Advances in Kriging-Based Modelling Approaches of Winter Weather Vehicular Collisions – A Region-Wide Geostatistical Investigation

By

Andy Ho Ming Wong

A thesis submitted in partial fulfillment of the requirement for the degree of

Master of Science

In

TRANSPORTATION ENGINEERING

Department of Civil and Environmental Engineering

University of Alberta

ABSTRACT

The winter season is known for brisk cold weather and beautiful snowfalls, but it is also known for deteriorated driving conditions to where the risk of collisions becomes a major issue that plagues many municipalities around the world. As such, agencies are tasked with and strive to prioritize weather-related collision prone locations for an efficient mobilization of winter road maintenance services and to improve winter safety of motorists. Finding these locations is known as network screening and can be a challenge at times, requiring either a lot of data, or statistical background. Presented in this thesis is an alternative network screening method known as Regression Kriging (RK) that has recently gained some traction due to its wide applicability and excellent performance in modelling regionalized random variables. This thesis has three objectives and that is to (1) demonstrate the applicability and usability of RK models, (2) further enhance its predictive ability by substituting in network distances, and (3) characterize the underlying spatial structure of the winter collisions at various zonal scales to check the spatial continuity assumption, otherwise known as a second order stationarity. This thesis employs a case study within the state of Iowa where RK was utilized to model winter collisions using large-scale datasets of the entire state of Iowa that were collected over five (5) winter seasons from 2013 to 2018. The Winter Collision (WC) ratio was used as a surrogate measure for winter collisions as it represents a value used for relative comparison of collision sensitivity to winter conditions.

The results from the case study resulted in some key findings. As an estimator, RK was shown to be a very effective providing predicted results that outperformed results from multiple linear regression and from ordinary kriging (OK), a precursor to RK. Five statistical measures were used to compare model performances with RK outperforming OK on all measures, though the predicted values were overestimated. In an exploratory analysis in an attempt to improve RK estimations, network distances were substituted into the kriging modelling process. Using the same five statistical measures, it was found that network distances provided marginal improvements to the predicted values, but the real improvement was in the level of uncertainty in those values. The model now underestimates the true value, but not to the extent that it had previously overestimated them, thus reducing the level of uncertainty of the values. As for the underlying spatial structure, it was found that the spatial variance that the model relies on was not continuous or stable, thus suggesting that models should be built on a zonal level to better capture unique regional spatial characteristics.

The main contribution of this thesis can be broken down into three parts. For the first time in literature, RK has been shown to perform well as a network screening tool over a very large temporal and spatial scale. Secondly, network distances have now been shown to improve kriging results within network screening. And finally, it was determined that the spatial structure is highly sensitive to the area and its size. Applying RK over an extremely large spatial scale could possibly overlook regional factors that can affect the spatial structure as zonal analysis often gave different, often better, results.

PREFACE

Part of the work presented in this thesis has been accepted for publication.

Peer-reviewed Journal and Conference papers

1. Wong, A.H., and Kwon, T.J. (2021) Development and Evaluation of Geostatistical Methods for Estimating Weather-Related Collisions – A Large Scale Case Study. Accepted for publication at the *Transportation Research Record: Journal of the Transportation Research Board*, National Research Council. Washington D.C. April 2021.

2. Wong, A.H., and Kwon, T.J. (2020) Development and Evaluation of Geostatistical Methods for Estimating Weather-Related Collisions – A Large Scale Case Study. *Presented at and Proceedings of the 100th Transportation Research Board Annual Conference,* Washington, D.C., United States, October 2020.

In preparation

1. Wong, A.H., and Kwon, T.J. *Advances in Kriging-Based Methods for Estimating Statewide Winter Weather Collisions – An Empirical Investigation*

ACKNOWLEDGEMENTS

I would like to express my thanks and gratitude to my graduate supervisor Dr. Tae J. Kwon who took a chance and was patient with me when I asked to be his graduate student at a chance encounter at a Canadian Society for Civil Engineering meeting. His encouragements and understanding during times of research uncertainty helped keep me on course. With his guidance and tutelage, I have been able to reach this personal academic milestone.

I would like to thank to the examiners, Dr. Tony Qiu and Dr. Wei Victor Liu for being part of my MSc examination committee, and to Dr. Yuxiang Chen for chairing the defense. Their constructive feedback and suggestions contributed significantly to the betterment of my thesis and was greatly appreciated.

My thanks also goes to NSERC for providing the funding that allowed me the time and resources for conducting this study and preparing this thesis. My regards extend to Iowa Department of Transportation and Iowa State University for maintaining their open database and the data used within this study.

I would like to thank fellow team member Minjian Wu who kindly assisted me by doing the more complex python coding of some algorithms. To my fellow team members, Lian Gu, Simita Biswas, Shuoyan Xu, Davesh Sharma, and Tasnia Nowrin, I enjoyed our collaborations and teamwork on the many projects that we were a part of. Being on the cutting edge of research with friends and peers was a joy during my Masters. And to my fellow office mates Andres Rosales, Kaleab Yirgu, and Mohammed Ahmed, cheers to you for helping make my time at the office more fun, at the expense of productivity.

Finally, I want to especially recognize my immediate and extended family whom all provided their unwavering support and encouragement to help see me through to the end of my Masters. My sister Cindy, and housemate Court have provided me guidance at home when I most needed it. Their constant presence at home made dealing with the situations stemming from the COVID-19 pandemic of 2020/21 easier, and for that I am truly grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS / ACRONYMS

AADT	Annual Average Daily Traffic
ASE	Average Standardized Error
BLUEs	Best Linear Unbiased Estimates
CAD	Computer Aided Design
GIS	Geographic Information System
KDE	Kernel Density Estimate
KSI	Fatal or Serious Injury (collision classification)
MLR	Multiple Linear Regression
MStdE	Mean Standardized Error
MSqE	Mean Squared Error
OK	Ordinary Kriging
RK	Regression Kriging
RMSE	Root mean squared error
RMSSE	Root Mean Squared Standardized Error
RSC	Road Surface Condition
RSI	Road Surface Index
RST	Road Surface Temperature
RTM	Regression to the Mean
SK	Simple Kriging
SOSA	Second Order Stationary Assumption
UK	Universal Kriging
VIF	Variance Inflation Factor
WC	Winter Collision
WRC	Winter Road Conditions
WRM	Winter Road Maintenance

Chapter 1 INTRODUCTION

1.1 Background

There is no question that operating and maintaining a safe transportation network is a challenge that all transportation agencies face every day. Unsafe roads lead to situations that impact society humanitarianly, economically, and environmentally due to collisions. Because of these types of losses, municipalities and transportation agencies strive to ensure that the roads are safe to use for the general public. Though it is a year-round task, this is made even more difficult during times of inclement weather, especially for those in winter climes. For those that must contend with winter conditions, continual and efficient maintenance now becomes a major factor for them when trying to maintain safe roads. As of 2016 in the US, upwards of 16% of traffic fatalities are attributed to weather related incidences (US DOT Federal Highway Administration, 2020) while in Canada in 2017, over 14,000 injury collisions occurred in December alone (Royal Canadian Mounted Police, 2019). Though the travelling public have a responsibility to drive to the conditions, it is the managing authorities' responsibility to ensure the roads are in the best condition possible to reduce the chances of an incident due to drivers' mistakes. There are as many strategies as there are agencies that work to manage their roads during the winter and yet they all have the same problem: how, when, and where to apply their resources for maximum benefit.

Winter collisions (WC) and winter road conditions (WRC) pose a unique challenge for road maintenance personnel as snow and ice can accumulate on the roads continuously during a weather event. This buildup causes persisting slippery conditions that can be extremely hazardous to all road users and are difficult to manage as they can deteriorate road conditions rapidly even after servicing. Municipalities have a duty of care to their citizens to ensure that their roads are serviced by their road authorities after a winter event. Road authorities strive to maintain and service roads in a timely manner to improve mobility and reduce the chance of weather induced collisions through the active process of servicing winter roads, known as Winter Road Maintenance (WRM).

There are many aspects to WRM that ranges from planning in the off season to crews doing the work, all striving to make the roads less dangerous and more accessible, while improving winter mobility. Aside from the need for timely maintenance executions for safety, these processes also come with operational costs that municipalities try to minimize. To achieve these goals, one solution would be targeting their operations at high-risk weather sensitive collision locations – a process commonly referred to as network screening and is a critical part of the safety management process used by transportation safety engineers. Figure 1-1 shows the 6-step cyclic process involved in the safety management process.

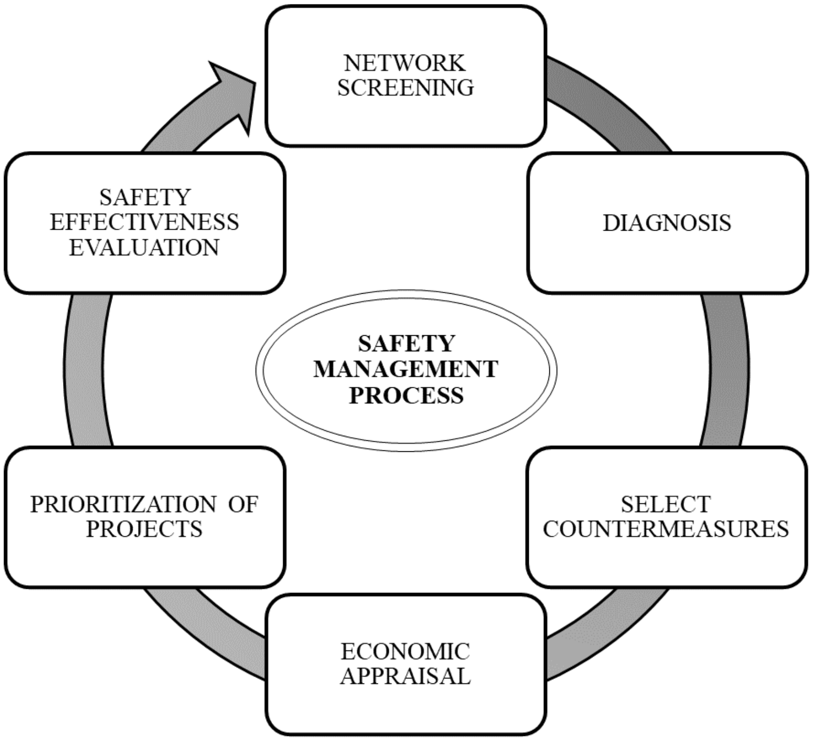

NETWORK SCREENING
DIAGNOSIS
SELECT COUNTERMEASURES
ECONOMIC APPRAISAL
PRIORITIZATION OF PROJECTS
SAFETY EFFECTIVENESS EVALUATION
SAFETY MANAGEMENT PROCESS

Figure 1-1 Safety Management Process adapted from the Highway Safety Manual (AASHTO, 2010)

Network Screening is the first and the foremost important step of the safety management process as it serves to identify and establish focus to sites for potential assessment or treatment (American Association of State Highway and Transportation Officials (AASHTO), 2010). This process provides rankings of high risk sites such that agencies are able to make budgetary and long term treatment plans. Many network screening methods have been developed over the decades in the transportation safety field and continues to be a heavily researched topic.

1.2 Problem Statement and Motivation

A US joint task force published the Highway Safety Manual which outlines many common network screening methods currently being used (AASHTO, 2010). One of the first renditions is also its simplest form, which is just a tabulation of the number of collisions at each location, known as the collision frequency. It comes with a lot of inherent assumptions that limit its accuracy, but it is an effective starting point for any transportation agency. Since then, efforts have been made to improve network screening analysis to better identify truly problematic locations for further attention. Collision rates were used as it accounts for traffic volume exposure and regression analysis brings in correlating variables to aid in identifying hot spots, but though these methods are relatively simple to implement and understand, they suffer from random fluctuations, over-dispersion of crashes, non-linear relationships to exposures, and to the phenomenon known as regression to the mean (RTM) (Retting, et al., 2003; Srinivasan, et al., 2016). RTM is a confounding phenomenon as cases where abnormally high or low numbers of random measurements or samples over a large area or period of time will trend towards the mean thereby providing misleading data interpretations of perceived trends (De Pauw, et al., 2014). More complicated methods were then developed to address these shortcomings such as negative binomial regression and the safety performance function, which have been used by transportation engineers for many years and have been proven to be effective, but they require a lot of data and do not provide a level of uncertainty for its estimates (Srinivasan, et al., 2016).

Recently, there has been an increasing body of research that has delved into the use of geostatistics for network screening as researchers show that collisions may display a level of spatial autocorrelation (Islam, et al., 2016; Thakali, et al., 2015). Of the various geostatistical methods being used, Kriging has been gaining notoriety for being a very adept estimator as it is able to use priori spatial relationships to make better estimations (Olea, 1999). It also provides a probability of uncertainty for the estimates generated thus adding a level of information previously not considered. With the growing body of research applying kriging to transportation safety, it still needs some level of benchmarking and validation of assumptions as there are few studies doing so in literature. In particular, Regression Kriging utilizes external covariates to improve estimations, but it has yet to be applied to winter collision analysis or for collisions in general. Additionally, kriging is normally applied using Euclidean distances between data points as the only factor for autocorrelation relationships. A fundamental assumption in kriging is that the autocorrelation of the data points between themselves is reliant on the spatial separation between them which is defined as their spatial structure (Olea, 1999). However, the true distance between points on a road network is dependent on the road network between them thus the spatial structure should rely on that instead, but has yet to be conclusively confirmed. Intuitively, the application of kriging in transportation engineering should use the Network Distance over Euclidean Distance, but few studies have done so. Figure 1-1 best illustrates the network vs Euclidean problem.

