VEHICULAR ACCIDENT OCCURRENCE ANALYSIS AND PREDICTION

by

Peter Devin Way

Mina Sartipi
Professor of Computer Science
(Chair)

Michael Ward
Professor of Computer Science
(Committee Member)

Farah Kandah
Professor of Computer Science
(Committee Member)

VEHICULAR ACCIDENT OCCURRENCE ANALYSIS AND PREDICTION

by

Peter Devin Way

A Thesis Submitted to the Faculty of the University of Tennessee at Chattanooga in Partial Fulfillment of the Requirements of the Degree of Master of Science: Computer Science

The University of Tennessee at Chattanooga
Chattanooga, Tennessee

May 2020

ABSTRACT

Vehicular accidents within Tennessee increased by 25% within 2009-2019 according to Tennessee's Integrated Traffic Analysis Network. Accidents rank in the top three causes of accidental death across all ages in the U.S, and in 2017 accounted for 11.9% of all deaths by injury, the National Vital Statistics Report and Center for Health Statistics report. Accidents represent a massive cost in economics with 12.5 million in damages within 2018 from statistics from National Safety Council. These statistics indicate need for thorough investigation into reduction of accidents in our society. This thesis focuses on that need, with introduction of a novel predictive model based on historical accident occurrence in Hamilton County, Tennessee. The use of weather forecasts, roadway geometrics, and aggregated variables aids in creation of predictions for future accident occurrence. Additionally, an application is presented for use by local law enforcement and emergency services to assist resource deployment based upon predictions.

ACKNOWLEDGMENTS

I would first like to thank my wife, for always being my editor, rubber ducky, and general cheerleader. I would have never gotten this far without you, and I can never thank you enough. I would also like to thank the faculty and staff at the University of Tennessee for everything that they've taught me, whether in the classroom or in the 'real-world'. Thank yous are also in order for my committee members, Dr. Michael Ward and Dr. Farah Kandah. An additional note of gratitude is due to Dr. Claire McCullough, who always encouraged me to do the work while also staying absolutely myself. Another thank you is in store for Grace McPherson, for introducing me to how fun and engaging computer science outside of class time could be.

I would also like to thank the members of the Chattanooga City IT team, Mr. Kevin Comstock - Smart City Director of Chattanooga, the wonderful officers of the Chattanooga Police Department for their time, feedback, and input, everyone at the Enterprise Center, and NSF US Ignite for funding this project through award number 1647161.

Of course, no proper listing of thank-yous would be complete without acknowledging the amazing research team I've spent the last three years of my life working with. Whether I met you back in the days of SCAL or more recently with the team at CUIP, I owe so much to you. A special thank you is in store for Jose Stovall and Dr. Mina Sartipi. Without my friendship with Jose, I would have never begun this amazing Data Science journey, and without Dr. Sartipi's direction, I would certainly not be penning this thesis. I am in eternal debt to Rebekah Thompson, for her guidance on this eventful thesis adventure based on her previous journey (There and back again...). Finally, a thank you to my research partner, Jeremy Roland. He has always made sure my writing doesn't get too loquacious, and my graphs not too detailed. Thanks Jeremy, I'll try to keep this one brief.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In areas without a strong public transit presence, daily transportation often involves the usage of passenger vehicles such as sedans, trucks, and vans. Statistics from the Center for Disease Control and the National Center for Health Statistics indicate that vehicular accidents are one of the two highest fatality risks across all age groups, causing more deaths than drug overdoses in the years 1999 to 2012 [1,2]. Accident occurrence in both Tennessee and the United States altogether has been escalating rapidly in recent years, as shown by Figure 1.1. Vehicular accidents placed within the top three causes of death for every age group tracked by the CDC in 2018, as shown by Figure 1.2. In 2017, motor vehicle accidents accounted for 11.9% of all deaths by injury [3]. Additionally, estimates from the US Department of Transportation place the number of vehicular accident injuries in 2018 alone at 1.6 million, with 4.1 million accidents resulting in property damage [4]. In 2018, the state of Tennessee reported 208,605 accidents that caused more than four hundred dollars in damages within its counties [5]. The total economic cost of accidents in 2018 within the United States exceeded 12.5 million dollars, equaling up to the annual incomes of over 208 households [6,7].

In light of the previously mentioned statistics, it is clear that vehicular accidents are one of the biggest threats within daily life. As accident occurrence continues to climb, particularly in Tennessee, the need to reduce or prevent as many accidents as possible continues to grow.

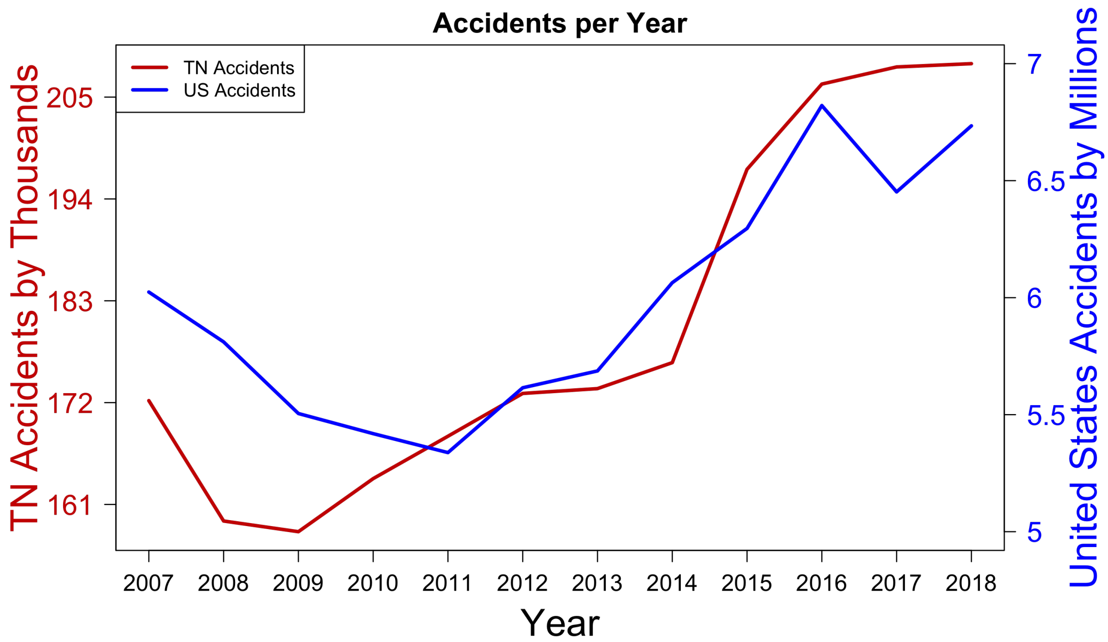


Figure 1.1 Total Vehicular Accident Counts for United States [4] and Tennessee [5] from 2007 to 2018. While both trend steadily upward, Tennessee totals experienced an alarming spike in 2015, and show no signs of decelerating. The cumulative United States totals reflect a slight dip in the most recent years.

1.1 Research Questions

The particular questions that arise from the above statistics that this work seeks to address are as follows:

- Where are accidents transpiring within the area of study?
- When are accidents ensuing within the area of study?
- Are there reoccurring areas of concern, historically?
- Is the occurrence of vehicular accidents able to be predicted?
- Are there specific actions able to be performed to mitigate accident occurrence?

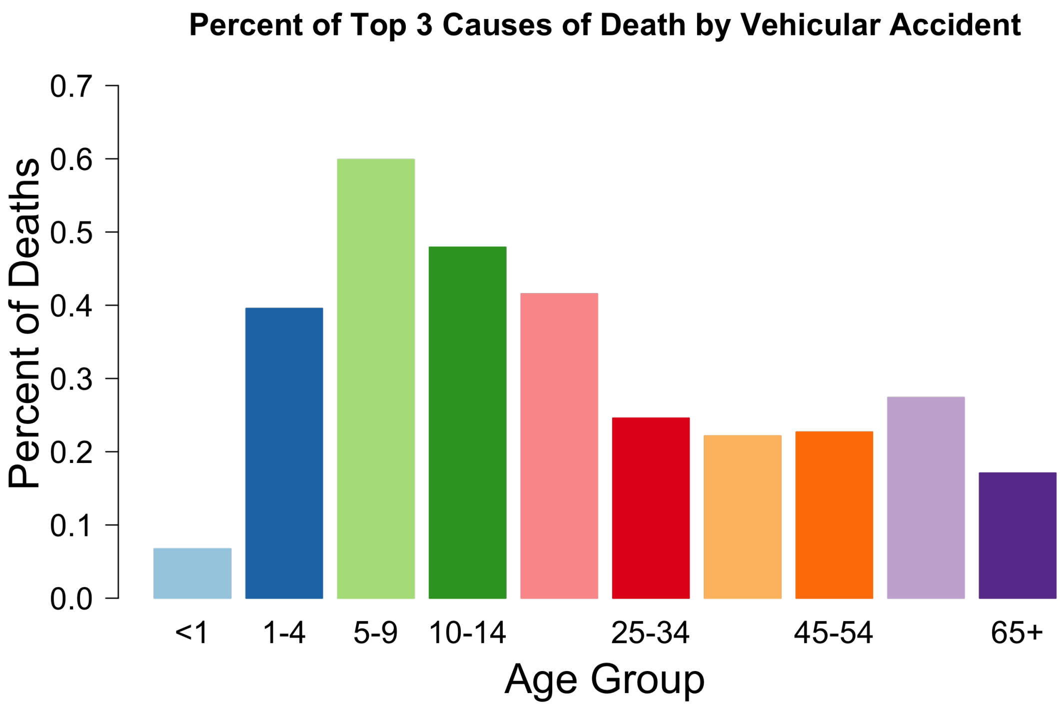


Figure 1.2 Percentage of the Top Three Accidental Death Causes from the Center of Disease Control [1] attributed to Vehicular Accidents, here shown divided by age group. Of particular concern is 59.89% of children aged 5-9 losing their lives in vehicular accidents, followed by 47.9% aged 10-14.

1.2 Motivations and Contributions

Therefore, this thesis seeks to address the needs outlined by the above research questions through the analysis, prediction and prevention of vehicular accidents within Hamilton County, Tennessee. Presented here is an MLP neural network model created by the author and the team at University of Tennessee, Chattanooga's Center for Urban Informatics and Progress in order to predict where and when accidents occur on any given day and time within the area of study. The model utilizes historical vehicular accidents and weather conditions, roadway geometrics, and other aggregated variables in creation of predictions of future accidents. Also introduced here is an application intended for local law enforcement use in resource deployment assistance based upon said model's predictions.

1.3 Organization

Within the following thesis, separate Chapters cover individual sections of analysis completed. Chapter 2 presents a brief introduction to the fields covered by this study. Chapter 3 presents works completed in this field of study previously. Chapter 4 explores the data used within this study. Chapter 5 presents how work was completed, as well as the architecture of the neural network utilized. Chapters 6 and 7 respectively explore the spatial and temporal distribution of accidents studying within this study. Chapter 8 discusses results arrived at through this study. Finally, Chapter 9 summarizes the contributions of this work, as well as possible future branches of this study.

CHAPTER 2

Background Information

For ease of understanding for the content presented within this thesis, this Chapter will explain key concepts used during the development of the accident prediction application presented in the following chapters. The concepts primarily covered in this section are: artificial intelligence and neural networks (Section 2.0.1), roadway geometric terms (Section 2.0.2, and weather terms (Section 2.0.3).

2.0.1 Artificial Intelligence and Neural Networks

Artificial Intelligence is defined by the Oxford Dictionary as :

> The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. [8]

Neural Networks can be considered akin to a computer brain required to perform those tasks. This brain is inspired by the layout of the real human equivalent, in that where brains have neurons, neural networks have nodes. The network takes variables in as input that will help inform the final decision. For example, if someone wanted to know if it was a good day to go to the park, the input variables might include Temperature, Weather Forecast, Parking, or Friends Available. The process of this decision making is shown in Figure 2.1.

There are a multitude of different neural networks, but the specific one discussed in this work is called a Multi-Layer Perceptron (MLP) Neural Network. The architecture of an MLP style neural network is very similar to the one described above, but includes at least

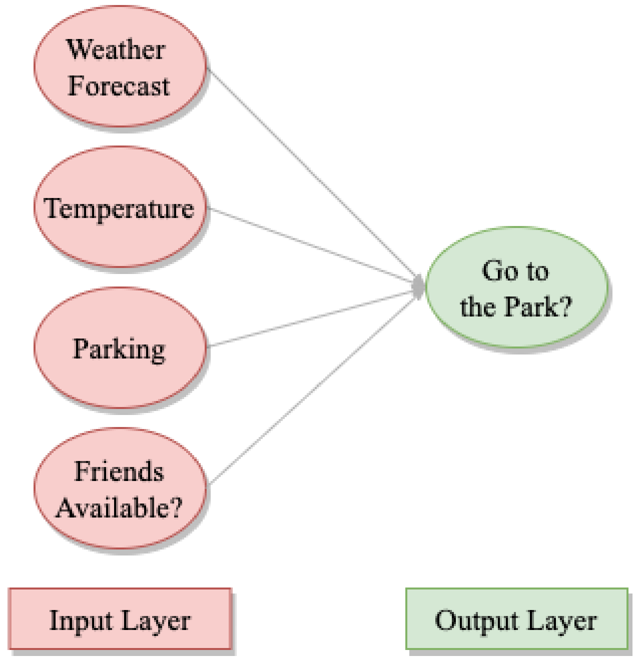


Figure 2.1 Basic Neural Network. Within neural networks, the possible determining factors of a decision are considered as input variables before a decision is derived, here termed output.

one hidden layer between input and output. Each layer of nodes is informed by the nodes of the previous layer, and in turn informs the next layer.

Extending the analogy from above, if that individual wanted to go to the park but they didn't want to get rained on, they may place more weight with (care more about) the forecast calling for rain more than if their friends were available, or if there was parking. This is exactly what MLP style neural networks do, too. Each input variable has a different weight assigned to it by the network that corresponds to how important that variable is to deciding the final output. The weights shift with each layer, with the output of a neural network just being the decision that one makes according to the inputs provided. This extension of classic neural networks is shown in Figure 2.2.

The second portion of an MLP architecture is Perceptron. Perceptrons date farther back than the artificial intelligence we know today, with the term first being coined in 1957 by psychologist Frank Rosenblatt [9], who was instrumental in the earliest days of artificial intelligence. Rosenblatt defined a perceptron as simply a linear algorithm for learning binary classifications. In the case of the park analogy above, one has a Yes or a No binary decision to make, although in most situations the question isn't quite so linear.

Perceptrons are especially used in what's known as Supervised Learning. This supervised learning involves giving a neural network both the input and the labelled answers, with the goal being to help the network learn how to best arrive at the correct answer.

Expanding the brain/neural network similarities more is the presence of Activation Functions. Activation functions are formulas that decide whether or not the node in question fires to pass the information along based off of a given threshold. For example, the weather forecast has a fifty percent chance of rain for the day one wishes to go to the park. Since the forecast can't say one way or another, there would be no reason to consider it at all. There are a multitude of activation functions for use in neural networks, all with their respective strengths and weaknesses. Sigmoid is one example, and is most suited for binary classification as in the case of deciding to go to the park. Sigmoid will be discussed further in depth in Section 5.2.

Finally, neural networks sometimes utilize a process called backpropagation. Essentially, this process is how neural networks analyze why a given output was wrong by calculating the error in respective to the weights assigned to any given node/layer. This analysis could be viewed like a game of pong with motions forward and back comparing the output to the correct answer. Forward passes are the neural network analyzing the given input to produce the output, and backward passes are the network checking their answers.

The process of backpropagation tells us which parameters could be adjusted to draw closer to the minimum error rate requested, since each forward pass and backward pass bring the output closer to the correct answer.

However, there does come a time that the neural network can no longer adjust the output toward the correct answer. This point is called convergence, and is considered the point where a network can no longer improve given its current architecture.

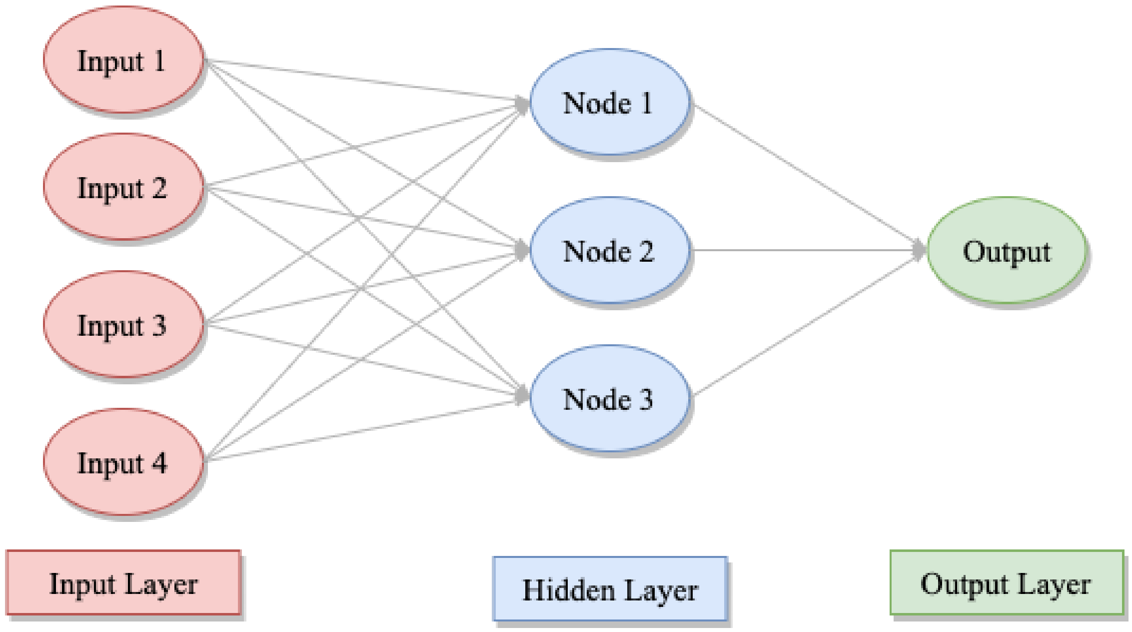


Figure 2.2 MLP Neural Network Architecture. This rudimentary example of a MLP network differs from the aforementioned standard neural network through the inclusion of a hidden layer, where the input variables are carefully weighted before being passed on to the output layer.

2.0.2 Roadway Geometrics

Many unique terms exist for the description of roadway geometrics, some of which are not known to those outside of the transportation sphere. Therefore, one must familiarize themselves with the vocabulary used before proceeding into accident analysis. Many of the following definitions were originally sourced from The Federal Highway Administration (FHWA) [10] to assist in understanding roadway data retrieved from the Tennessee Department of Transportation. A sampling of the terms referenced within this study are found in Table 2.1, and expanded upon here. Terrain type is defined as the type of land surrounding a given roadway and may include options such as rolling, flat, or mountainous. Grade is the term used for the height of an intersection or roadway segment. Options for grade may include below-grade, where one roadway dips below another or the surrounding terrain (such as a tunnel or underpass), or above-grade where one roadway rises above the surrounding terrain and roadways, such as in the cases of bridges or overpasses. Finally, 'at-grade' is the term used for cases where the roadways and surrounding terrain are all along the same height.

Function class is a term for the type of roadway. Function class types may include options such as rural local or urban interstate. Lane Number is quite self-explanatory, but includes the number of all lanes in a roadway segment including turn lanes. Average Annual Daily Traffic (or AADT) records the total volume of vehicular traffic for a given year, then divides that total by the number of days in that year. This provides a rough estimate for how much traffic specific roadways will experience.

Government Control explains which government entity controls the roadway segment in question. Entities range from private, to National Park Service or State Highway Agency. Administration system is quite similar to government control, but provides a finer view of the controlling entities. Examples of administration systems are NHS Airport Connector, State Park, and STP Urban. Note that explanations of roadway acronyms used here can be found in Table 2.2.

Access control reports how much access common citizenry have to the road in question. Control levels are none, partial, or full. Land use is another fairly simple term, describing the manner in that the surrounding portions of land are used. Options for land use include Commercial, Residential, Fringe/Mixed, Industrial, Public Use and Rural. Finally, operation describes the directional capability of a roadway, with either one-way or two-way.

Furthermore, different types of roadways are called by different classes based off of their traffic capability [11]. Freeways are the highest capacity roadways, designed for high speeds and the highest level of mobility freedom, but relatively limited access. Sometimes, freeways may be aggregated together with arterial roadways, which are termed as high capacity urban through roads. Collector roadways have a low to moderate traffic capacity, and serve shorter distances at a lower speed than arterials. Finally, local roadways are the lowest class, and feature the lowest speeds. This limits mobility, but with a related rise in access. This class contains the most roadways, as any roadways not defined as arterial or collector fall into this category [11].

Table 2.1 Sampling of Roadway Geometrics Terms

Term	**Description**
Terrain Type	Type of land surrounding roadway.
Grade	Height of intersection
Function Class	Type of roadway in question.
Lane Number	Number of lanes on given roadway. Numeric, whole numbers.
Government Control	Which government entity controls the roadway in question
Administration System	Similar to government control, albeit with a finer view.
Access Control	How much access average citizens have to the roadway
Land Use	The manner that the surrounding portions of land are used.
Operation	Directional capability of roadway.
Freeway	Highest capacity urban road, designed for high speeds
Arterial	High-capacity urban through road, connects freeways to collectors.
Collector	A low to moderate capacity urban through road, connects locals to arterials.
Local	Lowest capacity urban through road, lowest speeds

Table 2.2 Common Acronyms in Transportation

Acronym	**Meaning**
NHS	National Highway System
STP	Surface Transportation Program
TDOT	Tennessee Department of Transportation
FHWA	Federal Highway Administration
NHTSA	National Highway Transportation Safety Administration
AASHTO	American Association of State Highway and Transportation Officials
SPF	Safety Performance Functions
CBA	Cost Benefit Analysis
MMUCC	Minimum Uniform Crash Criteria
PCR	Potential for Crash Reduction
ADT	Average Annual Traffic
AADT	Average Annual Daily Traffic
CMF	Crash Modification Factors
HSIS	Highway Safety Information System
RAR	relative accident risk
HSIS	Highway Safety Information System
KABC0	K Class - Killed A Class - Incapacitating injury B Class - Evident injury C Class - Possible injury 0 Class - Property Damage Only

2.0.3 Weather Terms

Although some weather terms explored in this study are generally assumed as common knowledge, others may be considered obscure. As such, all weather terms will be defined and introduced here, and are shown in Table 2.3. Also discussed here are the effects each particular weather condition can have on drive-ability, as originally discussed by FHWA [12]. On average 5.8 million crashes occur each year, with roughly 21% of those accidents occurring during or due to adverse weather conditions. Nearly three fourths of all weather related accidents occur with slick or wet pavement, and almost half occur during rainfall [12].

