Left lane blocked •One lane blocked •Right and center lanes blocked •Right-hand shoulder blocked •Right lane blocked •Road closed •Shoulder blocked •Three lanes blocked •Two lanes blocked •Two left lanes are blocked •Two right lanes blocked
Severity	Shows the severity of the accident, a number between 1 and 4 where 1 indicates a minor injury, 2 is a moderate injury, 3 is a major and 4 is a fatal injury.
Location	Shows the latitude and Longitude coordinates of the selected location.
Traffic Speed	Shows the traffic speed when no accident is happening.

Table 3.1: Accident event profile parameters

also select an existing location or mark a new location on the map and capture the coordinates of the new location along with its physical address.

- Downloading Plots: this feature will allow the user to export a high-resolution plot of the map with these extensions: .jpeg, .png, .svg or .pdf.
- Downloading Data: this feature will allow the user to export prediction results of the model with these extensions: .xlsx, .csv.

Figure 3.8 shows the main interface page of the system, where it shows the selected

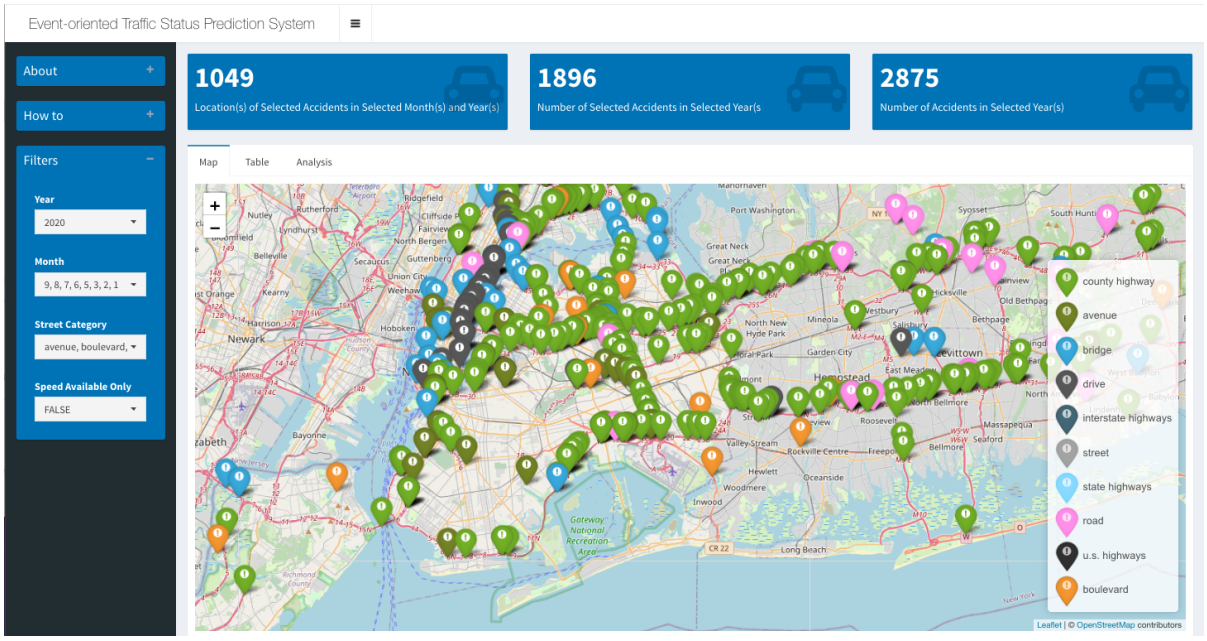

Figure 3.8: The main page of the system interface.

locations based on the street category that the user chose from the left bar. The user can zoom in and zoom out and cross the mouse on the top of the location marks to view the historical road accident of that selected location. Additionally, the system is able to display an analysis of the traffic status at the selected location as shown in Figure 3.9.

In Figure 3.10, we view how the user can select unseen spatial points and create an accident event at the new location. These locations are colored in red markers. After the user sets the markers for the new location, the coordinates of this location are captured, and the physical address is detected to pass the location information to the backend system.

After obtaining the new location information and saving it to our dataset, the user can start to define the accident event profile, such as the accident type, the time of the accident, and the severity of the accident, as shown in Figure 3.11. The prediction results are shown in the form of a table where it shows the accident duration in 15-minute intervals and the speed reduction until the accident time ends, then how the speed gradually goes back to normal speed (Figure 3.12).

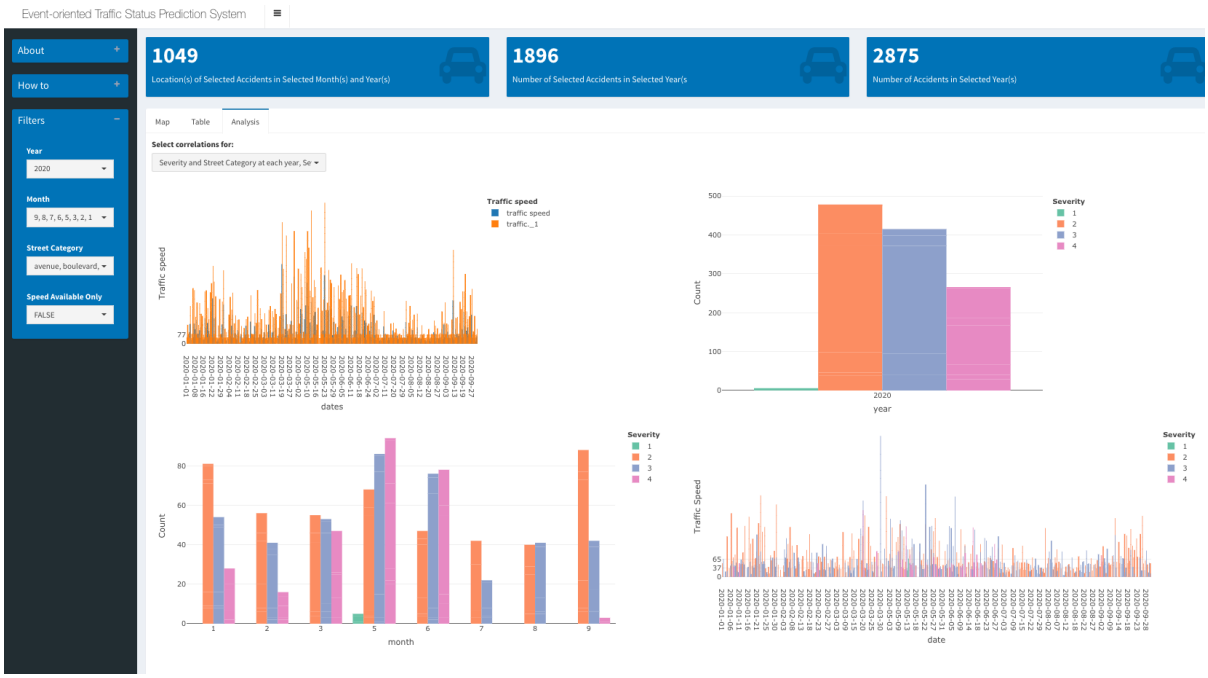

Figure 3.9: The analysis of the traffic data.

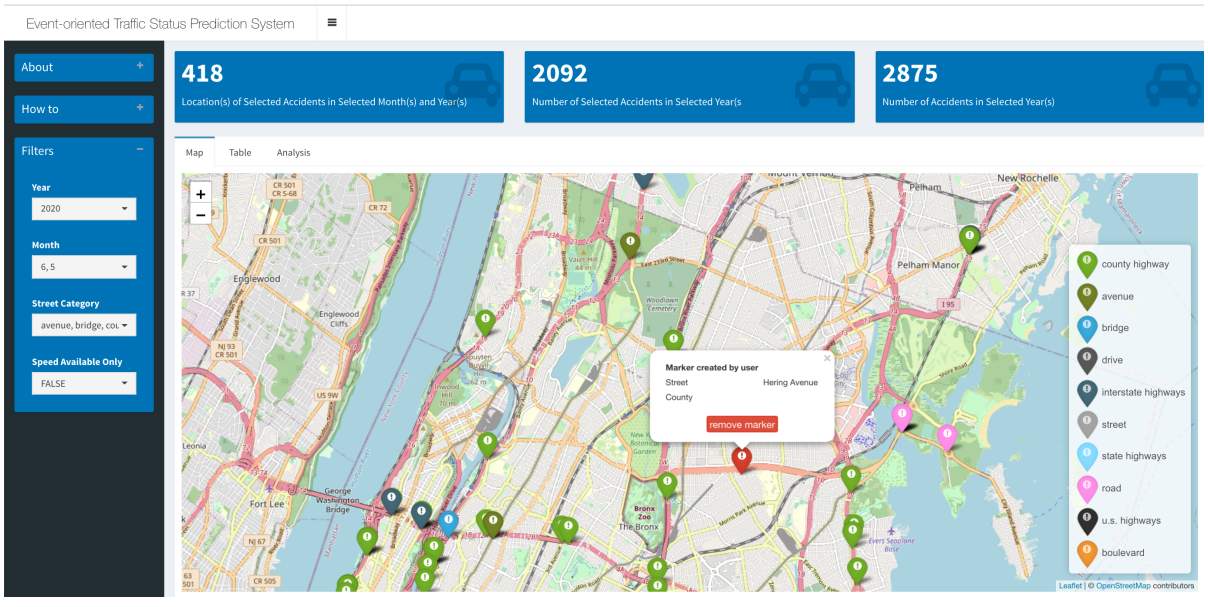

Figure 3.10: Creating a new location on the map.

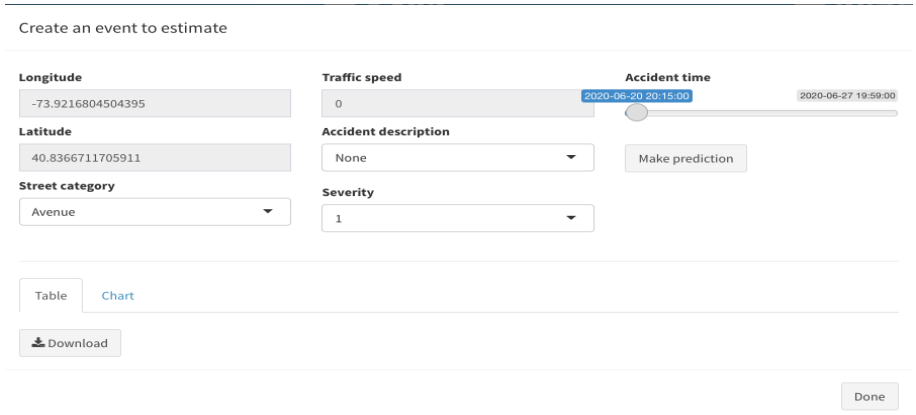

Figure 3.11: Creating an accident event profile.

3.3 Summary

In summary, the proposed system utilizes a number of highly sophisticated methods to simulate an accident impact and ensure accurate estimation of traffic prediction. The system is composed of three main components: the Gradient Decision Tree (GBDT) model, the boosted Linear Mixed-Effects Models(LMMs), and a visualization tool developed using Shiny in R. These components are integrated to simulate accident occurrences and predict traffic speed based on the simulated accident's influence on an interactive interface. The accident impact model, which we refer to as our accident impact profile computes the historical road accident data, the historical traffic speed data, and the user input parameters. We obtain two parameters from the first component: the traffic speed reduction based on the event characteristics and the duration in a 15-minute interval. These two parameters will be fed into the second component, which utilizes the LMMs and GBM models. The process of boosting the LMMs with GBM creates a series of intercepts and coefficient values that help determine the best value to use in order to achieve the highest prediction accuracy. Furthermore, simulating an accident at an unobserved spatial point would be performed using the Geolocation API address from Google Maps. This method retrieves the latitude and longitude coordinates of the selected location and converts them into its physical address. In the next chapter, we

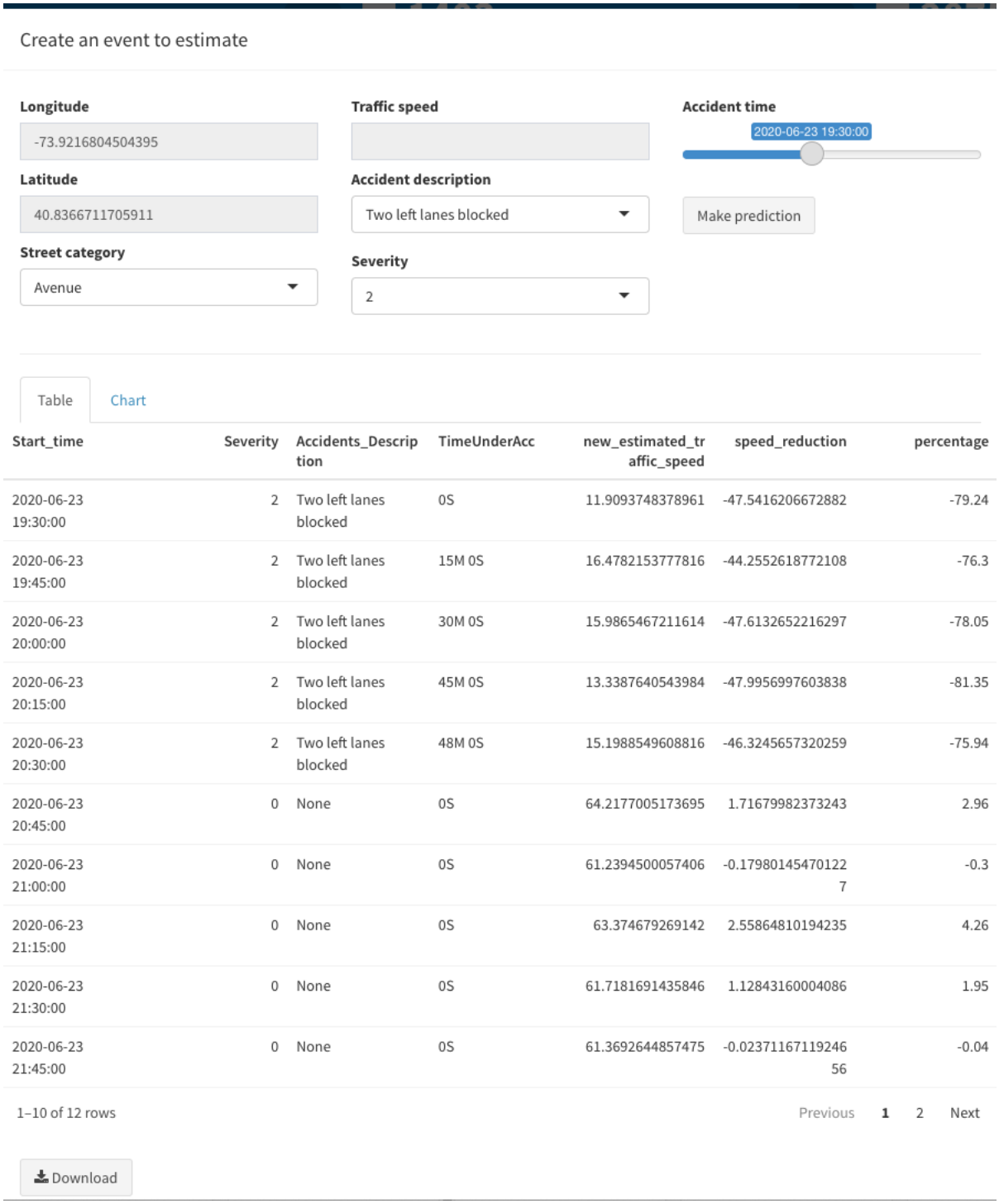

Figure 3.12: The output of our system shows in terms of speed and duration.

describe the experimental setup and the data acquisition method that is used to collect traffic data. We provide an insightful discussion of the results and the performance evaluation from different perspectives.

Chapter 4

Data Exploration and Analysis

4.1 Introduction

This chapter presents a preliminary exploratory analysis of traffic accidents and speed data. Section 4.2 begins with a brief description of the two datasets and user inputs, followed by an exploratory analysis that sheds light on the significant variables and steps needed to prepare the data for the models. This chapter's primary objective is to validate that the selected datasets are suitable for the proposed methodology.

4.2 Data Description

We provide a brief introduction of the three data we use in our system as well as we break each dataset into a number of components for modeling purposes and statistical analysis. In Section 4.2.4, a comprehensive analysis is carried out on the road accident data and the obtained traffic speed data.

4.2.1 Road Accident Data

To conduct our research, we use the US-Accident dataset [125], which contains approximately 2.8 million cases of traffic accidents that occurred in the 49 states between February 2016 and December 2021. This data was reported by the US and state departments of transportation and law enforcement agencies, in addition to using the distributed APIs within the road network, such as traffic cameras and traffic sensors. Table 4.1 shows the 39 variables in our road data after preprocessing and integrating other variables. This data is a good fit for our system because it provides a wide and varied set of data attributes to classify each accident record.

4.2.2 Traffic Speed Data

Our traffic data is derived from an open-source web-based data scraper tool called the RegTraffic [126]. This tool collects and exports usable traffic data from Google Maps. The tool extracts multiple features shown in table 4.2 such as time, coordinates, and congestion index, which refer to the average speed in kilometers per hour. It retrieves the traffic data as time series data, with each observation every 15 minutes. We use the tool to collect traffic speed data for the same spatial point as the accident data. We limit the data collection to the months of May and June of the year 2020.

4.2.3 User Input

The user input passed to our system's backend specifies seven parameters, four of which are fixed and whose values are derived directly from the map marker. The fixed parameters describe the location's coordinates (longitude and latitude), the street category, and the traffic speed. The other three parameters that the user specifies are the accident's occurrence time, the accident description, and the accident severity level. Table 4.3 displays the parameters that will be passed to our backend models.