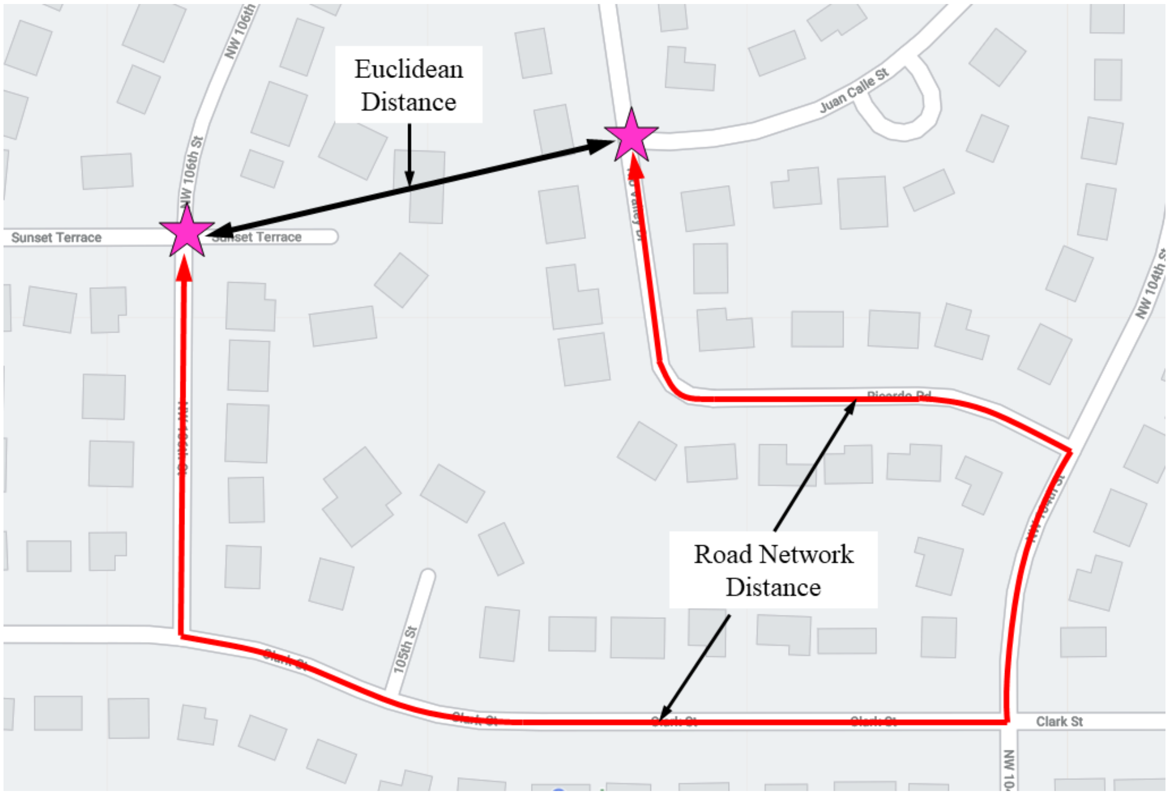


Figure 1-2 Euclidean vs Network Distance from City of Des Moines

Even with the use of kriging gaining notoriety for its performance for spatial estimation, rarely has its underlying spatial structure and its translation invariance been closely examined. The premise is that the spatial structure, which is the variance between any two points for a given separation distance, of the dataset is assumed to be the same throughout the whole area. This is an important premise as urban and rural road networks are spaced and used differently thus potentially having very different underlying spatial structures. Likewise, different counties may have different road planning and construction policies that again can affect the spatial characteristics of their road networks. Formally, this is known as the second order stationarity assumption (SOSA) and it implies that the interaction between any two variates is insensitive to the spatial translation (Olea, 1999). The spatial structure is represented by the semivariogram, which is the graphical representation of the change in similarity/dissimilarity between variates as the distance between them increases. If the SOSA holds, then it can be said that a single semivariogram is sufficient and reliable to be used in developing the kriging model. Otherwise, the underlying spatial structure cannot be represented by a single semivariogram, but instead each sub-region will require a semivariogram of their own (Olea, 1999). Outside of fundamental textbooks, the SOSA is rarely examined in research studies.

Identified above are three glaring gaps within network screening academic literature especially within the field of winter transportation safety analysis, thus providing the main motivation behind this thesis. This body of work will provide supportive justification for an alternative network screening method for locating and addressing roads that are disproportionally more collision prone under winter conditions helping municipalities maintain safer roads. By providing a more in-depth structural analysis of the methodology and providing a benchmark, kriging can become a more suitable method of network screening analysis in transportation safety.

1.3 Research Objectives

The overall objective of this thesis is to provide credibility and validity to geostatistics as an effective tool for winter collision modelling and estimation. The focus is on winter collisions because it is one of the more direct and immediate ways that agencies can reduce their annual collision statistic through WRM. Additionally, there is a lack of research into winter road safety for agencies to refer to when developing or evaluating their practices. Therefore, this will provide transportation planners and maintenance managers a new tool for them to make the roads safer during the winter months. While proving geostatistics abilities, this thesis has following three specific objectives:

1. Develop a hybrid geostatistical method known as regression kriging (RK) for estimating large-scale winter weather collisions by considering auxiliary traffic information and local road weather characteristics;
2. Enhance regression kriging estimates by utilizing Network distances for improved estimation accuracy; and
3. Analyze and characterize the underlying zonal spatial structures for regions with distinctively varying road network and environmental features and properties.

The findings of this thesis will provide justification and verification of RK to interested parties looking for an alternative method for modelling winter collisions. Furthermore, this study will have also benchmarked and validated the use of Network distance over Euclidean distances and the underlying spatial structure second order stationarity assumption (SOSA) for many other transportation engineering problems requiring of geostatistical interpolations and estimations.

1.4 Thesis Organization

The remainder of this thesis will be organized as follows:

Chapter 2 will provide a literature review of winter traffic challenges and safety, geostatistics, and regression kriging. This section will provide background into the work done and the gaps in knowledge that currently persists.

Chapter 3 provides a detailed overview of the case study area, data, and data pre-processing prior to utilizing it in the case study. Data filtering and descriptive statistics of the data will also be presented in this chapter.

Chapter 4 will outline the workflow and methodological framework for this thesis. It will provide an overview of the methodology, a specific process diagram for implementing RK, and the various statistics used to determine model goodness.

Chapter 5 will discuss the results from the case study, notable findings, challenges and solutions, and possible considerations.

Chapter 6 will summarize and conclude main findings and contributions, some areas that could be expanded, and future research directions that can be taken from this research.

Chapter 2 LITERATURE REVIEW

Transportation collision analysis has been extensively studied over many decades and there is a large body of literature on the subject from all over the world. Winter collisions is a specific topic that is often only taken upon by agencies and institutions that experience, and are regularly challenged by it, but it still has a significant number of literature on the subject. Likewise, geostatistics and kriging has itself a significant body of research in general, but its application into transportation engineering problems is relatively recent thus there are limited studies available.

This section will review the works that have already been done on winter transportation safety, factors affecting winter transportation safety, geostatistics and kriging, and applications of geostatistics and kriging to collision analysis. Limitations and gaps in the literature will be identified where the work of this thesis serves to address.

2.1 Winter Transportation Safety

Transportation safety is an important facet for any transportation system around the world. The World Health Organization's (WHO) 2004 report on traffic injury found that over 50 million people are injured and 1.2 million people are killed in traffic collisions world wide (Peden, et al., 2004). In 2016 the US federal highway administration reported a ten-year average of 5,891,000 collisions annually and approximately 21% of them were weather-related (US DOT Federal Highway Administration, 2020). They go on to find that 16% of all crash fatalities and 19% of all crash injuries were weather-related. Eisenberg and Warner's analysis of winter collisions within the 48 contiguous U.S. states over 25 years found that first-snow-days are extremely dangerous as fatal collision rates are relatively higher than on non-first-snow-day or dry days (Eisenberg & Warner, 2005). Furthermore, they found that property-damage-only (PDO) and non-fatal injury crashes are generally higher during snow days.

In Canada, the RCMP has reported that nearly 30% of all collisions happen on wet, snowy, or icy roads (Royal Canadian Mounted Police, 2019). They further state that the winter season going from November to February accounts for a third of all of those collisions putting a heavy strain on to emergency services, hospitals, insurance companies, and society as a whole. Pennelly et al (2018) found in their study of Edmonton roads that the number of property damage only (PDO) collisions (no injuries or fatalities) are higher over the winter months (Pennelly, et al., 2018). In

Andrey and Mills' (2003) study on Canadian roads, they surmise that Canadian drivers are accustomed to winter driving as the relative risk for fatal or major injury collisions in snowy conditions are less than in their ideal control conditions (Andrey & Mills, 2003). However, similar to Pennelly et al (2018), they found that the relative risk for PDO or minor injury collisions is greater, being 40% higher and is comparable to heavy rainfall events. Slippery conditions and snow build up tend to make roads less efficient as it impedes speeds by up to 64%, volumes by up to 44%, and road capacity by up to 27% (US DOT Federal Highway Administration, 2020).

As evident, there is no shortage of proof that winter conditions affect transportation safety significantly and it is for these reasons that municipalities strive to service their roads effectively and efficiently. One method to improve response time and maintenance efficiency would be targeted servicing whereby they address priority locations based on level of risk/hazard and the identification of these spots is known as Network Screening (AASHTO, 2010). The above mentioned studies are also all demonstrative of basic collision frequency analysis whereby collision hazards are discussed in terms of direct counts and ratios and how effective they are at communicating the core idea. However, this method does not account for traffic volume exposures or the phenomenon known as regression to the mean (RTM) leading to the development of more advanced analytical methods such as multiple linear regression, logistic regression, and Empirical Bayes (EB) method (Srinivasan, et al., 2016). As stated in chapter 1, geostatistics has seen an increase in interest and application, specifically kriging. Just like the EB and logistic regression methods, it also is not susceptible to RTM. But unlike them, it does not require as high of quality or quantity of data to have the best analytical results. The increasing number of research applying kriging to transportation engineering problems has shown it to be a powerful and effective analysis method.

2.2 Factors Affecting Winter Traffic Safety

Winter traffic safety can be heavily influenced and altered by many factors from environmental effects to human interventions. Many studies have looked into these various factors to try and understand what those factors are, how they affect traffic safety, how to control some of those factors, and what ways can it be done.

2.2.1 Effect of Winter Weather and Road Surface Conditions on Traffic Safety

There is no debate that slippery conditions follow winter weather events as they deposit snow onto the roads reducing the friction on them. One of the first studies to quantify the relationship between road surface conditions and traffic safety was done by (Norrman, et al., 2000). They developed a slipperiness classification system with 10 rankings and then matched them with collision rates, ultimately finding that 50 to 70% of winter time accidents were attributed to slippery conditions, based on the type of collision report. In a more recent government study, the federal highway administration (FWHA) found that across the US, 21% of all vehicle crashes, year round, are weather-related (US DOT Federal Highway Administration, 2020). In their thesis studying the effects of snowfall on the safety of Michigan's highways, Heqimi (2016) found that the risk of crash occurrences on freeways increases as the annual snowfall totals increase and that the majority of the crashes were PDO rather than an injury or fatal one (Heqimi, 2016). In either case, a significant proportion of winter collisions were due to slippery road conditions.

Asano and Hirasawa (2003) found that the majority of winter collisions on Japan's roads were both directly and indirectly a result of winter conditions such as ice and snow accumulation on the roads. They further noted that collisions from skidding (loss of traction) occurred more frequently at temperatures between –5°C to – 3°C than at any other temperature range. They suspect that slippery conditions develop more readily given the greater chance for a freeze-thaw cycle to occur (Asano & Hirasawa, 2003). A similar temperature trend was also found by Andersson (2010) where road surface temperatures (RST) between –4°C and –1°C saw greatest number of collisions from their study range of -15°C to +6°C. They also mention freezing rain, a phenomenon where rain would be super cooled during its decent from an upper warm air layer to a much colder lower air layer and then freeze upon contact with the surface (Andersson, 2010). This instantly adds a layer of ice on any surface causing extremely slippery conditions and thus this precipitation type was rated as the highest risk greatly affecting traffic safety.

Likewise, Usman et al. (2012) was able to show, for the first time empirically at the disaggregate level, that the road surface condition had a significant influence over the safety of the road segment. They defined a unique road surface index (RSI) value where road surface conditions (RSC) were used as a surrogate for road friction levels and found that for all of their sites and for all models, their RSI was a statistically significant factor. Environmental factors such as Air Temperature, wind speeds, visibility, and precipitation intensity were also found to be significant contributors to accident frequencies. Again, in Eisneberg & Warner's (2005) study they found that the first snowfall event of the season has a disproportionally high amount of collisions when compared to non-first snowfall events for the rest of the season. In conjunction with the first snowfall event, they found that age also contributes to the collision rates, with older drivers being more likely to be involved in a first-snowfall event collision over their younger counterparts.

There has also been a study that found that individuals of differing ages and gender will react very differently to deteriorated road conditions that affect collision rates (Morgan & Mannering, 2011). Their findings also found that the chance of injuries were highly correlated with age and gender under poor road surface conditions as opposed to ideal road conditions. This study only further supports the fact that deteriorated road surface conditions can severely affect collision risks and outcomes.