Temperature may not be of particular risk for an accident, but the combination of low temperatures and precipitation could cause icy or slick roadways [12]. Dewpoint is related to temperature, in that dewpoint is defined as the atmospheric temperature below which water droplets begin to condense and produce dew. This is of particular concern in the early morning hours of winter, when dew may create slick and icy areas along roadways [12]. Humidity also relates to temperature, in that it is the amount of water vapor in the surrounding air relative to the temperature. Here, relative humidity is specified, meaning it is a percentage.

Precipitation is another concern when analyzing accidents, and as such multiple variables are to be studied. The first is simply the type of precipitation forecast for a specific location at a given date-time. Generally precipitation type is a string or word, providing the name of the precipitation, like rain or snow. The next variable is Precipitation Intensity, or the amount of precipitation occurring at the given location and time, generally measured in inches of liquid per hour. High precipitation amounts can obscure vision and affect the driver's ability to maintain their lane. Finally, the last variable directly concerned with precipitation is the probability of precipitation occurring, given in a zero to one inclusive format. An example would be .8 equaling 80% probability of precipitation.

Pressure is the next variable, defined as air pressure at sea-level. Often studied in the analysis of accident occurrence due to its relation to tire pressure which directly affects the driver's control over the vehicle, pressure is measured in millibars. Too little tire pres-

Table 2.3 Explanation of Weather Terms

Variable	**Explanation**
Temperature	Current temperature at given location
Dewpoint	Atmospheric temperature below which water droplets begin to condense, producing dew
Humidity	Amount of water vapor in the surrounding air relative to the temperature (0-1 inclusive)
Precipitation Type	Generally a string variable, giving the name of the type of precipitation currently being experienced
Precipitation Intensity	Intensity of precipitation occurring at the given time, given in inches of liquid per hour
Precipitation Probability	Probability of precipitation occurring (0-1 inclusive)
Pressure	Air pressure at sea-level, given in millibars
Cloud Cover	The percentage of the sky currently covered by clouds (0-1 inclusive)
UV Index	Measure of ultraviolet rays at current time/place
Visibility	Visibility in miles due to current weather conditions
Wind Bearing	Direction of wind origin, in degrees (0 being True North)
Wind Gust	Sudden, short lived burst of high speed wind
Wind Speed	Speed of ordinary wind, given in miles per hour

sure means more of the tire's surface area is encountering the roadway, increasing friction. Increased tire friction can lead to overheating and premature wear and tear on the tire trends [13]. Conversely, too high of pressure in the tires leads to not enough tire surface area meeting the road, causing bounce and decreased friction. This could result in loss of control, or swerving [13].

Cloud cover is the amount of the sky currently covered by clouds. This is of use in the area of accident analysis due to the visibility of a driver possibly being compromised by sun glare if cloud coverage is quite low. On a related note, UV index is the measure of ultraviolet rays at the given time and place. While generally being related to sun-burns, UV index can also inform as to the intensity of the sun's total rays for the given time. Visibility is generally measured in miles, and may rise or fall depending on precipitation conditions, sun glare, or other air quality events.

The final category of weather variables of concern is wind. As noted in the table, wind gusts are sudden, short lived bursts of wind that reach much higher speeds when compared to sustained wind. Wind bearing also bears mentioning, although this entirely depends on the direction of the roadway (An example being if a roadway is North/South and winds are blowing East, vehicle stability and control may be compromised). While not as much of a concern in the mountainous areas of Tennessee, wind gusts and high speeds are dangerous in flat, low lying areas. Another concern with wind is blown precipitation. High winds may blow snow into drifts along the edges of lanes, or obscure vision. Vehicle performance may also suffer, as the driver struggles to keep the vehicle in their lane [12].

CHAPTER 3

Related Works

Previous studies sought how to mitigate the dangers of highways and expressways [14–18], as well as weathers' effects on our roadways [19–23]. Some research has combined these approaches, studying the interplay between the two types of factors [24–26]. Work has also been done into factoring driving habits, age, and other driver based factors into accident occurrence statistics [27–29]. Furthermore, studies have examined injury intensity as related to type of vehicular accident [30]. However, studies have only recently delved into the reduction of accidents before they can even occur [31–35].

Due to the subject matter at hand spanning several subdivisions of research, this literary survey will be divided into a number of subsections. Section 3.1 will delve into the existing literature concerning how roadway geometrics can affect accident occurrence, while Section 3.2 examines previous studies concerning weather conditions and accidents. Section 3.3 then explores studies that combined both of the two previously mentioned. Section 3.4 explores studies where the focus was more on understanding where accidents occur, regardless of causation, as well as driver/occupant geared studies. Finally, Section 3.5 presents previous works on actual accident prediction via usage of an assortment of methodologies.

3.1 Roadway Geometrics

The first completed study included here occurred in 1974 studying the vehicular accidents that had transpired in rural Kentucky between 1970 and 1972 [14]. Of particular interest with this research is the development of a Severity Index by the authors for the comparison of roadway types, such as two-lane, four-lane highways, interstates, etc. In the work, the severity index is defined as shown in Equation 3.2, where EPDO (shown in Equation 3.1, and utilizing the KABC0 levels of injury explained within Table 2.2) is equal to the weighted sum of fatalities (shown as F) and incapacitating injuries (A), which is then summed together with a lesser weight sum of non-incapacitating injuries (B) and possible injuries (C). This total is then summed to the number of property damage only accidents (PDO). The EPDO sum is then divided by the total number of accidents that had occurred on the given roadway type (shown as N_T). Therefore the complete formula results in an index used to calculate the severity of each individual roadway type [14].

$$EPDO = 9.5(F + A) + 3.5(B + C) + PDO \tag{3.1}$$

$$SeverityIndex(SI) = \frac{EPDO}{N_T} \tag{3.2}$$

Based off of this severity index, four lane undivided highways were found to have the highest average rate of severity, with toll roads having the lowest [14]. Once this index was adjusted to account for occupant injury rather than roadways, it was found that pedestrian involved accidents scored the highest index, followed closely by single vehicle accidents. Seatbelt usage was found to lower severity, with only .4% fatalities for seat-belted occupants compared to the 1.7% fatality rate of non-seat-belted occupants [14].

Roadway geometrics were also under study in Washington state, through the use of Negative Binomial regression, in the analysis of annual accident frequency of principle arterials of Washington [15]. Data was split into Eastern and Western portions due to drastically different roadway geography. The study concluded several relationships. For example, increases in vertical grade (steep hills) resulted in an increase in accident frequency [15]. It is suggested this is due to freight trucks slowing dramatically on the inclines and increased levels of risk taking behavior from passenger vehicles. It was also found that higher posted speed limits resulted in lower accident frequency, with this relation being contributed to better geometric conditions existing along highway corridors [15]. Furthermore, roadside shoulders of less than 1.5 meters in width were associated with higher frequencies of accidents. This is contributed to vehicles having less room for error correction and maneuverability. Finally, the study concludes with the statement that the Poisson distribution is a poor choice for accident analysis due to over-dispersion (that is, variance being greater than the mean) within accident data [15].

State-specific Safety Performance Functions (SPF) for rural interstates and rural 2-lane roads were used to identify the 20 segments of each type with the highest Potential for Crash Reduction (PCR) in a study within Kentucky [16]. A Cost Benefit Analysis (CBA) was then performed, using appropriate Crash Modification Factors (CMF) for the types of crashes occurring in an effort to make an index that normalized the safety benefit of all roadway classes based on implementation cost. Minimum Uniform Crash Criteria (MMUC) was used along with Knee Airbag Deployment models for identifying and classifying accident data [16]. Once road segments with the highest PCR values were identified another CBA was used to highlight which sites would provide a return on investment as well as ranking the segments deserving treatment.

A study was undertaken to develop SPF for sections of a four-lane National Highway in Uttarakhand, India [17]. Numerous road geometrics were considered, such as the curvature change rate, slope change rate, transverse slope, ADT, number of median openings per segment, number of sedans, number of sport utility vehicles (SUV), operating speed of traffic, etc. The dependent variable in this case was the number of accidents per 200 meter segment per year. Poisson Weibull distribution and ordered logit models were employed for analysis, with model results suggesting that the presence of SUVs was the most significant contributor to increases on crash occurrences by far [17]. Closely clustered in second and third were number of median openings and curvature change rate of the segments in question. It is suggested that the number of median openings leads to maneuvering traffic conflicts with opposing traffic streams. Higher speeds led to the most inverse for vehicular accidents. This is possibly due to the absence of non-motorized traffic, as well as lower numbers of vehicles on the roadway segment in question all together. The study also suggests that operating traffic speed may vary significantly due to congestion, poor weather or irregular geometric design on roadway segments [17].

Accident data regarding commercial semi-trucks was retrieved from HSIS for the testing of four regression models (Two conventional linear, and two Poisson) [18]. The limitations of both types were discussed, as linear regression techniques lack the ability to adequately describe randomness of accident events. Meanwhile, as mentioned before, Poisson style regression techniques are generally good statistically, but fail if the data is over-dispersed. This leads to severe under or over statement of likelihood. Nevertheless, in the testing of all four models horizontal curvature, vertical grade, and AADT were found to positively contribute to accident occurrence [18]. The inclusion of extremely short road segments (those measuring .05 miles or less) lead to the over inflation of accident counts when utilizing linear regression. The Poisson style models performed better on likelihood functions, as well as Akaike Information Criterion (AIC) values. Additionally, the first Poisson model required less computational effort than the second [18].

3.2 Weather Conditions

The effects of weather conditions on daily crash counts were analyzed using a discrete time-series model within the Netherlands [19]. In the project, an integer autoregressive model was introduced for modeling count data with time inter-dependencies. Then a model was built from daily vehicular accident data, meteorological data, and traffic exposure data from three cities in the Netherlands: Dordrecht, Utrecht, and Haarlemmermeer. Daily vehicle counts were collected for each road segment of the major road networks based on loop detector data. From this, day-to-day total amount of vehicle kilometers driven were calculated on the major road network of each city region. Weather data was also collected for the three cities and deaggregated into specific weather instances. This same type of deaggregation was done for wind, temperature, sunshine, precipitation, air pressure, and visibility. With all data then collected, an Integer-Valued Autoregressive model was used to find the significance of different weather conditions on accident occurrence. It was discovered that several weather variables were significant in relation to accident occurrence.

Weather conditions of Iowa were studied utilizing matched pairs analysis to conclude Relative Accident Risk (RAR) scores [20]. The statewide score during the period of study was found to be 1.69. However, the RAR scores drastically shifted due to liquid and frozen precipitation. The scores were also found to vary significantly by hour of day, with a RAR score close to 2 for 14:00, and 1.3 during early morning hours. It was also found that interstates and highways ranked higher on the RAR score than smaller roads, and precipitation affected areas were located. These two findings suggest interplay between precipitation and traffic volume/density (associated with hour of day) in regards to accident risk. Maximum RAR values (2.7) were observed during winter months, and the minimum (1.3) found in autumn and spring when variables were isolated. Temperature was also found to greatly affect RAR scores when isolated, with temperatures close to -5 degrees Celsius leading to a spike of 3.7. The effect of precipitation on accident occurrence was found to have a distinct lag, with roughly one hour lag for liquid precipitation, and up to 48 hour delay for frozen precipitation.

An extensive literature survey covering thirty-four papers as well as seventy-eight records was completed, spanning research completed from 1967 to 2005 [21]. The rate of accidents was normalized across studies in regards to effect size. Across the survey, it was found that accident rates increase during precipitation, with snow affecting accident rates more than rain. Rain tends to increase the accident rate by 71%, and the injury rate by 49%. Meanwhile, snow was found to increase crash rate by 84%, and 75% for injury rate. For example, snow led to a 100% increase in accident rates in the United Kingdom. As the intensity of precipitation increases, so does the risk of an accident. Slick or icy roadways compound upon this, leading to even more significant risks. Interestingly, it was found that the effect of snow on accident rates actually fell over the decades under study, from 113% from 1950-1979 to only 47% during 1990 to 2005. It was suggested that this change may be due to overall safety improvements, but perhaps winter maintenance methods could have contributed as well.

North Carolina accident data from the HSIS database was analyzed to understand injury severity through the use of ordered probit model estimations [22]. The HSIS database includes accident, traffic, roadway information, as well as other data. The majority of the records originate from rural areas, not local streets or secondary roads unless they are controlled by the state. The period of study was from 1990 to 1995, and included the counties of Dare, Graham, Pamlico, Swain, and Transylvania. All available records were cut down to only those including single vehicle ran-off-road and off-road object incidents, as well as two-vehicle side-swipe or rear-end incidents. It was found that traffic congestion (AADT per lane) was a significant push for two-vehicle accidents. Curved roadways also increased the relative likelihood of single-vehicle accidents (as compare to straight and level roadways). Grades and hills also were found to increase accident occurrence in two-vehicle accidents. Overall, wet surfaces increased accident risk compared to dry roadways. Adverse weather conditions led to a significant decrease in injury severity, perhaps due to increased driver caution in adverse conditions. It was suggested that AWS (Advanced Weather Systems) devices may be utilized to prematurely adverse drivers of slippery/slick roadways.

The effects of fog or smoke (FS) on vehicular accidents was explored in a study investigating Florida data from 2003 to 2007 [23]. A total of 994 FS accidents were discovered, with an additional 120053 accidents that did not involve fog or smoke being utilized as the control group. The study was completed in two steps, with the first devoted to examining FS accident characteristics in regards to temporal, influential factors, and accident types. The second step was the estimation of variable effects on injury severity with FS accidents. It was found that FS accidents most occurred in the early morning hours of December, January, and February. Also, FS accidents resulted in more severe injuries and involvement of more vehicles when compared to the control group. Head-on or Rear-end were the most common types of FS accident, with distribution of FS accidents mostly found along roadways featuring high speed limits, without dividers or sidewalks. Additionally, FS accidents resulted in more severe injuries when occurring at night, without illumination. The study concludes with the recommendation of the addition of street lights in high concentration areas of FS accidents.

3.3 Combined Weather and Roadway

The effects of traffic and weather characteristics on road safety were examined together by [24]. Gaps were identified, and needs for future research discussed as well. Roadway data such as average daily traffic (ADT), road density ratio, and speed were considered as well as weather data such as precipitation, fog, and sunshine. It was found that Logit models were the dominant means of analysis used for crash severity, and time series analyses were the dominant means of analysis used for weather data [24].

A study was performed in the early 1990s including data that spanned four European countries with concern originally placed on weather conditions [26]. Data was collected into a segmented database containing the monthly accident counts from the counties from each of the countries: Denmark (14 counties), Norway (19 counties), Finland (11 counties) and Sweden (24 counties). This database also included statistics for each month regarding weather conditions, duration of daylight, and either traffic counts for roadways or gasoline sales (used here as a proxy for traffic volume). The data collected spanned from the mid 1970s to the late 1980s, with slight variations in date coverage for each of the countries. A generalized Poisson regression test was completed regarding randomness, weather, daylight, exposure (traffic volume/gasoline sales), and changing routines or speed limits. It was found that 80 to 90% of the accidents could be accounted for by either randomness or the exposure variable. Additionally, the risk per exposure unit decreased by roughly .36% for each percent increase in traffic volume. Therefore, the effect of weather conditions on the occurrence of accidents was not found to be as concerning as the traffic volume. The study concluded by suggesting that the best way to decrease accidents was by decreasing traffic volume.

3.4 Monitoring and Quantification of Accidents

Seven years worth of vehicular accident records were retrieved from Bellevue, Washington in order to investigate 63 intersections slated for improvements during the time of study [27]. Roads were classified as principal, minor, collector arterials, or local streets. Each intersection was divided into four approaches, providing a division of the data due to the direction the accident took place within. In total, 1385 accident reports were collected, whose variables included number of lanes, speed limit, grade, sight-distance restriction, and others for 64 distinct variables all together. Accidents were provided a type from rear-end (26%), angle (30%), approach (32%), or other (12%). Negative Binomial Regression was completed on the data, with results indicating that left turn volume, right turn volume, and total traffic volume resulted in increases to accident occurrence of 2.28%, .92%, and 2.95% respectively. Intersection approaches that featured no control system led to lower number of accidents, although this could be contributed to lower volume intersections not requiring a control system. The same could be said for the local streets variable, which also led to lower occurrences. Higher approach speeds and restrictions on sight-distance also resulted in higher accident occurrences. The work concluded with a call for accident-reduction programs, as well as a less restricted study of intersections.

Crowd sourced crash alert data from Waze (a GPS navigation software that also allows users to report traffic congestion, accidents, and other roadway concerns) and California Highway Patrol (CHP) reports from the summer of 2018 were compared in Support Vector Machines to identify Waze alerts that corresponded to CHP reports [28]. The reports included time, location, type of incident, and user confidence in the report. Waze alerts were received 2 minutes and 41 seconds (mean) before the corresponding CHP report was initiated, with a precision measure of .87, and recall of .88. The earliest alert from Waze was received four minutes and three seconds before the corresponding report. Waiting for a second or third alert led to decreased advance warning compared to the CHP, without any corresponding improvement to precision (.87, .88 respectively) or recall (.88, .88). This study concluded that the use of crowd sourced alerts has the ability to decrease emergency response time by 20 to 60%, as well as providing advance warning to hospital staff or trauma teams.

A study investigating single-vehicle accidents (2729 total) in North Carolina wished to determine if recent speed limit increases led to more occurrence over a two year study period [36]. Paired Comparison and Ordered Probit models were both explored with similar results. The dependent variable was set as the highest level injury of the accident, from the KABCO scale (see Table 2.2 for explanation). Independent variables were split into two subsections: policy and external. Policy variables were created to classify accidents by occurrence on a study segment, as well as whether the accident occurred before/after the speed limit policy change. External are those variables that may influence the occurrence of accidents and are not related to the policy changes (i.e., roadway geometrics, weather, vehicle type). The Paired-Comparison model did not apply external variables, instead utilizing comparison sites that did not receive speed limit increases. Almost all injury levels and speed limit divisions were found to have increased likelihood odds of accident occurrence on study road segments after the speed limit increase when examined using the Pair-Comparison method. Similar results were presented by the Ordered Probit model, however external variables were included. It was found that higher occupant counts, vehicle flipping, alcohol involvement and striking an object all lead to increased odds for injuries. As presented, the study concluded that raising speed limits did in fact increase likelihood of sustaining all Class level injuries.

Another study utilizing HSIS data explored the effects of roll-over accidents on injury levels in Michigan and Illinois [37]. Three years of accident data was gathered from 1994-1996 for Michigan (35,447 accident records) and from 1993-1995 for Illinois (24296 accidents). Logistic Regression was utilized to understand the relationship between vehicle roll-over and injury severity. It was found that vehicle roll-over, neglect of seat-belts, alcohol use, and passenger cars (as opposed to pick-up trucks) all increased injury risk. Vehicle collision with an object before roll-over also increased injury occurrence. However, use of seat-belts provided a decrease of 15.6% in injury risk, and slick roadways provided a decrease as well, perhaps due to increased driver caution.

The effects of information and vehicle technology on injury severity of rear-end accidents was analyzed utilizing HSIS data, again from North Carolina [30]. Two and three vehicle rear-end accidents were under study, with controls in place for driver, vehicle, and roadway factors. Ordered Probit Models were created for each of the drivers in these records, with variables including the presence of Center Head Mounted Stoplights (CHMSL) and anti-lock braking systems (ABS), as well as vehicle age, which was chosen for its relationship to safety improvements over time. It is known that the presence of CHMSL systems can reduce the response time of following drivers by .11 seconds, and this was reflected in the results of the study. In two vehicle rear-ends, a newer model in vehicle one gave its driver an 8.7% reduction in injury occurrence. Similarly, if driver one is at the wheel of a large vehicle, it was found to provide an 11% reduction in KABC class injuries. If vehicle two was a large vehicle, then driver one was 5.8% more likely to sustain injuries. Ultimately this still leaves a net decrease in injury occurrence if both vehicles are of a larger model. In two vehicle accidents, the presence of CHMSL in vehicle one and ABS in vehicle two led to a reduction in injuries. The presence of a third vehicle changed the likelihood of injuries in both vehicle one and two. Driver one sustained injuries in 31.2% of two vehicle accidents, but only 23.6% in three vehicle accidents. The inverse is true for driver two, with injuries of 12.0% increasing to 37.8% in three vehicle incidents. Driver three was ultimately the least likely injured, with only 17.2% of third drivers sustaining injuries. Overall, the study of vehicle age and safety mechanisms found that technological improvements have had a material benefit to injury occurrence in vehicular accidents.

Accident data from 1993 to 1995 was provided by the Florida Department of Highway Safety and Motor Vehicles (DHSMV) and informed a study regarding driver age and injuries/death rates in Florida [31]. Resident and non-resident data was divided and studied separately, to glean more understanding into the injury risk to Florida citizens. Teen drivers between 15 and 19 were involved in 6.76 accidents (average) for every 100 teen drivers registered. Accident rates declined for each age bracket until drivers 70 to 74, with the lowest average at 1.38 per 100 registered drivers. The rate then rose again, to conclude at 1.8 for drivers 85 and older. Non-residents data was age-divided a bit differently, with 6.73 accidents per 10,000 visitors for ages 18 to 25. A similar U shape is presented in non-resident data as in resident data, with the lowest average found to be .91 per 10,000 visitors at ages 56 to 65. Visitors aged 65 and older were found to have an average of 2.05. Resident fatality rates were also investigated, with teen drivers at 2.33 fatalities per 10,000 registered drivers, then continuing into the usual U curve. Elderly drivers aged 80 and above were at the highest risk, with drivers 80 to 84 experiencing fatalities at 2.88 and drivers 85 and older experiencing 2.98 per 10,000 drivers. Data was also examined spatially, with teen drivers found to be at most risk in coastline counties and elderly drivers at risk in Dade (Miami), Seminole (Orlando), Gulf (Panama City), Dixie (Cross City), and Jefferson (Tallahassee) counties. All told, it was found that teen drivers and elderly Florida resident drivers are most at risk for both injuries and fatalities.