Table 4.1: Road Accident Data Variables

Component	Attribute	Description
Spatial Component	Longitude	Shows longitude in GPS coordinates of the spatial point.
	Latitude	Shows latitude in GPS coordinate of the spatial point.
	Street	Shows the street name in the address field.
	City	Shows the city name in the address field.
	County	Shows the county name in the address field.
	State	Shows the state name in the address field.
Temporal Component	Minute	Shows the minute when the accident occurred.
	Hour	Shows the hour when the accident occurred.
	Day	Shows the day when the accident occurred.
	Month	Shows the month when the accident occurred.
	Year	Shows the year when the accident occurred.
Accident Information	Accident.Description	Shows natural language description of the accident.
	Severity	Shows the severity of the accident, a number between 1 and 4.
	Duration	Shows the duration where the accidents took to declare the road.
	Traffic.Speed.Accident	Shows the traffic speed when accidents happen.
	Traffic.Speed.Normal	Shows the traffic speed when no accidents happened.
Weather Component	Temperature(F)	Shows the temperature (in Fahrenheit).
	Humidity(%)	Shows the humidity (in percentage).
	Visibility(mi)	Shows visibility (in miles).
	Precipitation(in)	Shows precipitation amount in inches if there is any.
	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Road Condition	Amenity	indicates the presence of amenity in a nearby location.
	Bump	indicates the presence of a bump or hump in a nearby location.
	Crossing	indicates the presence of crossing in a nearby location.
	Junction	indicates the presence of a junction in a nearby location.
	No_Exit	indicates the presence of no exit in a nearby location.
	Railway	indicates the presence of a railway in a nearby location.
	Roundabout	indicates the presence of a roundabout in a nearby location.
	Station	indicates the presence of a station in a nearby location.
	Stop	indicates the presence of a stop in a nearby location.
	Traffic_Signal	indicates the presence of a traffic signal in a nearby location.
	Turning_Loop	indicates the presence of a traffic loop in a nearby location.

Table 4.2: Traffic Speed Data Variables

Component	Attribute	Description
Spatial component	Longitude	The Longitude of the road segment.
	Latitude	The latitude of the road segment.
	Street	The street name of the road segment.
Temporal component	Minute	Shows the minute when the traffic speed is obtained.
	Hour	Shows the hour when the traffic speed is obtained.
	Day	Shows the day when the traffic speed is obtained.
	Month	Shows the month when the traffic speed is obtained.
	Year	Shows the year when the traffic speed is obtained.
Speed Information	Normal Traffic Speed	The traffic speed on a road segment

Table 4.3: User Input Data Variables

Component	Attribute	Description
Spatial component	Longitude	Shows longitude in GPS coordinates of the selected location.
	Latitude	Shows longitude in GPS coordinates of the selected location
	Street Category	Dominic Matteo Dominic Matteo Dominic Matteo Dominic Matteo
Temporal component	Accident time	Shows the start time of the accident at the selected location using the format yyyy-mm-dd HH:MM: SS
Accident Information	Traffic Speed	Shows the traffic speed when no accident is happening
	Severity	Shows the severity of the accident, a number between 1 and 4 where 1 indicates a minor injury, 2 is a moderate injury, 3 is a major and 4 is a fatal injury
	Accidents Description	•Left and center lanes blocked •Left lane blocked •One lane blocked •Right and center lanes blocked •Right-hand shoulder blocked •Right lane blocked •Road closed •Shoulderblocked •Three lanes blocked •Two lanes blocked •Two left lanes are blocked •Two right lanes blocked

4.2.4 Exploratory Data Analysis

The purpose of this analysis is to investigate patterns correlated with the severity levels of the road accident data and to define the relationship between the severity level, the accident type, and the duration of an accident so as to improve the selection of features for our hybrid GBDT-LMMs model. Further analysis is performed on the collected traffic speed data in order to examine the patterns of traffic speed at 15-minute intervals in various geographical areas.

4.2.4.1 Road Accident Data Analysis

We begin by viewing the accident across the US state as shown in figure 4.1. The map reveals the accident distribution over the states of the USA. As we see, the density of accidents on the east coast is relatively higher than on the west coast. The Middle States on the map seems to have a very low density of accidents.

Figure 4.1: Accidents locations across the US states.

On the other hand, we found that California has the highest record of accidents compared to other states. Further investigation on the accident distribution across the state is performed where we started by viewing the severity of the accident across all the states as well as the length of these accidents in each state. Figure 4.2 shows the dominant severity level, which is level 2, unlike level 3 severity which is scattered across the US states. Meanwhile, level 2 is denser on the east coast, and level 1 severity is almost not noticeable as well as the severity of level 4.

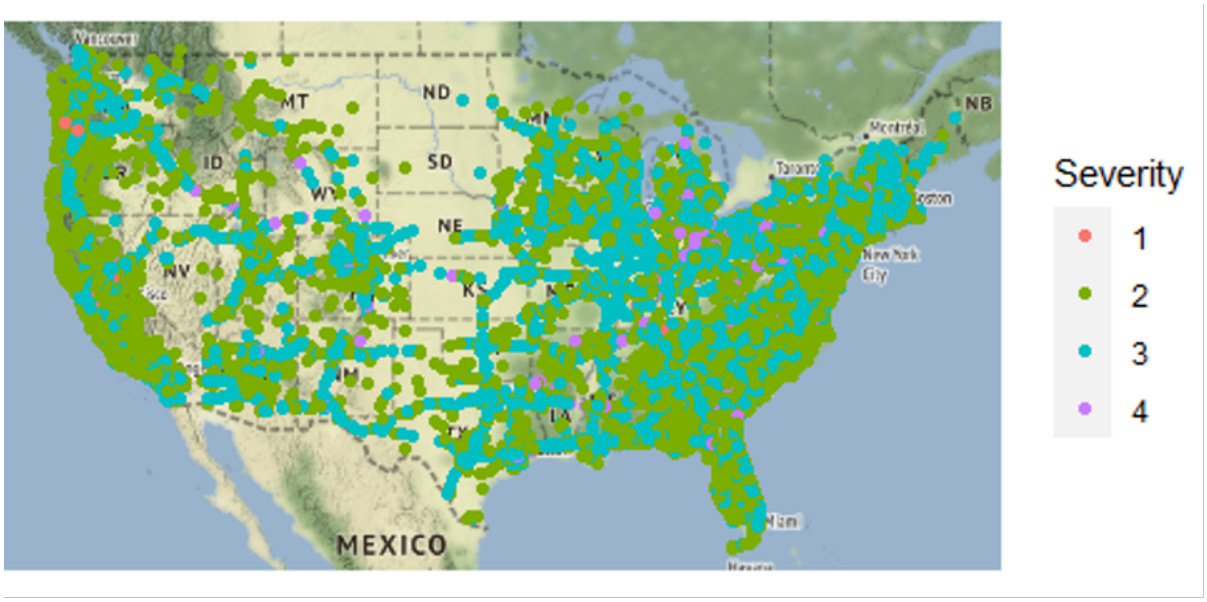

Figure 4.2: The accident distribution based on severity level.

In Figure 4.3, we view the accident distribution according to the accident description. We find that when an accident occurs, one lane is usually blocked with more than 180000 accidents, and this type of accident is typically moderate and does not involve a large number of vehicles. From that, we can conclude that there is a correlation between accident type and severity level and that the majority of accidents that block one lane result in moderate injuries. Also, most of the accidents happen on an Interstate Highway, with a significantly large number of accidents among other street categories, with more than 170000 accidents from 2016-2020. In addition, when an accident occurs, the right lane is typically blocked more than the left lane. This occurs more frequently on highways and

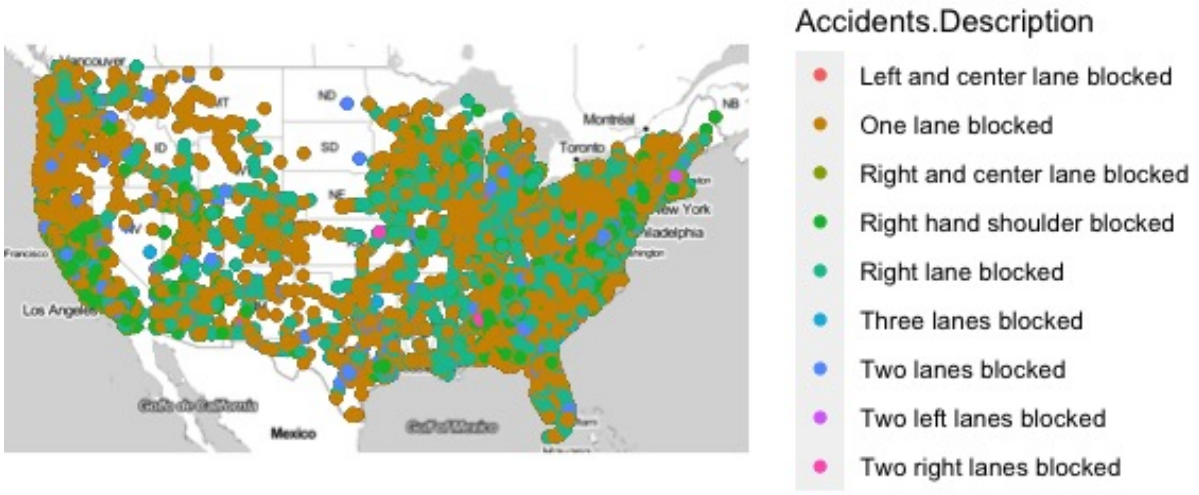

Figure 4.3: Accident distribution based on the accident type.

can result from drivers accelerating to merge in the merging lane, causing an accident. According to the United States Department of Transportation, the risk of accident occurrence increased in merging lanes compared to other lanes, with approximately 300,000 accidents occurring annually and 16.6% resulting in fatalities. Furthermore, we study the relationship between traffic speed and each accident type. Figure 4.4 shows the traffic speed distribution for each accident type, and we can see that there is an association between traffic speed and accident types.

This prompted us to investigate the duration of accidents and gain more information about our road accident data. Since each accident observation has a different starting time and ending time, we model the length of the accident by obtaining the accident duration variable. The duration of the accident is measured in minutes; however, it is difficult to view each accident with its duration time. Therefore, we create a duration category that classifies the duration into 4 intervals, as shown in Table 4.4. Based on the duration category, Figure 4.5 revealed that 24% of the accidents last between 30 minutes

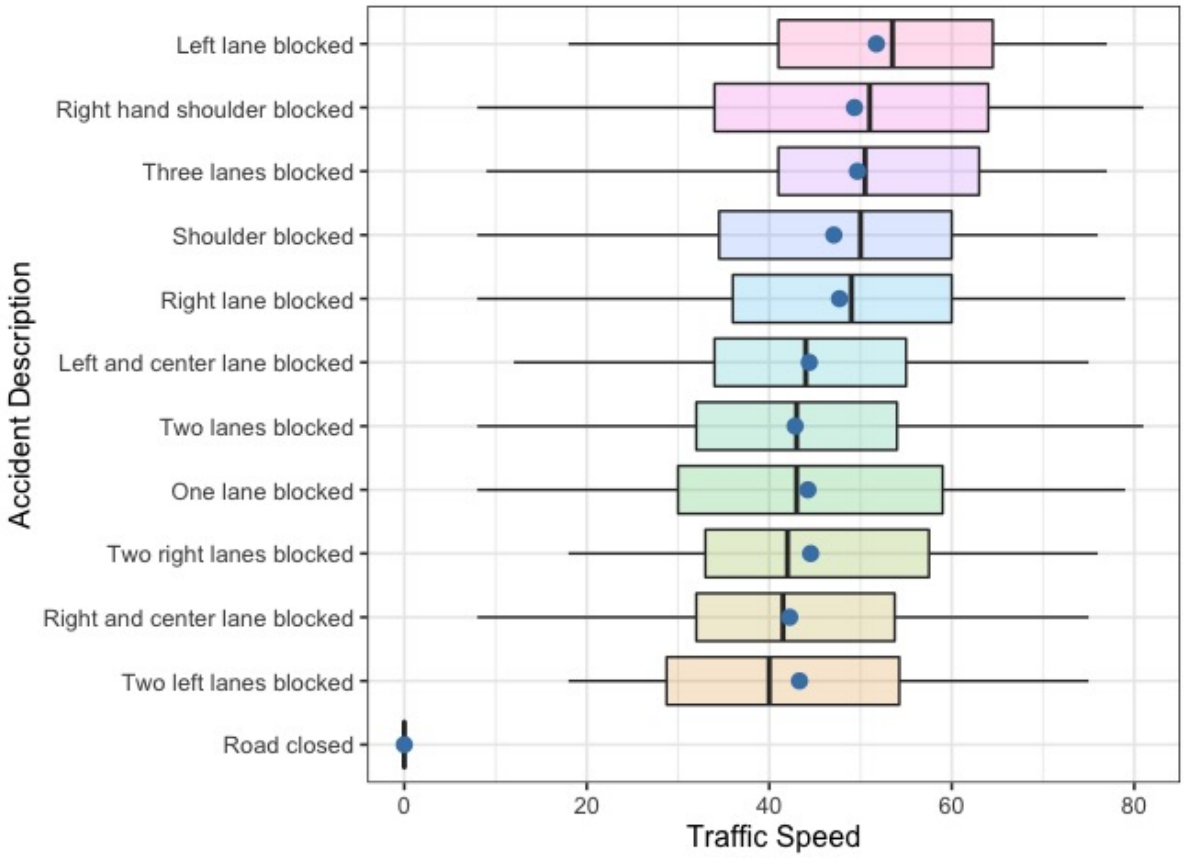

Figure 4.4: The Traffic Speed distributions based on accident type

and 60 minutes, and 35% of the accidents last more than 1 hour. The dominant category from our analysis was the accidents that lasted between 15 minutes and 30 minutes, with about 40% of the total accidents.

Table 4.4: The accident duration category.

Time Interval	**Category**
0 - 15 min	Short
15 - 30 min	Medium
30 - 60 min	Long
$<$ 60 min	Very Long

We also investigated the accident occurrences on a yearly, monthly, daily, and hourly basis by employing a number of time series analyses. Figure 4.6 illustrates the yearly accident rate from 2016 to 2020. The number of accidents increased rapidly until 2019, at which point the rate began to decline in 2020. Although the decline is not significant, where it fell by 1% in the US, the Canadian Motor Vehicle Traffic Collision Statistics

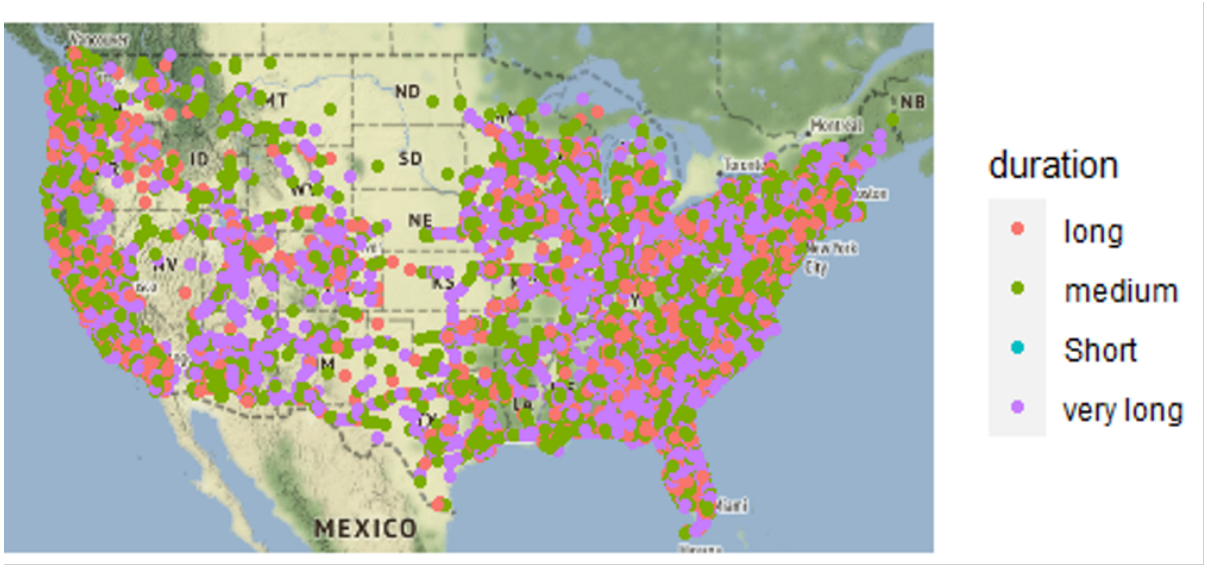

Figure 4.5: The accident distribution based on the duration

reported that the number of accident injuries and fatalities significantly decreased in 2020 compared to 2019. In 2020, the total injuries and fatalities decreased by 28% and 1%, respectively. This can be attributed to the COVID-19 effect, where fewer vehicles on the road during the lockdown and curfew lowered the number of accidents. Furthermore, we show the distribution of the accidents' occurrence from 2016 to 2020 in terms of their severity level as shown in Figure 4.7

In addition, we illustrated the number of accidents at each severity level for every month of the year in Figure 4.8. We discovered that the accident rate increases from the beginning of the year until October, when it reaches its peak, and then decreases until the end of the year in December. We noticed that in July, the accident rate is the lowest for all four severity levels, and this can be due to the weather conditions, unlike the winter period. Level 2 had the highest accident rate in October, while Level 3 had the highest accident rate from the start of the year to October.