2.2.2 Effects of Traffic and Road Features on Road Safety

Often not intuitive, but the traffic characteristics will also have an effect on traffic safety. As noted by Asano & Hirasawa (2003), despite its low traffic volume, Hokkaido has a disproportionally high traffic fatality rate compared to the rest of Japan due to the rural road setting and longer inter-city distances that tend to promote higher speeds. This was a factor that was also found by Usman et al (2010) where traffic volume, storm duration, and route length were found to contribute to the increase in collisions. In an interesting study by Pei et al. (2012) they found that on Hong Kong motorways, the risk of being in a fatal or serious injury (KSI) collision marginally increases as the speeds increase, but the chance of being in a collision in the first place decreases with speed. They state that it might be due to the fact that roads that are designed to carry higher speeds and volume tend to be more efficient with little to no intersections, thus limiting potential collision conflict points (Pei, et al., 2012). However, when they account for time-exposure (travel time spent on the roads) they found that both KSI and total number of collisions increase with speeds.

From Andersson's (2010) thesis, heavily trafficked roads tend to have less accidents due to slippery conditions, suggesting that traffic may prevent or prolong the formation of ice on the roads due to the constant breakup of forming ice. However, roads that have high traffic volumes during an ongoing snow storm event would see an increase in collisions numbers just due to the increased traffic exposure during deteriorating conditions (Usman, et al., 2012). Likewise in El-Basyouny &

Sayed's (2006) regression comparison study where in addition to unsignalized intersection density, AADT was a major factor in collision rates within Vancouver and Richmond, BC, Canada. In a more generalized Poisson regression analysis study it was also found that AADT and road features such as horizontal curvatures, lane dimensions, shoulder and median widths, urban/rural location, and section lengths all have direct and significant roles in contributing to collision frequencies, (Abdel-Aty & Radwan, 2000). Furthermore, they found that demographic factors such as age and gender were also significant when they noticed that more young divers tend to be involved in collisions while speeding on roadways with curves.

It is clear that traffic and road features such as AADT, speeds, and number of lanes play an important role in the outcome of safety and should be considered in any model related to traffic safety.

2.2.3 Effects of Winter Road Maintenance on Traffic Safety

Winter road maintenance (WRM) is one of the more visible factors that can change the level of traffic safety as its purpose is to make the roads easier and safer to traverse. Referring back to (Norrman, et al., 2000), they determined that increasing the frequency of WRM operations would result in a reduction in accidents. Though they also found that accidents will still occur during and after WRM thus they call for better public awareness about road conditions. Likewise, Usman et al (2010) found that WRM had a significant impact on the road surface condition and in turn, had a direct impact on the number of collisions that occurred. Though not directly linked, WRM can be said to have an impact on traffic safety levels.

In some cases, preemptive strategies are implemented to maintain a road network's safety and mobility during an event and can include depositing anti-icing chemicals or road salts to prevent the buildup and/or attachment of ice onto the road surface. A reactive strategy tends to address the road conditions after the event to bring it back to normal from a slippery and unsafe condition and typically involves snow plowing and friction sanding. In either strategy, (Fu & Perchanok, 2006) found that it will reduce the number of collision incidences. However, their study did not take into account exposure or road surface conditions in their analysis.

2.3 Geostatistics and Kriging

Geostatistics is a broad term that represents a myriad of numerical techniques used to characterize spatial attributes (Olea, 1999) and by using these different collections of methods, spatially or temporally correlated data can be analyzed and estimated (Einax & Soldt, 1999). As described by Hengl (2009), geostatistics is the science of which solutions are methodically developed to analyze and understand geospatially mapped data points or measurements. This was originally developed for, and matured in, the mining industry given the geospatial nature of their work. Over the years, geostatistics has been found to be applicable outside of mining and is being used in fields such as agriculture, meteorology, epidemiology (medicine), anthropology, oceanography, and engineering. In turn, the field geostatistics itself has also advanced and evolved in order to better apply to the unique problems from those fields and has gone from point data analysis to spatially continuous geographical information system (GIS) data, noise filtering, and spatial optimization mappings (Hengl, 2009). The most common application of this practice is in spatial prediction models where predictions and estimations are spatially calculated from the geospatially mapped measurements. Given that more and more transportation engineering data is being spatially mapped, it makes sense to use geostatistics for transportation analysis.

Though there are many techniques and methods for geostatistical interpolation, Kriging was one of the more popular methods and was often used as a synonym for it (Cressie, 1990). Kriging is the brain child of the South African statistician and mining engineer Dr. Danie G. Krige whom developed the original idea calling it the weighted moving average (Krige, 1981). It wasn't until French professor and mathematician G. Matheron formally derived the initial formulas thereby establishing the field of linear geostatistics and bestowed it the name Kriging in recognition of Dr. Krige's trail blazing of this field (Krige, 1981; Cressie, 1990). It is a type of Linear Statistical Probability Model (LSM), where it objectively estimates model parameters that follow probability theory (Krige, 1981; Hengl, 2009). The main benefit of this method is that it is able to provide the estimate with a prediction error value thus providing a more objective analysis (Olea, 1999; Hengl, 2009). This results in a more reliable and objective data map, an understanding of possible sources of errors, and potential problem zones that may require further investigation. However, in order to make use of this methodology, the data must often meet some strict statistical assumptions (Hengl, 2009).

While kriging itself has proven to be quite robust and reliable, a lot of effort has expended by many researchers and statisticians to expand it. Three of the most commonly used forms of kriging is Simple Kriging (SK), Ordinary Kriging (OK) and Universal Kriging (UK), which synonymously called Regression Kriging (RK). Olea, (1999)'s Introductory book on Geostatistics covers these three kriging forms extensively. As the simplest form, SK lives up to its namesake and is considered the most basic form of kriging in concept and formulation. However, SK makes a lot of assumptions that limits its accuracy and effectiveness. A major assumption unique to SK is that it assumes that the mean (m) is known and constant. OK is the progression from SK where it now assumes an unknown, but still constant mean value in its calculations. This is important as the true mean is never really known thus a sample mean is usually used. But through some algebra, the mean is removed from the algorithm thereby not having to use a mean value at all, thus any bias from using the sample mean is removed from the calculations and allows for varying means. UK was developed as a way to remove trends within the underlying spatial structure, and by extension considers the mean as a function rather than a value. This is known as detrending the data and the trend that UK focused on was on coordinate trends. This is done by conducting a regression analysis using only the coordinates as covariates, hence it is also commonly referred to as a special case of regression kriging. True regression kriging takes in other covariates that are suspected of influencing the variate, thus accounting for more sources of trends and errors (Hengl, 2009). Figure 2-1 shows how SK, OK, and RK interprets the use of their mean values.

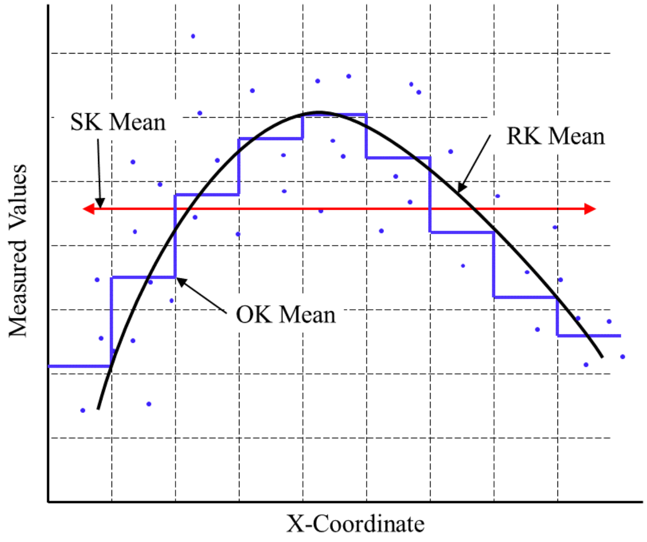


Figure 2-1 Example of SK, OK, and RK mean values

As a result, Regression Kriging (RK) has seen a consensus amongst many geostatistics professionals as being the Best Linear Unbiased Prediction (BLUP) model (Christensen, 1991; Hengl, 2009). As such, RK can be interpreted as not only being the most powerful kriging variant, but fundamental to geostatistics. Given the nature of winter transportation being affected by many external forces and influences, it naturally follows that RK would be a logical method to employ for winter traffic collision analysis. As with all forms of kriging, it makes use of Euclidean distances as its measure of separation between data points (Hengl, 2009). In an open field, this would make sense, however, on a road network, the true separation distance between points is bound by the road network. Therefore, there is the possibility for better estimate results by replacing Euclidean distances with network distances.

2.4 Applications of Geostatistics in Traffic Engineering and Safety

In Nicholson's (1999) paper, they noticed how accident distributions changed once an accident reduction plan was implemented suggesting a possible spatial influence. In turn they explored accident counts using clustering by quadrants and clustering by nearest neighbors and found that indeed there was a geospatial relationship present. Levine, et al (1995) looked into the use of the nearest neighbor clustering analysis within Honolulu. Black and Thomas (1998) looked into the network autocorrelation on Belgium's motorways using Moran's I statistic and found that accidents showed some autocorrelation along a number of simple linear networks and that it was an improvement to the existing methods that use point or count data. For the study done by Harirforoush and Bellalite (2019), they applied network kernel density estimation (KDE) to identify clusters of crashes on the streets of Sherbrooke and compared it to using aggregated crash data. Their study applied geostatistics over the entire road network of Sherbrooke rather than on a single length of road, showing the potential of KDE and geostatistics in modelling transportation engineering problems spatially. Overall, they found network KDE to be more effective in identifying potential hotspots over crash aggregation but more research into geostatistics is required.

Spatial models have the benefit of taking into account any spatial structure and kriging is one method that best utilizes this feature. Functionally, it combines both deterministic and stochastic analysis to make a more robust prediction model (Olea, 1999) and will be covered in greater detail in the next chapter. Kriging has been shown to be an adept modeling and prediction tool and has recently gained in popularity in the traffic/transportation engineering fields. The comparative study done by Thakali et al (2015) found that ordinary kriging (OK) performed better than the kernel density estimation (KDE). In Gu, et al (2018) study, they made use of regression kriging (RK) to interpolate and estimate road surface temperatures (RST) on Highway 16 in Alberta, Canada. They were able to demonstrate that RK was an effective interpolation tool for estimating RST between road weather information system (RWIS) stations as multiple linear regression was not sufficient enough. These studies, however, were limited in scale and were often isolated lengths of road, single neighborhoods, or within a municipality.

In Kwon, et al. (2019), they characterized and developed regional and zonal semivariograms for RWIS Network optimization. They developed semivariograms for different climate zones in Southern Ontario, Canada and found that the autocorrelation structure differed significantly from zone to zone. This implies that a single regional autocorrelation structure may not be optimal when developing an RWIS implementation strategy. This study has a much larger area and range than the others, but also shows that there is a limit to kriging with the autocorrelation structure and the size of area it is relevant in.

Few transportation studies have applied kriging on a large spatial scale but one of them would be Selby and Kockelman's (2013) study where they estimated AADT for the state of Texas. In their study, they compared the model results from using Universal Kriging (UK) to those from geographically weighted regression (GWR) and found that UK provided better results over GWR. In addition to their spatial method comparison, they also looked into the use of network distance between points over the standard Euclidean distance and again found that the former outperformed the latter. However, their study was limited to only one year's worth of data and were isolated to interstate roads only limiting its conclusiveness, yet still is suggestive that UK and network distances are better methods to employ within kriging analysis for transportation problems. In the study done by Zhang and Wang's (2014) they also applied network distances with kriging in analyzing subway ridership on New York's subway lines. They found that the use of network distance outperformed Euclidean distances, but their analysis was limited to only two short lines segments from a vast subway system which limits the conclusiveness of their findings, but is also suggestive of using network distances.

Although few prior studies employed spatial statistical methods in an attempt to conduct a hotspot analysis, they were mostly limited to covering relatively small areas within a short time span. More importantly, there are no large-scale studies currently available for evaluating the feasibility of using one of the most advanced and powerful kriging variants – regression kriging. Combining the fact that current collision analysis utilizes auxiliary variables with the notion that collisions may have a spatial component associated with it, then using a spatial analysis method that uses auxiliary components could prove beneficial. Therefore, in an effort to expand upon the use of geostatistical analysis methods on a larger scale, the primary objective of this study is to provide a framework and validation for implementing Regression Kriging (RK) to model winter collisions.

To better examine the effect of winter road conditions on weather related crashes, a ratio of winter collisions to all collisions is used in this study, as also suggested in previous studies (Khan, et al., 2008). Sites where certain weather warning types (snow/ice) and other meteorological factors (e.g., road surface temperature) that are shown to have an adverse impact can also be identified to help highway authorities make more informed decisions on implementing appropriate countermeasures.

2.5 Summary

As illustrated by various studies, winter conditions pose a challenge for jurisdictions tasked with keeping their roads safe during their winter seasons. Winter driving conditions have been shown to lead to an increased collision rates, traffic delays, and maintenance costs. There are many factors that can affect the level of safety upon the roads and planners must find way to handle them utilizing the many tools that have been developed to help maintenance crews clear the roads and prevent slippery conditions. Many studies have looked into ways to model and forecast some of these factors and also how they relate to winter collisions. However, most of them suffered in scale and scope with being restricted to small urban areas or select lengths of a single road. Regression kriging has not been applied to this context either. These are two glaring gaps in the current body of knowledge that will be addressed in this thesis.