3.5 Accident Prediction

A case study was conducted for prediction of traffic accidents by utilizing and comparing the results of four different classification models of prediction [32]. Those methods included: linear Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Deep Neural Networks (DNN). Here, a method of generating non-accident data was performed and called negative sampling. For each positive example (accident), the value of only one feature was changed among hour, day, and road ID. Afterwards, the resulting sample was checked for a positive (match found) or negative (no match found) result amongst the existing dataset. In the end, results dictated that the most optimal model was DNN for the application in discussion.

Convolutional Long-Short Term Memory (ConvLSTM) is a subsidiary of Long Short Term Memory involving the use of convolutional operations inside of the LSTM cell. The convolutional operations allow for multi-dimensional data such as radar or satellite imagery. This ConvLSTM setup was applied to a study concerning vehicular accidents in Iowa, between

2006 and 2013 [33]. Data included crash reports from Iowa DOT, rainfall data, Roadway Weather Information System (RWIS) reports, and further data provided by Iowa DOT such as speed limits, AADT, and traffic camera counts. This study mentions that no other previous work had fused such large heterogeneous datasets together before, nor had any included spatial structures of the roadway network. As such, a five kilometer square per block grid layout was constructed to cover the state for prediction forecasting. Training data included 2006 to 2012 reports, with 2013 being reserved for testing. Tests involved predicting locations for the next seven days based on data provided by the previous seven days. ConvLSTM results out performed all baselines in prediction accuracy. As well, the system correctly predicted accidents resulting from the case study of December eighth in 2013, where a significant snowstorm caused numerous accidents.

The accidents within the city of Montreal were studied by a team from Concordia University via the use of a Balanced Random Forest algorithm [34]. Their model included data from three open datasets. Accident data was retrieved from Montreal Vehicle Collisions, weather information was provided by the Historical Climate Dataset, and roadway segment information was retrieved from the National Road Network database, provided by the Canadian government. Four different models were tested, including BRF (Balanced Random Forest), RF (Random Forest), XGB (XG Boost), and a baseline model. Negative samples (that is, examples of non accident occurrence) were created. A total of two billion negative samples were possible, with the team electing to only utilize .1% of such. Predictions were for roadway segments by the hour, a highly specific definition. All together, the systems were able to predict 85% of Montreal accidents, with a False Positive Rate (FPR) of only 13%. It is notable that the datasets in use were open source, implying that the study could easily be shifted to another locale relatively easily, since no data restrictions were in place.

An initiative led by the USDOT seeks to partner crowd-sourced data and safety policy decisions [35] to help predict vehicular accidents. Prediction is completed through the use of Classification and Regression Trees (CART) and Random Forest models. The pilot for the study includes six months of accident data from Maryland, paired with the corresponding (if any) Waze alerts. Specific temporal and spatial event patterns created by the pilot model are quite similar to the actual accident records, albeit not identical. Additionally, the model tends to under predict accidents in early morning hours, while over predicting accidents for high-commute periods. Of note is the model's ability to predict Waze alerts for minor accidents, that is, those not serious enough to report but significant enough to inhibit standard traffic flow. The study continues with the partnership of state and local partners for implementation of multiple case studies of the Waze crash estimation model.

3.6 Summary

A wide variety of studies have been discussed, without a particular common thread between them. Based upon the work of previously mentioned studies, it can be said that weather and roadway conditions do indeed have an effect on the rate of accident occurrence regardless of the locale under scrutiny. All of the studies employed assorted data sets, with all having some sort of accident reports and most having weather or roadway data, if not both. Particularly in the case of accident prediction, it is critical to utilize as much data as possibly available. This fact informed the data employed in the study presented here, whose data will be discussed in the following Chapter.

CHAPTER 4

Data

Within this Chapter, the specifics of the data utilized within this study will be reviewed. Section 4.1 covers the three datasets in question. Accident reports are discussed within 4.1.1, temporal and spatial additions within 4.1.2, weather variables within 4.1.3 and concluding with roadway geometrics within 4.1.4. The Chapter concludes with process of negative sampling being discussed within Section 4.2.

4.1 Data Sources

4.1.1 Hamilton County Emergency District

Beginning in March of 2018, the previous day's 911 records regarding traffic accidents were delivered to the Center for Urban Informatics and Progress on a daily basis for study. This was in addition to the collective accidents from late 2016 to that date that were delivered together initially. The records resulted from citizens calling into 911 to report witnessing or being part of a vehicular accident. Some accidents were duplicated in cases where multiple people had called in to report the same accident and duplication removal was necessary to avoid skewing the data toward accidents with multiple reports. Data fields from the initial files are shown in Table 4.1.

Early records of the initially available set often did not include address or city, with some not even including latitude or longitude. As such, some early reports had to be dropped due to incompleteness. The time of the accident is considered to be the same as the response date, since there is no way to determining exact time of collision from the data available.

Table 4.1 Original 911 call report information fields

Field Name	Description
Response Date	DateTime call to 911 was placed
Fixed Time Call Closed	DateTime the responding officer closed the call report
Address	Address of accident occurrence
City	City of accident occurrence
Latitude	Latitude of accident occurrence
Longitude	Longitude of accident occurrence
Problem	Level of injury involved, as reported by caller.

4.1.2 Spatial and Temporal Additions

Throughout testing, two different types of spatial aggregating grid block setups were explored, with two different sets of predictive models and results. These grid layouts will be thoroughly discussed in Chapter 6. The grid block setup of both editions were factored into their respective models through the addition of multiple variables. Additionally, some later variables depended on grid block designation rather than the accidents themselves. Initial alpha-numeric indicator Grid_ID was reworked to be recorded as Grid_Num, Grid_Col and Grid_Row. This allowed for the preservation of the relationship between rows and columns of grid blocks, while also accounting for the limitations of neural network models. Also of note is the consideration of historical accident occurrence for each of the grid blocks. This was reported via the use of the Join_Count variable.

Table 4.2 Explanation of Spatial Terms in Use

Variable	Explanation
GRID_ID	Alpha-numerical identifier of the grid block in question
Grid_Num / Grid Block	Numerical identifier of the grid block in question, for hexagonal and fishnet grids respectively
Grid_Col	Numerical identifier of the column for the grid block in question
Grid_Row	Numerical identifier of the row for the grid block in question
Join_Count	Historic Number of Accidents per grid block.

Aggregated temporal variables were also utilized for simplification of the predictive process. The specific variables used are shown by Table 4.3, and include DayFrame, Hour, WeekEnd, WeekDay, Unix, and DayOfWeek. DayFrame was a variable created based off of historical trends demonstrated by the data. WeekDay and WeekEnd are binary variables indicating whether the accident occurred on Saturday/Sunday (WeekEnd being 1) or during the work week (with WeekDay being 1). Unix replaced the Response Date column originally delivered so that neural networks could comprehend the data. DayofWeek is the numerical representation of the specific weekday, with 0 beginning the week on Monday. Many of these variables will be further discussed in Chapter 7.

Table 4.3 Explanation of Temporal Terms in Use

Variable	**Explanation**
DayFrame	Section of the day the accident occurred within
Hour	Hour of the day the accident occurred (military time)
WeekEnd	Binary indicator of accident occurring during the weekend
WeekDay	Binary indicator of accident occurring during the work week
Unix	Numeric representative of seconds since January 1st, 1970, the Epoch time
DayOfWeek	Numeric representation of day of the week, with 0 beginning at Monday.

4.1.3 DarkSky Weather API

After the vehicular accident records had been received, they were then matched to the weather conditions happening at the given time and location via the use of the DarkSky API for Python, including many of the variables previously discussed in Section 2.0.3. The API only requires a date-time object (or Unix code) and latitude/longitude pair to return the optimal weather report for that given location, sourced from a variety of weather reporting services. For example, one weather source was selected from the weather reports available in Table 4.4 when a latitude and longitude position from Ooltewah, TN was provided.

4.1.4 E-TRIMS Roadway Geometrics

The roadway geometrics mentioned in Section 2.0.2 could not be used for each and every accident if data was to be aggregated into gridblocks. Thus, the mode of roadway geometricss within each gridblock was found in order to be utilized by the model. These variables are

Table 4.4 Weather Sources Referenced for given Weather Report from DarkSky

Source	**Description**
HRRR	NOAA's High Resolution Rapid Refresh Model, available in the continental USA
NWSPA	NOAA's Public Alert system, available in the USA.
CMC	NCEP's Canadian Meteorological Center ensemble model, available globally.
GFS	NOAA's Global Forecast System, available globally.
ICON	The German Meteorological Office's icosahedral nonhydrostatic, available globally.
ISD	NOAA's Integrated Surface Database, available near populated areas globally for events more than two weeks past.
MADIS	NOAA/ESRL's Meteorological Assimilation Data Ingest System, available near populated areas globally.
NAM	NOAA's North American Mesoscale Model
SREF	NOAA/NCEP's Short Range Ensemble Forecast,
DARKSKY	Dark Sky proprietary hyper-local precipitation forecasting system, backed by NOAA's NEXRAD system radar.

displayed in Table 4.5. The mode number of lanes variable represents exactly as its title suggests. This allows for a suggestion of traffic estimates without the availability of AADT for each roadway and grid block. The TY_TERRAIN variable was the mode of terrain types within the given gridblock, from three possible values of flat, rolling, and mountain. Of course, these values are provided in numeric form for the model's usages. Note that the majority of gridblocks were found to have rolling terrain, due to the hilly nature of Chattanooga. The next variable utilized from roadway geometrics was FUNC_CLASS, or the mode of the function class of roadways within the given gridblock. These values correspond to the same sorts of values as discussed in 2.0.2. This again provides a clue to the types of roadways within a gridblock and their respective volume levels.

Table 4.5 Explanation of Aggregated Roadway Geometric Terms in Use

Variable	**Explanation**
NBR_LANES	Mode Number of Lanes in roadways within grid block, discrete integer
TY_TERRAIN	Mode Type of Terrain within grid block, represented numerically (Examples: Flat, Rolling, Mountain)
FUNC_CLASS	Mode Function Class within grid block, represented numerically (Examples: Urban Local, Urban Freeway)

4.2 Negative Sampling

Concerning the process of negative sampling there is no one agreed upon definition. This is due to various fields of study utilizing the term within a wide variety of applications. However, for this particular study, let negative sampling be defined as: The creation of negative events based off of adjustments made to positive events that render them sufficiently dissimilar.

4.2.1 Negative Sample Creation

In this thought process, negative samples were created in order to understand how the given variables influence the occurrence of accidents in the area of this study. Two different types of negative samples were completed. They are:

1 Grid Fix - Wherein the location variable Grid_Num is fixed in place, with temporal variables Hour and Date being shifted to new values. The resulting negative is tested for matches against the existing positive records to guarantee a true negative status. This test was informed by the work completed by [33], in that spatial variables are given a preferred treatment. Grid_Num remains the same, while temporal variables are shifted. This creates a negative with the same location, but entirely different time.

2 True Random - Wherein Grid_Num, Hour and Date are all shifted to a random new value, and tested against the existing positives to determine if a match exists. This test could be seen as a completely random negative example in that no special treatment is given to any of the variables.

4.2.2 Negative Sample to Data Ratios

Various ratios of data to negative samples were testing to test model predictive power over multiple rarity levels of data. For example, negatives were initially created in a nine to one ratio, provided that this many negatives were available for creation in the given model type. This process was done to simulate the relative rarity of accident occurrence. After this initial ratio was sent through the model, two additional cuts were created. This cutting process was completed by finding the percentage of accidents within the existing dataset, and then dropping a given number of negatives in an organized fashion to match negatives to the desired percentage. For example, if 100 accidents existed with negatives originally numbered at one thousand and we wished for an even distribution of negatives, we would drop all but every tenth record in negatives. This preserves the original range of distribution while cutting to the number of negatives sought after. The first split was called the 75/25 Split. As the name suggests, three negatives were retained to every one accident record. The second split was called the 50/50 Split, having an even split between accidents to negatives. Both versions of the Grid Block layout involved this ratio testing. Specifics of the given ratio cut entry totals are displayed in Table 4.6. Model results for each of the given models are to be discussed in Section 8.3.1 for Hexagonal models and 8.2.1 for Fishnet models.

Table 4.6 Ratio Counts for Types of Model Input

Model Name	**Grid Block Layout**	**Ratio**	**Total Entries**
Grid Fix	*Fishnet*		586353
	Fishnet	75/25	179799
	Fishnet	50/50	89453
	Hex		387426
	Hex	75/25	219577
	Hex	50/50	107677
True Random	*Fishnet*		434011
	Fishnet	75/25	170183
	Fishnet	50/50	90063
	Hex		422604
	Hex	75/25	237166
	Hex	50/50	104710

CHAPTER 5

Methods

Within this Chapter, Section 5.1 explores the stages of work necessary to complete the project, with Section 5.2 presenting the specific neural network utilized in this study.

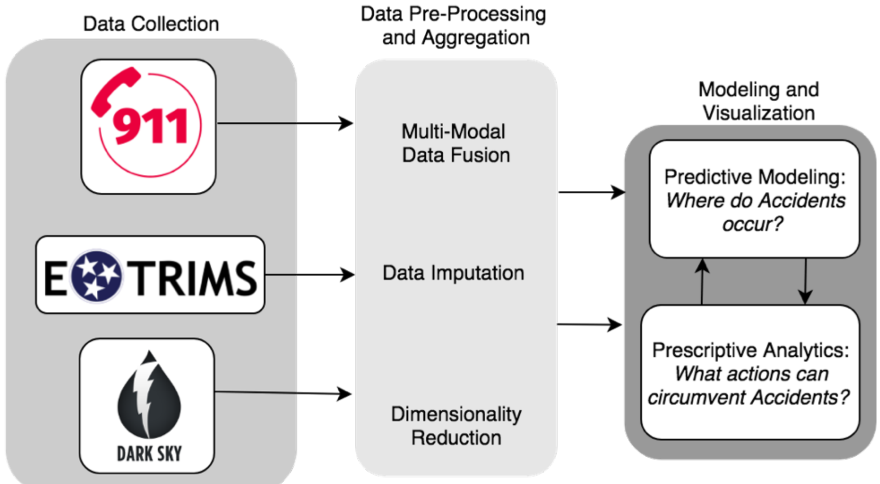


Figure 5.1 Workflow of Data Collection and Processing. Phase one of workflow includes gathering all datasets in use, shown by the Data Collection portion of the Figure. Phase two, Data Pre-Processing and Aggregation, involves fusing said datasets together, reducing redundancies where possible, and dropping variables that do not contribute to the possibility of prediction. Phase three concludes the workflow, with predictive modeling and prescriptive analytics, as shown by the right most portion of the figure, titled Modeling and Visualization. That is, this stage involves discovering where accidents historically have occurred, where they are likely to occur again, and what could be done to avoid these future accidents from occurring.

5.1 Workflow

Completion of the project at hand was completed within three distinct phases, as shown in Figure 5.1. Data collection (Phase 1) involves gathering all utilized data sources. Data Pre-Processing and Aggregation includes merging said data sources into one master file, reducing redundancies and determining which variables produce the most influence over the occurrences. The final stage involved in Modeling and Visualization, where spatial and temporal occurrence of accidents are visualized for understanding, as well as discussions breached to determine the best case of action for their eventual prevention.

5.1.1 Data Collection

First, data was collected via the previously mentioned three sources through various methods. Accident reports are received via email each morning for the previous day, although retrieving the reports in real time is currently under discussion. Roadway geometrics from the Tennessee Department of Transportation is retrieved via the E-TRIMS online database in the form of shape files. Since the data in question is not time sensitive, roadway geometrics are retrieved and updated roughly every three months. The final data source is Dark Sky, the previously discussed weather API for the Python programming language. Historical weather occurrence is retrieved and stored locally, while updated weather records are retrieved when accident reports are added into the main dataset.

5.1.2 Data Pre-Processing and Aggregation

The second Phase of workflow involves assembling the aforementioned individual data sets into a master file, while reducing any possible redundancies as well as eliminating unnecessary dimensions of variables. Decisions were made as to which variables to include, and which to drop. For example, initial Dark Sky weather reports include three seperate Temperature data points. These are Minimum Temperature, Maximum Temperature and Temperature, where the third variable included is the actual temperature at the time being referenced. Therefore, both minimum and maximum temperature were excluded in favor of the current temperature.

5.1.3 Modeling and Visualization

The final phase of workflow was the most involved, where spatial and temporal exploration began, as well as predictive modeling via neural networks. Much could be understood simply from the spatial and temporal distribution of accidents, before any addition testing had began. The understanding of spatial occurrence will be further discussed within Chapter 6, with temporal occurrence being reviewed within Chapter 7. Also, the combination of the two will be summarized within Section 8.4. Several different types of analysis were tested on the given dataset, with many under-performing or simply not working with the number of dimensions in question. Tested methods included logisitic regression, decision trees, variance thresholds, logit tables, and support vector classifiers. These early on tests will not be discussed in preference for allowing more exploration of neural network testing and results.

5.2 Neural Network

5.2.1 Architecture

Multi-Layer Perceptron style neural networks, as discussed in Section 2.0.1, were found to be the most beneficial tool for use in the given prediction study. The details of the given architecture are given by Table 5.1, with one input layer, three cascading node Dense hidden layers, and one node for the output layer. The number of nodes depending on the specific model editions number of variables (here referred to as X), and as such was found by setting the number of variables equal to X, then subtracting 5, then 10 from that model's unique value. These particular values were chosen to provide a gentle decrease in the number of nodes per layer. Note the use of a Dropout layer, which prevents over-fitting by randomly turning hidden nodes off (that is, the node is set to zero, or no) with a given probability. Here, that probability is set to 10%, meaning that there is a ten percent chance that any given node will be deactivated. This means that neighboring nodes will become more important when considering changing the weights within the backpropagation process.

Compilation was initially provided via binary cross entropy, due to its specific usefulness for binary classification results. However, as research continued it was found that Mean Squared Error (MSE) provided superior results. Across testing MSE compilation resulted in lowered loss levels, even when testing at minimal epoch counts. Optimization was provided by Nadam, an extension of the widely popular Adam style optimizer, with the incorporation of Nesterov Momentum instead of the standard RMS propagation of its parent optimizer. Where RMS divides the learning rate for a given weight by a running average of the magnitudes of the recent gradients for that particular weight, Nesterov Momentum instead makes a predicted gradient informed by the previous accumulated gradient, then measures the actual resulting gradient and makes changes accordingly.

Table 5.1 MLP Neural Network Architecture

Layer	**Location**	**Type**	**Node**	**Activation**
1	Input	Dense	X	Sigmoid
2	Hidden	Dense	X-5	Sigmoid
3	Hidden	Dropout (.1)	-	Sigmoid
4	Hidden	Dense	X-10	Sigmoid
5	Output	Dense	1	Sigmoid

As for activation, Sigmoid (shown in Equation 5.1) was selected due to its favored use in binary classification problems. Sigmoid provides a bounded, differentiable function of reals whose definition allows for any and all real input values with a non-negative derivative existing at each point. This results in a probability result that is easily understood, rather than an arbitrary integer or floating number.

$$S(x) = \frac{e^x}{e^x + 1} \tag{5.1}$$

CHAPTER 6

Spatial Exploration of Accidents

As discussed initially in Section 4.1.1, the accident reports provided span the entirety of Hamilton County. Throughout the time of study, the specific areas researched were reassessed twice. These two stints of study are discussed here, with the Fishnet styled block layout covered by Section 6.3 and the Hexagonal styled block layout by Section 6.4. However, before exploring the grid block layouts, it is important to understand why such extreme divisions of data were deemed acceptable, and to analyze the initial distributed spread of accidents across the entire county, as well as garnering a stronger understanding of Hamilton county itself. Thus, statistics concerning Hamilton County will be discussed in Section 6.1, and raw accident reports (i.e., before any grid layout was applied) will be discussed in Section 6.2.

6.1 Hamilton County Statistics

Hamilton County lies on Tennessee's southern border with Georgia, between 35.5 to 34.9 degrees North (Latitude) and 85.5 to 84.95 degrees West (Longitude). Upon its establishment in 1819, Hamilton became the 43rd county of Tennessee, with a population of 821 on the 1820 census. Hamilton covers 576 square miles, with just over 364,000 residents estimated in 2018 (population of 87.4 per square mile) [38]. Chattanooga, the main city of Hamilton County, accounts for 179,139 of the county's residents [39].

Hamilton County's residents travel roughly 21 minutes (20.7 minutes reported by Data USA, 21.7 minutes reported by US Census Bureau) to work each day, with 79.2% of the population above legal working age [7,39]. However, it is notable that over 25% of employed residents report more than thirty minutes commute time, with 7.44% reporting above forty-five minutes commute time [39]. Just under eighty percent of employees drive their own vehicles, another roughly ten percent utilizing ride share mobility or carpooling, with an additional seven percent working from home (Remaining 3.36 percent is divided between bicycle, public transit, walking, taxis, etc.) [7]. In total, Hamilton County residents commute for 56,256.76 hours average each day. Assuming for a 30 mile per hour speed limit, that adds up to over 1.69 million miles driven each day in Hamilton County.

6.2 Hamilton County Accident Distribution

The overall scatter of accidents within Hamilton County is shown by Figure 6.2a, with all accidents over the three years of study plotted onto the same map. This demonstrates that many accidents do occur in the southern half of Hamilton County, although individual roadways can be distinguished from the remaining scatter in the northern half.