Further analysis was performed to analyze the accident severity level over the day of the month and how it differs on each day of the month (See Figure 4.9). We found that accidents with severity levels of 3 and 2 are significantly more common than those with severity levels of 1 and 4. The monthly behavior of accidents with severity levels 3 and 2 is nearly identical. We observed that the rate of these accidents increased from day 3

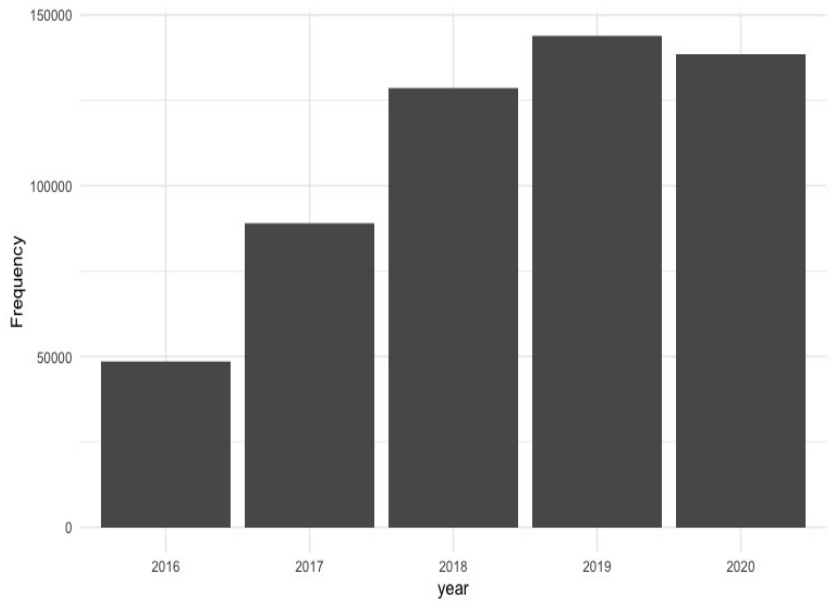

Figure 4.6: Accident number from 2016 to 2020.

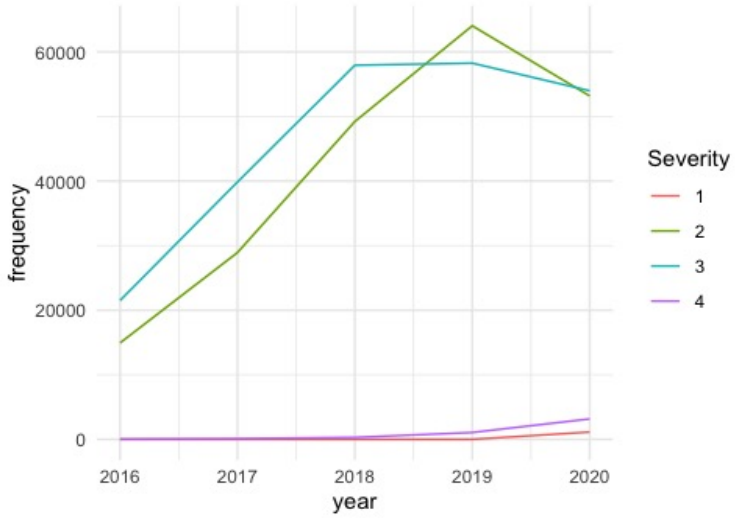

Figure 4.7: Accident number from 2016 to 2020 based on severity level.

to day 4 of the month, peaked on day 6 of the month, and then decreased by the end of the month.

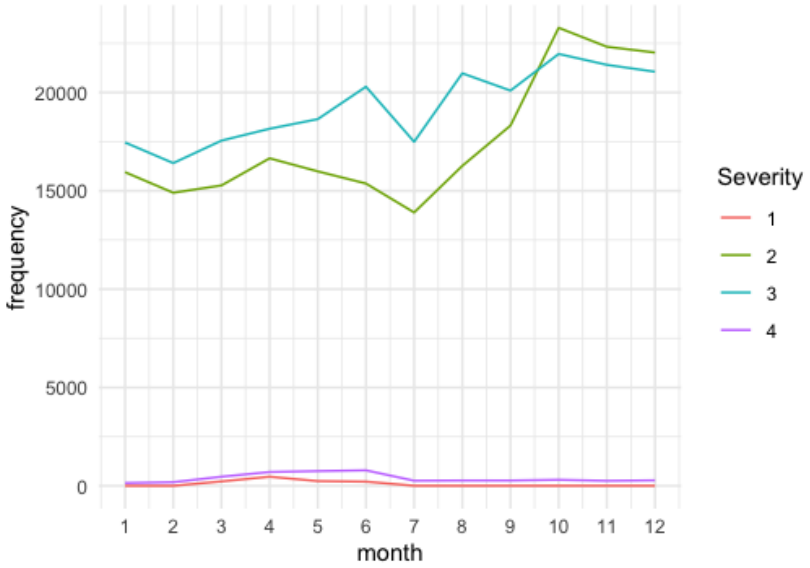

Figure 4.8: Accident distribution every month.

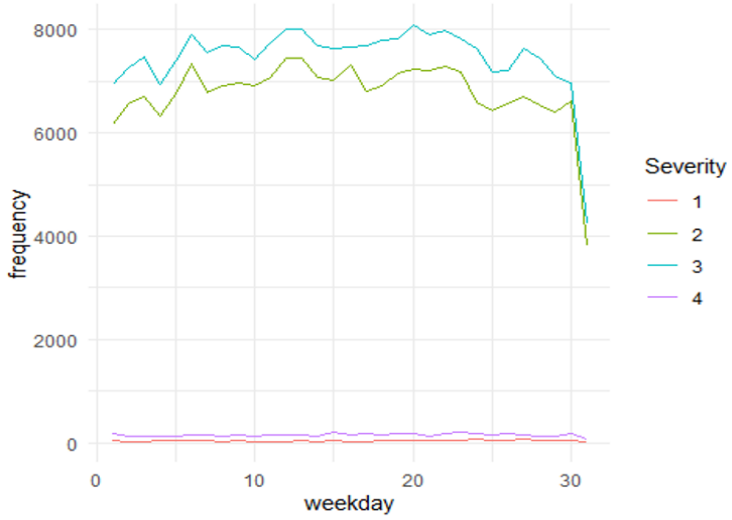

Figure 4.9: Accident distribution over a month period.

Although most accidents occur during rush hours, which are typically between 7 a.m. and 5 p.m., we analyze the accident severity level over a 24-hour period for better insights. This analysis aims to show the time series of accidents over the day and night for each severity level. Level 1 and level 4 do not have impressive records since they are much

lower than levels 3 and level 2. The peak for level 2 is at 8 a.m., with almost 20.000 accidents, and the second peak is at 5 p.m. Level 3 has the first peak at 7.30 am with about 17000 accidents, and the second peak lies at the same time as the second peak for level 2, which is at 5 p.m.

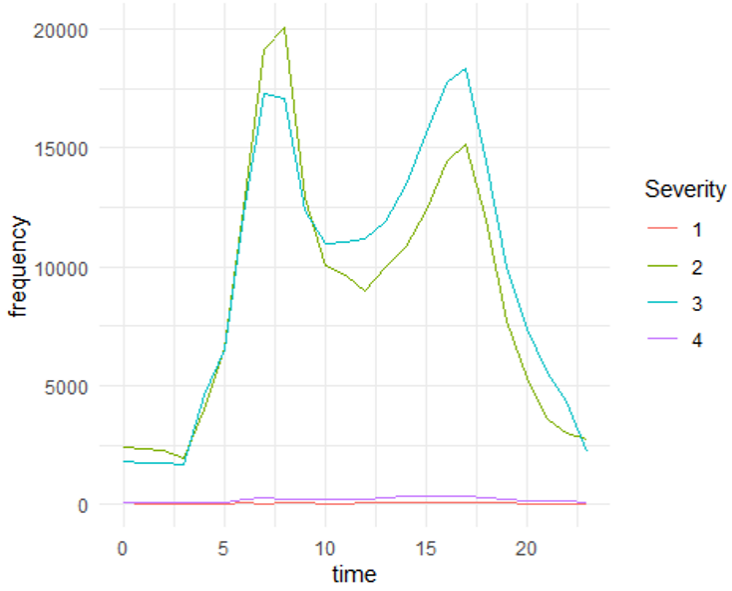

Figure 4.10: Accident distribution over 24 hours.

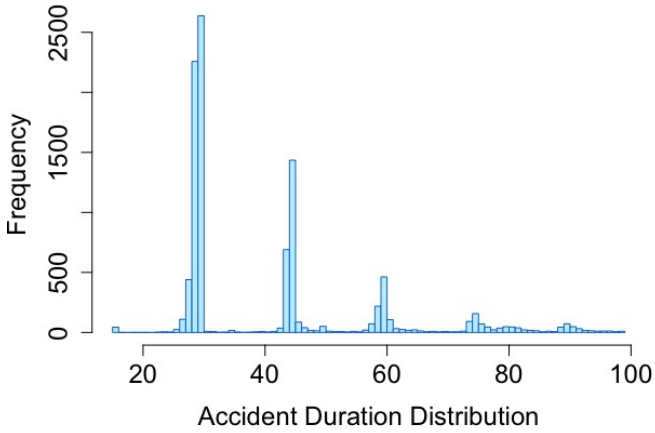

Figure 4.11: The distribution of the accident duration.

To better understand the accident's impact, we analyze the length of the accident.

Figure 4.11 shows the frequency distribution of the duration of our accident. Most accidents last between 25 and 30 minutes, and a few last longer than one hour.

4.3 Traffic Speed Data

We start by viewing the distribution of the location based on the street category where we have 9 street categories, and most of the observations are collected on county highway type as shown in Figure 4.12.

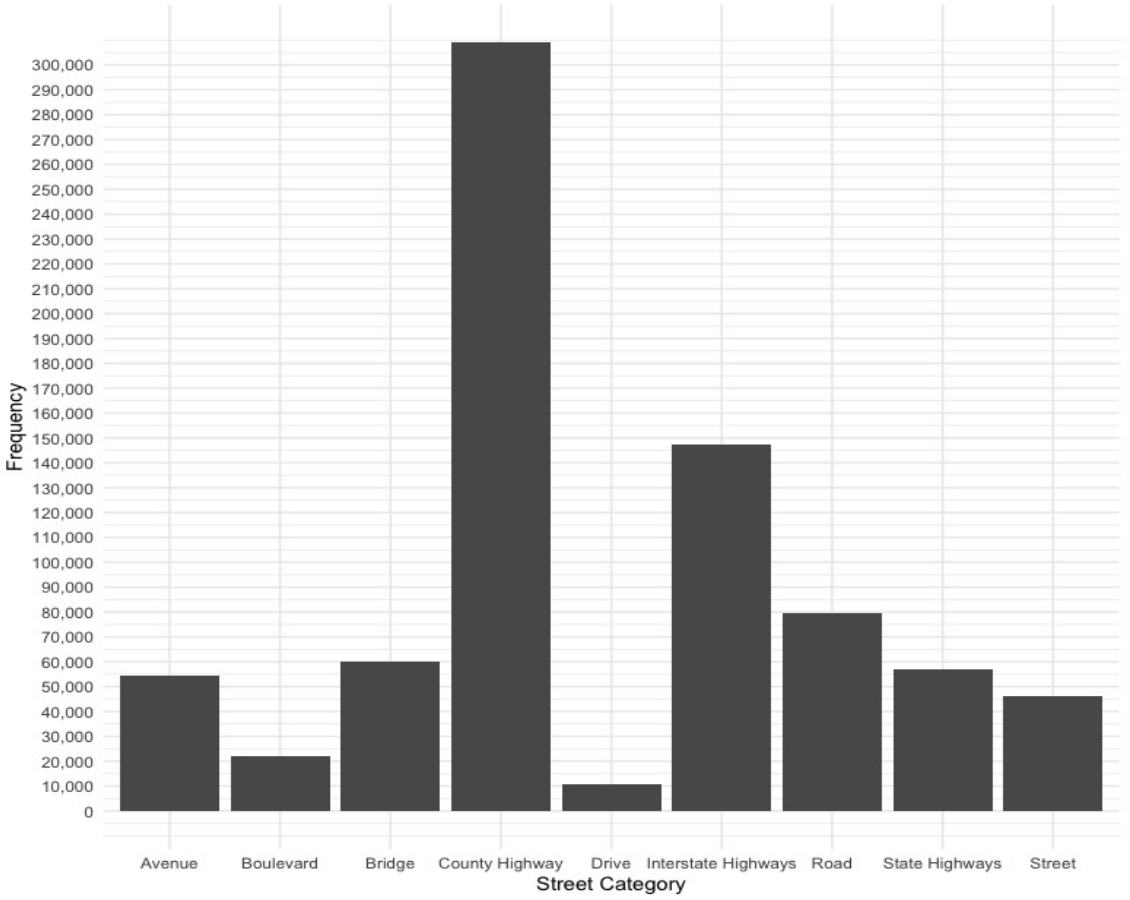

Figure 4.12: The observation distribution is based on the street category.

Further, we view the Traffic speed distribution for each street category in Figure 4.13, and we find that that street category explains about 30% of the variation in traffic speed and shows that there is an association between traffic speed and each street type.

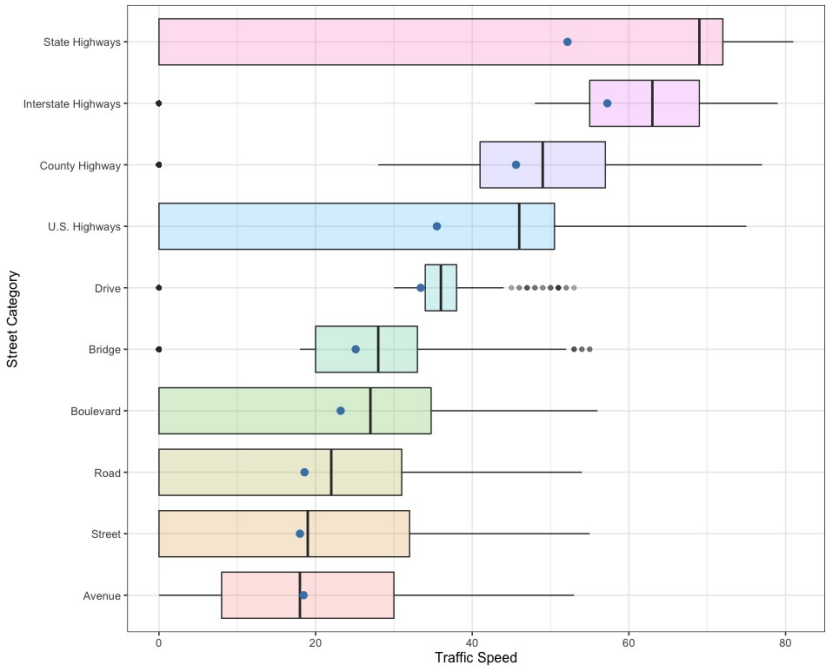

Figure 4.13: Traffic Speed distribution for each Street category.

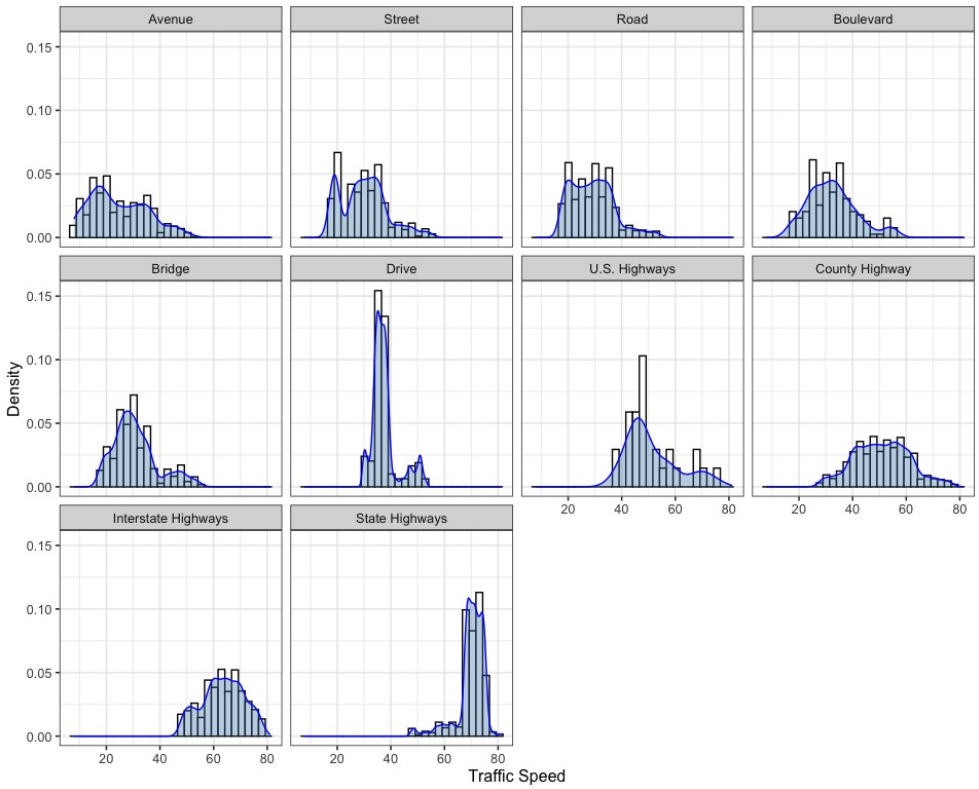

Figure 4.14: The median distribution of traffic speed for each street type

We will further investigate the traffic speed distribution in the below histograms that are sorted by median 4.14. The mean is seen to be gradually increasing, which also indicates an association. This analysis helps to select the important features in the model implementation and draw an assumption on our accident data and traffic data.