As a way to improve the level of service and response, various approaches have been developed to service the road network as quickly as possible and in order to apply their strategy, they must have reliable data on problematic roads. Many various network screening methods have been developed, but the emerging geostatistical method known as regression kriging is beginning to gain notoriety for its performance. However, as reviewed, it has not been applied to analyze winter collisions with correlated variables found by other studies. Given the number of possible factors that can affect collisions being present in the winter time, regression kriging is a logical method to employ as it can take into account many covariates. Building upon the literature reviewed, some the covariates of interest that will be considered in this thesis include, but are not limited to, annual average daily traffic (AADT), road surface temperature (RST), air temperature, lane numbers, speed, and snowfall amounts. Furthermore, RK estimates could possibly be enhanced by using road network distances over Euclidean distances. Finally, none of the studies looked into characterizing the underlying spatial structure zonally or as a whole – a prerequisite that needs to be tested for improved generalization potential.

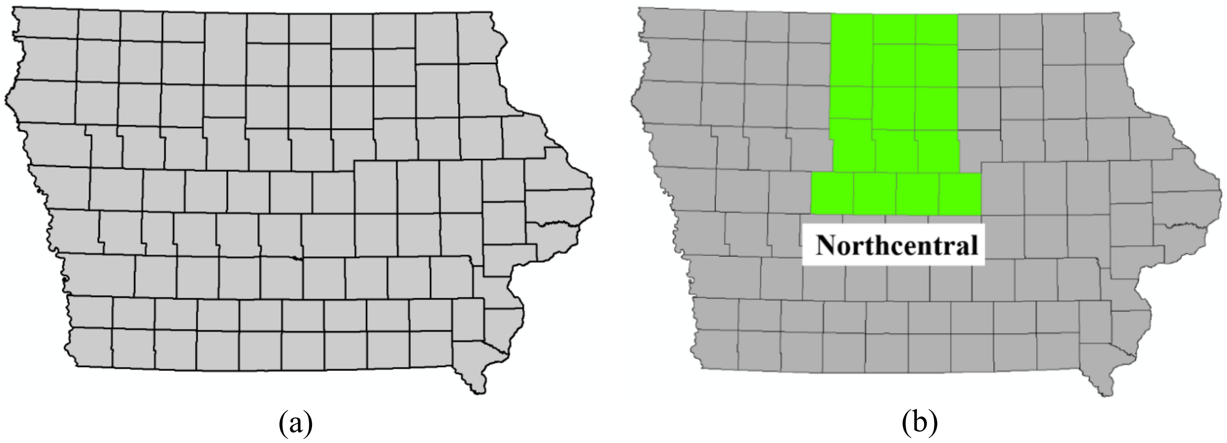
Chapter 3 STUDY AREA AND DATA DESCRIPTION

This chapter will detail the study area, describe the data collected, and the data preprocessing. The majority of data management, visualization, mapping, and pre-processing was done via ArcMAP 10.6.1 by ESRI (ESRI, 2011). Other software used in this stage include Microsoft Excel (Microsoft, 2016), and R (R Core Team, 2020).

3.1 Study Area and Period

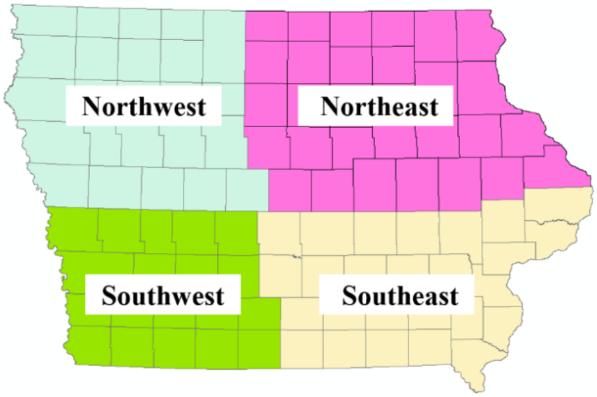
The study area encompasses the entire state of Iowa, US and sub-regions within it. Iowa was chosen for its openly accessible data, up to date databases, non-proprietary data formats, distinct winter weather events, gentle topography, and weather/RWIS station network. It covers an area of 145,700 km^2 over 99 counties and has a population of 3.15 million people as of 2019's federal census (U.S. Census Bureau, 2021). There are over 191 thousand kilometers of roads throughout the state that saw almost 290 thousand collisions from January 2013 to May 2018. They have 62 road weather information system (RWIS) stations and 128 reporting national weather service cooperative (NWS COOP) stations that provides an extensive state-wide coverage of hourly and daily weather conditions, respectively. Iowa is a data rich state, with little bureaucracy to obtaining datasets which makes for an ideal location for the intended studies.

Figure 3-1 shows the various zones that will be considered for thorough investigation in this thesis. The study area and its sub-regions were spatially mapped, partitioned, and projected into the working coordinate system (NAD 83 UTM 15N) using ArcMAP by ESRI.



(a)

(b)



(c)

Figure 3-1 Study area (a) Iowa (b) Northcentral Counties (c) Iowa Quadrants

Figure 3-1 (a) is the whole state of Iowa and will serve as the prime study area for first objective covered in Chapter 1. It will serve as a basis for benchmarking RK against MLR and OK. Figure 3-1(b) will serve as the study area for exploring the enhancement of RK using network distances. This region was mainly selected for its smaller road network density that also needed to be trimmed, and to run the Network RK modelling in a reasonable amount of time due to the computationally intensive process that increases with road network density. The areas denoted in Figure 3-1(c) serves as zonal regions for the comparison between the regional and the zonal spatial structures as part of the examination of the second order stationarity assumption (SOSA).

The purpose of this study is to show that geostatistics can be used to estimate winter collisions thus the time frame for the analysis will focus on the winter months and transitional months. According to the US National Oceanic and Atmospheric Administration (NOAA), Iowa's winter months are December, January, and February and its shoulder months are October, November, and March (NOAA, 2020). Typically, these months see snowfall and ice buildup that would constitute winter conditions for both the road and the environment. As covered in Chapter 2, winter time collisions can be caused by more than just slippery roads and can also be attributed to reduced visibility from falling or blowing snow. From here on, the winter season shall be defined as the months going from October to March. The other months were omitted given the lack of winter weather related collisions occurring in those months, often numbering less than 3 collisions all month even in April.

The time frame was limited by the number of weather and environmental data obtained and those spanned from October 2013 to April 2018. This time frame encompasses the five winter seasons (2013-14 to 2017-18) which will make up the study periods.

3.2 Data Collection, Processing, and Management

To complete the goals set forth in in Chapter 1, state and county boundaries, road information, traffic volumes, and environmental data are required. The source of the data is freely available from two online databases managed by the Iowa Department of Transportation (DOT) and Iowa State University (Iowa Department of Transportation, 2020; Iowa State University, 2020). The Iowa DOT database is managed by them using the ESRI ArcGIS open database system and all data from there can be downloaded in multiple formats, quality, and quantity. The Iowa mesonet database is managed and updated by the Iowa State University. Their weather information datasets are up to date to the day and they make available an extensive historical data library of RWIS, NWS COOP, and various other weather station systems. The sources are considered reliable and trustworthy as they are from a Government Entity and an internationally well-respected academic institution. From the raw data formats, subset datasets were generated for sub regions and zones.

Road and Traffic Data

The source of the road data is from Iowa's Department of Transportation (DOT) Open Data Portal that they maintain and is freely available to the public (Iowa Department of Transportation, 2020). The roads being used will be those classified under Federal Functions 1 to 5 for the majority of the comparisons being made. Table 3-1 provides a breakdown of the roads used:

Table 3-1 Road Network Details

Federal Function Number	Description	Length (km)
1	Interstate Highways	76
2	Principal Arterial Freeways and Expressways	3,134
3	Principal Arterial Other	9,962
4	Minor Arterial	8,862
5	Major Collectors	24,300
Total		46,334

These roads are federal interstate highways, principal arterial state highways, secondary state highways, major arterials, and main thoroughfares through cities and towns.

Since comparing Euclidean distance vs Network distance kriging requires the distances between every pairing of data points, it can become computationally prohibitive due to the number of data points. Therefore, given the computationally complex and intensive process of generating an Origin-Destination matrix between every pair of points for Network Distances, only roads that have a Federal Function of up to 4 will be used. This will then limit the roads to federal interstate highways, primary state highways, and divided secondary highways. Figure 3-2 below show the road networks that will be used for their respective analysis.

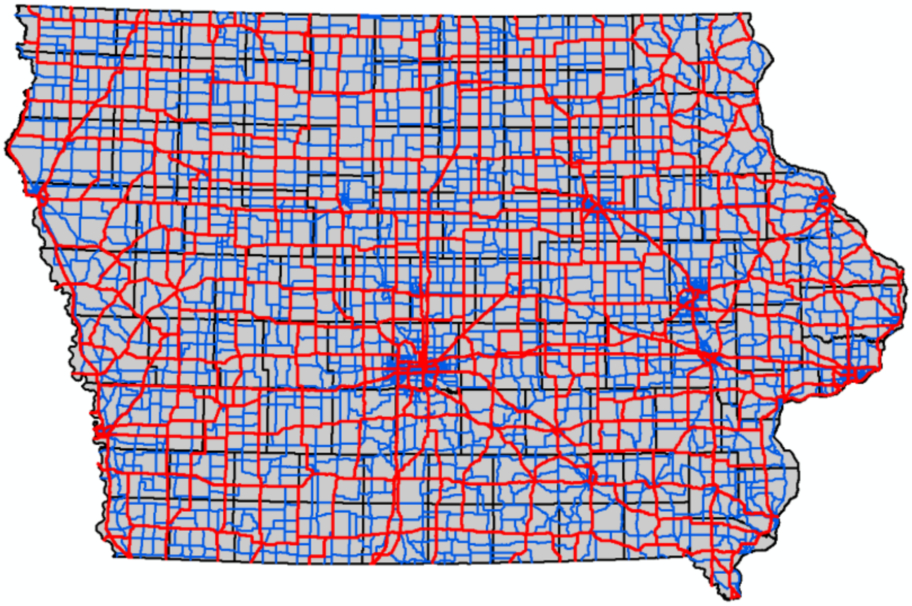


Figure 3-2 Road Networks used. Red Roads are Roads with Federal Function 1 to 4; Blue roads are roads with federal function 5

3.3 Data Preprocessing

Prior to utilizing the obtained dataset in kriging, they needed to be cleaned up, quality controlled, and if needed, transformed and projected into the same coordinate system. For this thesis, the coordinate system used is NAD 1983 UTM Zone 15N as that makes the use of metric units easier for calculations and aggregation. The majority of the data processing was done via GIS processing with ESRI ArcMAP. GIS processing takes in and spatially locates and places data features such as data points, lines, polygons, or shapes into the digital space for analysis.

Road Network

To facilitate analysis using the road network, it was divided into segments no longer than 5.0 km whereby shorter lengths are usually sections that were broken up by the presence of an intersection. This allows for the aggregation of collision and environmental data while maintaining a desirable level of fidelity. These road segments will act as data/measurement points for the study. Table 3-2 provides some details the road network shown in Figure 3-2 before and after the segmentation process described below.

Table 3-2 Road Network Statistics

	Unmodified Road Network from Shapefile		**Modified Road to max 5.0 km segment lengths**	
Road Network	**Federal Function Roads 1 to 5 (a)**	**Federal Function Roads 1 to 4 (b)**	**Federal Function Roads 1 to 5 (a)**	**Federal Function Roads 1 to 4 (b)**
Total Number of Segments	95,241	50,570	33,141	21,955
Total Length (km)	46,334	19,224	46,334	19,224
Min. Length (m)	0.002	0.050	0.02	0.50
Max. Length (km)	16.20	16.15	5.00	5.00
Average Length (m)	487.2	380.2	1,400.0	1,127.8
Std. Dev (m)	1,020.1	927.7	1,627.3	1,434.4

ESRI ArcMAP provided functions that in a non-direct way were used to segment the road network at a given length interval or at an intersection point. To create these segments, the road was first quality controlled to ensure that there are no micro-gaps (breaks in the line) throughout the entire road network that could disrupt network tracing and connectivity. These gaps in the virtual road network lines stem from imperfect importation of the line work from one CAD software to another as the algorithms for the digital representation of the lines are typically proprietary. For example, AutoCAD and MicroStation represent their lines and arcs differently thus importing/exporting line work from one to the other typically results in broken and sometimes slightly misaligned lines. To ensure that the gaps were not there erroneously, some of the larger one (about 1.0 to 3.0 m) were visually verified using Google Maps satellite and street view. Some of the gaps were valid as they would represent a closed road to prevent short-cutting traffic through a residential area. Gaps greater than 3.0 m were deemed to be purposeful ones and were left as is. During this process, over 2000 micro gaps were manually closed.

Once the road network has been checked for completeness, the road network was then dissolved to create one continuous line object. Points were then created on this line object at regular intervals, and in this case every 5.0 km. Then using the break-at-point function, the single line element was then broken into lengths of 5.0 km or shorter as the breaks also occurred at intersections. This will lead to some small lengths of road at noted in the table above as these are the "remainders" of roads that are just over 5.0 km long before hitting an intersection. Many of these micro lengths, those that are less than 1 meter, are kept in the road network for connectivity in order to facilitate a network route trace, but are not used as data points given their negligible and meaningless lengths in this context.