Despite the total of miles driven each day in Hamilton County, most of the accident occurrences are clustered into the southern half of the county, particularly around the downtown Chattanooga area, or along the four main interstates of the area. These certain roadways are shown by black lines in Figures 6.2a and 6.2b. Interstate 75 begins in the eastern border of the county and reaches to 85.2 West, 35 North where it turns south, with Interstate 24 beginning its path westward. Before this main split, there is a smaller split where Interstate 75 meets Highway 153, around 85.16 West, 35.05 North. Note the contrast between accidents occurring along Interstate 24 and 75 as compared to Highway 153. Interestingly, US Route 27 (shown extending north from just east of downtown Chattanooga) shares roughly the same amount of accident frequency shown by 153 rather than Interstate 75 or Interstate 24. This could be contributed to highway/route status as compared to interstate, but all four of the above mentioned routes feature similar speed limits and roadway geometric layouts, thus the only unaccounted for difference would be traffic volume. As expected, there is a drastic difference in the roadways' AADT averages. As shown in Table 6.1, Interstate 24 has the highest volume of traffic from the group, with Interstate 75 in second place. Highway 153 is quite behind Interstate 75, with State Route 27 ranked as the lowest of the four in traffic volume.

Table 6.1 Average Annual Daily Traffic for Highest Volume Roadways

Route	Average AADT 2018 (All Stations)
Interstate 24	105278.72
Interstate 75	82303.67
Highway 153	70452.05
State Route 27	56177.93

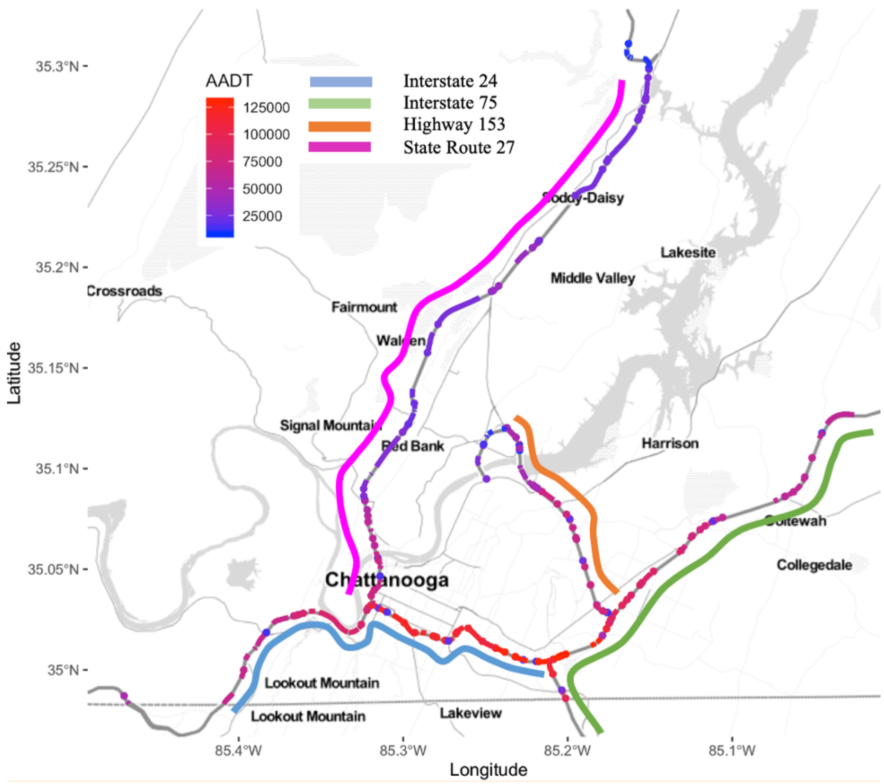
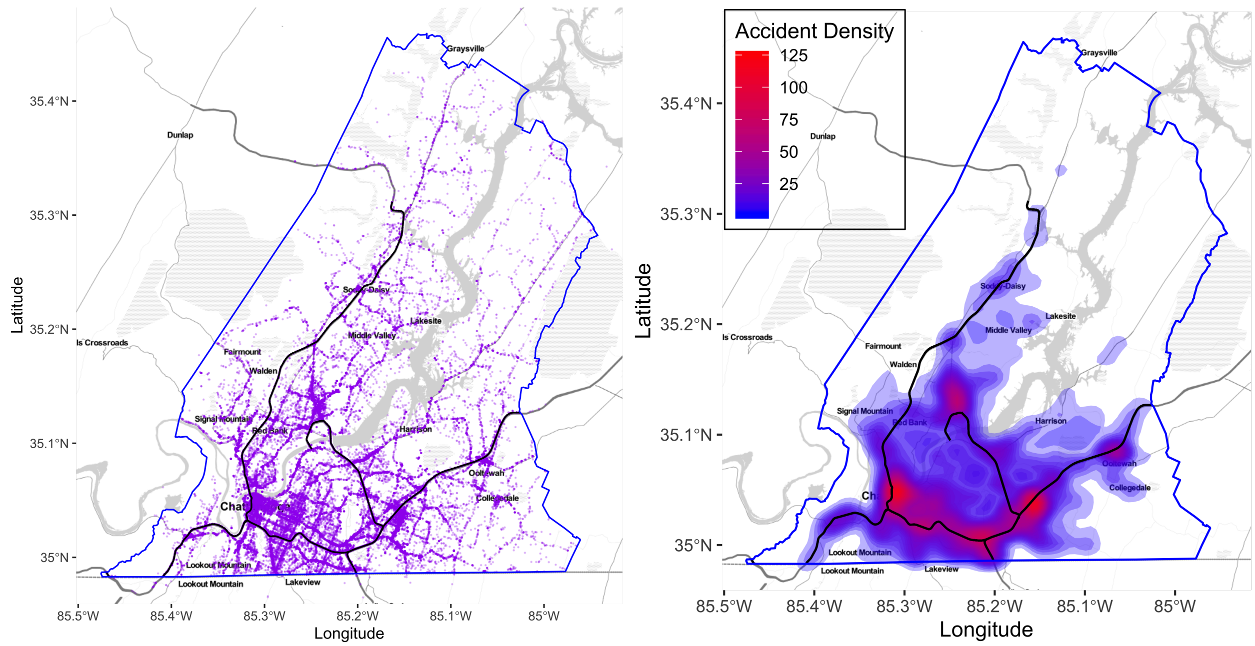


Figure 6.1 Local Highway AADT Stations. It is of note that although State Route 27 is the longest of the four arterials, it has the least traffic. This is due to the roadway passing through mainly rural areas, while Interstates 24 and 75, as well as Highway 153 mainly travel through urban areas. Interstate 24 in particular connects the high traffic areas of Hamilton Place/Gunbarrel Road and downtown Chattanooga.

Also of interest is the concentration of accidents at the end of Highway 153 in Hixson, TN. This particular density patch is located near Northgate Mall, Memorial North Hospital, and a troublesome railroad crossing currently under study by another project. The clustering of accidents in the downtown Chattanooga area is shown by the flare near 85.32 West, 35.05 North. The last non-Interstate or city center related hot spot is near the largest mall in the area, Hamilton Place, shown at 85.15 West, 35.04 North.

Traffic volume as experienced by individual grid block will be discussed further within Section 6.4, where it was used in a standardizing capability paired with total number of accidents within each individual grid block.



(a) Overall Scatter

(b) Density Distribution

Figure 6.2 Accident Distribution in Hamilton County. Figure 6.2a displays the total scatter of accidents across the county, with more clustering toward the Southern half. Figure 6.2b simplifies this scatter, displaying accident density, as well as the location of area interstates.

6.3 Fishnet Grid

In order to simplify the analysis of accident occurrences, the area of study was initially aggregated into 906 grid blocks of .2 mile width and height, in a fishnet styled pattern that closely followed the roadway network of Chattanooga. This layout is shown in Figure 6.4a. Although available accident data covers the entire county, limiting the area of study to the fishnet grid reduced the dataset by only 2.7%, as major hot spots of accident occurrence mainly arose in the downtown area, or along interstate corridors in the southern half of the county. This concentration is shown by Figure 6.3, which was created early on in the project in part to assist in determining the optimal coverage for the fishnet grid network.

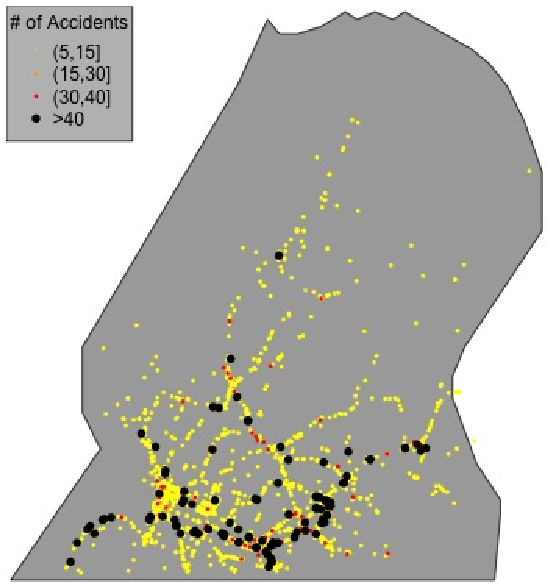
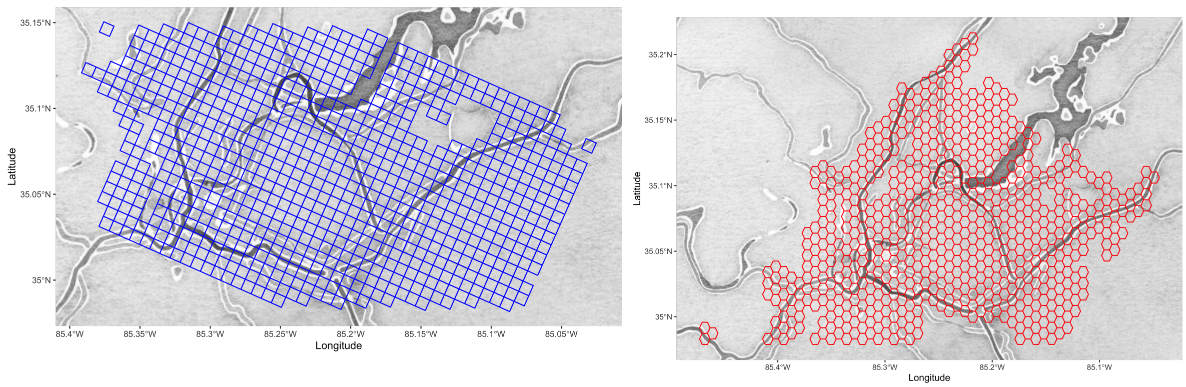


Figure 6.3 Concentration of Accidents over Hamilton County. Of note here is the heavy concentration of accidents in the Southern half of the county, where most of the urban areas are located. Locations of area interstates are also quite clear, shown here by heavy clustering of black points.

The grid block system itself was employed to aggregate accidents into more easily predicted points, as 906 unique locations is more digestible by predictive modeling systems as opposed to attempting to predict for infinitely many discrete GPS coordinates.

However, the beta version of the fishnet styled grid block layout was not originally oriented in this diagonal fashion. Rather, it was oriented horizontally across the area and featured 1100 .2 square mile grid blocks, as shown by Figure 6.6. This fishnet layout was replaced by the superior version already discussed due to the later properly lining up with the local roadway network. This switch circumvented odd segmentation of roadways and intersections that prevented the predictive model from providing satisfactory results. Said results of all Fishnet styled grids will be further discussed in Section 8.2.

Once the fishnet layout was applied, accident counts within each given grid block could be assessed. This division of accidents can be observed within Figure 6.7. Note that the



(a) Original Fishnet Style

(b) Updated Hexagon Style

Figure 6.4 Contrasting GridBlock Layouts from Project. The original layout (Figure 6.4a) was originally proposed, covering 97.3% of all accident reports. Later, the layout shown by Figure 6.4b was created for usage within Chattanooga city limits. This reduced the dataset to 62.9% of the available reports from across the entire county.

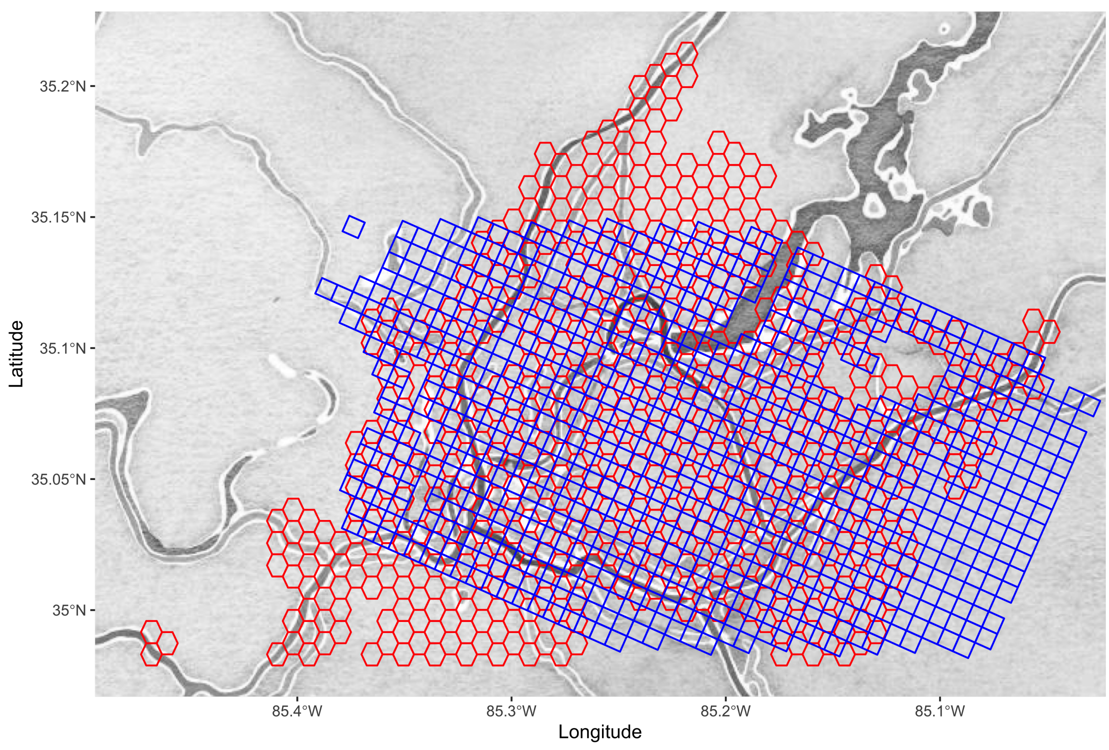


Figure 6.5 Illustration of Overlap with Grid Layouts. While both layouts cover the downtown Chattanooga area, interstates, and the high traffic areas around Hamilton Place Mall (the largest in the county), the Fishnet layout covered much more residential areas in the Eastern portion of Hamilton County than the Hexagon layout. The Hexagon layout in turn covers more branches of the area interstates, extending further North and South-west.

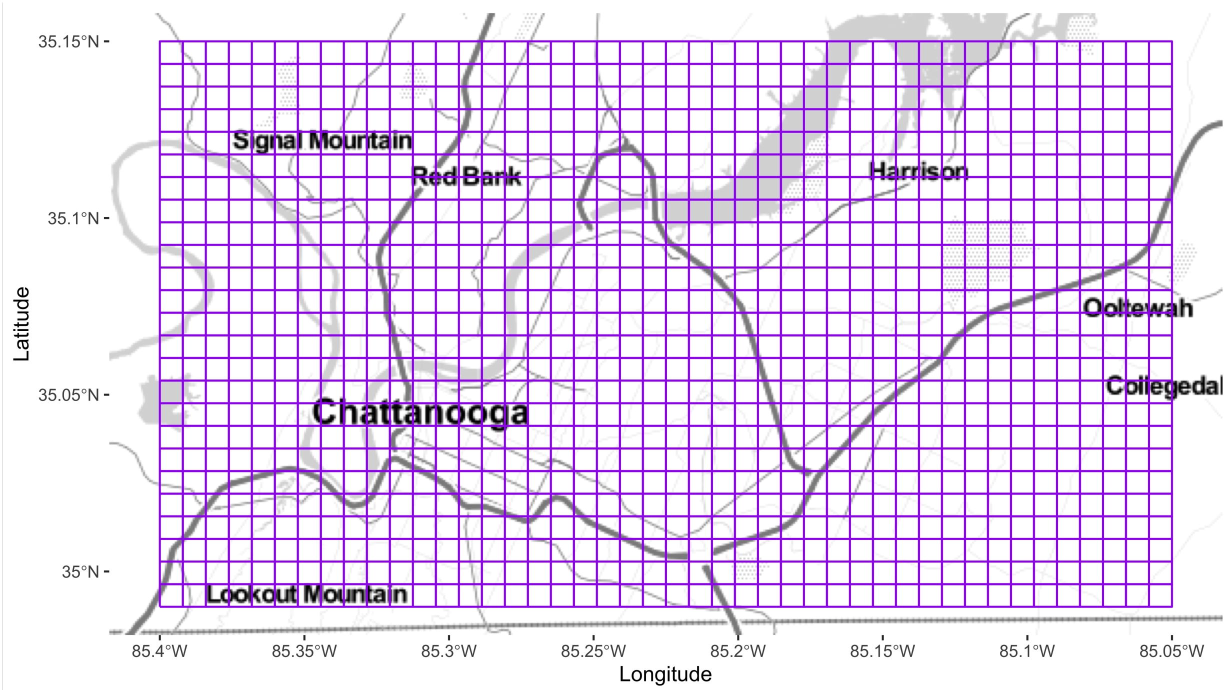


Figure 6.6 Beta Version of Fishnet Grid Layout. This version was abandoned in favor of an updated layout oriented to avoid bisecting as many roadway intersections, as well as removing some residential areas that had no accidents reported.

path of the interstates within the study area can easily be discerned, as well as the downtown area. The highest number of accidents within any given block was Grid ID N-21, located in the downtown Chattanooga area with 738 accidents. This particular block is shown by Figure 6.8, with the main feature being the entrance/exit ramps to Interstate 24 central to the block's coverage.

6.4 Hex Grid

As the study continued, opportunity arose to partner with Chattanooga Police Department in efforts to produce a predictive application for the Chattanooga area in hopes of allocating law enforcement resources to best prevent accident occurrence. This once again led to a restriction of the study area, this time confined to the Chattanooga city limits. Thus, the previous Fishnet style grid was set aside in order to focus the project solely on Chattanooga city limits. In this addition, hexagonal grid blocks were selected due to their preferred status

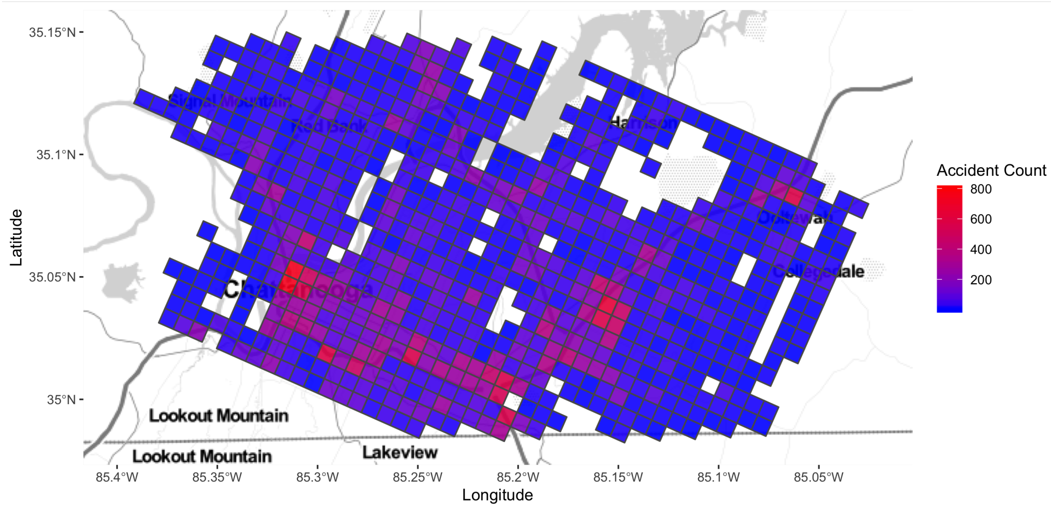


Figure 6.7 Accident Counts within Fishnet Grid Layout. Clearly standing out are the gridblocks containing the interstates, as well as the downtown area of Chattanooga, and the area surrounding the county's largest mall, Hamilton Place.

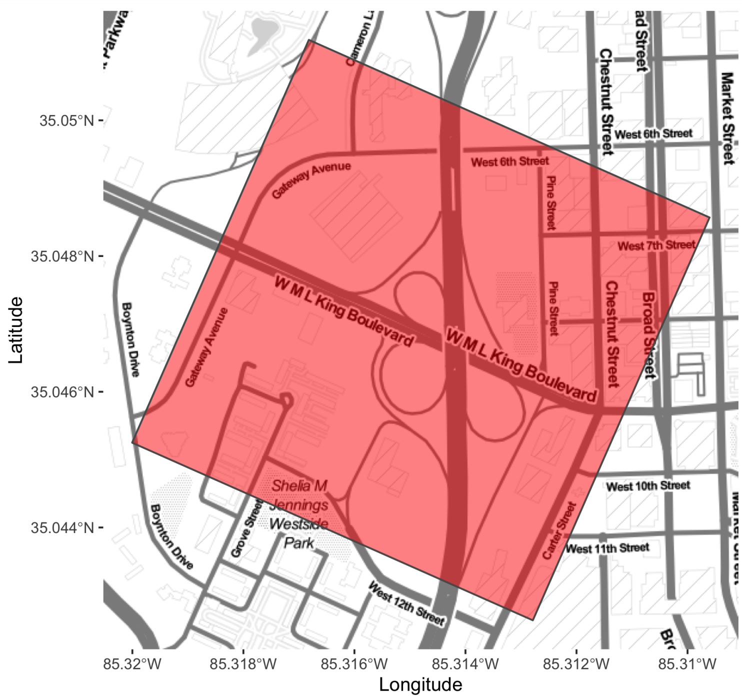


Figure 6.8 Highest Accident Count Block from Fishnet Grid Layout. Of particular concern in this gridblock is the Martin Luther King Boulevard entrance and exit ramps to Interstate 24. These ramps are historically problematic in accident occurrence, and are currently under construction for improvements as of the writing of this thesis.

by geo-spatial scientists and researchers. This resulted in a significant reduction in accident input for the predictive model, cutting the input to 62.9% of the initially available reports. However, this spatial reduction led to a significant boost to model performance power, as will be discussed in Section 8.3. The hex grid setup of 694 equally sized blocks (once again, each covers .2 square miles) is shown by Figure 6.4b. Any missing hex blocks within the body area indicate areas with no roadways, such as parks or bodies of water. A comparison of the land covered by the two layouts is illustrated by Figure 6.5, where the different study areas can easier be discerned.