4.4 Summary

This chapter provides an exploratory analysis of road accident data as well as traffic speed data. The accident data we used to conduct our research is the US-Accident dataset, which contains approximately 2.8 million cases of traffic accidents that occurred in 49 states between February 2016 and December 2021. On the other hand, we use an open-source web-based data scraper tool called RegTraffic to collect the traffic speed data from Google Maps. We investigated the patterns of both datasets in a variety of geographical areas and found the following:

- Most of the accidents happen on an Interstate Highway, with a significantly large number of accidents among other street categories.
- Accelerating to merge in the merging lane increases the risk of accident occurrence compared to other lanes.
- About 40% of the accidents lasted between 15 and 30 minutes, and the accident rate increases from the beginning of the year until October, when it reaches its peak and this can be due to the weather conditions.
- Since each accident observation has a different starting and ending time, modelling the duration of the accident in minutes will be challenging. Therefore, creating a duration category that classifies the duration into 4 intervals (short, medium, long, and very long) is recommended. This will help us gain more information about our road accident data and how the type of accident could contribute to the length of the accident.

Both datasets were prepared to classify their attributes in terms of their spatial, and temporal components. For the road accident data, defining the road condition, and the weather condition is essential to examine their impact on the accident occurrence. Furthermore, when computing the user inputs, three categories of data are classified: spatial, temporal, and accident information components. The system will be able to model these parameters based on their types, such as time, coordinates, or numerical variables. In the next chapter, we conduct our experiments and evaluate the system's performance in terms of prediction accuracy and computational complexity.

Chapter 5

Experiments and Findings

In this chapter, we present the results of the experiments to evaluate the performance of the proposed system. First, we describe the experimental setup and the performance metrics for assessing the performance. We describe the study area and the results of our case studies. We experiment on existent and unseen locations to evaluate the system's performance. We removed the location from the dataset and performed the prediction for model performance validation for the unseen location.

5.1 Computation Requirements

Despite that the proposed system primarily runs offline, access to real-time streaming traffic would improve the system's performance with higher prediction accuracy and the ability to capture short-lived events as well as fine-grained traffic status in real-time. The real-time operation mode would allow the system to get real-time traffic status from traffic sensors and adapt the prediction of the actual status, as opposed to estimating the prediction based on historical data alone. For offline mode, we divided the data into 70% for training and 30% t for testing. With a Lambda machine (specs: 256 GB of memory and a 16-core AMD CPU), it took approximately 7 hours to train 70% of 2.8 million observations. We anticipate that the training time would be reduced if the models

were trained on a 32-core or higher machine. The prediction time is roughly between 30 seconds and 1.50 minutes, which we consider acceptable because we are still predicting for the 15-minute window.

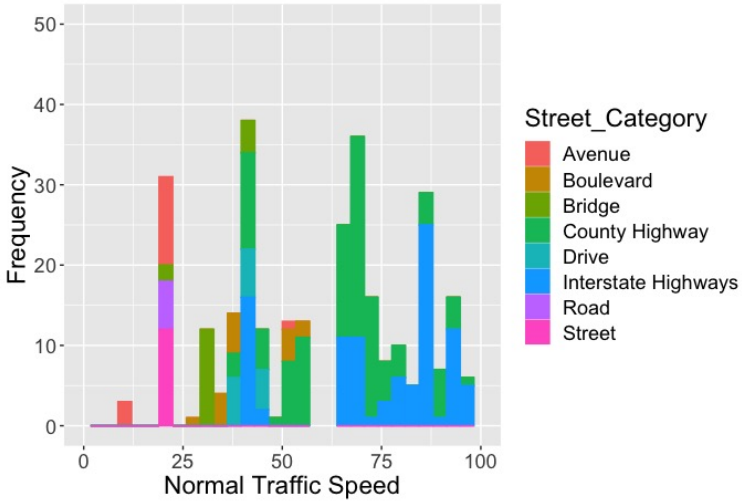

Figure 5.1: The traffic speed distribution based on street category.

5.2 Experimental Setup

In our experiment, we set the collected normal traffic speed data to cover 160 unique geospatial locations in New York State at 15-minute intervals for the entire month of May and June 2020. The normal traffic speed data provides the average speed of a road link in kilometers per hour (kph). The histogram figure 5.1 shows that the speed rate is heavily clustered into two bands dependent only on the street category variable. Each band only features a few different values as the numbers are rounded to the nearest integer. On the other hand, Figure 5.2 shows that the accident duration data is heavily right-skewed, and the median accident duration is 44 minutes. The road accident data cover a much larger time span from 2016-03-26 to 2020-12-31 than the normal traffic speed data, with 12,552 accidents in the dataset. Due to these issues with the normal traffic speed data, a number of statistical adjustments are performed to adjust our datasets. We estimate the

percentage reduction in speed during an accident based on outside estimates dependent on accident type to give a more realistic depiction of accident impact. Additionally, traffic speed is jittered to avoid over-fitting and make it more normally distributed for the LMM as it is rounded to the nearest traffic speed value. We perform this process using a random number generator with a uniform distribution from -0.5 to 0.5 when fitting LMM. The experiment setup is important to achieve a practical, accurate prediction, as in our case, there is a trade-off between the prediction accuracy and the variation in modeled data.

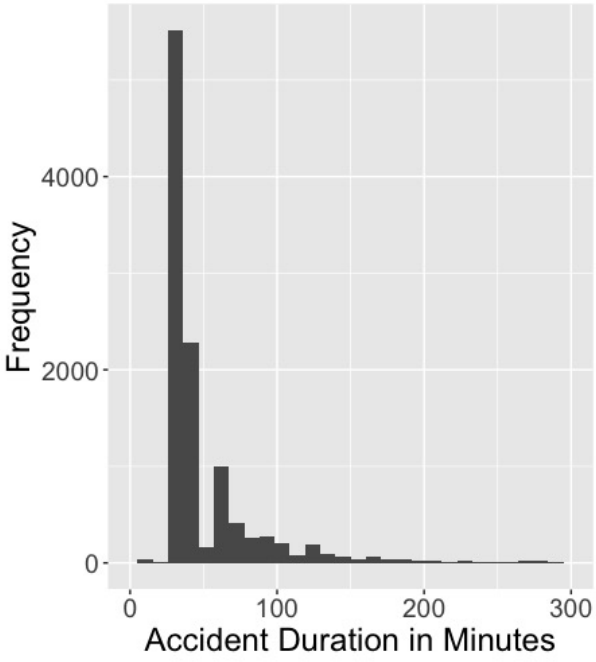

Figure 5.2: The median accident duration from 2016-2020.

5.3 System Performance Evaluation

We evaluate the system's efficiency using quantitative and graphical methods. In the quantitative methods, we compare the predicted traffic status and the ground truth we built from our collected/historical data using various evaluation metrics such as distance and error metrics.

The graphical methods compare the time series charts of the predicted and observed

series. Both graphical and quantitative methods demonstrate how well predictions agree with observations when comparing long series of predicted data.

5.3.1 Error Metrics

A number of prediction error measurements are widely used to measure prediction accuracy, such as the mean absolute error (MAE), root means squared error (RMSE), and the coefficient of determination R^2. In this section, we apply these errors measurement to evaluate our models' performance.

5.3.1.1 MAE

The Mean Absolute Error is the most used method for measuring the average magnitude of errors in a given set of predictions, and it is simple to interpret. The MAE is useful for comparing prediction approaches applied to a single continuous variable or multiple continuous variables with the same units (in this case, km/h) due to its ease of interpretation and computation. The MAE formula is defined in Equation 5.1 where $\hat{y}_i \in \mathbb{R}^N$ represents the predicted traffic speed value, and $y_i \in \mathbb{R}^N$ represents the observed traffic speed value.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{N} \tag{5.1}$$

5.3.1.2 RMSE

The Root Mean Square error, also known as the standard deviation of the residuals, estimates the average of the absolute difference between predicted and actual values and takes the square root of the mean of the residuals to cater for positive and negative differences.

$$\text{RMSE } = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{N}} \tag{5.2}$$

5.3.1.3 R^2

R squared, commonly known as the coefficient of determination, is a quantitative measure of the variation in the dependent variable $\hat{y}_i$ that can be directly attributed to the independent variable y_i.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y_i})^2} \tag{5.3}$$

5.3.2 Experiment (1): Simulating an accident on existent Locations

In this experiment, we have chosen one of Staten City's existing locations, which is a borough of New York City. The location chosen represents the urban Interstate 278 (I-278) with an average traffic speed of 95 kph. According to our historical data, a number of accidents have occurred at this location at various times, with the most recent incident occurring on June 1, 2020. This location was chosen to illustrate the accident's effect on a major interstate highway that goes over Verrazano Bridge to connect New York City neighbors such as Staten Island and Brooklyn. It has seven lanes, with three in each direction and one that is a reversible lane. When there is traffic congestion, the reversible lane increases the road's capacity and helps reduce traffic in either direction. Figure 5.3 shows the location of the selected spatial point and the last recorded accident information at this location which was two lanes blocked with the severity level 4.

We show the user a piece of information about the selected location; however, in our model, we incorporate the roadside of the accident location and the accident impact on the spatial points on all highways in general and the same high in particular. We chose

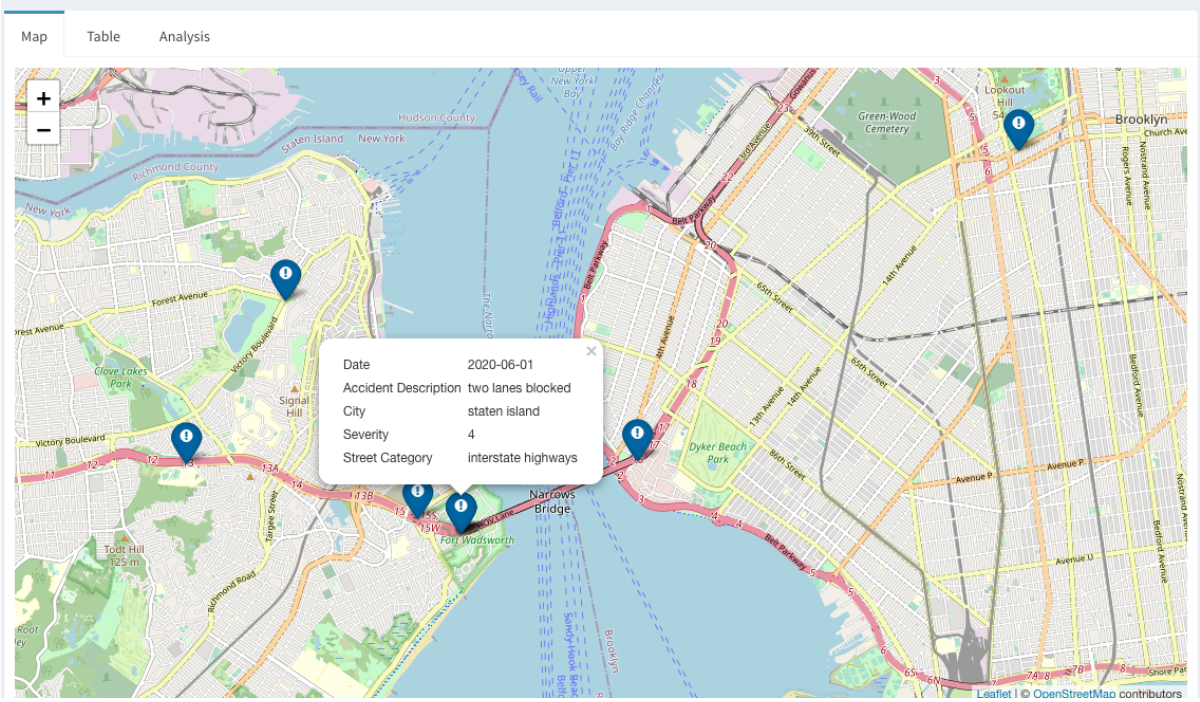

Figure 5.3: Information of the selected location- Experiment 1.

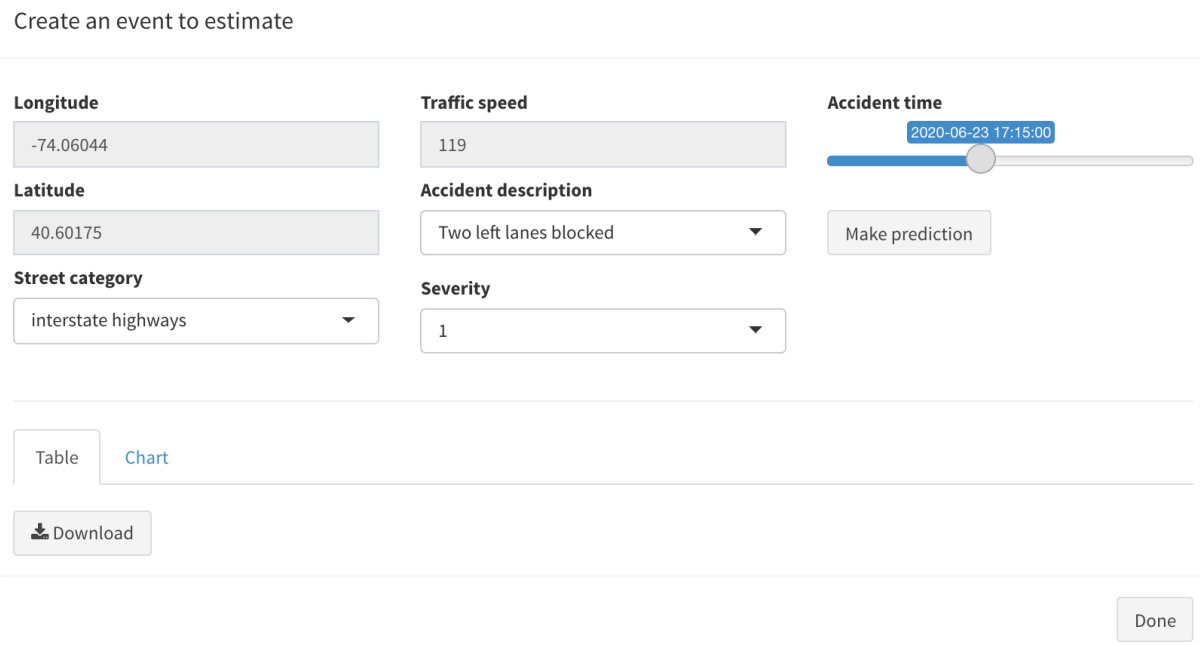

Figure 5.4: Accident profile creation.

this location due to the complexity of bridge accident occurrences. The confined space on a bridge causes a significant reduction in speed and a longer wait time until the accident is cleared. The accident we simulated is described as blocking two left lanes at 5:15 p.m., which is considered a peak hour. The accident's severity level is 1, and the average traffic speed at this location from 2016 to 2020 is usually 97 kph.

5.3.2.1 Experimental (1) Results

The results of the simulated accident scenario can be shown in Figure 5.5. The model shows that the accident duration of the simulated accident is 39 minutes, and the speed reduction in the first 15 minutes is reduced by 40 kph which causes a massive drop in the speed from 119 kph to 77 kph. Starting from the second 15 minutes, the speed rate slowly increased to 81 kph, possibly due to the traffic congestion caused by accidents. Also, the confined space on a bridge makes it difficult for the volume of traffic to pass through, so an accident on a bridge significantly impacts speed compared to an accident on regular roads. In the last 15 minutes of the entire accident duration, the speed rate was around 91 kph, and then after the accident scene was cleared, the speed went back to above 100 kph.

5.3.3 Experiment (2): Simulating an accident on existent Locations

Another experiment was conducted by selecting a different spatial point on Manhattan's Interstate 95, which is connected to multiple roads. The distinct characteristic of this location is that its traffic flow originates from the east side of the Alexander Hamilton Bridge, which is approximately 2 kilometers far. It connects the Bronx to Manhattan and has four lanes in each direction. We set the time of the simulated accident to coincide with rush hour, 5:15 p.m., and the level of severity to 4. This location's average traffic speed is approximately 97 km/h, and there have been more than 37 accidents between

Table Chart

Start_time	Severity	Accidents_Description	TimeUnderAcc	new_estimated_traffic_speed	speed_reduction	percentage
2020-06-23 17:15:00	1	Two left lanes blocked	0S	77.1626357875679	-40.8306302649565	-34.31
2020-06-23 17:30:00	1	Two left lanes blocked	15M 0S	81.9485982006816	-34.7576886128854	-29.71
2020-06-23 17:45:00	1	Two left lanes blocked	30M 0S	82.9429940094001	-34.381336580406	-29.39
2020-06-23 18:00:00	1	Two left lanes blocked	39M 0S	91.8991955852523	-25.0971990188084	-21.45
2020-06-23 18:15:00	0	None	0S	89.6668573861345	-26.4699123333552	-22.82
2020-06-23 18:30:00	0	None	0S	100.224348795731	-18.5929461848901	-15.62
2020-06-23 18:45:00	0	None	0S	103.632568555659	-14.068340533818	-11.92
2020-06-23 19:00:00	0	None	0S	107.931521803653	-7.28769847657603	-6.34
2020-06-23 19:15:00	0	None	0S	112.69190153252	-4.29492539433174	-3.67
2020-06-23 19:30:00	0	None	0S	112.343391519948	-5.92221564670581	-4.98

1–10 of 12 rows Previous 1 2 Next

Figure 5.5: Results of Experiment 1.

2016 and 2020. Figure 5.6 demonstrates the location of the selected spatial point on the map as well as the most recent accident data recorded at this location. According to the National Highway Traffic Safety Administration (NHTSA), I-95 is one of the major accident hotspots for truck accidents, especially around the area at the end of the Alexander Hamilton Bridge.