The road network also provides several covariates that will be explored along with the other suspected environmental covariates to see if things such as geometry or traffic volume affects winter collisions more than the weather itself. As reviewed in chapter 2, traffic speed and volumes, and the number of travel lanes have been found to be significantly correlated with collision rates (Asano & Hirasawa, 2003; Andersson, 2010; Abdel-Aty & Radwan, 2000; Usman, et al., 2012). Since the availability of these factors were conveniently part of the road network dataset obtained it would be a severe oversight to not include them into the analysis given their prevalence in past collision studies. Table 3-3 provides a summary of the road covariates considered in the study.

Table 3-3 Road Network Based Covariates

Covariate	Min	Max	Mean	Std. Dev
Number of Lanes	1	13	2.3	0.70
Speed Limit (km/h)	0	113	76.0	23.5
AADT	0	140,300	8013.0	12,067.1

Traffic Data

Annual average daily traffic (AADT) was used to represent the traffic volume demands in this thesis. Fortunately, the AADT values for the road network came with the road shapefile thus it required little data pre-processing. However, it required GIS processing after the road network was segmented. Some of the newly generated 5.0 km study segments overlap more than one base segment from the raw file and rather than choosing one or the other, the average AADT was calculated and then projected onto the new road segments. This was done using the join-by-location toolset with the variables being averaged as part of the ArcMAP software environment. There were some road segments with a reported AADT value of Zero. These road segments were typically park, military, or side roads that were to indicate a turn off from a main road, or were just misclassified. These roads were removed from the road network set.

Given how the AADT is a highly skewed dataset and is significantly larger than the WC ratio, this needed to be transformed. Regression analysis using the native values of AADT would result in extremely small coefficients (if statistically significant) and would not be meaningful. Therefore, a logarithmic transformation is done to reduce the skewness and bring down the values to a more manageable and meaningful scale, relative to the WC ratio values, for regression analysis.

Collision Data

The collision data was also obtained from the Iowa DOT Open Data portal and at the time of download spanned just over 10 years from January 2008 to June 2018. From this main set, collisions that occurred in study period and that have occurred on the road segments selected were isolated for this study. The collision data was geocoded to the NAD83 UTM 15N coordinate system by default at the time of reporting thus it was simple to spatially map them all to the proper roads using ArcMAP's import function. The collision data is point based and records a single incident in time and space along with its severity, injuries, damage value, road surface condition, and environmental condition. As such, collision statistics are often aggregated onto road segments or intersections to provide a spatially continuous and long term picture of the roads' hazard risk. Figure 3-3 shows all the collisions that met the filtering criteria. The overall collision descriptive statistics are shown below in Table 3-4.

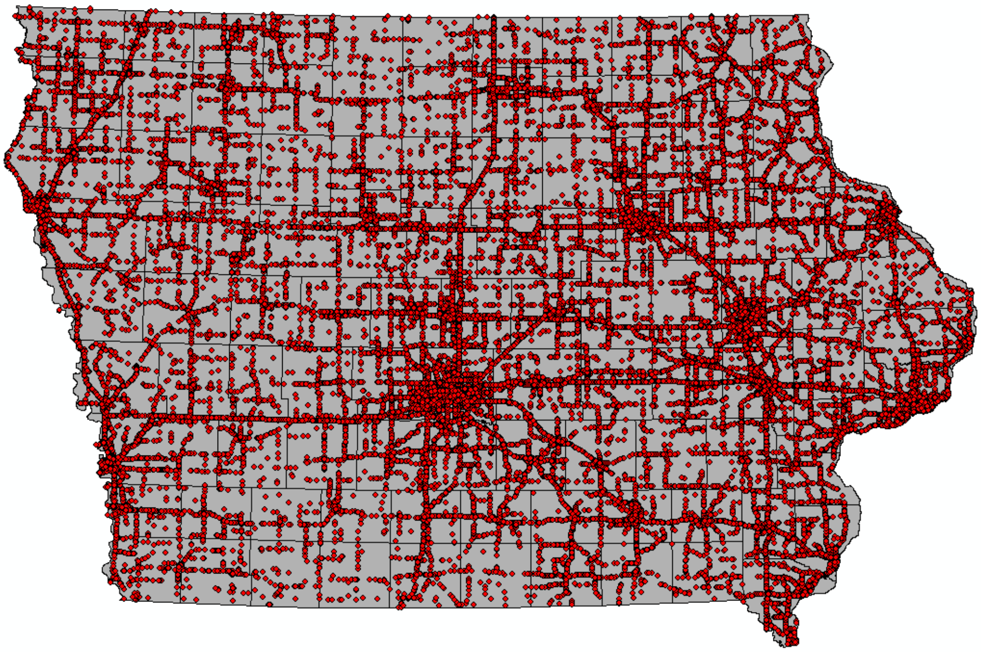


Figure 3-3 Collisions within the study area and winter periods

As introduced in Chapter 2, to properly model how winter conditions affect collisions on the road network, each road segment will have a Winter Collision Ratio (WC ratio) calculated. This is the ratio between all collisions that occurred to those that occurred under winter road and environmental conditions and follows the implementation used by Khan et al. (2008) when they utilized it to study the effects of various adverse driving conditions on the proportion of collisions they contribute to. One benefit of using the WC ratio is the ease it provides in understanding the relative effect that WC has on the safety of a road segment. For instance, high WC ratios indicates that a road segment is riskier during adverse weather/surface conditions than segments with lower WC ratios under the same weather/surface conditions. Since this study mimics Khan et al. (2008), but with a focus on winter conditions and utilizing kriging, the use of WC ratio was adopted. Simply put, it is represented by:

$$WC\ Ratio = \frac{X_{WC}}{X_{total}} \tag{1}$$

Where:

- X_{WC} = total winter condition collisions on the road segment
- X_{total} = total of ALL collisions that occurred on the road segment

Table 3-4 Seasonal Collision Descriptive Statistics

Seasonal Collision Statistics	2013-14 Season	2014-15 Season	2015-16 Season	2016-17 Season	2017-18 Season	5-year Seasonal Totals	5-year Seasonal Average	Seasonal Std. Dev
Total Collisions	22,178	21,529	22,821	22,264	22,907	111,699	22,340	557.4
Total Winter Collisions	7452	4911	4440	3912	5052	25767	5153	1360.4
Winter Collision Proportion	33.6%	22.8%	19.5%	17.6%	22.1%	23.1%	23.1%	6.2%
Total Fatal Collisions	95	106	101	127	93	522	104	13.6
Total Major Injury Collisions	422	401	390	415	357	1985	397	25.6
Total Minor Injury Collisions	1744	1541	1700	1694	1673	8352	1670	76.8
Total Possible Injury and PDO Collisions	19,917	19,481	20,630	20,028	20,784	100,840	20,168	535.6

To calculate a WC ratio for each road segment, it needs to have a count of all the collisions on that road segment and a total count of all collisions that were deemed winter condition collisions. Each collision recorded also has the road and weather conditions at the time of the incident. Using these two fields, each collision is then labelled as either occurring under winter weather conditions or not. This follows the method set forth by the FHWA where crashes that occur under adverse weather and/or slippery pavement conditions are classified as "weather-related" (US DOT Federal Highway Administration, 2020). In this case, a collision will be classified as a winter collision (WC) based on the road or weather conditions listed in Table 3-5 below on the collision report.

Table 3-5 Reported Conditions for Winter Condition Collision Classification

Road Surface Condition	Weather/Environmental Condition
Snow	Snow
Wet	Freezing rain/drizzle
Ice/frost	Blowing Snow
Slush	Sleet, hail

In order to ensure no collision was double counted on any segment of road, each collision was tagged with the road ID of the nearest road segment. Using ArcGIS, each road segment then got a count of all of the collisions by using their road ID. This also provided a count of how many of them were WCs. With each road segment having a total and WC count, the WC ratio for each road segment was then calculated. There will be sections of road with no collisions at all and thus this will throw a null error as dividing by zero (0) is undefined. Additionally, a road with a zero WC ratio is not conceptually or mathematically the same as a road with no collisions on it. A no collision road would be akin to non-data point rather than a location where winter collisions make up no portion of all collisions that occur there. These "null WC ratio" road segments are removed from the point dataset. Furthermore, the road segments were then collapsed to the mid-point of the road as kriging makes use of point data over line geometries. Figure 3-4 shows all of the "non-null" road segments and their center points totaling 19,591 valid segments.

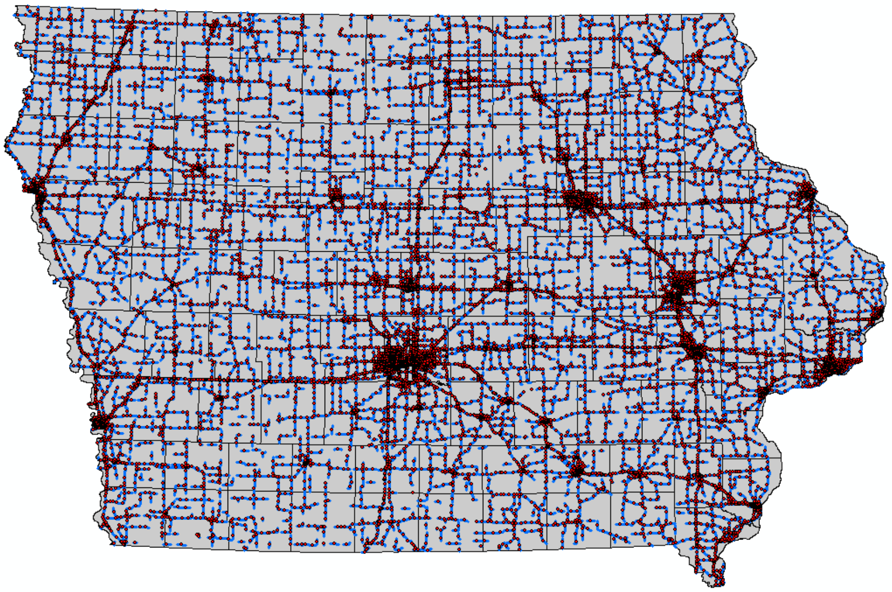


Figure 3-4 Non-Null WC Ratio Road Segments and their Mid-Points

Road Weather and Surface Conditions Data

The environmental and road surface data are collected via two common methods; via the Road Weather Information System (RWIS) and by the National Weather Service Cooperative Observer Program (NWS COOP). A RWIS is a combination of environmental sensors, detectors, imagers, and communication systems that is typically installed alongside the road and records the weather and road surface conditions at its location. This information is then relayed to a central data center where it is recorded and made available to maintenance agencies and the general public. The NWS COOP is a series of locations where a locally stationed personnel would make notable environmental and climatological recordings at regular intervals and then submit that information to a central hub where it is recorded. This system provides a good, long term historical environmental history that is often used by climate researchers.

Together, these two sources of data provide semi-continuous or daily, respectively, environmental and surface condition measurements that are used by maintenance agencies tend to use these sources of information for planning their operations around weather events. For this thesis, the environmental data used was collected from the mesonet database as maintained by the University of Iowa (Iowa State University, 2020). From there, the RWIS data and NWS COOP data was downloaded, screened, aggregated, and eventually interpolated.

For each station, the data is screened for completeness (error rates and long-term outages) where the data quality may be unreliable. In this case, a data completeness rate of 70% was the cutoff, meaning if the station’s total dataset has missing, incomplete, or erroneous data more than 30% of the time over the study period, then these stations were omitted. Once this problem stations are removed, the data is then aggregated into seasonal averages at each point which makes up the weather data for this study. Figure 3-5 shows the location of all the valid stations used. Table 3-6 provides a summary and descriptive statistics of the covariates used that were deemed relevant and from which station network they were derived from.

From these two station types, the environmental covariates were aggregated by season and averaged to obtain seasonal averages. A surrogate for road surface conditions are the road warning messages from the RWIS stations. The road condition warning messages were colour coded based on their level of severity for winter roads in a similar fashion to the Minnesota DOT (Minnesota

Department of Transportation, 2020) and many other state DOTs. A summary of those codes is illustrated in Figure 3-6.