Another spatial exploration of accidents within the hexagonal grid involved utilizing the average of AADT in each given hex block to normalize the number of accidents that had occurred within. Not all hex blocks contained AADT reporting stations, leading to their elimination for this particular step of analysis. The Exposure Rating (ER) is found by the following equation:

$$\text{Exposure Rating (ER)} = \frac{\text{Accident Sum of Hexagon Block}}{\text{Average AADT within Hexagon Block}} \tag{6.1}$$

Figure 6.9 illustrates the entire set of 522 hex blocks that had AADT reports available. Many of the eliminated hex blocks (shown solely by black outlines) featured residential areas with low accidents counts over the study period, and as such would have fallen under any given threshold. Most hex blocks Exposure Risk ratings were below .02, with the minimum Exposure Rating roughly .000056 compared to the highest ER at .0893. In total, the Exposure Rating contained a range of 0.089244. It was very clear that certain hexagons experienced much higher accident counts per AADT than others. Therefore, the top ten at risk hexagons are shown in Figure 6.10. Note that the highest Risk score within these ten was .08931 (hexagon 282), with the lowest score being .05985 (hexagon 195). The closest following this tenth hex block was at .04728, a difference of .01257.

Many of the ten worst rated hexagon blocks were clustered in high density areas, such as downtown Chattanooga (circled in pink), or Hamilton Place Mall (circled in blue). The hexagon with the highest rating was number 282, located from the bend of the Tennessee

River to just before the intersection of Cherry Street and Georgia Avenue. A total of 1020 accidents occurred in this hexagon over the course of the study, with an average AADT of roughly 11420 and an average speed limit of just over 30 miles per hour.

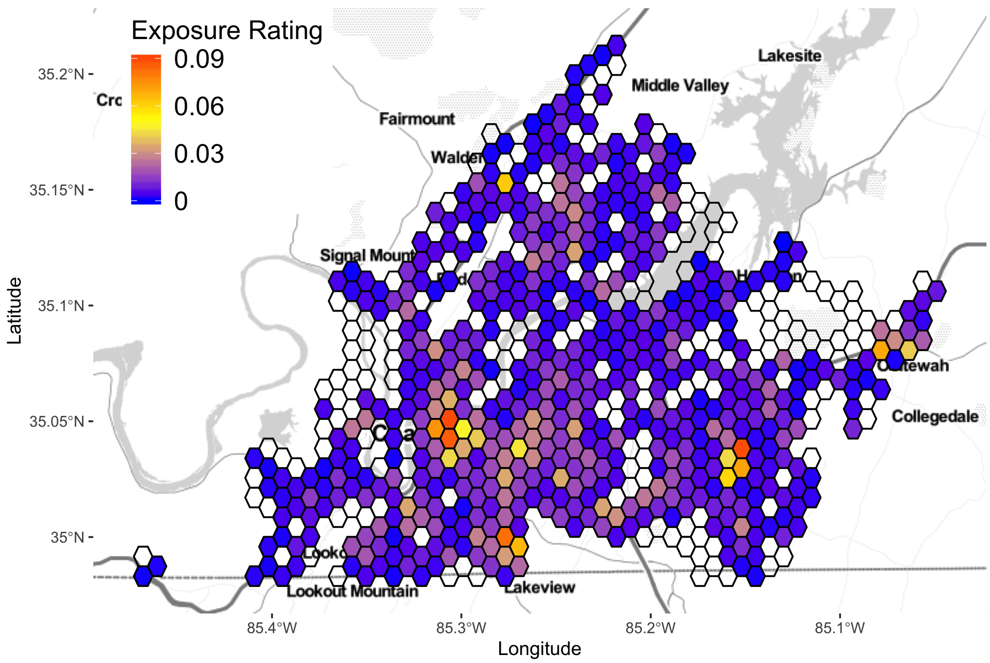
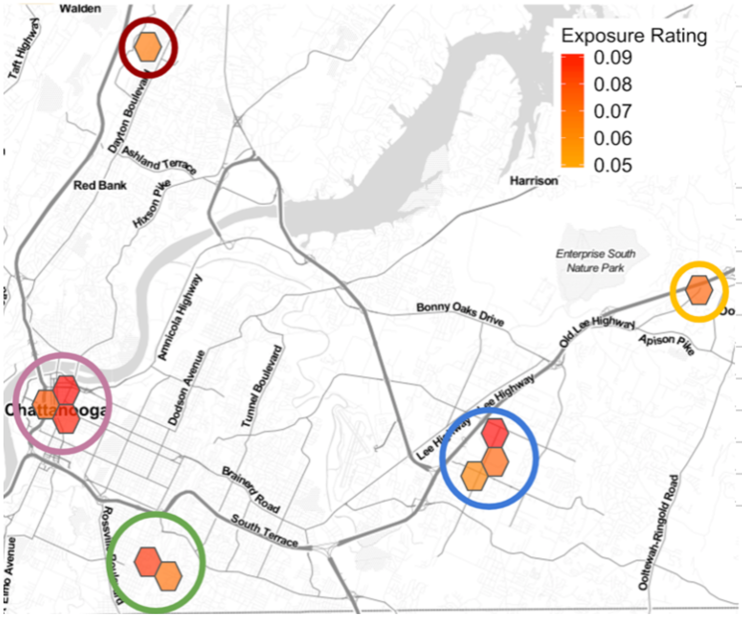


Figure 6.9 Standardized Risk Distribution in Chattanooga. The Exposure Rating variable was found by dividing the sum of accidents historically occurring within each hexagon block by the corresponding block's average of roadway AADT within. Many blocks demonstrated quite low Exposure Ratings, but particular problem areas can be immediately discerned. Not all areas had AADT available, and as such are shown solely by outlines. Most of omitted areas are residential, with very low accident counts.



Hex	Accident Count	AADT	Average Speed Limit	Exposure Rating	Cluster
282	1020	11420	30.4	0.0893	
233	1411	15872	37.5	0.0889	
246	558	6482	31.4	0.0861	
65	61	740	30	0.0824	
245	1283	17263	35.9	0.0743	
196	671	9485	37.5	0.0707	
423	328	4676	39	0.0701	
44	49	740	30	0.0662	
638	5	80	25	0.0625	
195	681	11378	32.5	0.0599	

Figure 6.10 Ten Highest Exposure Rating Hexagon Blocks, clustered by color. Note that three of the top five are all located in downtown Chattanooga, where many smaller streets are packed rather closely together. All of the worst ratings demonstrate lower average speed limits when compared to the amount of traffic passing through them each day.

CHAPTER 7

Temporal Exploration of Accidents

When considering accident occurrence, it is not enough to only consider the locations most affected, but also which times lead to the most occurrence. Therefore, this chapter explores temporal distribution throughout a range of divisions. The first exploration is by hour of the day, in Section 7.1. Next explored is monthly distribution, in Section 7.4. Finally, accidents will be examined in their yearly totals in Section 7.5.

7.1 Hourly Distribution

The first temporal exploration of accidents presented is the number of accidents per hour of the day. This distribution is shown by Figure 7.1, where one can observe the trends of an average day. Note that occurrences are relatively low in the overnight hours, before beginning to sharply increase at roughly five in the morning, coinciding with the morning rush. Around the eight A.M. hour, there is a slight decline into the nine A.M. hour. Then another crest begins, subsiding slightly at thirteen hundred hours. Then the largest peak of all begins, with the highest number of accidents being reported in at fifteen hundred hours at roughly seven thousand accidents reported across the study period. From fifteen hundred hours onward, there is a marked decline in the number of accidents as the night progressed.

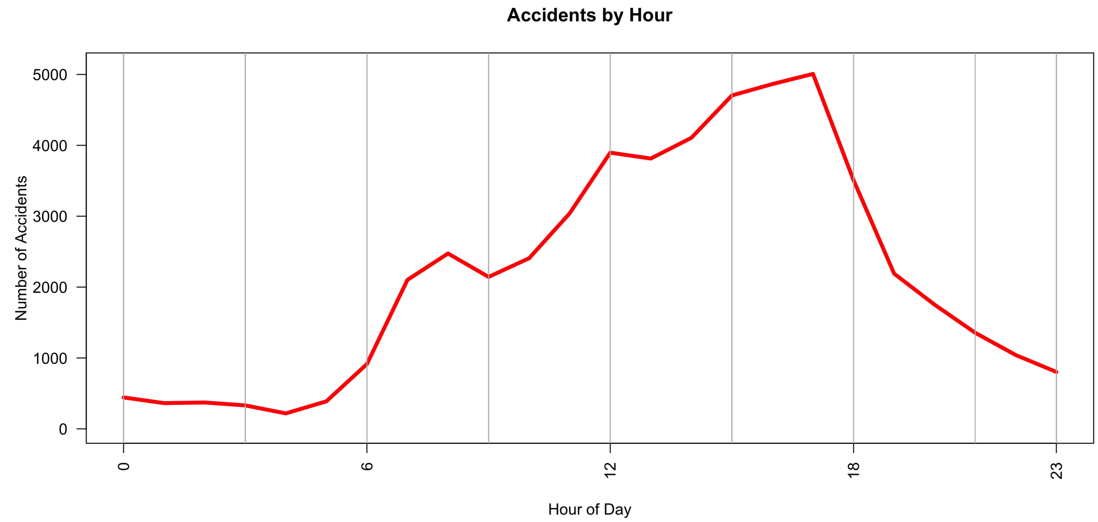


Figure 7.1 Base Hourly Breakdown of Accident Occurrence. Note that overnight hours have comparably low numbers when compared to daylight hours. Particular trends can be discerned throughout the day, within the morning and evening rushs clearly visible.

7.2 DayFrame Distribution

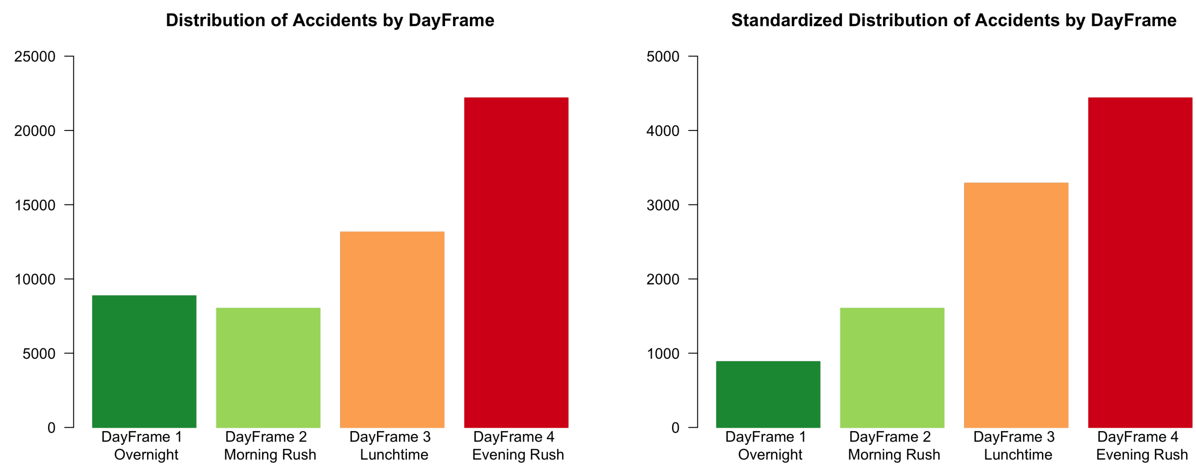
As was mentioned above, there are definite areas of time during the day that have specific trends of accidents. For example, from nineteen hundred hours across into 4 A.M. tends to be a very low occurrence time frame when compared to thirteen hundred hours to seventeen hundred hours. Therefore, it was decided to divide the hours of the day up into four distinct DayFrames for predictive purposes. This division is shown by Table 7.1. The distribution of accidents into the four DayFrames is shown by Figure 7.2, with Figure 7.2a showing total number of accidents occurring within each DayFrame. Note that although DayFrame 1 involves the largest number of hours, it is almost even with the number of accidents involved in DayFrame 2 (8867 in DayFrame 1 compared to 8022 in DayFrame 2), which has half the number of hours. Although DayFrame 3 only has four hours contained within it, it ranks second in terms of number of accidents at 13159. Finally, DayFrame 4 involves the most number of accidents, at 22191.

The uneven distribution of hours within each DayFrame does have the ability to skew the perceived total number of accidents within them. Therefore, Figure 7.2b was created. This figure demonstrates the number of accidents divided by the number of hours within each

DayFrame. This properly illustrates the division of accidents, with DayFrame 1 ranking solidly in last at 886.70 average accidents per hour, and DayFrame 2 at an average of 1604.4 per hour. This is just a bit under double the amount, despite DayFrame 2 having half the number of hours within it. DayFrame 3 and 4 retain their previous rankings at 3289.75 and 4438.2 average accidents per hour, respectively.

Table 7.1 DayFrame Breakdown

DayFrame	**Hours Covered**	**Description**	**Total Hours**
DayFrame 1	0 - 4 and 19 - 23	Overnight	10
DayFrame 2	5 - 9	Morning Rush	5
DayFrame 3	10 - 13	Lunch Hours	4
DayFrame 4	14 - 18	Evening Rush	5



(a) Accident Occurrence by DayFrame

(b) Standardized Accident Occurrence by DayFrame

Figure 7.2 Distribution of Accidents by DayFrame. Figure 7.2a reflects raw totals within this timeframes, with Figure 7.2b illustrating the totals divided by the number of hours within each DayFrame. Note that Table 7.1 reports the exact division of hours within DayFrames.

7.3 Week Distribution

The number of accidents that occur varies significantly based on the day of the week, as shown by Figure 7.3. Sunday begins the week with a very low number of accidents (7190), with occurrence increasing each day until Friday, the highest day for accident occurrence at 13,821 accidents reported. Saturday, albeit not as low in occurrence as Sunday, ends the week with relatively low numbers to the rest of the previous days.

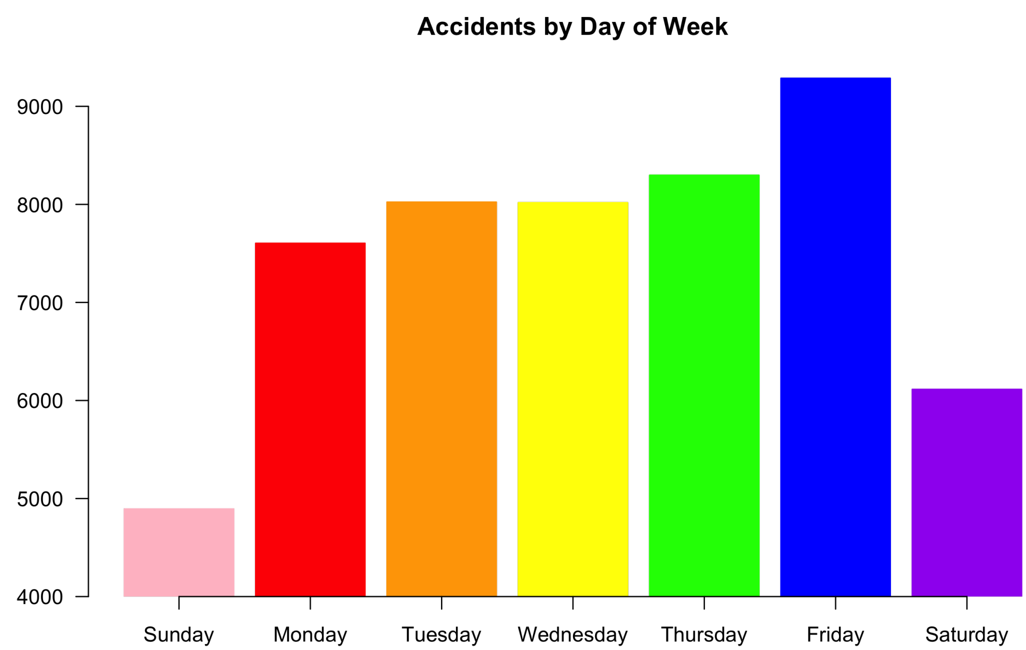
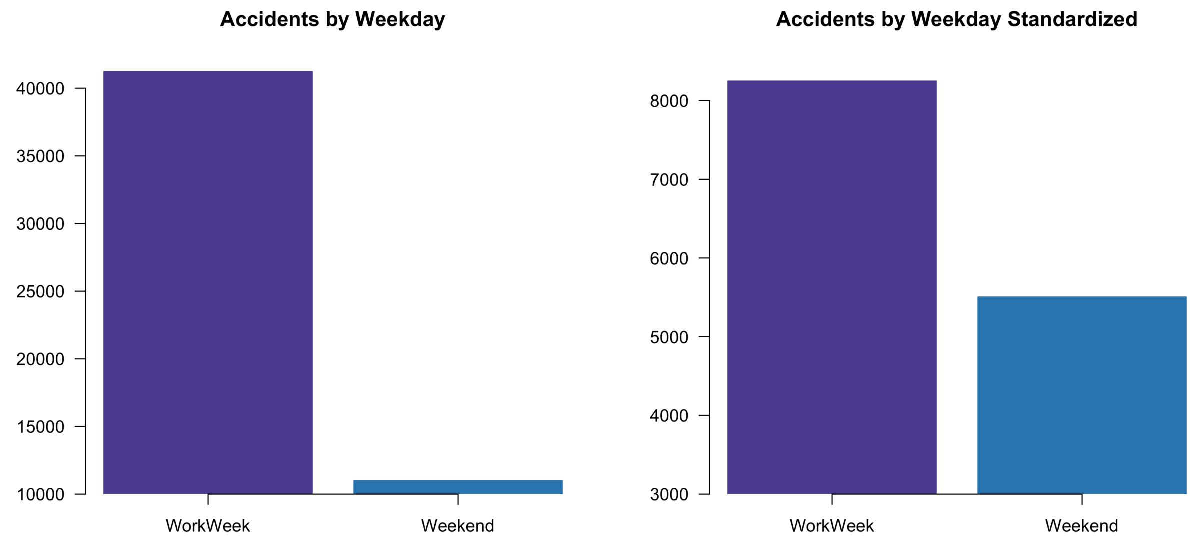


Figure 7.3 Accident Occurrence by Day of Week. Sunday begins with a very low number of accidents, with accidents steeply increasing on Monday. The highest accident count occurs on Friday, after which accidents are more than halved on Saturday.

Exploration of Figure 7.3 lead to the investigation of total number of accidents as divided into workweek and weekend. This investigation was supported by Figure 7.4. Figure 7.4a displays the total of accidents that occurred within the workweek as compared to weekend, with a tremendous difference. But, it is also important to explore the number of accidents across the days as an average. Therefore, Figure 7.4b displays the number of accidents divided by the number of days within each category (five for workweek, two for weekend). This leads to a different division, although with the workweek again taking the lead.



(a) Accident Occurrence by Weekday

(b) Standardized Accident Occurrence by Weekday

Figure 7.4 Distribution of Accidents by Weekday. Figure 7.4a demonstrates total accidents within the two timeframes, while Figure 7.4b shows the distribution of accidents divided by the number of days within each category. The Workweek includes five days (Monday through Friday) with Saturday and Sunday within the Weekend.

As shown by Figure 7.5, the traditional workweek days have a very particular trend, quite similar to the trend explored above in Section 7.1. However, the weekend displays quite a different trend. Accidents near midnight begin much higher than during the workweek and peak about 2:00, but then decline until roughly 4:00. Then, there is a very slow and gradual increase in number of accidents as the day progresses, with the peak occurring around eleven on Saturday, and 13:00 on Sunday. Then, there is a plateau remaining for several hours, with a gradual decline beginning at 18:00. Again, the weekend days have a greater total of accidents in the overnight hours when compared to the workweek.

7.4 Monthly Distribution

The sum of accidents also displays varying trends based on the month of the year. As shown by Figure 7.6, early months of the year tend to have lower occurrences of accidents with January and February containing 5133 and 5318 respectively. Accidents then increase

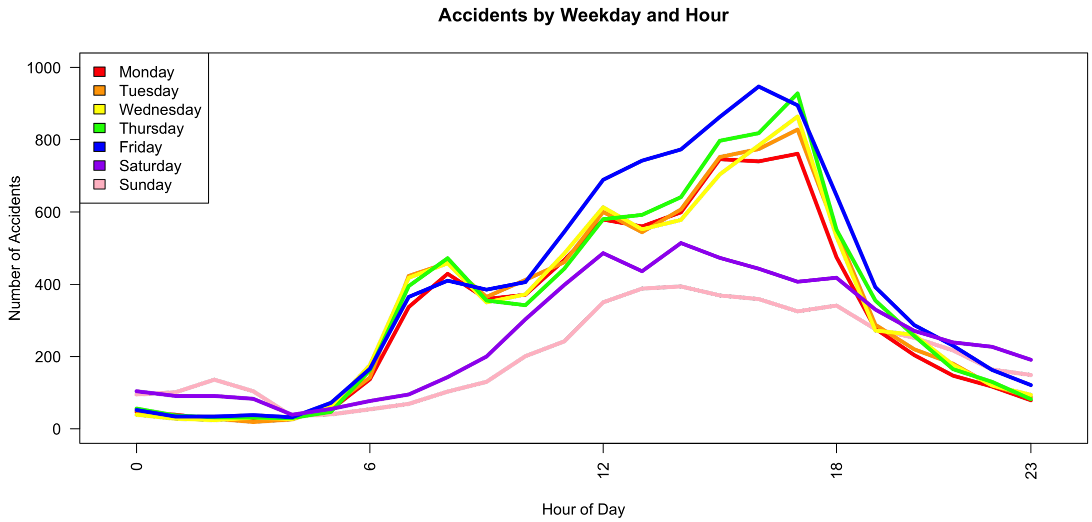


Figure 7.5 Hourly Breakdown of Accident Occurrence by Weekday. Workweek and Weekend accidents demonstrate two very different trends. Accident occurrence during the Workweek rises and falls in relation to traditional worktime commute hours, while weekend accidents only have one rough peak throughout the day.

from March to May, with another decrease into June and July. A second peak occurs in October and November, with almost equal accident sums at 6968 and 7144 (most from the year) accidents respectively. December finishes the year with a comparatively small 6329 accidents.