5.3.3.1 Experimental (2) Results

The accident duration of the simulated accident scenario can be shown in Figure 5.7 where the accident impact lasts for 41 minutes and causes a 23.69% reduction in speed during the first 15 minutes. The speed rate decreased from 116 kph to 89 kph and then gradually increased until the speed reduction reached 10.51% after 30 minutes, bringing the speed back to 103 kph. In this experiment, the decrease in speed rate was not as

severe as in the previous one, and this is due to the increased number of lanes that allow traffic to flow. In addition, the traffic flow at this location can branch onto the nearby roads or proceed directly to the exit of the interstate highway.

Figure 5.6: Information of the selected location- Experiment 2.

5.3.4 Experiment (3): Simulating an accident on unseen Locations

This experiment demonstrates our system's ability to simulate an accident in a location that does not exist in our data and predict traffic speed based on the accident effect. The system will consider the selected spatial point characteristics and model them with the existing spatial point characteristics. When a user clicks on a location on the map, the Geolocation API address from Google Maps captures the location's latitude and longitude coordinates. The coordinates are then converted to their physical address, and from that, we detect the spatial point street name to classify the street type. The selected location is highlighted in red in Figure 5.8, which is on Woodhaven Boulevard in Queens City. The nearest locations to our new location are highlighted in blue and are around

Table | Chart

Start_time	Severity	Accidents_Description	TimeUnderAcc	new_estimated_traffic_speed	speed_reduction	percentage
2020-06-23 17:15:00	4	Two left lanes blocked	0S	89.1619593310491	-27.716975021948	-23.69
2020-06-23 17:30:00	4	Two left lanes blocked	15M 0S	93.7814828208144	-21.9199421677008	-19.06
2020-06-23 17:45:00	4	Two left lanes blocked	30M 0S	96.8858417187689	-20.5021189105146	-17.67
2020-06-23 18:00:00	4	Two left lanes blocked	41M 0S	104.844199514746	-12.0904287975536	-10.51
2020-06-23 18:15:00	0	None	0S	103.02122471043	-14.2773824012551	-12.1
2020-06-23 18:30:00	0	None	0S	106.300761961205	-9.40665067538025	-7.9
2020-06-23 18:45:00	0	None	0S	111.353884549007	-6.19796806489178	-5.34
2020-06-23 19:00:00	0	None	0S	111.648902958928	-5.20532128739356	-4.49
2020-06-23 19:15:00	0	None	0S	113.874283611476	-3.84783922616479	-3.26
2020-06-23 19:30:00	0	None	0S	114.03492081035	-2.03405791108074	-1.71

1–10 of 12 rows Previous 1 2 Next

Figure 5.7: Results of Experiment 2.

5 kilometers away from our new location. We have set the simulated accident to occur at 5:15 p.m., with a severity level of 1. Also, we specify the accident type to block two right lanes. This Queens location was chosen due to its proximity to Woodhaven Station and the intersection of Queens Boulevard. When conducting this experiment, the system will incorporate random effects from nearby locations with similar characteristics, such as intersections, roundabouts, traffic signals, stations, etc.

5.3.4.1 Experimental (3) Results

Figure 5.9 shows the system's interface of our simulated accident at an unseen location. This location's average traffic speed cannot be determined because this observation does not exist in our dataset. However, the system uses historical data to model spatial points

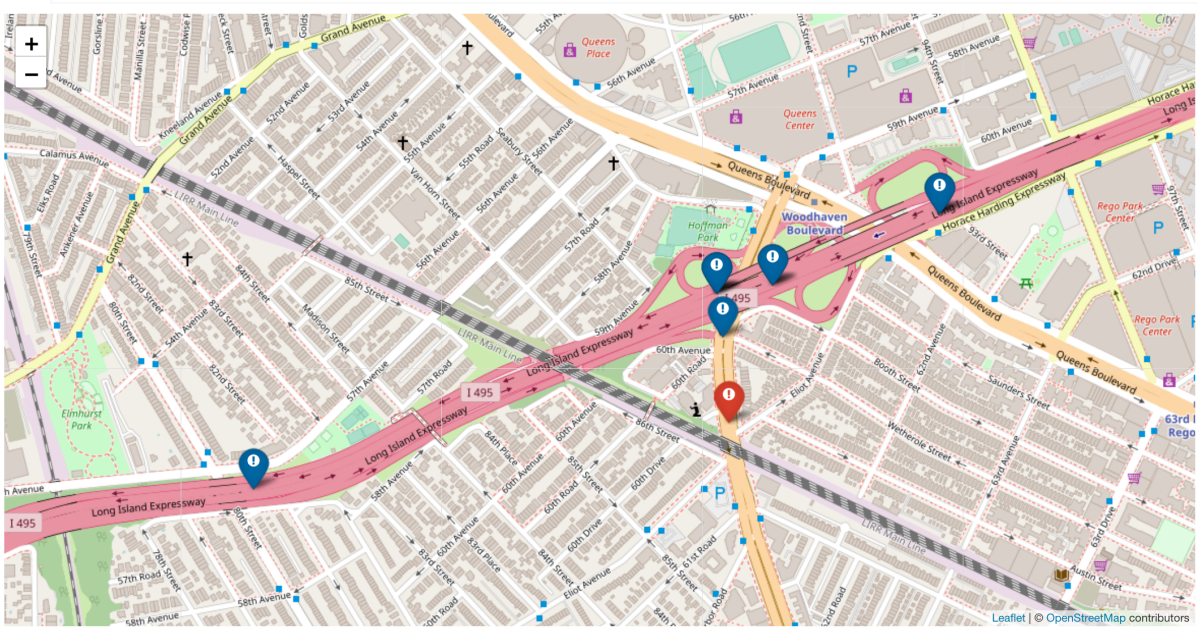

Figure 5.8: Information of the selected location- Experiment 3.

near our new location to estimate the average traffic speed, which is approximately 55 kph. After simulating the accident, the speed rate decreased by 40.61% to around 30 kph. The accident at this location lasted 46 minutes, and the traffic speed increased gradually, starting at 5:45 pm.

The selected location uses our dataset's existing locations' characteristics to model the accident impact and predict the traffic speed. The system models the distance between these locations and all the nearby locations using Euclidean distance. This method assists us in overcoming challenges when simulating accidents in unseen locations where there are no accident records or historical data about the average traffic speed. In addition, the system model can predict the normal traffic speed data at existing locations and the normal average traffic speed at unseen locations.

As can be seen in Figure 5.10, the speed begins to return to normal after 6:15 and is typically between 59 and 67 kph. When creating the accident profile in the interface, the user can investigate further to determine the normal traffic speed without simulating any accident scenarios by selecting the "None" feature. Figure 5.10 illustrates the average

speed of traffic at the same location and time point. It demonstrates that the average speed when there are no accidents is between 59 kph and 62 kph. The user is able to conduct various experiments at various locations and observe how the speed value varies based on the characteristics of the selected location.

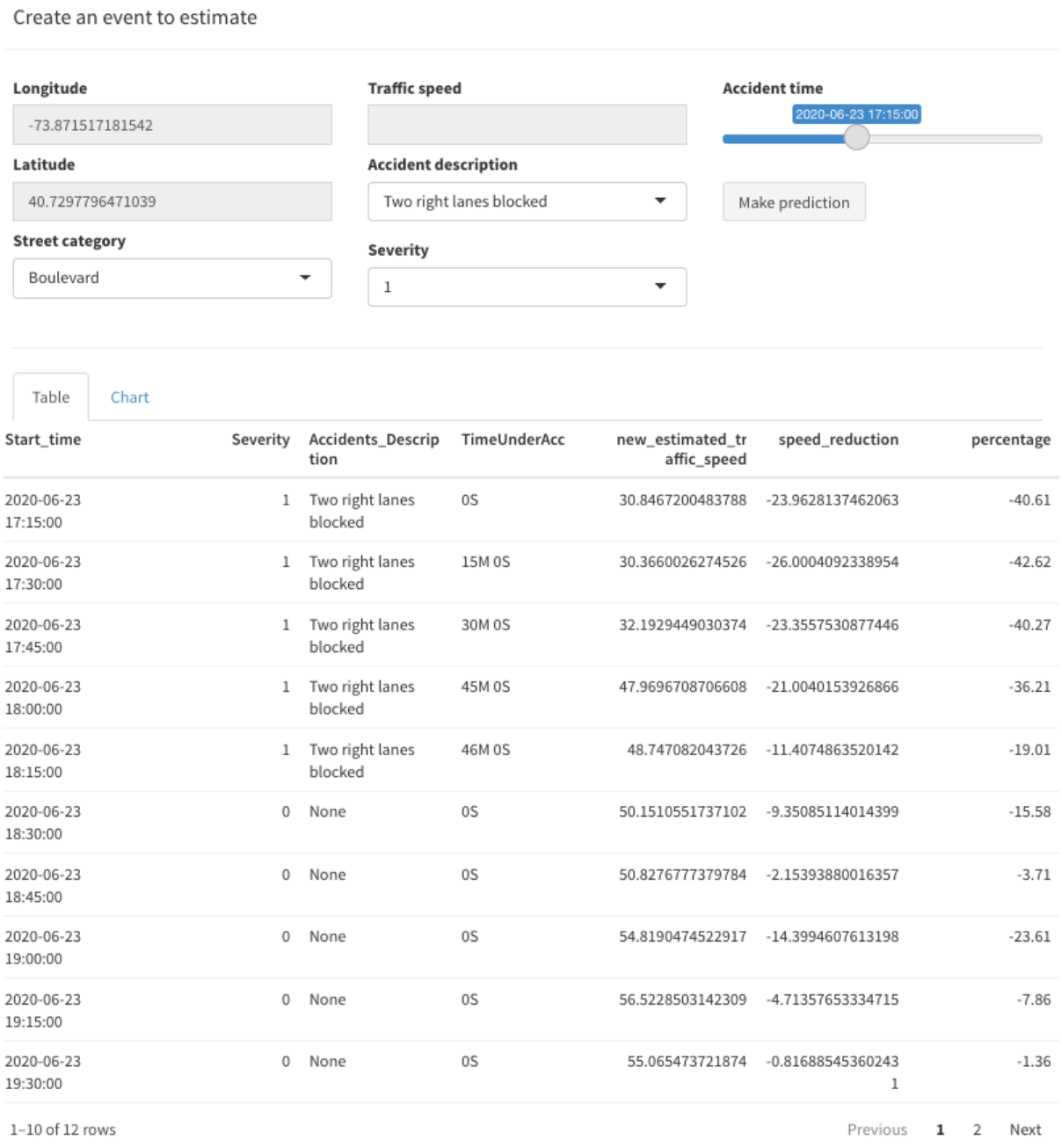

Figure 5.9: Results of Experiment 3.

Start_time	Severity	Accidents_Description	TimeUnderAcc	new_estimated_traffic_speed	speed_reduction	percentage
2020-06-21 16:30:00	1	None	0S	59.3680927700696	-0.658187643100703	-1.13
2020-06-21 16:45:00	1	None	15M 0S	62.0681967863888	-0.982351746967439	-1.61
2020-06-21 17:00:00	1	None	30M 0S	60.6874404562259	-0.849269014116352	-1.42
2020-06-21 17:15:00	1	None	45M 0S	61.2200981492725	0.794294157645105	1.37
2020-06-21 17:30:00	1	None	48M 0S	62.0108391140147	-0.187055823258717	-0.31
2020-06-21 17:45:00	0	None	0S	61.2850417575073	-0.16334025916278	-0.28
2020-06-21 18:00:00	0	None	0S	61.7299650034915	-0.329968155252963	-0.56
2020-06-21 18:15:00	0	None	0S	62.0565179076464	-0.155752105325469	-0.27
2020-06-21 18:30:00	0	None	0S	61.8078614752106	-0.00948986402504204	-0.02
2020-06-21 18:45:00	0	None	0S	61.6490485748469	-0.017332162059374	-0.03

1–10 of 12 rows Previous 1 2 Next

Figure 5.10: System prediction results with no accident.

5.4 Discussion and Results

It is worth mentioning that a comprehensive analysis of both road accident data and normal traffic speed data is significantly important to explain the unexpected results. According to our analysis that is conducted in Chapter 4, the impact of accident severity is the most counterintuitive. More severe accidents had lower associated street blockage times. Our assumption is that more severe accidents lead to faster response times due to a sense of urgency. In addition, an accident that causes full road closures results in a significantly longer accident duration than other types of accidents, with an average duration of over 90 minutes. This is not surprising given the assumption that accidents resulting in full road closures will be larger and take longer to reopen. Furthermore, accidents that only block the left lane also take longer, which may be realistic given that

drivers tend to use the left lane as the fast lane. According to the National Highway Traffic Safety Administration (NHTSA), left-lane accidents result in more severe injuries and deaths due to higher rates of speed [127].

Despite that the normal traffic speed and road accident datasets have a complex structure, our novel system can achieve high accuracy in less computational time. We discuss and evaluate the statistics derived from the test data for the GBDT and the boosted LMMs models. The Boosted LMMs were fitted using the residual maximum likelihood (REML) criterion. The model summary shows that the REML value is 58 indicating a better-fitting model. The intercept of our levels in the boosted LMMs for the county, lane, and spatial points levels are 49.3722, 11.3513, and 94.17339, respectively. This indicates how the random effects are attributed to the nested effect. We can observe that the influence of spatial point random effect alone is significant, with a value of 94%. The boosted LMMs fixed-effects, and random-effects hyperparameters are shown in table 5.1. The standard error of our parameters illustrates how the error rate is negligible in our samples. The residual standard deviation shows how our data is close to the mean with a value of 1.393, indicating how good our model is in fitting our dataset. As we mentioned previously in Chapter 3, the boosted LMMs compute the normal traffic speed for a location in four time periods of 15-minute intervals as follows: NTS_{L1} NTS_{L2} NTS_{L3}, and NTS_{L4}. Also, it models the average of the NTS_L values for all locations on a roadside RS_{L1}, RS_{L2},RS_{L3}, and RS_{L4}. The estimates of the variance of the random-effects parameters indicate how each random-effect parameter can be viewed between the LMMs hierarchy. The estimates of the fixed effects parameters are also shown in table 5.1. The constant (intercept) describes the slopes of each fixed effect of the dependent variables when all the predictors are set to zero.

Figure 5.11 shows the distribution of the residual errors of the random-effects parameters following a normal distribution, satisfying normality assumptions. The boosted LMMs achieved high performance on the test data with an R^2 of 0.9190 and an R^2 of

0.9291 on the full fitted dataset, explaining more than 92.9% of the variation in the data in both cases. The boosting process in the LMMs model estimates the hyperparameters of 1000 trees with an interaction depth of 7 starting from the root until the end of the split nodes among the predictors. Our model performs the same based on our observation with an interaction depth of 7 and no more than 1000 trees.

Table 5.1: The random and fixed effects parameters in the boosted LMMs model

	Parameters	**Variance**	**Standard Error**
Random-effects	Municipality (Intercept)	49.3722	2.22
	NTS_{L1}	6.41	0.080
	NTS_{L2}	5.77	0.042
	NTS_{L3}	7.26	0.026
	NTS_{L4}	5.36	0.023
	Accidents Description	1.59	12.63
	Parameters	**Intercept**	**Standard Error**
Fixed-effects	Severity	4.62	2.10
	TimeUnderAcc	4.49	1.72
	RS_{L1}	3.13	3.33
	RS_{L2}	4.85	2.27
	RS_{L3}	1.09	2.93
	RS_{L4}	2.78	2.88
	NTS_{L1}	2.79	6.78
	NTS_{L2}	8.47	3.94
	NTS_{L3}	5.25	3.04
	NTS_{L4}	14.72	2.80
	$NTS_{L_{day}}$	2.64	1.44

Table 5.2: Accident duration prediction model error measurements

Model	**MAE**	**RMSE**
Accident duration predicting model	0.24	0.53

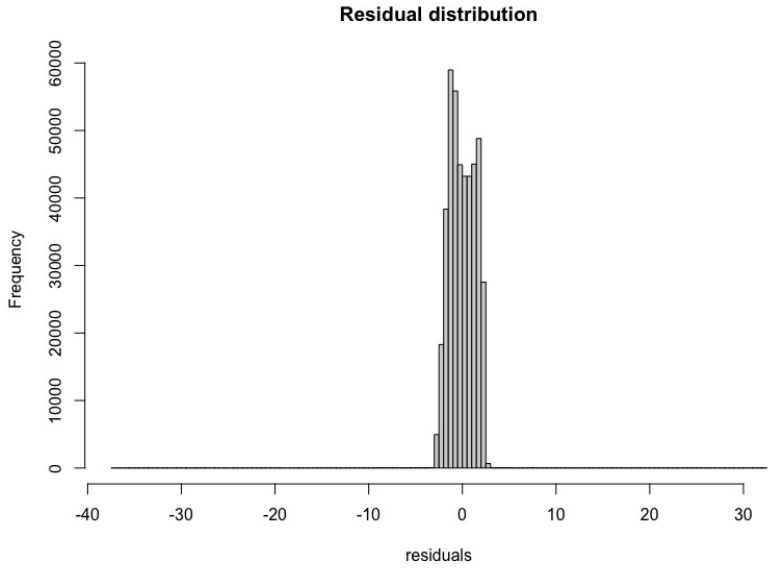

Figure 5.11: The residual distribution of the LMMs model

Table 5.3: Comparison between LMMs and boosted LMMs models

Model/ Data	**MAE**	**RMSE**	R^2
Boosted LMM on test data- Month of May	0.27	0.80	0.9292
Boosted LMM on test data- Month of June	0.24	0.79	0.9291
LMM alone on test data- Month of May	1.31	2.39	0.8931
LMM alone on test data- Month of June	1.13	1.29	0.8999

The accident duration model has an R^2 on the logged test dataset of 0.24, meaning that it explains about a quarter of the variation in accident lengths. The hyperparameters that are used in the accident duration model are 1,355 trees with an interaction depth of 10. The response was modeled using a T distribution with 4 degrees of freedom to account for outliers that were present even after logging accident time. In other words, the GBM accurately predicts the average speed at the first 15 minutes with a less computational time. We evaluate the accident duration prediction model using the MAE and the RMSE as shown in Table 5.2. The error rates in our event profile model are negligibly small, validating its efficiency. Table 5.2 shows the MAE and RMSE for the test data as 0.24 and 0.53, respectively, which shows that the model errors are extremely small. Another

evaluation was performed on test data from a different time period where we conducted a second experiment utilizing a separate dataset for the month of May. This data has 128 locations, and 70% of these locations were also present in the June data. The R^2 on these test data was 0.9292, the MAE was 0.27, and the RMSE was 0.80, indicating that the fit to the May data was excellent and that the model fit to the June data extrapolates well to the May data.