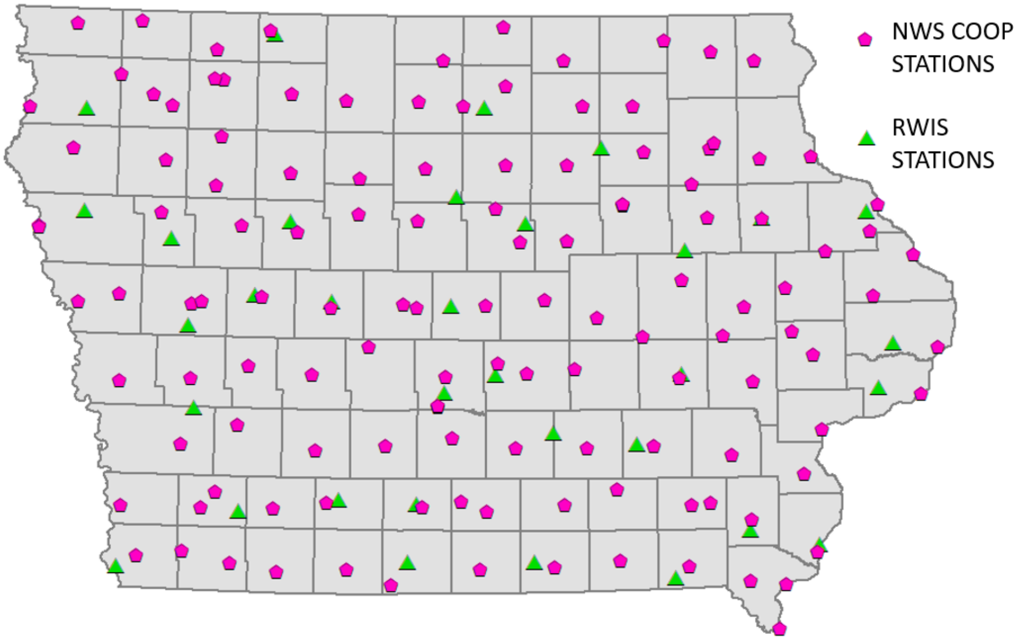


Figure 3-5 RWIS and NWS COOP Station Location Map

Table 3-6 Meteorological and Environmental Covariates

	Avg Mthly Road Surf Temp	Avg Mthly Air Temp	Avg Mthly Red Warnings	Avg Mthly Orange Warnings	Avg Mthly Yellow Warnings	Snowfall Totals	Avg Daily High Temp	Avg Daily Low Temp
UNIT	°C	°C	Count	Count	Count	cm	°C	°C
MIN	-7.9	-8.5	0	0	0	0	-10.2	-21.6
MEAN	2.7	1.0	100	875	30	5.1	5.4	-5.5
MAX	14.9	14.8	990	2481	213	34.7	21.5	10.7
STD DEV	6.1	6.2	133	632	33	5.6	7.5	6.8
STATION TYPE	RWIS					NWS COOP		
No. OF STATIONS	33					128		

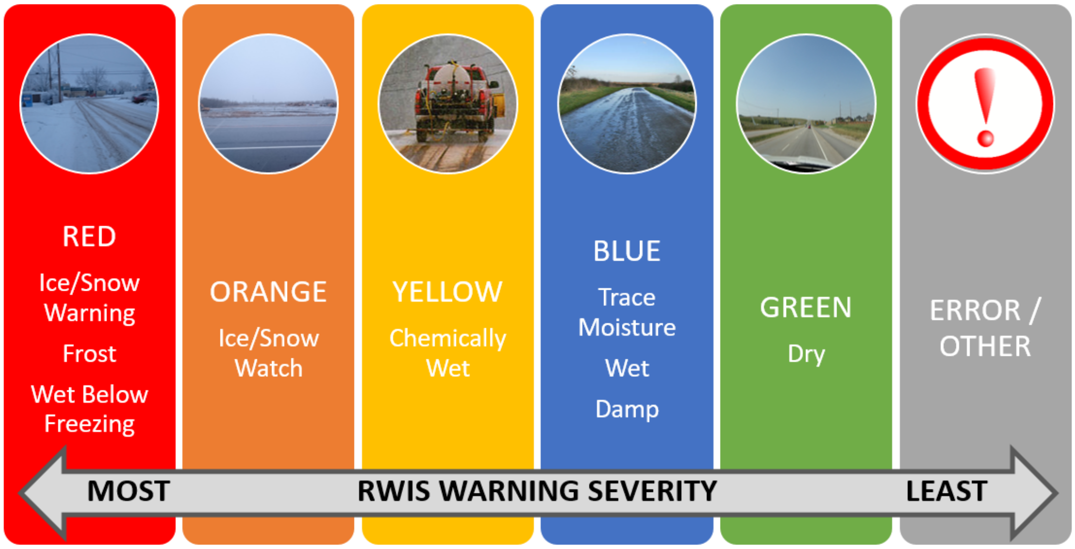


Figure 3-6 RWIS Road warning colour codes

In this study, the three top warning levels (i.e., red, orange, and yellow) are the ones of interest as they are indicative of winter road surface conditions and can be associated with WCs. Red warnings are the most severe as it is given when there is a very high chance for the presence of ice or buildup of snow on the roads make it extremely treacherous thus it is meant to alert drivers to adjust to the conditions. Yellow warnings were included as it was considered the lowest indicator of a non-ideal road surface condition. What is meant by chemically wet is that there is either anti-icing solution on the roads from treatment trucks or that there is a salt brine solution that is a result of salt melting the ice on the roads. In either case, it signifies the presence of a wet surface at a cold temperature. Other environmental factors used are Road Surface Temperatures (RST), daily average ambient air temperatures, snowfall totals, and daily high and low air temperatures. Recalling the studies from section 2.2.1, it was found that temperatures both from the air and the road surface plays a significant role in their contribution to collision frequency. It mostly is related to ice formation on the roads as cold RST promotes ice formation and cold air temperatures can lead to frost depositing onto the roads. These conditions can cause an increase in the number of collisions thus increasing the WC ratio for the road segment. But changes in air temperature are also associated with snowfall events. As was mentioned by Heqimi (2016) snowfall amounts also play a role in collision occurrences. For these reasons, these covariates were chosen to be included in the models.

Given the large spatial and time frame of this study, this data needs to be aggregated for data management and calculation purposes. Additionally, not all stations were installed near a road therefore requiring the weather data to be spatially interpolated in order to have a spatially continuous map of the environmental data that covers all road segments.

There are many studies that looked into the interpolation of weather data and it was found that Kriging performed quite well. Kwon and Gu (2017) showed that the road surface temperature (RST) interpolated between RWIS stations using kriging gave reliable results. Eguia et al. (2016) demonstrated that using kriging for the spatial interpolations of meteorological data can be done with confidence. Weather events such as precipitation can also be interpolated using ordinary and indicator kriging as was done for mapping precipitation amounts in Switzerland (Atkinson & Lloyd, 1998). The results from spatially mapping the weather events are covered in Chapter 5, section 5.1.

3.4 Summary

The study area of Iowa State was chosen for its distinct seasonality and the quality and quantity of data that is freely available. The main administrators of the datasets and the databases they reside in are by the Iowa DOT and by Iowa State University. The data and their sources are deemed trustworthy and reliable as they are a reputable government and academic, respectively, organizations with a long history in transportation engineering research.

The road network data serves as the base for which all the other obtained data is assigned to for analysis. The raw form of the road network required substantial data cleanup and quality control before it was segmented into segments no longer than 5.0 km and then turned into data points to facilitate kriging analysis. The road dataset came with several covariate values, namely the AADT, number of lanes, and speed limits, that were utilized in the case study. The collision data was provided in 10-year increments and spanned from 2008 to 2018. From this set, only the winter months for the latest five (5) winter seasons were carefully screened, cleaned, and validated before they were aggregated onto the road network for analysis. Once the collisions were projected onto the roads, the WC ratios for each road segment was calculated and became the variate of study for this thesis.

Following previous studies, several environmental covariates were selected and then obtained from the Iowa mesonet database as administered by the Iowa State University. These covariates were the RST, Air Temperatures (Daily average, highs, and lows), snowfall totals, and road warning counts. The road warning counts will serve as a surrogate to the road surface index (RSI) as that value was not provided natively from any of the RWIS stations. Following the methodologies completed by previous weather researchers, the environmental data was interpolated via OK in order to ensure that all roads have environmental data associated to it. With all the data now processed, mapped, and projected, the analysis may proceed.

Chapter 4 METHODOLOGY

The core of this thesis is to show that regression kriging is a valid and effective estimation tool for to geospatially estimate winter collisions and any influential auxiliary variables on a large spatial scale, something that has yet to be done in existing literature. As mentioned in Chapter 2, most studies have done comparative analysis using kriging, but they were often limited in spatial and temporal scope. The enhancement of RK using network distances has also not been done before in existing literature and thus will be conducted here for the first time. Finally, the examination of the underlying spatial structure is often overlooked when kriging is applied to case studies. Here, the spatial structure is thoroughly examined to see if a single spatial structure is representative of the entire region, or if zonal structures are more appropriate. By expanding the spatial and temporal scope, the results shall be more robust and conclusive. Figure 4-1 below provides a high level overview of the workflow for this thesis.

This chapter will detail the regression kriging methodology, how it may be enhanced with network distances, and how the regional and zonal spatial structures will be characterized and analyzed. Section 4.1 will go over the multiple linear regression process used. Section 4.2 covers the construction and analysis of the underlying spatial structure using the semivariogram. Section 4.3 details the Regression Kriging formulation and methodology followed by Section 4.4 where it is enhanced using network distances in place of Euclidean distances. Section 4.4 details the methodology and logic behind the comparisons. Section 4.5 describes how the spatial structure will be characterized and examined and what it means by checking the second order stationarity assumption. Section 4.6 will summarize the processes and methods covered in Chapter 4.

The hardware used throughout this process is a University provided Dell Workstation Desktop running Windows 10 Education on an Intel Core i7-8700 with 16.0 GB of single threaded RAM. Software and coding environments utilized in this study were ArcGIS 10.6.1 (ESRI, 2011), MSOffice Excel 2016 (Microsoft, 2016), Python 3.1 (Van Rossum & Drake, 2009), and R via RStudio 1.1.456 (R Core Team, 2020).

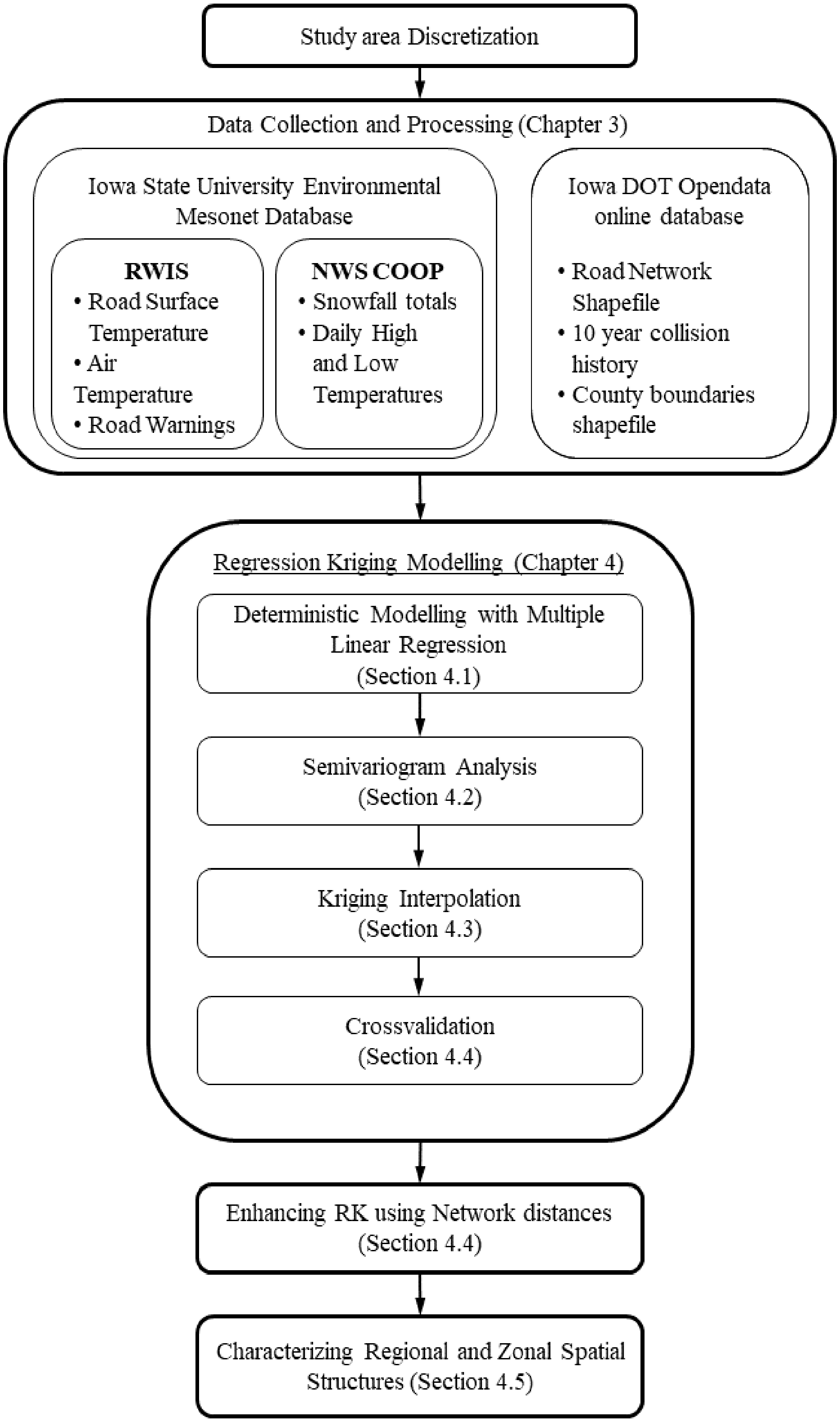

Study area Discretization
Data Collection and Processing (Chapter 3)
Iowa State University Environmental Mesonet Database
RWIS
• Road Surface Temperature
• Air Temperature
• Road Warnings
NWS COOP
• Snowfall totals
• Daily High and Low Temperatures
Iowa DOT Opendata online database
• Road Network Shapefile
• 10 year collision history
• County boundaries shapefile
Regression Kriging Modelling (Chapter 4)
Deterministic Modelling with Multiple Linear Regression (Section 4.1)
Semivariogram Analysis (Section 4.2)
Kriging Interpolation (Section 4.3)
Crossvalidation (Section 4.4)
Enhancing RK using Network distances (Section 4.4)
Characterizing Regional and Zonal Spatial Structures (Section 4.5)

Figure 4-1 Overall Project Workflow

4.1 Deterministic Modelling via Multiple Linear Regression

Regression is a popular form of basic deterministic statistical analysis in transportation safety studies by utilizing suspected correlated independent variables to estimate collisions (AASHTO, 2010). It is also a key step in RK thus it should be completed first. Given the number of covariates at hand, the multiple linear regression is done to determine which of them are best correlated with the variate. Correlation between the dependent and independent variables do not need to be high, for if they were, then regression would be sufficient enough and there would be no need for geostatistics. This is to say, that the R-squared value, the measure of the model's goodness of fit, does not need to be high in order for it to be considered a sufficient model for RK.