Although all the months exhibit the same general trend of hourly accidents one can observe, via Figure 7.7, particular months that exhibit higher numbers of accidents than others. For example, January and February have the lowest peaks at rush hour from all of the months, with May having the highest. December has a delayed peak of accidents at rush hour, with the highest count occurring at 18:00 instead of the usual 16:00. Interestingly, June features the highest number of accidents from the lunchtime hours, with November having the highest count for the earlier morning rush hours, while August has a later peak, but a greater number of accidents.

As explored earlier, different days of the week feature one of two trends of accident occurrence. This is also demonstrated when accidents are considered by the month and day of week, as shown in Figure 7.8. Once again, Sunday and Saturday have the lowest number of accidents across the board, although Sunday still ranks lower than Saturday. While Sunday

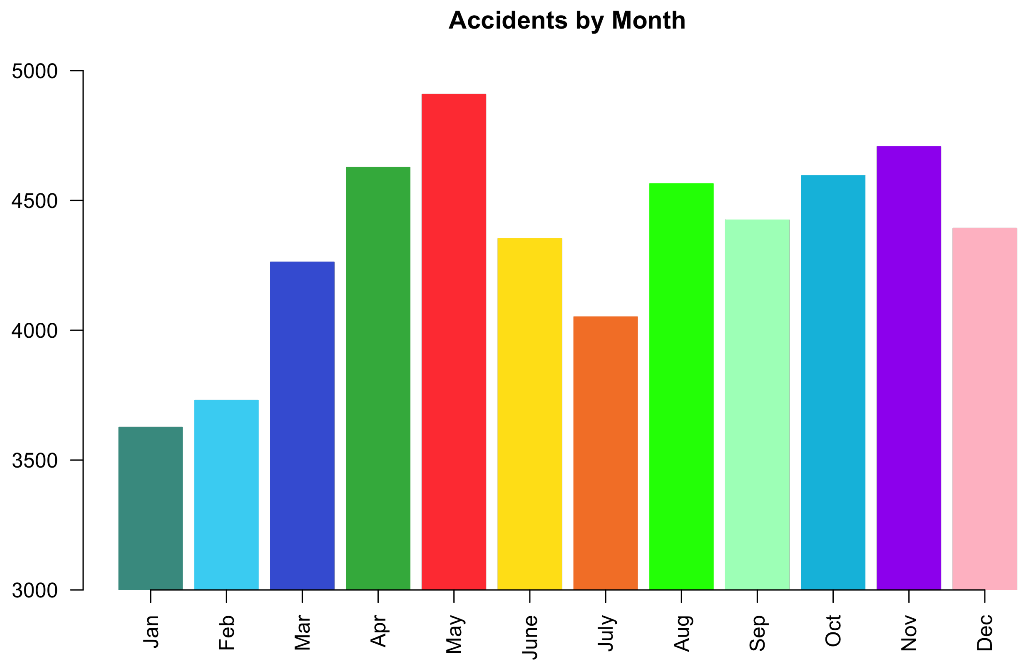


Figure 7.6 Accident Occurrence by Month from January 2017 through December 2019. The highest number of accidents occur in the month of May, followed by November. Of note is the low accident counts of July.

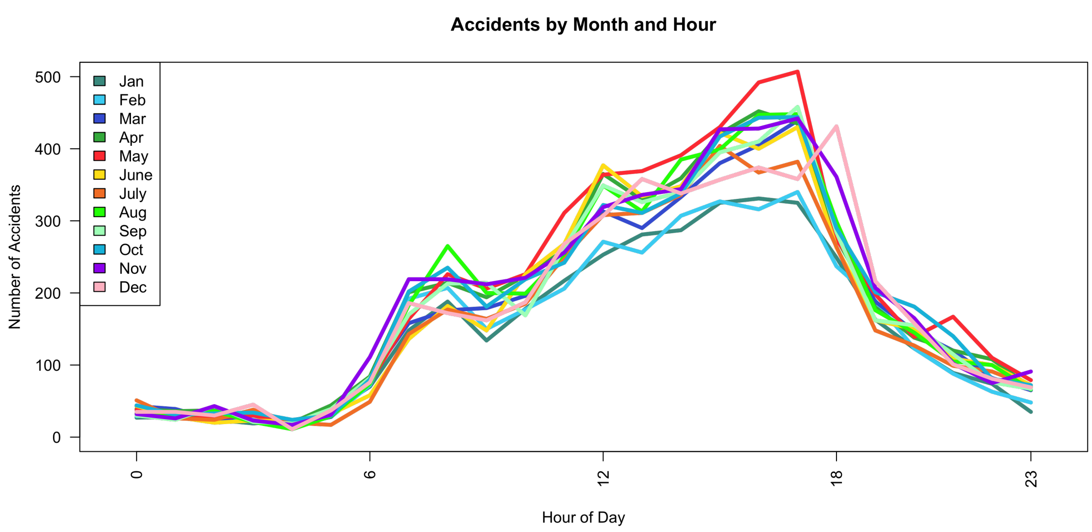


Figure 7.7 Hourly Breakdown of Accident Occurrence by Month. While the months roughly follow a similar trend, there is a unique late Evening Rush spike in December. December's Evening Rush accidents happen mostly within the 6PM hour, peaking above all other months at this time. This is converse to other months, whose peaks occur closer to 4PM or 5PM.

features a peak and valley within the summer months, Saturday gradually experiences a single slope in the same time frame, with the highest point in April and slowly decreasing into August. Despite Friday previously holding the highest overall accident occurrence, Thursday is the leader in August, as well as in being equal to Friday in May. Friday experiences the most accidents in November, lending reliability to the assumption that holidays (such as Thanksgiving and Black Friday) would contribute to accident counts. Interestingly, Tuesday has the most accidents in October.

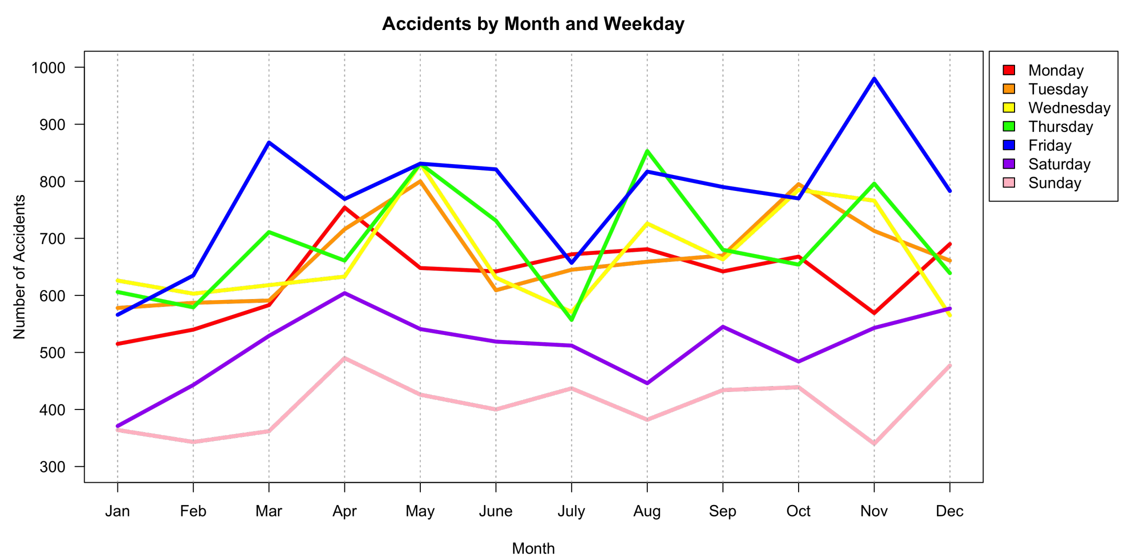


Figure 7.8 Monthly Breakdown of Accident Occurrence by Weekday. Across all months of the year, Saturday and Sunday experienced the lowest number of accidents, echoing the findings of Figure 7.5. The highest number of accidents occurred on Fridays, with two note-worthy peaks occurring in March and November.

7.5 Yearly Distribution

Another way to explore accident occurrence in the time of study is dividing the study into years. Three years (2017, 2018, and 2019) are currently covered by the project, with each having their own particular trends of accidents. For a basic understanding, Figure 7.9 was created. The study began with 17424 accidents in 2017, peaking with 17483 accidents in 2018, then dropped again in 2019 to 17332 accidents.

The hourly distribution of accidents was also observed for the years of the study, as shown

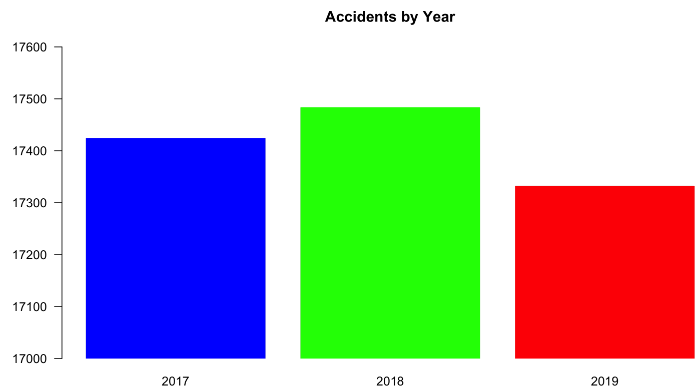


Figure 7.9 Total Accident Occurrence by Year. There was a significant increase in the number of accidents within 2018, but 2019 followed with a total dipping below both 2017 and 2018.

in Figure 7.10. Although all three years follow the general trends already discussed, it is interesting to see that the significant difference from 2018's total accidents is made up of slight increases across the mid-day hours. As shown, 2018 experienced a much higher number of accidents around noon, and continued to rank higher than both other years until the peak around 16:00. Note that 2018 also experienced higher accident counts at 22:00 as well.

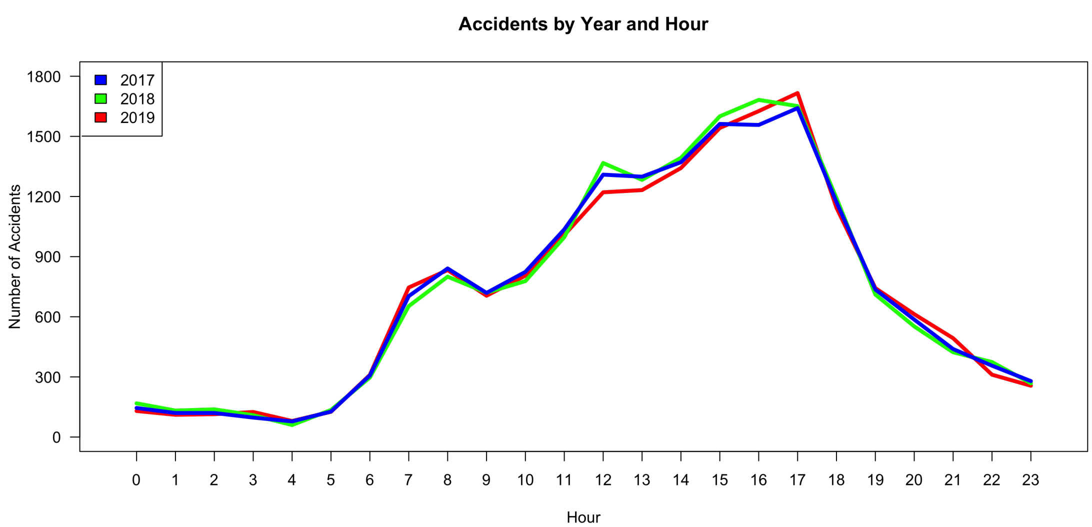


Figure 7.10 Hourly Breakdown of Accident Occurrence by Year. Note that there is more deviation between the three years within the Lunch time and Evening Rush (DayFrames 3 and 4) hours then at any other time of day. Overall, 2019 experienced less accidents than either other year, despite having a higher (and later) spike within DayFrame 4.

The most interesting time frame breakdown within this section was the distribution of

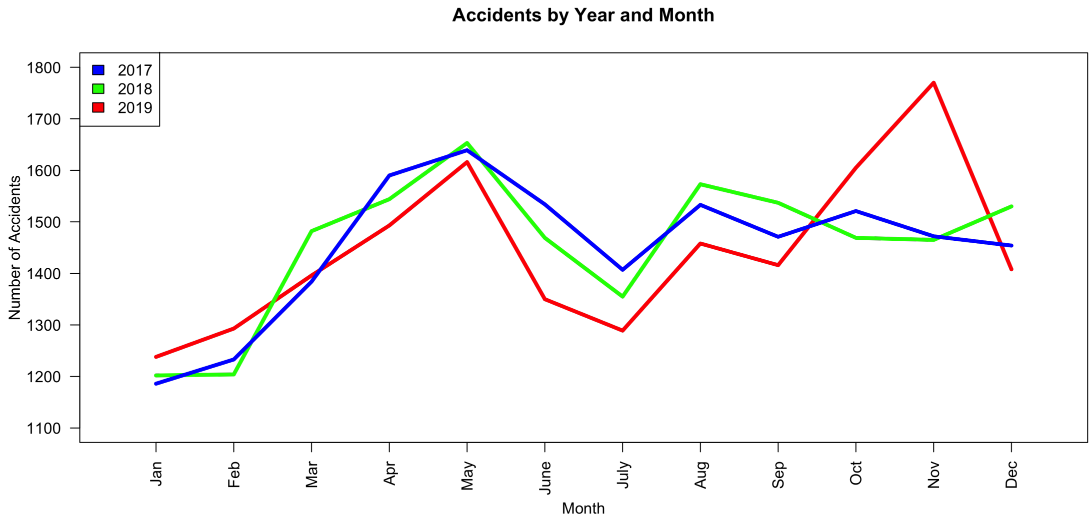


Figure 7.11 Monthly Breakdown of Accident Occurrence by Year. While all three years follow a loose pattern, November of 2019 demonstrated a deviation from this trend, reporting the highest number of accidents across the entire time of study. This reasons behind phenomenon have as of yet not been discovered.

accidents across months for all three years. Although 2018 ranked the highest amongst the three in total, January and February of 2018 ranked below 2019. March of 2018 broke above both other years after a slight dip in February. Then, the roles were swapped in April, where 2017 ranked the highest and 2019 the lowest. All three years experienced a similar number of accidents in May. The summer months saw a very similar dip in accidents for 2019 and 2018 (particularly in July), with 2017 seeing the least amount of decline amongst the three. For the rest of the year, 2017 and 2018 exhibited similar counts to one another, albeit with 2018 totals normally higher than 2017. However, 2019 did not follow this trend, instead peaking with a dramatic total of 1770 accidents in November, only to drop once again below both other years in December.

CHAPTER 8

Results

Within this Chapter, the results of the study are discussed. The Chapter begins within Section 8.1, where specifics of performance metrics are discussed. Section 8.2 reviews the results of the various fishnet layout models training and testing, while Section 8.3 does the same for the various hexagon layout models. Section 8.4 reviews how understanding of where and when accidents occur has been improved by this study. Section 8.5 concludes the Chapter by discussing the actual prediction power of each of the layouts, and then comparing them together.

8.1 Performance Evaluation Metrics

When considering the results of a rare event predictor, it is unwise to rate performance solely upon the Accuracy metric. Accuracy considers both the number of correct negative and correct positive events predicted. This skews the actual performance rating of the predictor where positive events are uncommon since negative predictions should vastly outweigh positive predictions. A more fitting performance metric for rare event predictors is the Recall. Recall is the measure of true positive predictions divided by the sum of true positive and false negative predictions. As such, recall is considered a measure of correctly identified positive events, and is vastly more useful in rare event prediction. Also of concern is the number of false positives, presented by the False Positive Rate (FPR). FPR is found by dividing the number of false positives by the sum of false positives and true negatives. A high FPR indicates the model is over predicting positive events, which in this particular study is not as problematic as it may sound. A False Positive in terms of accident prediction simply means that an accident was predicted, but none transpired. This is not as troublesome as a False Negative, where an accident occurred, but was not predicted. While a particularly high FPR rating would render a prediction useless, one must strike a balance between the Recall and FPR performance of a rare event predictor carefully and determine a threshold for an acceptable FPR.

Additionally, Specificity is the measure of correctly identified negative events. The equation for Specificity is the number of True Negatives divided by the sum of True Negatives and False Positives. A high Specificity score indicates a low number of False Positives, and as such a good predictor of negative occurrences. Therefore, when considered with the Recall score, it is beneficial to seek high Specificity and high Recall scores. When both scores rank highly, this indicates a model able to correctly identify both classes of events.

8.2 Fishnet Model Performance

The first grid block layout tested was the Fishnet grid, initially discussed in Section 6.3. This grid layout's final setup involved a total of 906 grid blocks spread across the southern half of Hamilton County. A final total of 59887 accidents were included in this layout. Interstate accidents were included in this grid's testing, and accounted for roughly 7% of the total number of records. The inclusion of these particular records led to the model learning that the blocks containing the interstate roadways were much more likely to experience accidents. While true, this did lead to under-prediction within other grid blocks and over-prediction for interstate locations.

8.2.1 Model Editions

As discussed in Section 4.2, both editions of the grid layout were tested using two different types of negative sampling. These two types once again are Grid Fix, and True Random.

The performance of the given layout's models fluctuated drastically according to which type of negative was utilized, as well as the ratio of negative to positive records. Both versions of the fishnet models will be discussed here, with their respective ratio cuts.

In general, the models from the fishnet model performed fairly well in terms of accuracy while in training and testing, with specific metrics for all versions shown in Table 8.1. Note that GF is shorthand for Grid Fix, and TR short for True Random. These abbreviations are used simply for space considerations within tables.

It is of note that the True Random model editions performed better in Recall than the Grid Fix editions, with the lowest score for True Randoms at 61% when compared to the highest Grid Fix at 68%. This is also reflected in the raw number of True Positives for the Grid Fix editions, with each of the Grid Fix models totaling much less than their True Random counterparts. The best performing model overall would have to be True Random with a 50-50 split, with an 81% Recall. Although True Random with no split ranked higher in Recall, there was a significant drop in Precision, or the number of True Positives compared to False Negatives. Such a low Precision indicates a vast amount of false positives, leading to a model that over predicts accident occurrences.

Table 8.1 Results from Training and Testing for Fishnet Models

Model Type	**Train Acc.**	**Train Loss**	**Test Acc.**	**Test Loss**	**AUC**	**TN**	**FP**	**FN**	**TP**	**Precision**	**Recall**	**Specificity**	**FPR**
GF 50-50	75.08	0.17	75.18	0.17	0.83	11173	2443	4217	9003	0.79	0.68	0.82	0.18
GF 75-25	85.13	0.11	84.96	0.11	0.86	39463	1171	6939	6367	0.84	0.48	0.97	0.03
GF No Split	94.62	0.05	94.62	0.05	0.84	160690	1966	7499	5751	0.75	0.43	0.99	0.01
TR 50-50	81.58	0.13	81.36	0.13	0.89	11344	2528	2507	10640	0.81	0.81	0.82	0.18
TR 75-25	85.48	0.11	85.44	0.11	0.90	35523	2305	5129	8098	0.78	0.61	0.94	0.06
TR No Split	67.97	0.04	68.01	0.21	0.84	100992	50147	2428	10800	0.18	0.82	0.67	0.33

The Recall scores of each of the six model editions is shown by Figure 8.3a throughout the training/testing process. Of interest is the recall score of the True Random model with the original split, which begin at quite a bad Recall and finished with the best from all six tests. This is also despite having the lowest Accuracy from all shown models. The True Random 50-50 split is still considered a better model, since the Recall began at a high score and remained there.

8.3 Hex Model Performance

The most recent grid layout was the Hexagon grid, previously introduced in Section 6.4. This grid layout featured 694 total hexagons, all contained within the city limits of Chattanooga, Tennessee. Interstate accident records are not included within this layout, since city officers do not patrol said areas. This lead to a reduced total number of accidents, with 52239 accidents considered here. This lead to a reduced number of positive records within the models and a lower Recall score, but did not necessarily lead to reduced performance when actual predictions were attempted.

8.3.1 Model Editions

The performance of the hexagon layout's models fluctuated even more drastically than the fishnet version based upon which type of negative was utilized, particularly when comparing ratio cuts. Both versions of the hexagon models will be discussed, along with their respective ratios. The particular statistics of the Hex grid models are shown in Table 8.2. Again, GF is used as shorthand for Grid Fix, and TR for True Random. It is interesting to note that the Recall scores for the Hex models are not as high ranking as those within the Fishnet models, but their actual prediction models perform quite better. However, the predictions from the models will be discussed later within this Chapter.

Table 8.2 Results from Training and Testing for Hexagon Models

Model Type	**Train Acc.**	**Train Loss**	**Test Acc.**	**Test Loss**	**AUC**	**TN**	**FP**	**FN**	**TP**	**Precision**	**Recall**	**Specificity**	**FPR**
GF 50-50	71.90	0.18	71.55	0.18	0.72	11620	5079	4110	11494	0.69	0.74	0.70	0.30
GF 75-25	81.60	0.14	81.79	0.13	0.62	49962	432	11563	3916	0.90	0.25	0.99	0.01
GF No Split	89.33	0.09	89.34	0.09	0.82	100624	110	12276	3218	0.97	0.21	0.82	0.18
TR 50-50	82.23	0.13	81.85	0.13	0.82	13115	2965	2735	12598	0.81	0.82	0.82	0.18
TR 75-25	86.77	0.09	86.82	0.09	0.77	52475	3219	6162	9294	0.74	0.60	0.94	0.06
TR No Split	91.38	0.06	91.20	0.06	0.70	109239	1890	9269	6383	0.77	0.41	0.98	0.02

The hex grid model recall scores throughout training history is shown by Figure 8.3b, where one can observe True Random 50-50 split scoring the highest Recall score throughout the entire training and testing process. Recall was especially bad with both Grid Fix 75-25 and original split, but 50-50 split scored almost as high as its True Random counterpart.