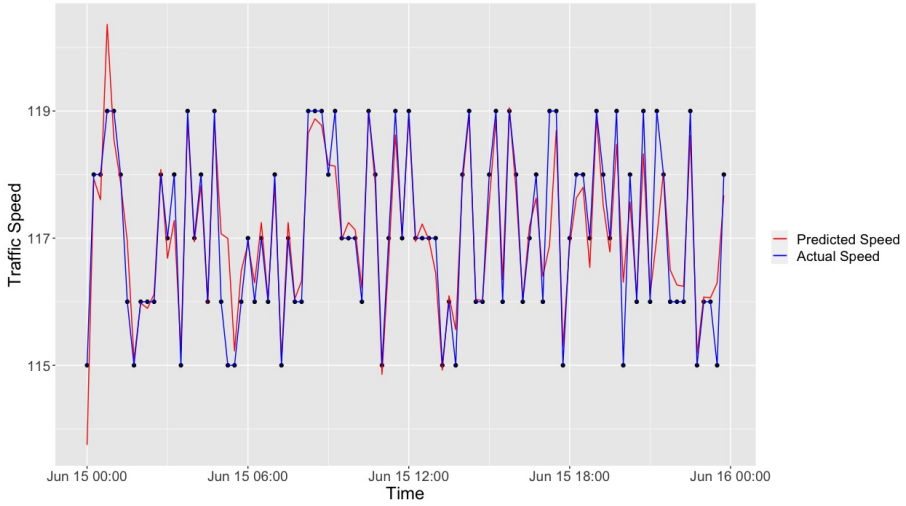

Figure 5.12: Actual traffic speed and the predicted traffic speed.

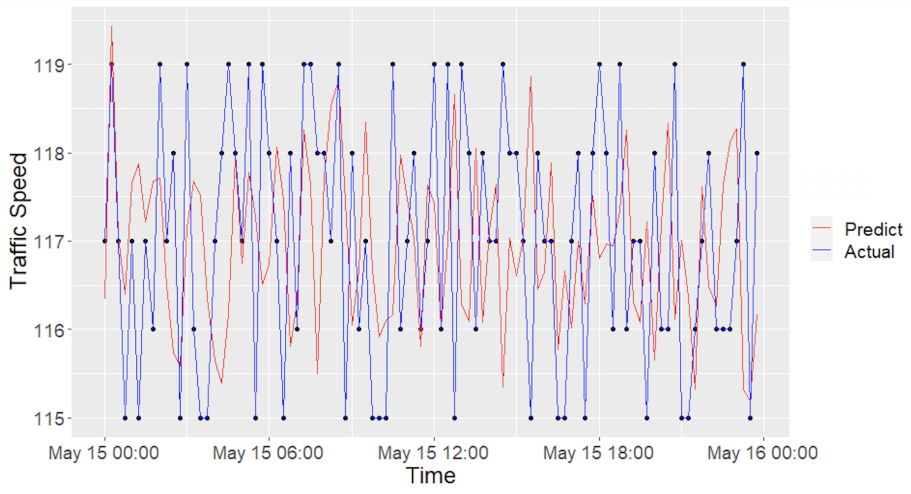

Figure 5.13: Actual traffic speed and the predicted traffic speed - New Location.

Figure 5.12 shows the predicted traffic speed in red over the next 24 hours, and it is roughly equal to the actual traffic speed over the time period. We can see that the

proposed approach successfully predicts the traffic speed when an accident happens and gradually increases to reach the actual average traffic speed. To evaluate our approach in predicting when modeling unseen locations, we removed a location from the dataset and performed the prediction for that exact location. Figure 5.13 shows the predicted traffic speed in red and the actual speed for that location over the next 24 hours. We can say that the developed model successfully predicts the traffic speed for the unseen location.

5.5 Threats to Validity

The main threat to our approach could be that our system predicts the traffic status every 15 minutes. Therefore, short-term changes in traffic status, such as 5 or 10 minutes, would not be detected by the system. In the field of traffic research, the absence of minute-by-minute traffic data presents a significant obstacle. Moreover, if traffic data can be collected every minute, it is recommended to aggregate the data into 5-minute intervals to reduce computational complexity. However, this may pose a negative impact on the validity of predicting short-lived events/fine-grained traffic status changes. Another threat to our approach is that by predicting traffic every 15 minutes, we are unable to capture the smoothness of returning to the actual traffic status following a road event.

5.6 Comparison of the proposed system with similar approaches

This section compares the proposed approach with other approaches described in the literature. We divided this comparison into two sections; a comparison based on the implementation analysis and a comparison based on the functionality, usability, and viability.

5.6.1 Implementation Analysis

We face challenges in comparing our models to state-of-the-art to evaluate the prediction accuracy and the computational time due to the different implementation structures that model spatiotemporal data differently. Qiang et al., [128] presented a hybrid deep learning approach that utilized the bidirectional long short-term memory (Bi-LSTM) and long short-term memory (LSTM). The approach models text data using the natural language description of the accident as input. Their approach lacks spatial and temporal components where it relies on modeling the weather condition, accident description, and severity level. They model the features in three phases and limit the scope of the accident duration to 90 minutes. Although this approach predicts the accident duration, missing the spatiotemporal component requires significant modification in the implementation to fit our data for comparison. On the other hand, Lina et al., [129] proposed multiple XGBoost binary classifiers to predict the accident duration. In their approach, they categorized the duration into multi-binary classification tasks that model the period as the following categories: 10-20 minutes, 20-30 minutes, 30-4- minutes, 40-50 minutes, and finally, more than 60 minutes. The purpose of modeling the periods into multiple categories is to improve their prediction accuracy. The final classifier is selected after integrating all the classifiers using ANN. They define the latitude and longitude to represent one block, and each block is colored based on the accident duration category. An experiment performed applying our dataset to their model to predict the duration is shown in table 5.4. The experiment shows how our proposed model slightly has better accuracy than the XGBoost binary classifiers.

5.6.2 Designing Analysis

The main three components of our system: traffic speed prediction, traffic simulators, and accident simulation and prediction are compared to respective approaches due to the lack

Table 5.4: Comparison between the accident duration event profile model and XGBoost binary classifier

Model	MAE	RMSE
Accident duration event profile model	0.24	0.53
XGBoost binary classifier	2.9	3.82

of existing systems that integrate the three components. We evaluate each component with its respective state-of-the-art in terms of functionality, usability, and viability. For road event simulation, existing traffic simulation tools are used to simulate various traffic conditions, including traffic flow, traffic volume, and vehicle movement patterns such as SUMO. Additionally, other event simulators, specifically accident event simulators, are used to simulate the collisions, and crashes that occur during an automobile accident, such as the LS-DYNA simulator. These simulators do not perform identically to our simulator, but we provide a comparison to highlight their conceptual, functional, and performance differences.

5.6.2.1 Traffic Simulators

SUMO (Simulation of Urban MObility) is one of the common traffic simulators that uses time as the fundamental independent variable for drawing conclusions about traffic conditions. We conducted a comparison between our traffic system simulator and the SUMO simulator, despite the fact that both simulators do not serve the exact objective. In our proposed system, the objective is to simulate an accident, whereas the objective of the SUMO simulator is to simulate traffic flow. In their simulation, they both utilize Open Street Map (OSM) and use statistical distribution for the input data. In our proposed system, input data consists of a list of geometry variables stored in both shapefile and CSV files. The shapefile stores the geometric location and attribute information of geographic features, including the traffic speed values at various time points. We use these values to define the study area for our accident simulation. The network geometry

in SUMO is generated in XML format using either netgenerate or OSM Web Wizard, which are network generators for the microscopic level. In our system, the user is not required to do anything in order to visualize locations on the map, as they are imported directly from the shapefile and the CSV file. However, SUMO must be modified in order to specify the network locations where the event simulation is to occur.

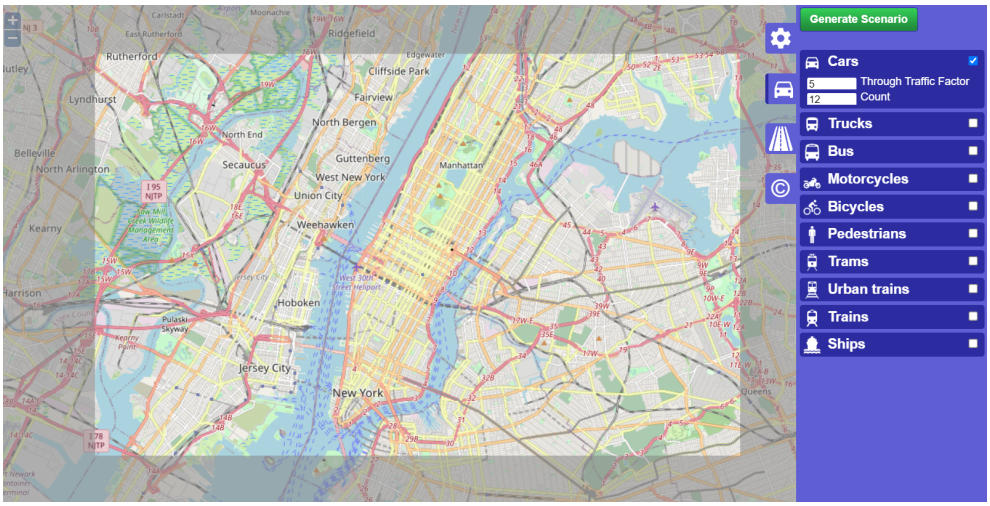

Figure 5.14: OSM Web Wizar interface.

Both simulators utilize the time component as one of the independent variables; however, SUMO focuses primarily on simulating the vehicle count to predict the traffic flow, whereas our simulator models the accident effect to predict the traffic speed. In contrast to our system, which uses spatial points as nodes on the map, the data in the SUMO simulator is defined in Origin-Destination traffic data format to visualize the flow's movement. The available scenarios in SUMO are limited to the vehicle types illustrated in Figure 5.14. The user will specify the number of vehicles to be simulated and visualize the traffic flow based on the number of lanes. In our system, we simulate the impact of an accident based on the number of lanes and the position of the lanes, whether they are on the right, the left, the center, or the shoulder. In Figure 5.14, the user specifies the study area on the map and assigns vehicle counts using the OSM Web Wizard. Then, we set the time frame for this scenario and click the Generate Scenario button. For the

study area in our Figure, the OSM Web Wizard consumes around 12 minutes to extract the generated network for about 12 vehicles. Then, an XML file will be generated to import all routes and the OD-matrix, which describes the movement of the 12 vehicles from one district to another within a given time frame.

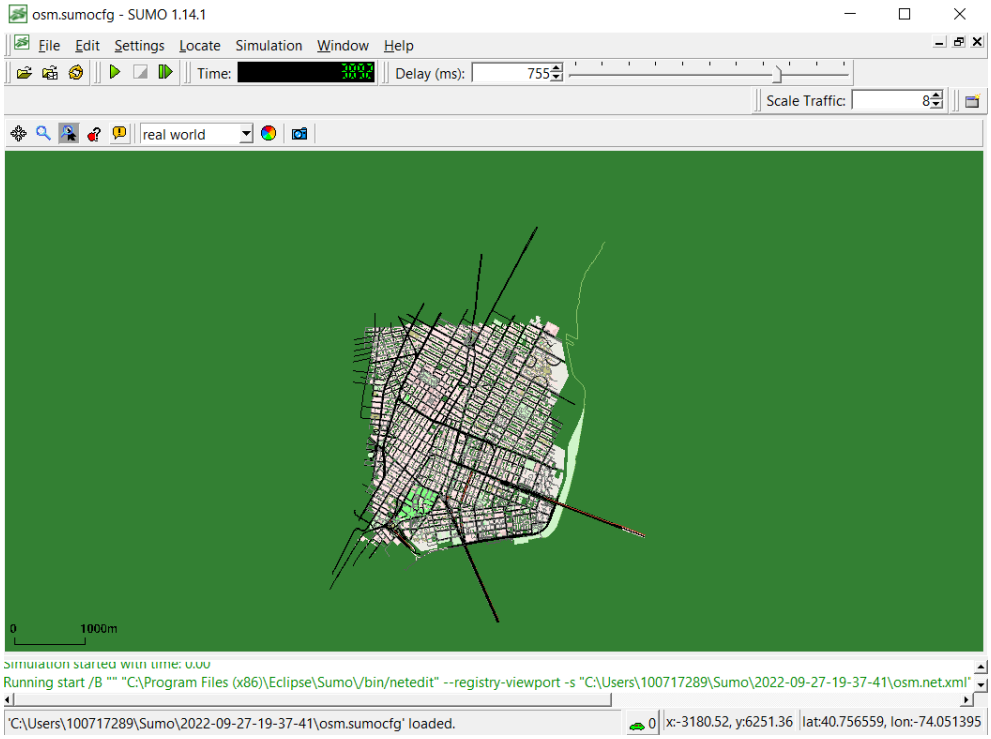

Figure 5.15: SUMO simulator networks.

The computation process in the SUMO simulator can be summarized as follow:

- Define the explicit routes in the Network
- Import the OD-matrix to fill the traffic flow in the network
- Compute the trip within a given time-frame
- Generate trip table/trip list for the simulation

After computing the trip table, a NETCONVERT runs in the command prompt to convert the output of OSM into an XML file and visualize the XML contents SUMO simulation. Figure 5.15 shows the network of the selected area in Manhattan in New York City with the defined routes ready to run the simulator. In SUMO, lane-changing and car-following algorithms are used to perform the experiment. In Table 5.5, we see

the different features between SUM and our proposed system in terms of models, input, output, and rum time.

Table 5.5: SUMO simulator features and capabilities.

Features/Simulator	**Proposed System**	**SUMO**
Run Time	Short	Long
User- friendly	Yes	No
Maps	Open Street Map , Google Map	Open Street Map
Functionality	Analysis, Simulation, Prediction	Analysis, Simulation
Scenarios	Accidents	vehicle count
Models	LMMs , GBDT models	Lane-changing algorithm, Car-following algorithm
Data Format	CSV. Shapefile	XML
output	Text, Graph	Graphs
output variables	Traffic Speed, Accident duration	Traffic Speed
Import Map	Yes	Yes
Programming Language	R	C++, VB, Matlab, Python
Flexibility in infrastructure Development	Flexible	Limited
Coding	Easy	Difficult

5.6.2.2 Accident Simulation and Prediction

As we previously mentioned in Chapter 2, most of the proposed approaches to accident simulation and prediction are mainly implemented to detect the accident before it happens. Therefore, a fair comparison here is challenging to conduct. Still, few proposed systems are used to simulate an accident using the Finite Element Method (FEM), which is a numerical calculation method used for nonlinear analysis and problems. The characteristic feature of the FEM method makes it suitable for applications that involve crash analysis scenarios. LS-DYNA is one of the accident simulator software that uses the FEM method along with the Equivalent Static Loads (ESL) method to analyze and com-

pute the accident impact on vehicles. The simulation involves high-speed dynamics of a number of moving objects to observe the behavior when a crash occurs. In other words, LS-DYNA can simulate the response to smashes and crash situations using three main algorithms: the Standard Penalty Formulation algorithm, the Soft Constraint Penalty Formulation algorithm, and the Segment-based Penalty Formulation. These algorithms are used to formulate the impact of a crash on different surfaces. Figure5.16 shows how the LS-DYNA simulator defines the vehicle's material to simulate an accident where a vehicle crashes into a wall.

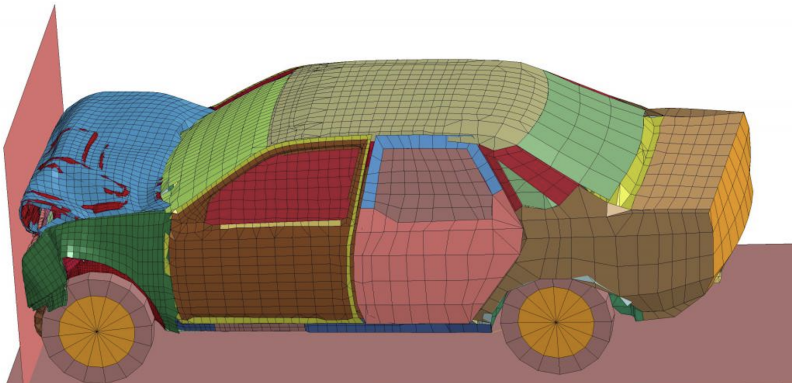

Figure 5.16: LS-DYNA simulator of accident impact.