Prior to regression, a collinearity and multicollinearity check are done to ensure a robust model. A correlation matrix is done for each zone's dependent variable and its potential covariates to ascertain any highly correlated covariates and to also determine if any covariates have a high pairwise correlation. Here, highly correlated pairs will have a correlation of 0.60 or higher. From this pair of covariates, the one with the highest correlation to the dependent variable is kept, while the other one is removed from the model. Interaction terms and polynomial models are not considered as they do not maintain a sense of parsimony, the recognized ethical and best practice of regression analysis by using the simplest model possible based on the knowledge of the data and problem (Montgomery, et al., 2001).

To determine the best regression model, a backwards elimination stepwise selection analysis can be done. This method is often the preferred method of many analysts since it starts out with all the variables thus ensuring nothing is missed (Montgomery, et al., 2001). This can be done semi-automatically using many commercial software including R, which was used in this thesis. The function *stepAIC(direction = backwards)* used is part of the R base package.

Once the initial analysis is done, a multicollinearity check must be done. If two or more independent variables happen to be dependent on at least one other regressor, then that relationship will inflate the variance for that term thus also inflating the regression coefficients (Montgomery, et al., 2001). This effect is known as the variance of inflation (VIF) and can be used to detect multicollinearity for adjustments. For more detailed explanations and formulation for this process, the reader is referred to any introductory book to linear regression analysis such as the one referenced here by Montgomery, et al (2001). It is calculated based on the following formula:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad \textbf{(2)}$$

Where:

- R_j^2 Is the coefficient of determination as obtained when variable x_j is regressed against the remaining $n - 1$ regressors.

VIF values are provided as part of many linear regression algorithms such as those found in SAS, R-script, and MatLAB. The VIF in this study is calculated using the VIF() function from the *car* library package in R. The higher the VIF value, the more inaccurate the coefficient value is due to the inflated variance within. A VIF value of 1.0 means near ideal independence of the regressor from all other variables as it graphically represents a perfect orthogonality of the variable to all others (Montgomery, et al., 2001). VIF values greater than 5 but below 10 would be a cause for concern, while values above 10 would indicate a serious issue with multicollinearity. In this thesis, the multicollinearity issue is handled by iteratively removing variables with VIF values higher than 10 and then generating a new model without that variable until the VIFs of the remaining variables are all under the value of 10 (Montgomery, et al., 2001). The variable with the highest initial VIF is not always the one that is left out of the final model. The final regression model will then be of the form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \qquad \textbf{(3)}$$

Where:

- β_0 is the model's y-intercept value
- β_n is the coefficient multiplier for variable x_n

From the regression results, if the R^2 value is low, then the precedence is set for using kriging to improve upon the estimates. Here, an R^2 value below 0.60 is used as it implies that more than one third of the variability in y is unaccounted for by the model.

4.2 Quantifying Spatial Structures via Semivariogram

Geostatistics is based on the assumption that the data is autocorrelated with respect to the spatial distance between each data point in order to generate estimates or predictions. This assumption plays a significant role in the development and use of Kriging models. Determining the spatial relationship and structure of the measured data points is the deterministic part of kriging in general and is done via a semivariogram analysis (Hengl, 2009).

The semivariogram is a plot of the level of dissimilarity between points as the distance between increases and can be done either by point to point or more commonly via lags. This is done in lags in order to smooth out the normally noisy data and makes the plot easier to visualize. In essence, the semivariogram helps define the level of dissimilarity given a lag distance between pairs of data points, and how it changes as the lag increases (Olea, 1999). The plotted points are known as the experimental/sample semivariogram and these points are the result of the measured points. Each point represents the average semivariance between points with a certain lag distance, otherwise known as the sample variance at lag h. It is calculated using equation (4) below:

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{k=1}^{m(h)} [Z(x_k) - Z(x_k + h)]^2 \tag{4}$$

It is important to note that if the data is not normally distributed or is highly skewed, then it should be transformed. Another factor affecting the accuracy of a semivariogram model would be directionality which states that if the data shows some form of strong spatial association in a particular direction, then it is said to have an anisotropic tendency (Olea, 1999; Oliver & Webster, 2015) and may require additional investigation. Figure 4-2 is an illustrative example of how the semivariogram can be directionality dependent.

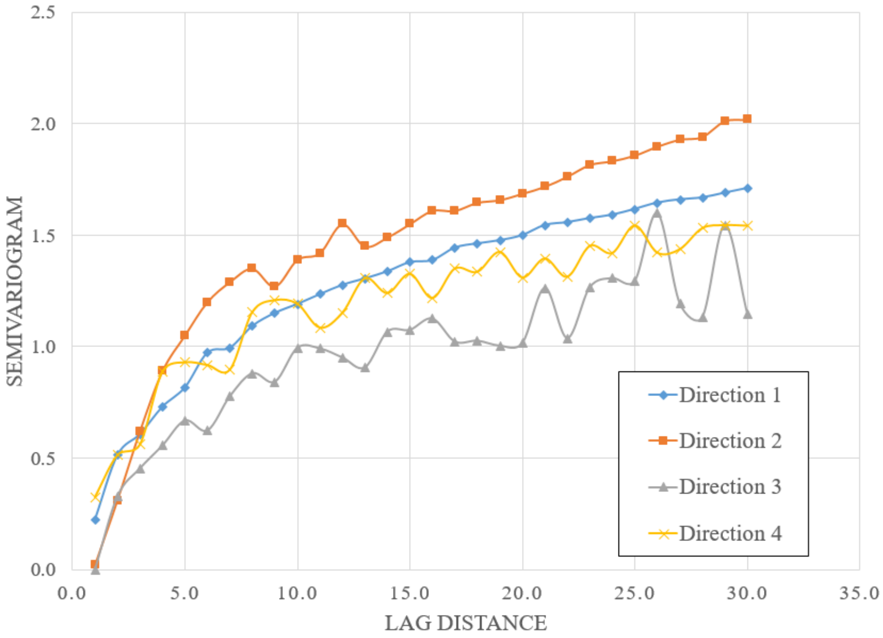


Figure 4-2 Example of Semivariogram directionality

If there is no spatial association with a particular direction, then it is said to be isotropic and that the spatial structure is omnidirectional and that there is no spatial directionally to the variance structure in the semivariogram. For this thesis, the data is assumed to be isotropic meaning that there is no directionality to the data points and that the data sets are all omnidirectional.

Once the sample variogram points are calculated, they are plotted and then used to fit a mathematical semivariogram model that best fits the data structure. This is known as the fitted or functional semivariogram model. Regardless of the mathematical model used, the fitted semivariogram model provides three important values: the nugget, the range, and the sill. The nugget is the measurement error or variations associated with the data collection and manifests as the y-intercept value, even though it should idealistically go through the origin. Data collection and measurements are seldom perfect thus the nugget will usually have an effect on the value of the sill, shifting it upwards. The range is the distance value where the level of dissimilarity flattens out or reaches a plateau. The sill is the value of this plateau and is the point where the level of dissimilarly no longer changes. It is at this point where the dissimilarity between data points is at its maximum and points that are separated by more than the range value are no longer considered spatially correlated. Figure 4-3 illustrates both the experimental and functional semivariograms and where the nugget, range, and sill values come from.

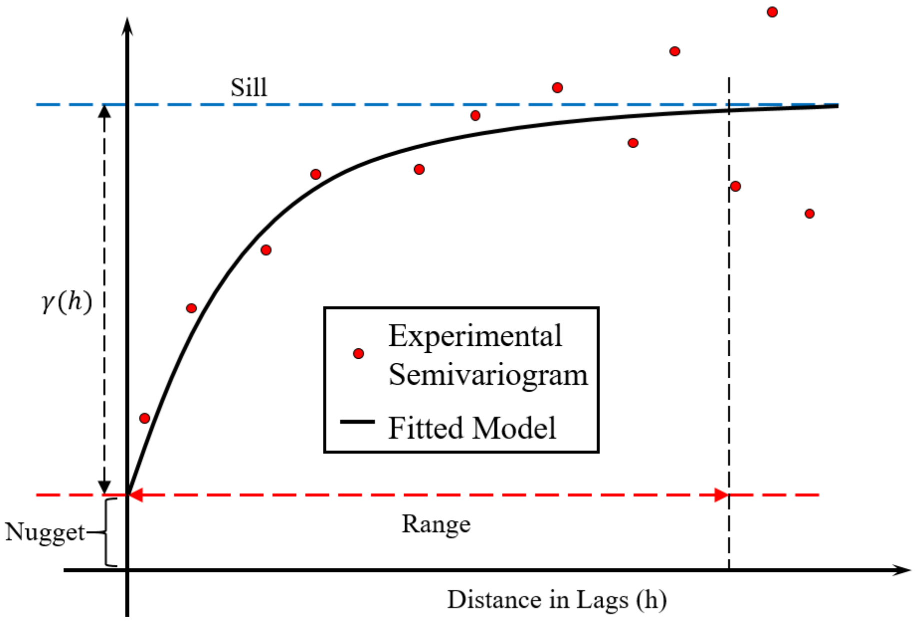


Figure 4-3 Example of a typical semivariogram plot

As logic dictates, the level of dissimilarity between two data points increases with their separation distance and is often reflected in both the empirical semivariogram plot and fitted model as illustrated in Figure 4-3.

Of the many functions that have been developed to fit the experimental semivariogram, the three most common functions are spherical, exponential, and Gaussian. Other less-commonly used functions are the power, cubic, sine-hole, and pentaspherical (Olea, 2006).

Table 4-1 below lists the commonly used semivariogram models used and their formulation.

Table 4-1 commonly used Semivariogram models and their formulas

Spherical	$\gamma(h) = \left\{\begin{array}{ll} C\left(\frac{3h}{2a} - \frac{1}{2}\cdot\left(\frac{h}{a}\right)^3\right) & ,0 \le \lvert h \rvert \le \lvert a \rvert \\ C & ,\lvert a \rvert \le \lvert h \rvert \end{array}\right\}$
Exponential	$\gamma(h) = C\left(1 - e^{-\frac{3h}{a}}\right)$
Gaussian	$\gamma(h) = C\left(1 - e^{-3\left(\frac{h}{a}\right)^2}\right)$
Cubic	$\gamma(h) = \left\{\begin{array}{ll} C\left(7\left(\frac{h}{a}\right)^2 - 8.75\left(\frac{h}{a}\right)^3 + 3.5\left(\frac{h}{a}\right)^5 - 0.75\left(\frac{h}{a}\right)^7\right) & ,0 \le \lvert h \rvert < \lvert a \rvert \\ C & ,\lvert a \rvert \le \lvert h \rvert \end{array}\right\}$
Power	$\gamma(h) = \alpha h^{\beta}\ , \qquad 0 < \beta < 2$
Sine hole	$\gamma(h) = C\left(1 - \frac{\sin\left(\pi\frac{h}{a}\right)}{\pi\frac{h}{a}}\right)$
Pentaspherical	$\gamma(h) = \left\{\begin{array}{ll} C\left(\frac{15h}{8\alpha} - \frac{5}{4}\left(\frac{h}{a}\right)^3 + \frac{3}{8}\left(\frac{h}{a}\right)^5\right) & ,0 \le \lvert h \rvert < \lvert a \rvert \\ C & ,\lvert a \rvert \le \lvert h \rvert \end{array}\right\}$

It is important to note that the accuracy of the semivariogram is highly dependent on the number and quality of the data points. There is no consensus for the minimum number of points and varies a lot in literature and can be as low as 25, but a minimum of 100 points is suggested though it can be lower if the data density is well dispersed (Oliver & Webster, 2015).

4.3 Regression Kriging: Formulation & Implementation

There are many variants of Kriging and primarily vary in the assumptions being made and computational complexity. However, all of them combine deterministic and stochastic statistical analysis whereby the deterministic results are used to reduce the uncertainty with the stochastic estimation results (Hengl, 2009). Kriging has expanded from being a linear predictor to now include non-linear attributes resulting in a family of kriging methods (Cressie, 1990). The more common kriging methods include, in the order of complexity, Simple, Ordinary, Universal, and Regression Kriging.

Simple kriging (SK) is the most basic form of kriging that requires many assumptions that limit its accuracy and effectiveness (Olea, 1999). The main assumption made is that sampled values are partial solutions to the random function and that this function is second order stationary which implies that any two variate points are dependent to each other based solely on the Euclidean distance between them (Olea, 1999). Importantly, SK is the only version that assumes that the mean (m) is known and constant which may lead to a biased estimator.

Ordinary kriging (OK), often considered the progression from SK, while still being one of the more simple implementations of kriging and is a functional part of the regression kriging (RK) process. Given the use of OK and RK in this thesis, the details of these two variants of kriging will be covered extensively in this chapter. The mathematical formulation of OK is shown in Equation (5) and the estimator is shown in Equation (6) below (Olea, 1999; Hengl, 2009; ESRI, 2011):

$$Z(x) = \mu + \varepsilon(x) \tag{5}$$

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i \, Z(x_i) + [1 - \sum_{i=1}^{n} \lambda_i]\mu \tag{6}$$

Where $\hat{Z}(x_0)$ is the estimator at the unmeasured location x_0, x_i are measured locations, and λ_i are the weights for the OK estimator that minimizes the variance of the estimator and the mean squared error (MSqE).