8.4 Spatial and Temporal Understandings

Overall, both model versions were able to properly understand when accidents occur across the area of study. All versions of models were able to discern that DayFrame 4 received more accidents than others, with DayFrame 3 following behind it. Fishnet models understood that the presence of the interstate would lead to higher numbers of accidents, and the Hex models understood the impact of areas with high ER scores, as discussed in Section 6.4. Although some models did over-estimate the number of accidents that occurred, models with an even split of accidents and negative samples tended to perform the best, often receiving the same levels of accuracy as their more rare counterparts. This was especially true with the Hex grid layouts, where both Grid Fix and True Random 50-50 splits had better Recall scores than their 75-25 equivalents.

8.5 Accident Prediction Capabilities

Both grid layouts were employed in the pursuit of accident prediction. Prediction capability as demonstrated by the Fishnet grid layout will be discussed in Section 8.5.1, with an example prediction run being discussed, as well as the average performance of the fishnet models. Hex grid layout performance will be discussed in Section 8.5.1, where an example prediction run will be examined, as well as the average performance of the hex models. Finally, comparisons between the two will be discussed within Section 8.5.3.

8.5.1 Fishnet Model Predictions

In the case of the Fishnet layout, areas of prediction were considered where the probability output from the network was at least 50%. Prediction areas were found by supplying the model with a record for each grid block and DayFrame combination, to further simplify the model's architecture. Ten dates were tested against the model to gauge ability to properly predict where and when accidents occurred. These tests led to Table 8.3, which illustrates each models abilities to discern accident occurrence within a grid block/DayFrame identifier. That is, if an accident occurred within grid block 34 at 9:35 in the morning and the model predicted an accident within grid block 34 in DayFrame 2, a True Positive would be recorded. This setup allows for a simplified confusion matrix compared to attempting to predict for each given hour and grid block combination. Although Accuracy and Specificity are quite high for a number of the listed models, the Recall displayed is quite unsatisfactory. Once again, Recall is considered the most important indicator of predictive power when considering vehicular accidents, since this is where an accident would be predicted, and an accident indeed occurred. The best Recall demonstrated here would be the even split edition of Grid Fix testing, with a value of only 50%.

Table 8.3 Averages across Ten Random Dates for Fishnet Model Predictions

Model Type	**TP**	**TN**	**FP**	**FN**	**Accuracy**	**Precision**	**Recall**	**FPR**	**Specificity**
Grid Fix No Split	4.67	3060	198.33	37	0.93	0.02	0.13	0.06	0.94
Grid Fix 50-50	20.86	2061.14	1198.43	19.57	0.63	0.02	0.5	0.37	0.63
Grid Fix 75-25	9	2685.33	574.67	31	0.82	0.01	0.24	0.18	0.82
True Random No Split	3.67	3189.67	73	33.67	0.97	0.05	0.12	0.02	0.98
True Random 50-50	18.29	2201.57	1058	22.14	0.67	0.02	0.44	0.32	0.68
True Random 75-25	12	2415.86	843.71	28.43	0.74	0.02	0.33	0.26	0.74

An example run of the fishnet model predictions is shown by Figure 8.1, where the date predicted for would be August 16, 2018 within the evening rush. As mentioned, the model predicted the highest probabilities of accident occurrence along the interstate areas. For this particular prediction, true negatives numbered 802, with true positives at only 5. False negatives totalled 19, and false positives were 80. Accuracy was 89% on this test, with a Precision value of 6%. Recall was quite low, at 21%. The False Positive Rate (FPR) was quite good, at 9%. Specificity was also quite good, at 91%. Overall however, this particular prediction demonstrated a high number of False Positives, with quite a high number of False Negatives. As such, its performance was considered unsatisfactory.

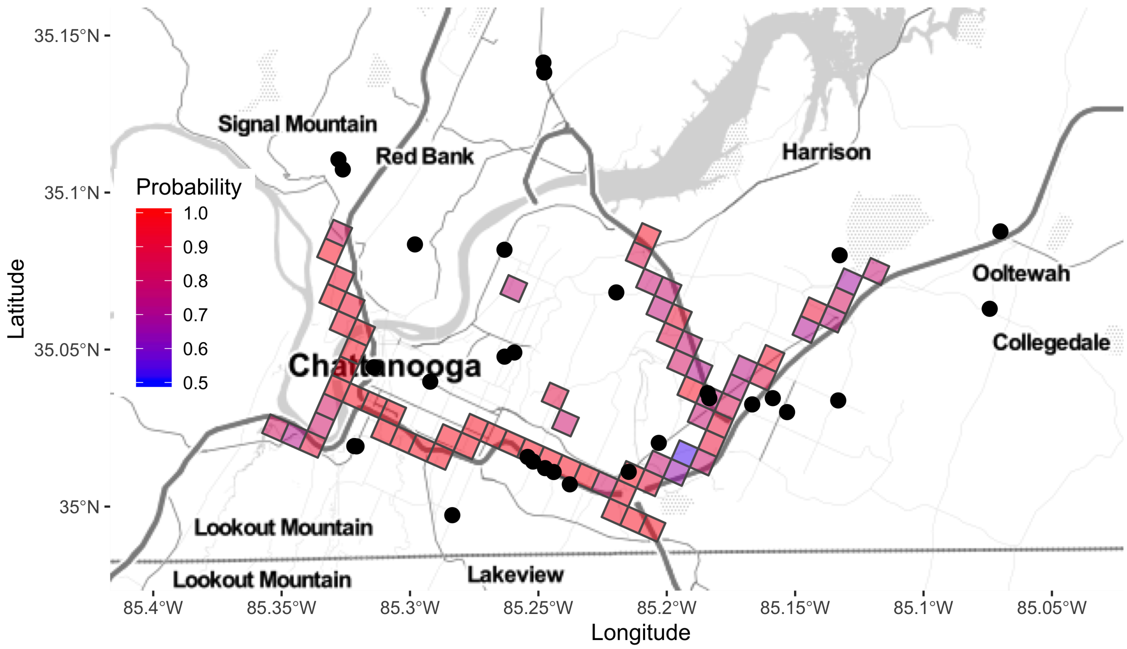


Figure 8.1 Fishnet - Grid Fix No Split Model Prediction. Data from before August 16, 2018 (the first report of which was from January 2017) was utilized to create a predictive model. This particular prediction was for the Evening Rush hours (DayFrame 4), selected due to the higher number of accidents traditionally present in this time-frame.

8.5.2 Hexagonal Model Predictions

Initial testing of the hexagonal grid prediction power was completed on the same ten dates as the fishnet layout, in order to compare the two directly. The results from these tests are shown in Table 8.4. The best performing model in Accuracy was the True Random 50-50 split, which also ranked the highest in Specificity and Precision. False Negatives were lowest when utilizing the Grid Fix model with the original number of negatives. This particular model also scored the highest Recall, with 72%.

Table 8.4 Averages across Ten Random Dates for Hex Model Predictions

Model Type	**TP**	**TN**	**FP**	**FN**	**Accuracy**	**Precision**	**Recall**	**FPR**	**Specificity**
Grid Fix 50-50	23.1	952.9	1639.8	12.2	0.37	0.01	0.67	0.63	0.37
Grid Fix 75-25	21	1001.1	1591.6	14.3	0.39	0.01	0.6	0.61	0.39
Grid Fix No Split	25.4	901.6	1691.1	9.9	0.35	0.02	0.72	0.65	0.35
True Random 50-50	21.4	1834.5	758.2	13.9	0.71	0.03	0.59	0.29	0.71
True Random 75-25	14.3	1650.1	942.6	21	0.63	0.02	0.43	0.36	0.64
True Random No Split	15	1395.3	1197.4	20.3	0.54	0.01	0.46	0.46	0.54

For actual future predictions completed by the hex grid layout, a shifting threshold was assigned to each model and DayFrame combination. This was to assist in reducing general noise in predictions, as well as simplifying model output. This shifting threshold system certainly fulfilled this task, as shown in Figure 8.2. The threshold for this particular model and DayFrame combination was .85, where any block with 85% probability or above was considered.

Also displayed is the historic risk of the accident location, shown by arrows and squares. Downward pointing arrows indicate hex grid blocks with low historic risk, or less than 25 total accidents within the time of study. Squares indicate areas of medium risk, or blocks that experienced 25 to 54 accidents within time of study. Finally, upward pointing arrows demonstrate areas of high historic risk, or where 55 accidents or more had occurred during time of study.

This particular run of the predictive model had 596 True Negatives, and 29 True Positives. False Negatives numbered 29, with False Positives equalling 40. These values resulted in an accuracy of 90%, with a Precision of 42.1%. Specificity was calculated at 93.7%, and the FPR was only 6.3%. Recall for this predictive run was 50%.

8.5.3 Comparison of Prediction Methods

The Recall scores for all models previously discussed are shown by Figure 8.3. In both fishnet and hexagonal layouts the best Recall was demonstrated by the True Random 50-50 split, although within the Fishnet methods the True Random with the original data split managed to improve drastically over the history of model training. For both editions, the Grid Fix 75-25 and original split models under-performed in Recall, beginning at 0% and improving to only roughly 20%.

Between the two predictive forecasts outlined above, certain observations can be made, as shown by Table 8.5. While the hex layout did employ the use of shifting thresholds, the accuracy at 85% probability and above ranked higher than that of the fishnet layout at simply 50%. Precision was higher on the hex edition by 36.1%, a significant margin. Specificity

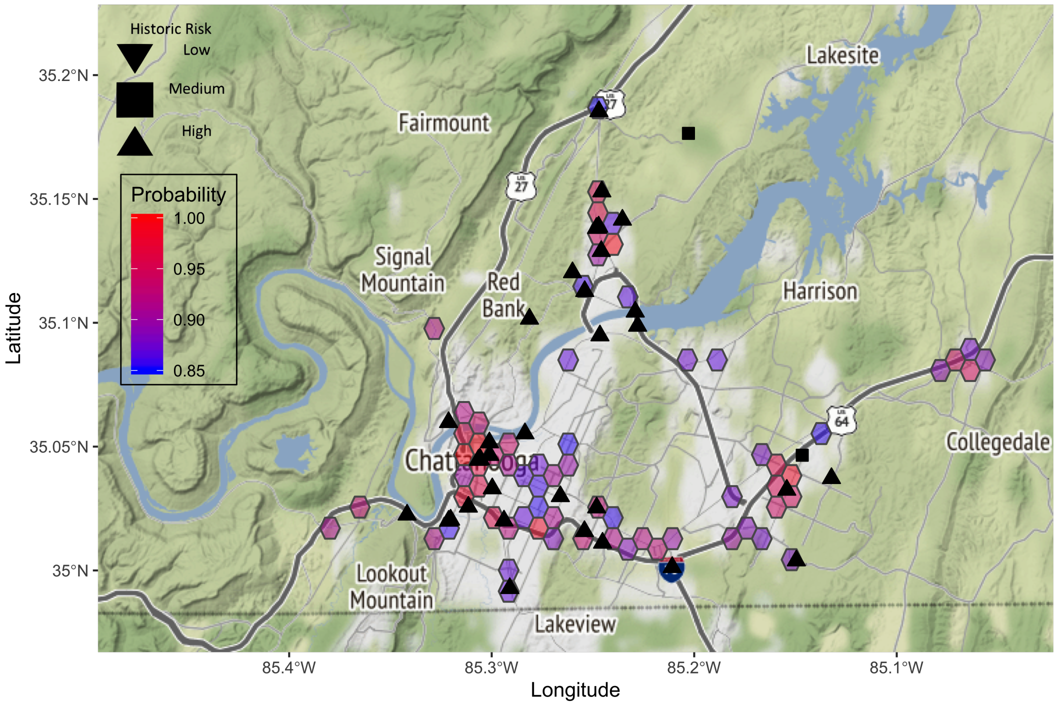
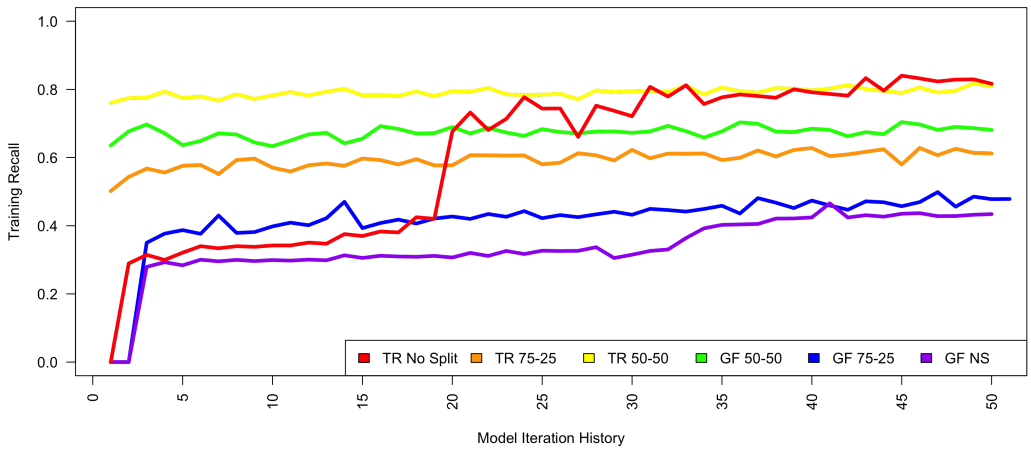


Figure 8.2 Hex Layout- True Random 50-50 Split Model Prediction. Data from January 2017 to December 2019 was utilized in creating the predictive model, and a random future date was selected outside of this time-frame to truly test the predictive abilities of the model. (January 23, 2020 Evening Rush)

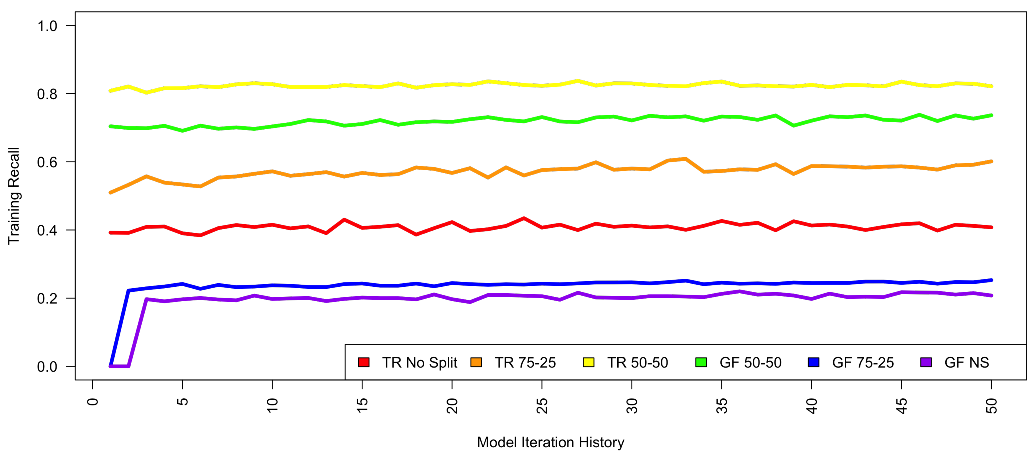
from the hex layout was also superior, at 93.7% compared to the fishnet layout specificity score of 91%. The false positive rate of the fishnet grid was 6.3%, compared to 9% with the fishnet. Recall, as mentioned above, was quite low on the fishnet layout, but did improve with hex. Overall, even though the hex layout predicted ten more false negatives than the fishnet layout it performed better overall, predicting 29 true positives when compared to only five predicted by the fishnet. The hex layout also halved the amount of false positive, partially due to the usage of shifting thresholds.

Table 8.5 Predictive Model Results from Fishnet and Hexagon Layouts

	TP	***FN***	***FP***	***TN***	***Accuracy***	***Precision***	***Recall***	***FPR***	***Specificity***
Fishnet	5	19	80	802	89	.06	.21	.09	91
Hexagon	29	29	40	596	91.7	.421	.5	.063	93.7



(a) Fishnet Grid Recall Scores



(b) Hex Grid Recall Scores

Figure 8.3 Comparison of Recall Scores from Both Grid Layouts. Note that the rare event (90-10) split of data demonstrated much lower Recall across both layouts, when compared to either the 75-25 or 50-50 split data. In both layouts, it was determined that the best training/testing was found in the True Random model, with a 50-50 split of accidents and negative samples.

CHAPTER 9

Conclusions

9.1 Understanding and Reduction of Accidents

The introduction to an accident prediction application was presented here, designed for use by the local emergency service district. This application would provide first responders with informed predictions of where and when accidents would occur on a given day. Better distribution of resources could be achieved, placing ambulances and patrol units closer to predicted potential areas of need. Of course, any final resource allotment would be considered by the end users, our emergency services district. Solutions could include shifting or construction of emergency medical technician stations closer to the areas of consistent concern. Ambulance routing could be optimized to avoid the particular areas of concern on any given day based upon the weather and historical risk. Temporary speed deterrent systems could be utilized to discourage aggressive driving in high risk locations. Infrastructure changes could also be implemented if a need exists.

This final benefit has in fact already been employed, concerning a telephone pole in a residential North Chattanooga. The area in question presented a high number (88 total) of accidents within its hexagon block, with 23 accidents along a certain roadway. The roadway in question is presented by Figure 9.1. In researching the particular cluster of accidents, it was discovered that a single telephone pole was positioned dangerously close to the road. Many of the accidents at this location involved vehicle rollover, those occurrence often indicates a higher level of injury [37]. Our team recommended the telephone pole's placement be adjusted, further back from the roadway. No accidents have been noted along that stretch of road since June 21, 2019, before which accidents had occurred roughly very few weeks.

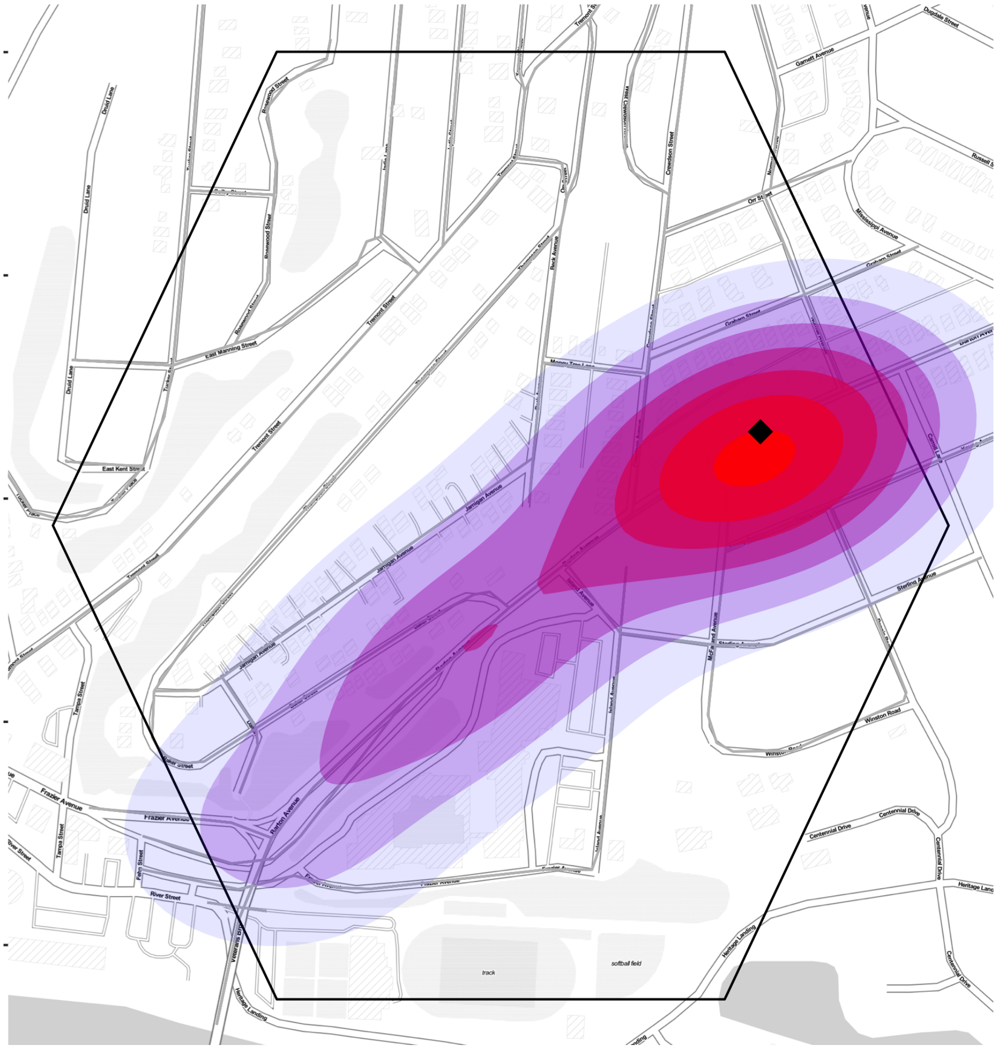


Figure 9.1 Hexagon Block featuring a clustered accident pattern. The diamond symbol indicates the position of a telephone pole positioned far too close to the roadway. 23 accidents occurred along the roadway in question, with the concentration of such shown by the intensity of color. It was recommended the pole be re-positioned, and no accidents have occurred since June 21, 2019.

9.2 Final Thoughts

Within this thesis, a project was discussed concerned with the understanding and reduction of accidents within the Hamilton County and Chattanooga areas. Although this project is not as-of-yet fully completed, the groundwork has been firmly laid and plans for completion have been compiled. Furthermore, present results indicate a favorable end product for the emergency district responder end users.

9.3 Future Research

The presented project's goals must of course include the conclusion of development for the proposed accident prediction application, whether that be in the form presented here, or another. Short-term goals include research into the injury levels of accidents reported, as well as a further investigation into interstate accidents. Long-term goals currently include the completed release of the intended application, as well as oncoming upkeep to the application whether that be through direct interaction or automated.