The LS-DYNA simulator defines different materials within the material models that play the main component of this simulator in order to evaluate the accident impact. However, our proposed system evaluates the accident impact through the predicted speed and the predicted duration. The LS-DYNA does not predict the duration of the accident scene or the speed values during the accident duration. Furthermore, the LS-DYNA takes into consideration the pressure on the interface to evaluate the damage of the accident when an accident occurs. In contrast, our proposed system looks into the accident records and the nearby location to assess the accident damage in terms of speed and delay. Table 5.6 shows a summary of a number of features between our proposed system and the LS-DYNA accident simulator.

Table 5.6: LS-DYNA simulator features and capabilities.

Features Simulator	**Proposed System**	**LS-DYNA simulator**
User- friendly	Yes	Yes
Functionality	Analysis, Simulation, Prediction	Analysis, Simulation
Scenarios	Accidents	objects/vehicle damages
Models	LMMs , GBDT models	FEM, ESL methods
output	Text, Graph	Graphs
output variables	Traffic Speed, Accident duration	prototype response
Import Map	Yes	No
Programming Language	R	C
Flexibility in infrastructure Development	Flexible	Limited
Coding	Easy	Difficult

5.6.2.3 Traffic Speed Prediction

For our traffic speed prediction component, it is challenging to find an approach that predicts traffic speed based on both historical data and user input. Therefore, we chose to compare the traffic speed prediction approach with one of the state-of-the-art approaches that predict traffic speed based on the road environment. The candidate approach uses long-short-term memory (LSTM) to predict the traffic speed at a given spatial line every 30 minutes using speed data from the nearby area and weather data in particular rainfall [130]. We treat the weather data similarly to our accident data in our approach and build our comparison based on the type of inputs, outputs, functionality, and accuracy.

Their proposed method requires modeling a sequential pattern of the road event variable in order to predict traffic speed. In contrast, our method excels at predicting unanticipated events, which in our case are road accidents. In the LSTM method, the effects of unexpected events may reduce the accuracy of predictions. One of the main dif-

ferences between both approaches is that they predict the speed of the road accident but not specific spatial points, as we propose in our approach. Both approaches incorporate the characteristics of adjacent roads and predict the road environment before estimating the traffic speed. Furthermore, we both use the loss function to measure the error between predicted and actual speeds. We did not compare the two approaches using statistical measures due to the fact that both the approach and the modeled datasets are implemented differently. Nonetheless, we have summarized the differences and similarities between these two approaches in Table 5.7 for a comprehensive overview.

Table 5.7: Comprehensive comparison between the proposed system and similar approaches

Features/ Simulator	**Proposed System using LMMs and GBDT**	**Road Speed Prediction approach using LSTM**
Data source	Historical data and user input	Historical data
Functionality	Analysis, Prediction	Analysis, Predict
Spatial component	Spatial point	Spatial line
Temporal component	Timestamp	Timestamp
Models	LMMs , GBDT models	LSTM , ANN
Input	Traffic speed data and accident data	Traffic speed data and weather data
output variables	Traffic Speed, Accident effect	Traffic Speed, Rainfall effect
Predicted Time-window	The 15 minutes for the next 24 hours	The next 30 minutes
Programming Language	R	Python
Flexibility in implementation	Flexible	Flexible
Event	Accident	Rainfall
Event patterns	Non sequential pattern	Sequential pattern

5.7 Summary

This chapter discusses the findings of our experiment and evaluates the performance of the proposed system. We discussed the experimental setup and the performance metrics used to measure performance. In order to evaluate the performance of the system, we conduct the experiment at both actual and nonexistent spatial points. We eliminated the spatial point from the dataset and performed the prediction for the given spatial point in order to validate the prediction results. Various evaluation criteria, such as distance and error metrics, have been adopted to assess the performance of the system. The findings demonstrated that our innovative system was capable of achieving high accuracy in less computational time. The boosted LMMs achieved high performance on the test data with an R^2 value of 0.9191 and on the full-fitted dataset with an R^2 value of 0.9291. Both the MAE and the RMSE values indicate that our proposed system successfully predicts the traffic speed when an accident happens with values of 0.27, and 0.80, respectively. We also compared different components of the proposed system, such as the traffic prediction model and the simulator, to respective state-of-the-art based on implementation, functionality, usability, and viability.

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis, highlights how we fulfilled a number of the research objectives and answered research questions. A future research plan on a number of system components will be discussed to achieve the rest of the thesis research objectives.

6.1 Conclusion

In chapter 1, we introduced the spatiotemporal traffic modeling concept and how it is used to analyze and predict traffic behavior, which plays a big role in traffic engineering and assessing road traffic facilities' performance. Additionally, we described the traffic simulation tool and how it differs from the traffic prediction method. We further discussed different traffic problems, such as the risk of road accidents and how these road accidents heavily impact traffic status. After acknowledging the effects of road accidents on traffic status, we proposed developing a hybrid spatiotemporal traffic speed prediction system capable of simulating a fabricated accident and predicting the time elapsed from the occurrence of the accident to the accident clearance time. Recognizing these challenges to developing such a system, we formulated the thesis research question and how this thesis contributed to the ongoing efforts to guarantee better decision-making and road management under unexpected road circumstances.

In Chapter 2, we presented a comprehensive literature review of the research related to traffic modeling studies and traffic modeling studies that incorporate event modeling. We reviewed some of the most frequently used spatiotemporal prediction models that have been proposed to analyze and predict traffic status. Having discussed the state-of-the-art of currently applied spatiotemporal prediction models, we then focused on the literature that proposes either Bayesian-based, ANN-based, or ST-Kriging-based approaches. We constructed a comparison to evaluate different aspects of these approaches to help decide which is more suitable for our research study. Additionally, we briefly compared the time intervals used in spatiotemporal traffic prediction to demonstrate the challenges in long-term and short-term traffic modeling prediction. A comprehensive comparison of the available traffic simulation tools and their limitations was conducted. The provided comparison was conducted based on seven features and their strengths and weaknesses. Also, we listed several critical challenges in these traffic simulators that conflicted with our research goal. We have also illustrated the structure of the literature review section for better understanding. To conclude Chapter 2, we reviewed the research conducted on predicting traffic status incorporating future events such as weather conditions and traffic accidents. Most recent research has concentrated on developing methodological methods to analyze the association between these events and traffic congestion status. After reviewing the state of the art, we described the structure of the proposed system in Chapter 3.

In Chapter 3, we started by describing the structure of the proposed system and highlighting the characteristics of each of the three components in the proposed system. Our proposed system leverages three components: the boosted Linear Mixed-Effects Models(LMMs), the Gradient Boosting Decision Tree (GBDT model), and the Shiny with R interface. In this chapter, we walk through the prediction steps in order to model the impact of the time of the day on accident duration. Furthermore, we described the boosting process using the LMMs with GBM to optimize the prediction accuracy and reduce

the prediction errors for the LMMs' output. This process generates a sequence of intercepts and coefficient values and identifies the optimal value that maximizes prediction accuracy.

Chapter 4 presents a preliminary exploratory analysis of traffic accidents and speed data. We began with a brief description of the two datasets and user inputs, followed by an exploratory analysis that sheds light on the significant variables and the steps needed to prepare the data for the models. This chapter's primary objective is to validate that the selected datasets are suitable for the proposed methodology. We provided a brief introduction of the three datasets we use in our system, as well as how we break each dataset into a number of components for modeling purposes and statistical analysis.

In Chapter 5, we presented our experiment's results and evaluated the proposed system's performance. In this chapter, we first described the experimental setup and the performance indicator for evaluating the performance. We described the study area and the results of our case studies. Furthermore, we perform the experiment on existent and nonexistent locations to evaluate the system's performance. We removed the location from the dataset and performed the prediction for model performance validation for the nonexistent location. We measured the model's efficiency using quantitative methods to demonstrate how well predictions agree with observations when comparing a long series of predicted data. Different evaluation criteria have been adopted for evaluating the system's performance, such as distance and error metrics. The results in Chapter 5 proved that our novel system could achieve high accuracy in less computational time, where the boosted LMMs achieved high performance on the test data with an R^2 of 0.9190 and an R^2 of 0.9291 on the full fitted dataset. The MAE was 0.27, and the RMSE was 0.80, indicating that the fit of our data was excellent. The final GBDT model had hyperparameters of 1000 trees with an interaction depth of 7, a shrinkage parameter of 0.4, and a bag fraction of 0.5. The accident duration model had an R^2 on the logged test dataset of 0.24, meaning that it explains about a quarter of the variation in accident

lengths. The evaluation results conclude that our proposed system successfully predicts the traffic speed when an accident happens and gradually increases the speed to reach the actual average traffic speed.

We concluded the thesis proposal in Chapter 6 and described how we fulfilled our research objectives that were mentioned in Chapter 1. Later, we listed a number of limitations that are beyond the scope of our system and the challenges we face so that we couldn't overcome these challenges. Lastly, we provided a future research plan on a number of system components and discussed how to achieve them through some recommendations.

6.2 Limitations

Although we were able to answer the research questions posed in this thesis and the proposed system was able to achieve the stated objectives in Chapter 1, it's worth mentioning that there are some limitations mainly related to the initial scope of the thesis. First, the scope of the work was restricted to unplanned occurrences, specifically accident occurrences on a particular type of street. Events such as weather conditions, sporting events, and holiday events are not modeled in our system due to the lack of availability of such data in the spatiotemporal structure. To incorporate these events, it is required to model them on multiple levels, taking into account the accuracy of the predictions and the complexity of the model.

Furthermore, our system primarily predicts spatial points; however, it models the available observations on an entire lane and a whole county. Predicting traffic speed on a single route in order to observe traffic behavior on this route is beyond our scope. Also, our system does not support predicting the traffic status at nearby locations or updating the traffic speed for a specific range within the selected spatial point. Our system is limited to predicting the traffic speed at 15-minute intervals for the time series

prediction. Even though the prediction results indicate a significant increase in traffic speed after the accident has been cleared, we believe that a 15-minute interval is sufficient for such an increase. To investigate the increase in traffic speed, the time intervals can be reduced to 5-minute intervals; however, the obtained traffic speed data limited our system to perform the prediction at 15-minute intervals. Although there are several open-source traffic speed datasets, the obtained dataset has significantly supported our thesis' original objective. This dataset was utilized in our method due to its flexibility in selecting desired locations at different time points and the complex structure of its spatiotemporal component. Moreover, in its current implementation, the system only models historical traffic speed data. This can be modified to model a stream of real-time traffic speed data. Due to the unavailability of such real-time data streams, we passed the traffic speed data through our backend system. From the implementation perspective, assumptions are made, such as excluding accidents that entirely close roads since their predicted traffic speed will be zero. The accident duration for such an accident is not predicted; however, this can be incorporated into future work enhancements.

Another limitation of our system implementation is modeling the distance between a new location and an existing one. Exploring this area and implementing new methods to compute the characteristics of the new spatial point in our system without compromising system performance is outside the scope of this thesis. Although our hybrid system yielded satisfactory results, we considered incorporating an additional predictive sequential statistical method, such as a state space model. However, due to time constraints, we limited our implementation to the current models, and much more research can be carried out to increase the system's ability to incorporate additional models. These limitations can be considered to expand the system's capabilities and take advantage of the most advanced spatiotemporal models available. Therefore, future work and recommendations are discussed in the following section

6.3 Future Work

The proposed system could be extended in the future to include a number of improvements to accident simulation strategy, accident impact prediction, data integration, and sequence traffic speed prediction. We summarize the potential future work and the recommendations as follows:

- Defining the Origin-Destination (OD) parameter: Through the system interface, the user can specify the starting spatial point and the ending spatial point for the desired trip. After defining the route trip, simulating an accident could be applied to any spatial point on the formulated route.
- Predicting the impact region: The accident impact prediction can be extended to predict the affected region. This will give the system's user an overview of the area likely to be affected by the accident.
- Predicting the accident impact on neighboring regions: The accident impact prediction on neighboring areas that are not on the trip route is recommended, and further investigation is required.
- Real-time traffic speed data modeling: We can shift the proposed system to be a real-time system and model traffic speed data in real-time. Expanding the research on ensuring connectivity between the real-time data stream component and the system backend component is recommended.
- Traffic speed Modeling: The traffic speed is autoregressive, of a sufficiently high order to account for daily, weekly, and seasonal periodicities. A suggested method to optimize data modeling is to include holiday effects. This could be somewhat idiosyncratic to the location, depending on data size. An alternative way is to aggregate across regions or nations.
- LMMs parameters estimation: To reliably estimate parameters using the LMMs, we propose including covariates to overcome the issue when the random effects are

latent variables and not observed.

- Modeling Irregular time series: The non-linearity in the road accident makes it challenging to model the time series accurately. However, a log-link function can be used. Although this might introduce both numerical and modeling interpretation issues, it can be modeled as a stochastic point process to capture random effects.

Bibliography

[1] Chao Wang. *The relationship between traffic congestion and road accidents: an econometric approach using GIS*. PhD thesis, © Chao Wang, 2010.

[2] Chao Wang, Mohammed A Quddus, and Stephen G Ison. Impact of traffic congestion on road accidents: A spatial analysis of the m25 motorway in england. *Accident Analysis & Prevention*, 41(4):798–808, 2009.

[3] Gang-Len Chang and Hua Xiang. The relationship between congestion levels and accidents. Technical report, 2003.

[4] INRIX Roadway Analytics. Analyzing the most dangerous roads in the u.s., 2021. URL `https://inrix.com/learn/inrix-roadway-analytics/`.

[5] Goverment of Canada. Canadian motor vehicle traffic collision statistics, 2018. URL `https://tc.canada.ca/en/road-transportation/motor-vehicle-safety/canadian-motor-vehicle-traffic-collision-statistics-2018`.

[6] The World Health Organization (WHO). Road traffic injuries, 2021. URL `https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries/`.

[7] Thangamani Bhavan. The economic impact of road accidents: the case of sri lanka. *South Asia Economic Journal*, 20(1):124–137, 2019.

[8] Jiuh-Biing Sheu, Yi-Hwa Chou, and Liang-Jen Shen. A stochastic estimation

approach to real-time prediction of incident effects on freeway traffic congestion. *Transportation Research Part B: Methodological*, 35(6):575–592, 2001.

[9] Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, and Mihaela Van der Schaar. Mining the situation: Spatiotemporal traffic prediction with big data. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):702–715, 2015.

[10] Md Ebrahim Shaik, Md Milon Islam, and Quazi Sazzad Hossain. A review on neural network techniques for the prediction of road traffic accident severity. *Asian Transport Studies*, 7:100040, 2021.

[11] Chao Wang, Mohammed Quddus, and Stephen Ison. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the uk. *Transportmetrica A: Transport Science*, 9(2):124–148, 2013.

[12] Chiara Bachechi and Laura Po. Traffic analysis in a smart city. In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, pages 275–282, 2019.

[13] Joaquim Barros, Miguel Araujo, and Rosaldo JF Rossetti. Short-term real-time traffic prediction methods: A survey. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 132–139. IEEE, 2015.

[14] Antonella Ferrara, Simona Sacone, and Silvia Siri. *Freeway traffic modelling and control*. Springer, 2018.

[15] S Algers, E Bernauer, M Boero, L Breheret, C Taranto, K Fox, et al. A review of micro-simulation models, smartest: Simulation modeling applied to transport european scheme tests. *Institute for Transport Studies, University of Leeds*, 1996.

[16] Paolo M Ejercito, Kristine Gayle E Nebrija, Rommel P Feria, and Ligaya Leah Lara-Figueroa. Traffic simulation software review. In *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–4. IEEE, 2017.

[17] Nesrine Ghariani, Sabeur Elkosantini, Saber Darmoul, and Lamjed Ben Said. A survey of simulation platforms for the assessment of public transport control systems. In *2014 International Conference on Advanced Logistics and Transport (ICALT)*, pages 85–90. IEEE, 2014.

[18] Mustapha Saidallah, Abdeslam El Fergougui, and Abdelbaki Elbelrhiti Elalaoui. A comparative study of urban road traffic simulators. In *MATEC Web of Conferences*, volume 81, page 05002. EDP Sciences, 2016.

[19] John A Sokolowski and Catherine M Banks. *Principles of modeling and simulation: a multidisciplinary approach*. John Wiley & Sons, 2011.

[20] Simon Kwoczek, Sergio Di Martino, and Wolfgang Nejdl. Predicting and visualizing traffic congestion in the presence of planned special events. *Journal of Visual Languages & Computing*, 25(6):973–980, 2014.

[21] Nicolas Tempelmeier, Stefan Dietze, and Elena Demidova. Crosstown traffic-supervised prediction of impact of planned special events on urban traffic. *GeoInformatica*, 24(2):339–370, 2020.

[22] Aniekan Essien, Ilias Petrounias, Pedro Sampaio, and Sandra Sampaio. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*, 24(4):1345–1368, 2021.

[23] B Gültekin Çetiner, Murat Sari, and Oğuz Borat. A neural network based traffic-flow prediction model. *Mathematical and Computational Applications*, 15(2):269–278, 2010.

[24] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[25] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020.

[26] Arief Koesdwiady, Ridha Soua, and Fakhreddine Karray. Improving traffic flow prediction with weather information in connected cars: A deep learning approach. *IEEE Transactions on Vehicular Technology*, 65(12):9508–9517, 2016.

[27] Manish Agarwal, Thomas H Maze, and Reginald Souleyrette. Impacts of weather on urban freeway traffic flow characteristics and facility capacity. In *Proceedings of the 2005 mid-continent transportation research symposium*, pages 18–19, 2005.

[28] Yunlong Zhang, Larry E Owen, and James E Clark. Multiregime approach for microscopic traffic simulation. *Transportation Research Record*, 1644(1):103–114, 1998.

[29] Arief Koesdwiady. Large-scale traffic flow prediction using deep learning in the context of smart mobility. 2018.