9.3.1 Short-term Research Goals

Additional research is in store for the injury status of accidents in the study area, since the level of injury included in initial reports has not been studied in depth as of yet. This could consist of quick analysis in temporal and spatial distribution, or something akin to previous in depth studies as reviewed in Chapter 3.

Also of interest would be the study of interstate and highway accidents. Although they only represent roughly 7% of the reports, their concentrations are much more intense (Revisit Figure 6.3 or 6.2 for details), clearly portraying specific areas of concern. These areas could be explored further to determine what commonalities they hold, if any. Of course, the interstate and highway accidents are not necessarily possible areas for the prevention of accidents for the local police department, since they do not patrol those roadways.

9.3.2 Long-term Research Goals

Long term goals for this project would include the completed release of the proposed application for use by the emergency district, determining the threshold for accident predictions for the application, as well as further training/testing iterations of models to account for any shifting trends of accidents.

1 The release of the proposed application would of course involve upkeep of whatever form the release consists of (i.e. web application, installed program, etc). Of additional concern long term is the timing of the prediction reports. The predictions could be assessed each night for the following day, or could be continuously available for update, provided that the end user updates the application (a bit like a refresh for web pages) or just on a revolving basis (every hour, half hour, etc). This second format would involve a bit more in depth programming, but would provide a more real-time prediction, as weather could be fetched nearly instantaneously and as such would be as accurate as possible.

2 Another point to consider would be a High Alert indicator for users of the application if the forecasting would be done in the revolving basis mentioned above. If a particular grid block presented a probability of an accidents above a given threshold, the user monitoring that area would be given an alert via the application. The given threshold would need to be tested for determination for this usage so that false alarms would be minimized, but this could assist the end users to be in an area preferably before an accident would occur.

3 Periodically, the model back-end of the application would need to be assessed and re-trained, to determine if accident hot spot trends had shifted or flared in additional areas. This would involve the collection of accident records, weather, and roadway information up to the date of reassessment to retrain the model using the most recent incidents. This would allow for the model to understand trends in a timely fashion, and to remain relevant to the purpose at hand.

REFERENCES

[1] "10 Leading Causes of Death by Age Group, United States," 2017. [Online]. Available: https://www.cdc.gov/injury/wisqars/pdf/leading_causes_of_death_by_age_group_2017-508.pdf

[2] M. R. Henry Olaisen, PhD, P. Lauren M. Rossen, P. Margaret Warner, and P. Robert N. Anderson, "Unintentional Injury Death Rates in Rural and Urban Areas: United States, 1999 to 2017," National Center for Health Statistics, Tech. Rep. 343, July 2019, nCHS Data Brief. [Online]. Available: https://www.cdc.gov/nchs/data/databriefs/db343-h.pdf

[3] K. Kochanek, S. Murphy, J. Xu, and E. Arias, "Deaths: Final Data for 2017," *National Vital Statistics Reports*, vol. 68, no. 9, June 2019. [Online]. Available: https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06.pdf

[4] NHTSA, "Traffic Deaths, 2009-2018," United States Department of Transportation, Tech. Rep., 2018. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812749

[5] TITAN, "Tennessee Traffic Crashes by Year, Time of Day and County, 2007-2019," Tennessee Department of Transportation, Tech. Rep., 2019. [Online]. Available: https://www.tn.gov/content/dam/tn/safety/documents/TimeOfDay.pdf

[6] "Guide to Calculating Costs," Online. [Online]. Available: https:/injuryfacts.nsc.org/all-injuries/costs/guide-to-calculating-costs/data-details/

[7] "Hamilton County Quick Facts," Online. [Online]. Available: https:/www.census.gov/quickfacts/fact/table/US/PST045219

[8] O. Dictionary, *The Oxford Dictionary of Phrase and Fable*. Oxford University Press, 2020. [Online]. Available: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960

[9] F. Rosenblatt, "The Perceptron: A Probabilistic Model For Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, no. 6, 1958.

[10] "Planning Glossary," United States Department of Transportation, Federal Highway Administration, October 2017. [Online]. Available: https://www.fhwa.dot.gov/planning/glossary/

[11] J. F. Garvey, *Functional Classification*. Federal Highway Administration, 2017, ch. 3. [Online]. Available: https://www.fhwa.dot.gov/environment/publications/flexibility/flexibility.pdf

[12] FHWA, "How Do Weather Events Impact Roads?" United States Department of Transportation, Tech. Rep., 2018. [Online]. Available: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

[13] W. H. West, "Alarm valve for pneumatic tires," U.S. Patent 2 166 384, 1937. [Online]. Available: https://patentimages.storage.googleapis.com/f9/0a/16/4d299ffc8f73d1/US2166384.pdf

[14] K. R. Agent and R. C. Deen, "Relationships between Roadway Geometrics and Accidents," *Transportation Research Board*, November 1974. [Online]. Available: https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=2431\&context=ktc_researchreports

[15] J. Milton and F. Mannering, "The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies," *Transportation*, vol. 25, pp. 395,413, 1998. [Online]. Available: https://link.springer.com/article/10.1023\%2FA\%3A1005095725001#citeas

[16] E. C. Davis, E. Green, N. Stamatiadis, G. A. Winchester, R. R. Souleyrette, and J. Pigman, "Highway Safety Manual Methodologies and Benefit-Cost Analysis in Program-Level Segment Selection and Prioritization," US DOT, Tech. Rep., 2015. [Online]. Available: https://stc.utk.edu/wp-content/uploads/sites/43/2018/01/MRI-1-UK-Year-1-Report-HSM-BC.pdf

[17] N. K. ChikkaKrishna, M. Parida, and S. S. Jain, "Calibration of safety performance function for crashes on inter-city four lane highways in India," *Cogent Engineering*, vol. 2, no. 1, 2015. [Online]. Available: http://doi.org/10.1080/23311916.2015.1031929

[18] S.-P. Miaou and H. Lum, "Modeling vehicle accidents and highway geometric design relationships," *Accident Analysis and Prevention*, vol. 25, no. 6, pp. 689 – 709, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000145759390034T

[19] D. K. Brijs, Tom and G. Wets, "Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model," *Accident Analysis and Prevention*, vol. 40, no. 3, pp. 1180–90, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0001457508000067

[20] J. D. Tamerius, X. Zhou, R. Mantilla, and T. Greenfield-Huitt, "Precipitation Effects on Motor Vehicle Crashes Vary by Space, Time, and Environmental Conditions," *Weather, Climate, and Society*, vol. 8, no. 4, p. 399 to 407, 2016. [Online]. Available: https://doi.org/10.1175/WCAS-D-16-0009.1

[21] L. Qiu and W. A. Nixon, "Effects of Adverse Weather on Traffic Crashes: Systematic Review and Meta-Analysis," *Transportation Research Record*, vol. 2055, no. 1, pp. 139–146, 2008. [Online]. Available: https://doi.org/10.3141/2055-16

[22] A. J. Khattak, P. Kantor, and F. M. Council, "Role of Adverse Weather in Key Crash Types on Limited-Access: Roadways Implications for Advanced Weather Systems," *Transportation Research Record*, vol. 1621, no. 1, pp. 10–19, 1998. [Online]. Available: https://doi.org/10.3141/1621-02

[23] M. Abdel-Aty, K. Choi, A.-A. Ekram, and H. Huang, "A study on crashes related to visibility obstruction due to fog and smoke," *Accident Analysis and Prevention*, vol. 43, no. 5, pp. 1730 – 1737, 2011.

[24] A. Theofilatos and G. Yannis, "A Review of the Effect of Traffic and Weather Characteristics on Road Safety," *Accident Analysis and Prevention*, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0001457514001900?via\%3Dihub

[25] V. Shankar, F. Mannering, and W. Barfield, "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies," *Accident Analysis and Prevention*, vol. 27, no. 3, pp. 371 – 389, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000145759400078Z

[26] L. Fridstrøm, J. Ifver, S. Ingebrigtsen, R. Kulmala, and L. K. Thomsen, "Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts," *Accident Analysis and Prevention*, vol. 27, no. 1, pp. 1 – 20, 1995. [Online]. Available: KCDC

[27] M. Poch and F. Mannering, "Negative Binomial Analysis of Intersection-Accident Frequencies," *Journal of Transportation Engineering*, vol. 122, no. 2, pp. 105–113, 1996. [Online]. Available: https://ascelibrary.org/doi/pdf/10.1061/\%28ASCE\%290733-947X\%281996\%29122\%3A2\%28105\%29

[28] S. D. Young, W. Wang, and B. Chakravarthy, "Crowdsourced Traffic Data as an Emerging Tool to Monitor Car Crashes," *JAMA Surgery*, vol. 154, no. 8, pp. 777–778, August 2019. [Online]. Available: https://doi.org/10.1001/jamasurg.2019.1167

[29] M. Miller and C. Gupta, "Mining Traffic Incidents to Forecast Impact," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp 12. New York, NY, USA: Association for Computing Machinery, 2012, p. 33 to 40. [Online]. Available: https://doi.org/10.1145/2346496.2346502

[30] A. J. Khattak, "Injury Severity in Multivehicle Rear-End Crashes," *Transportation Research Record*, vol. 1746, no. 1, pp. 59–68, 2001. [Online]. Available: https://doi.org/10.3141/1746-08

[31] M. A. Abdel-Aty, C. L. Chen, A. E. Radwan, and P. A. Brady, "ANALYSIS OF CRASH-INVOLVEMENT TRENDS BY DRIVERS' AGE IN FLORIDA," *Institute of Transportation Engineers Journal (ITE)*, Feburary 1999. [Online]. Available: https://pdfs.semanticscholar.org/2993/a5ef3545a9612328a537e7748d96e72cabd0.pdf

[32] Z. Yuan, X. Zhou, T. Yang, J. Tamerius, and R. Mantialla, "Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study," in *Urban Computing*. Halifax, Nova Scotia: In Proceedings of 6th International Workshop on Urban Computing, August 2017. [Online]. Available: https://www.semanticscholar.org/paper/Predicting-Traffic-Accidents-Through-Heterogeneous-Yuan-Zhou/fdeaa1a8d25518b0424a66d98c04cf7bde4982ba

[33] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data," in *ACM*

SIGKDD. In Proceedings of 24th ACM SIGKDD International Conference, July 2018, p. 984 to 992. [Online]. Available: https://www.kdd.org/kdd2018/accepted-papers/view/hetero-convlstm-a-deep-learning-approach-to-traffic-accident-prediction-on-

[34] A. Hébert, T. Guèdon, T. Glatard, and B. Jaumard, "High-Resolution Road Vehicle Collision Prediction for the City of Montreal," *arXiv*, May 2019. [Online]. Available: https://arxiv.org/pdf/1905.08770.pdf

[35] D. F. Flynn, M. M. Gilmore, and E. A. Sudderth, "Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot analysis of 2017 Waze data and Police Accident Reports in Maryland," VOLPE, Tech. Rep., November 2018. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/37256

[36] H. Renski, A. J. Khattak, and F. M. Council, "Effect of Speed Limit Increases on Crash Injury Severity: Analysis of Single-Vehicle Crashes on North Carolina Interstate Highways," *Transportation Research Record*, vol. 1665, no. 1, pp. 100–108, 1999. [Online]. Available: https://doi.org/10.3141/1665-14

[37] K. A. Krull, A. J. Khattak, and F. M. Council, "Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes," *Transportation Research Record*, vol. 1717, no. 1, pp. 46–54, 2000. [Online]. Available: https://doi.org/10.3141/1717-07

[38] "Hamilton County Community Information," Online. [Online]. Available: http:/hamiltontn.gov/community/Default.aspx

[39] "Hamilton County, TN," Online. [Online]. Available: https:/datausa.io/profile/geo/hamilton-county-tn#about

APPENDIX A
Data Statistic Appendix

	Total	**min**	**max**	**range**	**sum**	**median**	**mean**
Accident	59887	1	1	0	59887	1	1
Latitude	59887	34.965041	35.431784	0.466743	2099757.651	35.042901	35.06199427
Longitude	59887	-85.469212	-84.946964	0.522248	-5103868.854	-85.231114	-85.22498796
Hour	59887	0	23	23	810620	14	13.53582581
Temperature	59887	7.96	94.84	86.88	3789776.53	64.28	63.2821235
Dewpoint	59887	0.88	76.75	75.87	3073935.29	54.51	51.32892431
Humidity	59887	0.12	1	0.88	40946.07	0.7	0.683722177
Month	59887	1	12	11	375219	6	6.265449931
Weekday	59887	0	6	6	168009	3	2.805433567
Visibility	59887	0	10	10	355255.24	6.38	5.932092775
Cloud_Coverage	59887	0	1	1	33012.09	0.55	0.551239668
Precipitation_Intensity	59887	0	0.131	0.131	478.2106	4.00E-04	0.007985215
Precip_Intensity_Max	59887	0	0.7437	0.7437	2968.8714	0.0044	0.049574555
Clear	59887	0	1	1	11768	0	0.196503415
Cloudy	59887	0	1	1	39422	1	0.658273081
Rain	59887	0	1	1	7057	0	0.117838596
Fog	59887	0	1	1	1617	0	0.027000852
Snow	59887	0	1	1	23	0	0.000384057
RainBefore	59887	0	1	1	5615	0	0.093759915
Terrain	59887	1	3	2	119379	2	1.993404245
Land_Use	59887	0	9	9	187292	2	3.127423314
Access_Control	59887	0	2	2	22557	0	0.376659375
Operation	59887	1	2	1	118670	2	1.981565281
Thru_Lanes	59887	0	8	8	200936	2	3.355252392
Num_Lanes	59887	0	10	10	209775	3	3.502847029
Ad_Sys	59887	0	31	31	785628	13	13.11850652
Gov_Cont	59887	0	67	67	159261	4	2.659358458
Func_Class	59887	0	19	19	919922	16	15.36096315
Pavement_Width	59887	0	60	60	1571107	24	26.23452502
Pavement_Type	59887	0	18	18	905499	15	15.12012624

Table A.1 Fishnet Accident Data Statistics, Basic

	SE.mean	CI.mean.0.95	var	std.dev	coef.var
Accident	0	0	0	0	0
Latitude	0.00026713	0.000523575	0.004273436	0.065371519	0.001864455
Longitude	0.000324519	0.000636058	0.00630685	0.07941568	-0.000931836
Hour	0.020210579	0.039612807	24.46189236	4.945896517	0.365393038
Temperature	0.066854343	0.131034753	267.6651365	16.36047483	0.25853233
Dewpoint	0.069379826	0.135984708	288.269679	16.97850638	0.330778535
Humidity	0.000832195	0.001631105	0.041474652	0.203653263	0.297859672
Month	0.014456997	0.028335765	12.51666738	3.537890244	0.564666589
Weekday	0.007714646	0.015120734	3.564220551	1.887914339	0.672949223
Visibility	0.007784427	0.015257504	3.628990382	1.904990914	0.321133028
Cloud_Coverage	0.001154673	0.002263163	0.079845539	0.28256953	0.512607395
Precipitation_Intensity	6.05E-05	0.00011851	0.000218942	0.014796695	1.85301135
Precip_Intensity_Max	0.000332598	0.000651893	0.006624787	0.081392792	1.641825969
Clear	0.001623732	0.003182521	0.157892459	0.397356841	2.022137079
Cloudy	0.001938117	0.003798716	0.224953388	0.474292513	0.720510267
Rain	0.001317514	0.002582332	0.103954397	0.322419598	2.736112011
Fog	0.000662342	0.001298193	0.026272244	0.16208715	6.00303845
Snow	8.01E-05	0.000156931	0.000383916	0.019593763	51.01789952
RainBefore	0.001191153	0.002334664	0.084970412	0.291496847	3.108970912
Terrain	0.000879222	0.001723278	0.046294538	0.215161654	0.10793679
Land_Use	0.010690856	0.020954116	6.844748901	2.616247102	0.836550361
Access_Control	0.003194682	0.006261588	0.6112062	0.781796776	2.075606843
Operation	0.000549687	0.001077388	0.018095182	0.134518333	0.067884886
Thru_Lanes	0.006735319	0.01320125	2.716745239	1.64825521	0.491246266
Num_Lanes	0.007745781	0.015181759	3.593047867	1.895533663	0.541140863
Ad_Sys	0.023191104	0.045454647	32.20886283	5.675285264	0.43261672
Gov_Cont	0.006431861	0.012606471	2.477455708	1.573993554	0.591869648
Func_Class	0.012682388	0.024857526	9.632402559	3.103611213	0.202045352
Pavement_Width	0.036622873	0.071780963	80.32253108	8.962283809	0.341621729
Pavement_Type	0.006493792	0.012727856	2.525395115	1.58914918	0.105101581

Table A.2 Fishnet Accident Data Statistics, Extended

	Total	**min**	**max**	**range**	**sum**	**median**	**mean**
Accident	52239	1	1	0	52239	1	1
Unix	52239	1483250400	1577833200	94582800	8.00E+13	1530590400	1530837101
Join_Count	52239	0	1359	1359	18542415	256	354.953483
Grid_Num	52239	1	694	693	15756000	258	301.6137369
NBR_LANES	52239	1	9	8	129366	2	2.476425659
TY_TERRAIN	52239	1	3	2	104767	2	2.005532265
FUNC_CLASS	52239	6	19	13	918399	19	17.58071556
Hour	52239	0	23	23	718685	14	13.75763319
hourbefore	52239	1483246800	1577829600	94582800	8.00E+13	1530586800	1530833501
cloudCover	52239	0	1	1	25221.32	0.44	0.482806332
dewPoint	52239	0.34	78.52	78.18	2752800.55	56.31	52.69627194
humidity	52239	0.12	1	0.88	36618.93	0.73	0.700988342
precipIntensity	52239	0	0.5174	0.5174	415.2934	0	0.007949873
precipProbability	52239	0	1	1	5593.61	0	0.107077279
pressure	51731	984.5	1043.2	58.7	52663048.8	1017.9	1018.01722
temperature	52239	13.07	97.69	84.62	3334314.4	66.53	63.82806715
uvIndex	52239	0	11	11	92180	0	1.764582017
visibility	52239	0.241	10	9.759	393091.707	8.606	7.524870442
windBearing	51113	0	359	359	9304593	182	182.0396572
windGust	52234	0	42.38	42.38	374827.93	5.7	7.175937703
windSpeed	52239	0	17.47	17.47	160072.56	2.57	3.064234767
Rain	52239	0	1	1	6566	0	0.125691533
Cloudy	52239	0	1	1	23692	0	0.453530887
Foggy	52239	0	1	1	1700	0	0.032542736
Snow	52239	0	1	1	9	0	0.000172285
Clear	52239	0	1	1	20265	0	0.387928559
Longitude	52239	-85.469479	-85.044879	0.4246	-4452617.026	-85.244924	-85.23549504
Latitude	52239	34.982202	35.214267	0.232065	1831361.92	35.045492	35.05736939
RainBefore	52239	0	1	1	6126	0	0.117268707

Table A.3 Hexagonal Accident Data Statistics, Basic

	SE.mean	CI.mean.0.95	var	std.dev	coef.var
Accident	0	0	0	0	0
Unix	119433.013	234089.8279	7.45E+14	27297433.46	0.017831704
Join_Count	1.390042435	2.724496237	100937.1346	317.7060507	0.895063905
Grid_Num	0.752545442	1.47499614	29584.23301	172.0006774	0.570268049
NBR_LANES	0.004626057	0.009067115	1.117935568	1.057324722	0.426955971
TY_TERRAIN	0.000378722	0.000742298	0.007492652	0.086560109	0.043160666
FUNC_CLASS	0.009807095	0.019221998	5.024300564	2.241495163	0.127497379
Hour	0.020282573	0.039754035	21.49022539	4.635755104	0.336958766
hourbefore	119433.013	234089.8279	7.45E+14	27297433.46	0.017831746
cloudCover	0.001669455	0.003272147	0.145594211	0.381568095	0.790312946
dewPoint	0.072185321	0.141483908	272.2028326	16.49857062	0.313088004
humidity	0.000843011	0.00165231	0.037124604	0.192677461	0.274865429
precipIntensity	0.000132357	0.000259421	0.000915146	0.030251374	3.805265237
precipProbability	0.001142451	0.002239214	0.068181997	0.261116826	2.438582935
pressure	0.025626647	0.05022848	33.97304196	5.828639804	0.005725483
temperature	0.070858792	0.138883899	262.2903719	16.19538119	0.253734476
uvIndex	0.011151644	0.021857326	6.496397826	2.548803215	1.444423206
visibility	0.012569194	0.024635738	8.252959328	2.87279643	0.38177354
windBearing	0.456433233	0.894613882	10648.43752	103.1912667	0.566861464
windGust	0.025462629	0.049906992	33.86567665	5.819422364	0.810963334
windSpeed	0.011808616	0.023144999	7.284384578	2.698959907	0.880794101
Rain	0.001450414	0.002842826	0.109895275	0.331504563	2.637445457
Cloudy	0.002178176	0.004269245	0.247845366	0.497840703	1.097699667
Foggy	0.000776336	0.001521626	0.031484309	0.177438184	5.452466656
Snow	5.74E-05	0.000112552	0.000172259	0.013124736	76.18034148
Clear	0.002131983	0.004178706	0.237444537	0.487282811	1.256114816
Longitude	0.000307955	0.000603594	0.004954146	0.070385694	-0.000825779
Latitude	0.000191763	0.000375857	0.001920988	0.043829075	0.00125021
RainBefore	0.001464858	0.002871135	0.112094872	0.334805722	2.855030381

Table A.4 Hexagonal Accident Data Statistics, Extended

VITA

Pete graduated from the University of Tennessee at Chattanooga in May 2018 with a Bachelors degree in Information Security and Assurance, shortly after getting married to the best person ever. During his Bachelor's studies, Pete was introduced to the field of data science through an introduction to the Smart Communications and Analysis Lab (SCAL) at UTC. This led to his pursuit of a subsequent Masters degree in Data Science, around which time SCAL evolved into the Center for Urban Informatics and Progress (CUIP). Working for CUIP granted him the opportunity to travel to Montreal, Canada to present his work as well as helping him become a proficient presenter. He graduated from the University of Tennessee at Chattanooga in May 2020.