[30] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43:3–19, 2014.

[31] Xiaolei Ma, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17(4):818, 2017.

[32] Afshin Abadi, Tooraj Rajabioun, and Petros A Ioannou. Traffic flow prediction

for road transportation networks with limited traffic data. *IEEE transactions on intelligent transportation systems*, 16(2):653–662, 2014.

[33] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

[34] Zhanguo Song, Yanyong Guo, Yao Wu, and Jing Ma. Short-term traffic speed prediction under different data collection time intervals using a sarima-sdgm hybrid prediction model. *PloS one*, 14(6):e0218626, 2019.

[35] Ibai Lana, Javier Del Ser, Manuel Velez, and Eleni I Vlahogianni. Road traffic forecasting: Recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine*, 10(2):93–109, 2018.

[36] Valdério A Reisen, Bartolomeu Zamprogno, Wilfredo Palma, and Josu Arteche. A semiparametric approach to estimate two seasonal fractional parameters in the sarfima model. *Mathematics and Computers in Simulation*, 98:1–17, 2014.

[37] Aditya R Raikwar, Rahul R Sadawarte, Rishikesh G More, Rutuja S Gunjal, Parikshit N Mahalle, and Poonam N Railkar. Long-term and short-term traffic forecasting using holt-winters method: A comparability approach with comparable data in multiple seasons. *International Journal of Synthetic Emotions (IJSE)*, 8(2):38–50, 2017.

[38] Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, and Amaury Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18): 2861–2869, 2007.

[39] Gregorio Gecchele, Riccardo Rossi, Massimiliano Gastaldi, and Armando Caprini. Data mining methods for traffic monitoring data analysis: A case study. *Procedia-Social and Behavioral Sciences*, 20:455–464, 2011.

[40] J Kihoro, RO Otieno, and C Wafula. Seasonal time series forecasting: A comparative study of arima and ann models. 2004.

[41] Xiaokun Wang and Kara M Kockelman. Forecasting network data: Spatial interpolation of traffic counts from texas data. *Transportation Research Record*, 2105 (1):100–108, 2009.

[42] Zhe Jiang. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1645–1664, 2018.

[43] Nectaria Tryfona and Christian S Jensen. Conceptual data modeling for spatiotemporal applications. *GeoInformatica*, 3(3):245–268, 1999.

[44] Yi Yin and Pengjian Shang. Forecasting traffic time series with multivariate predicting method. *Applied Mathematics and Computation*, 291:266–278, 2016.

[45] Qianlong Wang, Yifan Guo, Lixing Yu, and Pan Li. Earthquake prediction based on spatio-temporal data mining: an lstm network approach. *IEEE Transactions on Emerging Topics in Computing*, 8(1):148–158, 2017.

[46] Brent Selby and Kara M Kockelman. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29:24–32, 2013.

[47] Kennedy John Offor, Lubos Vaci, and Lyudmila S Mihaylova. Traffic estimation for large urban road network with high missing data ratio. *Sensors*, 19(12):2813, 2019.

[48] Matthew Hawes, Hayder M Amer, and Lyudmila Mihaylova. Traffic state estimation via a particle filter over a reduced measurement space. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8. IEEE, 2017.

[49] Yongze Song, Xiangyu Wang, Graeme Wright, Dominique Thatcher, Peng Wu, and Pascal Felix. Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles. *Ieee transactions on intelligent transportation systems*, 20(1):232–243, 2018.

[50] Haixiang Zou, Yang Yue, Qingquan Li, and Anthony GO Yeh. An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, 26(4):667–689, 2012.

[51] Jeremy K Hackney, Michael Bernard, Sumit Bindra, and Kay W Axhausen. Predicting road system speeds using spatial structure variables and network characteristics. *Journal of Geographical Systems*, 9(4):397–417, 2007.

[52] Hidetoshi Miura. A study of travel time prediction using universal kriging. *Top*, 18(1):257–270, 2010.

[53] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

[54] John Skilling et al. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

[55] Peter M Lee. *Bayesian statistics*. Oxford University Press London:, 1989.

[56] Ricardo A Daziano, Luis Miranda-Moreno, and Shahram Heydari. Computational bayesian statistics in transportation modeling: from road safety analysis to discrete choice. *Transport reviews*, 33(5):570–592, 2013.

[57] Parimal Mukhopadhyay and Parimal Mukhopadhyay. Bayes and empirical bayes prediction of a finite population total. *Topics in Survey Sampling*, pages 43–92, 2001.

[58] Lai Zheng and Tarek Sayed. Bayesian hierarchical modeling of traffic conflict extremes for crash estimation: a non-stationary peak over threshold approach. *Analytic methods in accident research*, 24:100106, 2019.

[59] Faustino Prieto, Emilio Gómez-Déniz, and José María Sarabia. Modelling road accident blackspots data with the discrete generalized pareto distribution. *Accident Analysis & Prevention*, 71:38–49, 2014.

[60] M Kumar, NS Raghuwanshi, and R Singh. Artificial neural networks approach in evapotranspiration modeling: a review. *Irrigation science*, 29(1):11–25, 2011.

[61] Zahraa E Mohamed. Using the artificial neural networks for prediction and validating solar radiation. *Journal of the Egyptian Mathematical Society*, 27(1):1–13, 2019.

[62] Wei Chen, Fangzhou Guo, and Fei-Yue Wang. A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):2970–2984, 2015.

[63] Romi Satria, Jonathan Aguero-Valverde, and Maria Castro. Spatial analysis of road crash frequency using bayesian models with integrated nested laplace approximation (inla). *Journal of Transportation Safety & Security*, pages 1–23, 2020.

[64] Cynthia Taylor and Deirdre Meldrum. Freeway traffic data prediction using neural networks. In *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*, pages 225–230. IEEE, 1995.

[65] Somnath Chaudhuri. *Spatio-temporal modeling of traffic risk mapping on urban road networks*. PhD thesis, 2020.

[66] Laura C Dawkins, Daniel B Williamson, Kerrie L Mengersen, Lidia Morawska, Rohan Jayaratne, and Gavin Shaddick. Where is the clean air? a bayesian decision

framework for personalised cyclist route selection using r-inla. *Bayesian Analysis*, 16(1):61–91, 2021.

[67] Lyle D Broemeling. Bayesian methods for medical test accuracy. *Diagnostics*, 1 (1):1, 2011.

[68] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[69] Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC press, 2008.

[70] Rens Van De Schoot, Joris J Broere, Koen H Perryck, Mariëlle Zondervan-Zwijnenburg, and Nancy E Van Loey. Analyzing small data sets using bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, 6(1):25216, 2015.

[71] Gary Charness and Dan Levin. When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *American Economic Review*, 95(4):1300–1309, 2005.

[72] AR Khaz'ali, A Emamjomeh, and M Andayesh. An accuracy comparison between artificial neural network and some conventional empirical relationships in estimation of relative permeability. *Petroleum science and technology*, 29(15):1603–1614, 2011.

[73] Alireza Khotanzad and Nayyara Sadek. Multi-scale high-speed network traffic prediction using combination of neural networks. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 2, pages 1071–1075. IEEE, 2003.

[74] Alfréd Csikós, Zsolt János Viharos, Krisztián Balázs Kis, Tamás Tettamanti, and István Varga. Traffic speed prediction method for urban networks—an ann approach. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 102–108. IEEE, 2015.

[75] Djukic Tamara, van Lint Hans, and Casas Jordi. Seta: An open, sustainable, ubiquitous data and service ecosystem for efficient, effective, safe, resilient mobility in metropolitan areas. In *H2020-ICT*, pages 85–90. CORDIS EU Research Results, 2015.

[76] US Department of Transportation. "types of traffic analysis tools", 2020. URL `https://ops.fhwa.dot.gov/trafficanalysistools/type_tools.htm`.

[77] David Wilkie, Jason Sewall, Weizi Li, and Ming C Lin. Virtualized traffic at metropolitan scales. *Frontiers in Robotics and AI*, 2:11, 2015.

[78] SM Sohel Mahmud, Luis Ferreira, Md Shamsul Hoque, and Ahmad Tavassoli. Micro-simulation modelling for traffic safety: A review and potential application to heterogeneous traffic environment. *IATSS research*, 43(1):27–36, 2019.

[79] Jeffrey Miller and Ellis Horowitz. Freesim-a free real-time freeway traffic simulator. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 18–23. IEEE, 2007.

[80] Manuel Lindorfer, Christian Backfrieder, Christoph F Mecklenbräuker, and Gerald Ostermayer. Modeling isolated traffic control strategies in traffsim. In *2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)*, pages 143–148. IEEE, 2017.

[81] Panpan Cai, Yiyuan Lee, Yuanfu Luo, and David Hsu. Summit: A simulator for urban driving in massive mixed traffic. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4023–4029. IEEE, 2020.

[82] District 6 Annual MeetingInstitute of Transportation Engineers (ITE) Institute of Transportation Engineers 2000. "practical applications of traffic simulation using siftraffic", 2000. URL `https://trid.trb.org/view/671329`.

[83] Hafiz Usman Ahmed, Ying Huang, and Pan Lu. A review of car-following models and modeling tools for human and autonomous-ready driving behaviors in micro-simulation. *Smart Cities*, 4(1):314–335, 2021.

[84] Richard W Rothery. Car following models. *Trac Flow Theory*, 1992.

[85] Johan Janson Olstam and Andreas Tapani. *Comparison of Car-following models*, volume 960. Swedish National Road and Transport Research Institute Linköping, Sweden, 2004.

[86] MF Aycin and RF Benekohal. Comparison of car-following models for simulation. *Transportation research record*, 1678(1):116–127, 1999.

[87] Haiyang Yu, Rui Jiang, Zhengbing He, Zuduo Zheng, Li Li, Runkun Liu, and Xiqun Chen. Automated vehicle-involved traffic flow studies: A survey of assumptions, models, speculations, and perspectives. *Transportation research part C: emerging technologies*, 127:103101, 2021.

[88] Da Yang, Liling Zhu, Yalong Liu, Danhong Wu, and Bin Ran. A novel car-following control model combining machine learning and kinematics models for automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):1991–2000, 2018.

[89] Sara Moridpour, Majid Sarvi, and Geoff Rose. Lane changing models: a critical review. *Transportation letters*, 2(3):157–173, 2010.

[90] Peter G Gipps. A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological*, 20(5):403–414, 1986.

[91] Mizanur Rahman, Mashrur Chowdhury, Yuanchang Xie, and Yiming He. Review of microscopic lane-changing models and future research opportunities. *IEEE transactions on intelligent transportation systems*, 14(4):1942–1956, 2013.

[92] Rahmi Akçelik. A review of gap-acceptance capacity models. In *CONFERENCE OF AUSTRALIAN INSTITUTES OF TRANSPORT RESEARCH (CAITR), 29TH, 2007, ADELAIDE, SOUTH AUSTRALIA, AUSTRALIA*, 2007.

[93] K Ahmed, M Ben-Akiva, H Koutsopoulos, and R Mishalani. Models of freeway lane changing and gap acceptance behavior. *Transportation and traffic theory*, 13: 501–515, 1996.

[94] G Kotusevski and KA Hawick. A review of traffic simulation software. 2009.

[95] PTV GROUP et al. Ptv vissim is the world's most advanced and flexible traffic simulation software, 2020.

[96] Jaume Barceló et al. *Fundamentals of traffic simulation*, volume 145. Springer, 2010.

[97] Jordi Casas, Jaime L Ferrer, David Garcia, Josep Perarnau, and Alex Torday. Traffic simulation with aimsun. In *Fundamentals of traffic simulation*, pages 173–232. Springer, 2010.

[98] Muhammad Rehmat Ullah, Khurram Shehzad Khattak, Zawar Hussain Khan, Mushtaq Ahmad Khan, Nasru Minallah, and Akhtar Nawaz Khan. Vehicular traffic simulation software: A systematic comparative analysis. *Pakistan Journal of Engineering and Technology*, 4(01):66–78, 2021.

[99] Johannes Nguyen, Simon T Powers, Neil Urquhart, Thomas Farrenkopf, and Michael Guckert. An overview of agent-based traffic simulators. *Transportation research interdisciplinary perspectives*, 12:100486, 2021.

[100] Andreas Horni, Kai Nagel, and Kay W Axhausen. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.

[101] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[102] Daniel Krajzewicz. Traffic simulation with sumo–simulation of urban mobility. In *Fundamentals of traffic simulation*, pages 269–293. Springer, 2010.

[103] Larry E Owen, Yunlong Zhang, Lei Rao, and Gene McHale. Traffic flow simulation using corsim. In *2000 Winter Simulation Conference Proceedings (Cat. No. 00CH37165)*, volume 2, pages 1143–1147. IEEE, 2000.

[104] Pete Sykes. Traffic simulation with paramics. In *Fundamentals of traffic simulation*, pages 131–171. Springer, 2010.

[105] Laron Smith, Richard Beckman, and Keith Baggerly. Transims: Transportation analysis and simulation system. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 1995.

[106] Rob Hranac, Emily D Sterzin, Daniel Krechmer, Hesham Rakha, Mohamadreza Farzaneh, et al. Empirical studies on traffic flow in inclement weather. Technical report, United States. Federal Highway Administration. Road Weather Management Program, 2006.

[107] Darcin Akin, Virginia P Sisiopiku, and Alexander Skabardonis. Impacts of weather on traffic flow characteristics of urban freeways in istanbul. *Procedia-Social and Behavioral Sciences*, 16:89–99, 2011.

[108] Hesham Rakha, Mohamadreza Farzaneh, Mazen Arafeh, Robert Hranac, Emily Sterzin, and Daniel Krechmer. Empirical studies on traffic flow in inclement weather. *Final Report–Phase I*, 385, 2007.

[109] Jean Andrey, Brian Mills, Mike Leahy, and Jeff Suggett. Weather as a chronic hazard for road transportation in canadian cities. *Natural hazards*, 28(2):319–343, 2003.

[110] Chung-Cheng Lu and Xuesong Zhou. Short-term highway traffic state prediction using structural state space models. *Journal of Intelligent Transportation Systems*, 18(3):309–322, 2014.

[111] II Ismagilov and SF Khasanova. Algorithms of parametric estimation of polynomial trend models of time series on discrete transforms. *Academy of Strategic Management Journal*, 15:20, 2016.

[112] Mike West and Jeff Harrison. Polynomial trend models. *Bayesian Forecasting and Dynamic Models*, pages 208–233, 1997.

[113] Zhixiao Xie and Jun Yan. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of transport geography*, 31:64–71, 2013.

[114] Miaomiao Yan and Yindong Shen. Traffic accident severity prediction based on random forest. *Sustainability*, 14(3):1729, 2022.

[115] Milica ekić. The use of video detection as a function of traffic safety. *Tehnika*, 66 (3):471–475, 2011.

[116] Vaishnavi Ravindran, Lavanya Viswanathan, and Shanta Rangaswamy. A novel approach to automatic road-accident detection using machine vision techniques. *International Journal of Advanced Computer Science and Applications*, 7(11), 2016.

[117] V Machaca Arceda and E Laura Riveros. Fast car crash detection in video. In *2018 XLIV Latin American Computer Conference (CLEI)*, pages 632–637. IEEE, 2018.

[118] Tanja Verster and Erika Fourie. The good, the bad and the ugly of south african fatal road accidents. *South African Journal of Science*, 114(7-8):63–69, 2018.

[119] Ludovic Gicquel, Pauline Ordonneau, Emilie Blot, Charlotte Toillon, Pierre Ingrand, and Lucia Romo. Description of various factors contributing to traffic accidents in youth and measures proposed to alleviate recurrence. *Frontiers in psychiatry*, 8:94, 2017.

[120] Banishree Ghosh, Muhammad Tayyab Asif, Justin Dauwels, Wentong Cai, Hongliang Guo, and Ulrich Fastenrath. Predicting the duration of non-recurring road incidents by cluster-specific models. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1522–1527. IEEE, 2016.

[121] Wei Chiet Ku, George R Jagadeesh, Alok Prakash, and Thambipillai Srikanthan. A clustering-based approach for data-driven imputation of missing traffic data. In *2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS)*, pages 1–6. IEEE, 2016.

[122] Zinat Ara and Mahdi Hashemi. Identifying the severity of road accident impact on traffic flow by ensemble model. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 115–122. IEEE, 2021.

[123] Douglas Bates. Linear mixed model implementation in lme4. *Manuscript, University of Wisconsin*, 15, 2007.

[124] Xin Yan and Xiaogang Su. *Linear regression analysis: theory and computing*. world scientific, 2009.

[125] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409*, 2019.

[126] Sifatul Mostafi and Khalid Elgazzar. An open source tool to extract traffic data from google maps: Limitations and challenges. In *The 2021 International Symposium on Networks, Computers and Communications*, pages 1–8. IEEE, 2021.

[127] Federico Vaca, Herbert G Garrison, Mary Pat McKay, and Catherine S Gotschall. National highway traffic safety administration (nhtsa) notes. *Annals of emergency medicine*, 47(2):203, 2006.

[128] Qiang Shang, Tian Xie, and Yang Yu. Prediction of duration of traffic incidents by hybrid deep learning based on multi-source incomplete data. *International journal of environmental research and public health*, 19(17):10903, 2022.

[129] Lina Shan, Zikun Yang, Huan Zhang, Ruyi Shi, and Li Kuang. Predicting duration of traffic accidents based on ensemble learning. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 252–266. Springer, 2018.

[130] Jongtae Lim, Songhee Park, Dojin Choi, Kyoungsoo Bok, and Jaesoo Yoo. Road speed prediction scheme by analyzing road environment data. *Sensors*, 22(7):2606, 2022.