In OK, the mean μ is assumed to be unknown but constant and through some algebra, it ultimately filtered out with the weights' constraint which results in the estimator being Equation (7) with its estimation error variance function being Equation (8).

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i \, Z(x_i) \tag{7}$$

$$\sigma^2(x_0) = 2\sum_{i=1}^{n} \lambda_i \, \gamma(x_i, x_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i, x_j) \tag{8}$$

The values of the weights become an optimization problem that requires the use of a Lagrangian function as the objective function (Olea, 1999). For OK, the Lagrange multiplier is shown in Equation (9) below, where $\sigma^2(x_0)$ was defined in equation (8).

$$L(\lambda_1, \lambda_2, \dots, \lambda_n; \mu) = \sigma^2(x_0) + 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{9}$$

Then with the Lagrangian function defined, then the weights will be the solution to equation (10) with $\gamma(h)$ being any of the negative definite semivariogram functions chosen (some are listed in Table 4-1):

$$\left.\begin{array}{c} \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_1) - \mu = \gamma(x_1, x_0) \\ \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_2 - \mu = \gamma(x_2, x_0) \\ \dots \\ \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_n) - \mu = \gamma(x_{\mathrm{n}}, x_0) \end{array}\right\} \tag{10}$$

Subject to: $\sum_{i=1}^{n} \lambda_i = 1$

The biggest difference between SK and OK is found in the assumption of the mean value and the way the weights are calculated. The weights applied to the measured data is constrained such that all weighting values λ_i must sum to 1. This constraint “filters out” the mean from the estimator due to the restriction on the weights summation resulting in the simplified equation 3 (Olea, 1999; Hengl, 2009). This greatly reduces the complexity of the process but still assumes that the mean is constant despite being unknown and filtered out (Olea, 1999). Naturally, OK shares many of the properties of SK but does benefit from the estimator not being biased given the weight constraint.

However, over a large space the assumption of a constant mean does not account for possible zonal means. Therefore, rather than assuming a constant mean, the mean is assumed to be constantly changing and can be represented by a function resulting in a kriging variant where there is the assumption that there is an inherent drift or trend in the data that needs to be accounted for. It goes by the terms Universal Kriging (UK), Regression Kriging (RK), and kriging with external drift (KED) and all represent the same technique, but with minor differences between them. This often causes some confusion with terminology, but RK can be considered the general term that encompasses UK and KED, where the true differences between them actually exist (Hengl, et al., 2004). Another way to understand it is that UK and KED are special cases of RK where UK models its drift or trend as a function of the coordinates only, while KED will incorporate or use other auxiliary values that may or may not be locational in nature. When combining both aspects of UK and KED is it then truly RK (Hengl, 2009).

The drift of the deterministic values can be modeled by using a linear combination of functions, usually polynomial functions of locational attributes (Olea, 1999). Using polynomial functions, the weights for UK are determined by minimizing the mean square error by employing Lagrange multipliers to determine the weights (Olea, 1999; Hengl, 2009). RK makes use of regression modelling to construct a regression function to detrend any external drift using as many covariates as deemed relevant by the modeler. Formulaically it is represented by equation (11) as follows:

$$z(x_0) = m(x_0) + e(x_0) \qquad \textbf{(11)}$$

Where $m(x_0)$ is the predicted value from linear regression and $e(x_0)$ is the residuals that are interpolated using ordinary kriging (Hengl, et al., 2004; Hengl, 2009). Expanding Equation (6) gives Equation (12) (Hengl, et al., 2004)

$$\hat{z}(x_0) = \sum_{i=0}^{n} \widehat{\beta_\iota} \cdot q_i(x_0) + \sum_{i=0}^{n} \lambda_i(x_0) \cdot r(x_i) \quad \textbf{(12)}$$

Where:

$\widehat{\beta_\iota}$ = model coefficients
$q_i(x_0)$ = auxiliary variables
$\lambda_i(x_0)$ = covariance weights
$r(x_i)$ = regression residuals

From the regression model, the estimates and residuals are calculated, interpolated, and then added back into the estimated values to obtain the spatially fitted values (Zhu & Lin, 2010). This method serves to spatially correct or detrend some model variance (i.e., removing external influences) resulting from the regression model in order to reduce the mean square error value thus optimizing the model (Oliver & Webster, 2015). This method also allows for the inclusion of auxiliary variables for study within kriging's spatial modeling method thus providing insight into the influence of covariates. It is for this reason RK can be used in winter collision modelling while gaining additional information and understanding of the influences from auxiliary variables such as weather conditions, road surface conditions, etc. This use of covariates and spatial correlation combines the best attributes of both a deterministic and stochastic estimation process.

Figure 4-4 outlines the regression kriging process utilized as a flowchart and was adapted from Peng et al (2013) in their study of spatial distribution of organic soils (Peng, et al., 2013).

The validation set in this case would be each data point via the leave-one-out crossvalidation process. These calculations were handled within R using functions and algorithms found in the following packages. Package *gstat* was used for the variogram and semivariogram analysis, OK interpolation, and leave-one-out crossvalidation interpolation (to be further discussed in the next section). The *sp* package has built-in functions pertaining to spatial data that are required of by gstat. In this case it was used to define coordinates and location data and set the imported data matrix into a spatial dataframe. The *MASS* package is a set of quality-of-life supportive functions that make statistical analysis simpler and more intuitive to implement. This package was primarily used to simplify basic commands to run multiple linear regression and output results into text or csv files. The *car* package is known as the companion to applied regression package and was used to calculate the VIF values. The *MLmetrics* package is a supplementary machine learning package with additional statistics and was used to calculate the RMSE value.

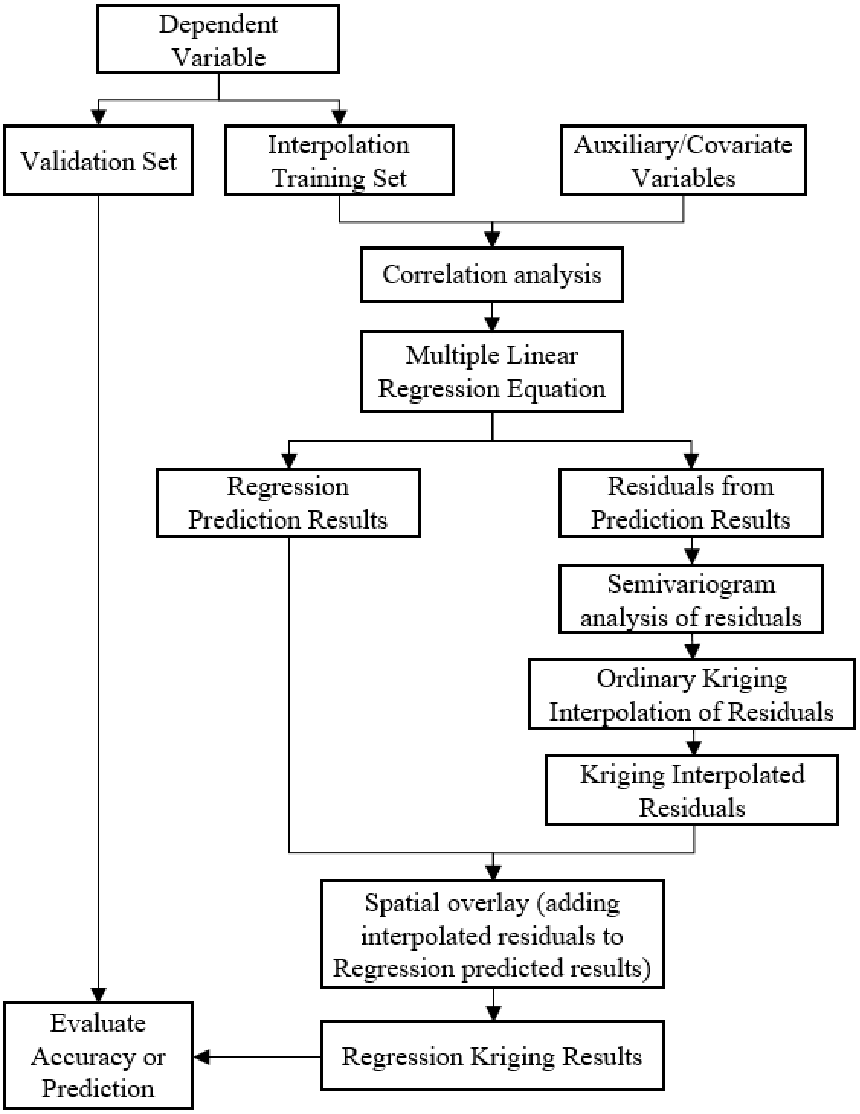

Dependent Variable
Validation Set
Interpolation Training Set
Auxiliary/Covariate Variables
Correlation analysis
Multiple Linear Regression Equation
Regression Prediction Results
Residuals from Prediction Results
Semivariogram analysis of residuals
Ordinary Kriging Interpolation of Residuals
Kriging Interpolated Residuals
Spatial overlay (adding interpolated residuals to Regression predicted results)
Evaluate Accuracy or Prediction
Regression Kriging Results

Figure 4-4 Regression Kriging Process Flowchart as adapted from Peng, et al (2013)

The Correlation and MLR process was covered in Section 4.1 the results from this part is used in the key part of regression kriging. Recall that the MLR is a linear approximation of the dataset and the deviation from this line results in the residuals of the MLR model. It is these residuals that are what will be spatially modelled. The newly modeled residuals are the predicted residuals and these are then added back into the residual estimates to obtain the final regression kriging estimated value. Figure 4-5 illustrates this concept.

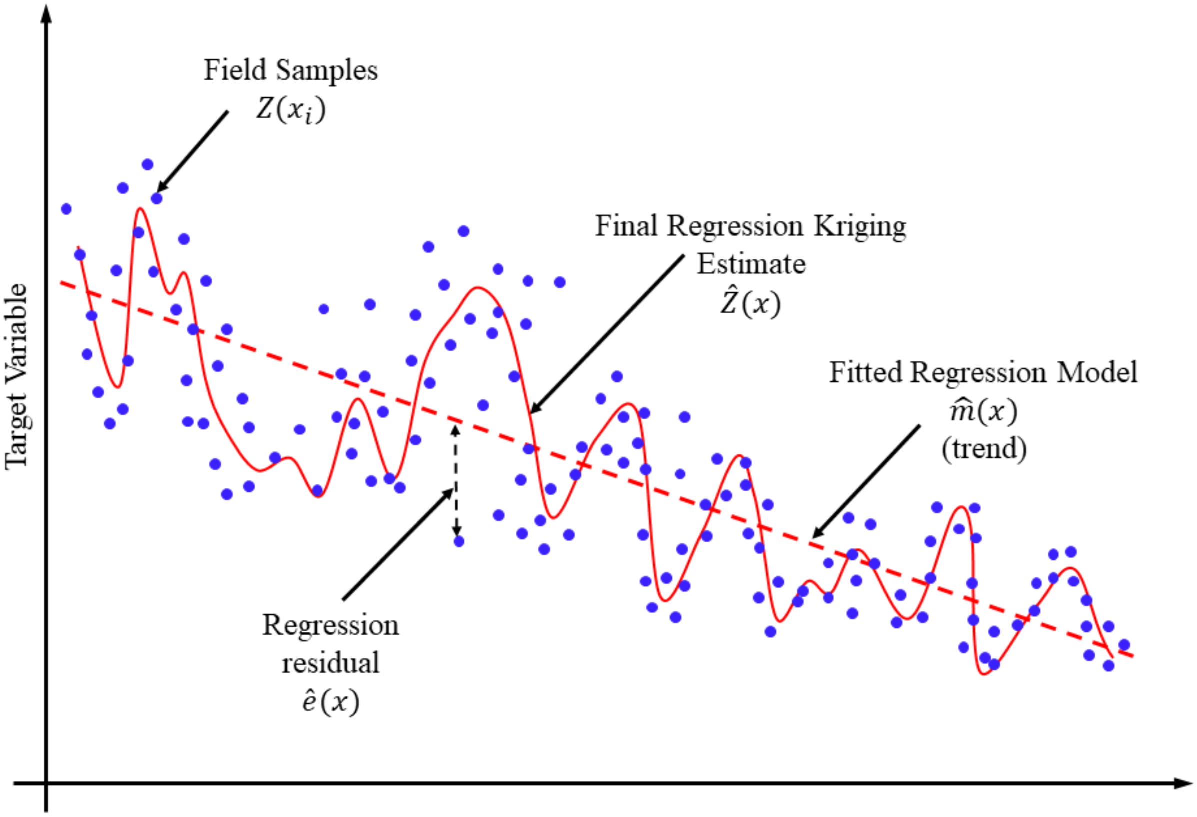


Figure 4-5 Illustration of Regression Kriging

Crossvalidation

Crossvalidation is often used as a measure of how accurate and reliable the generated model is and is how the model is evaluated. There are many ways to conduct a crossvalidation and the Leave-One-Out (LOO) methods is popular with kriging modeling. LOO is done by removing one data point and then using the model and the surrounding variables to calculate an estimate at that point, and also the associated variances (Oliver & Webster, 2015). This is repeated for all data points in the dataset and the results are then used to calculate comparison statistics using five error metrics as summarized in Table 4-2 (Hengl, 2009; Oliver & Webster, 2015; ESRI, 2011).

The MSqE and MStdE are often used to measure the quality of an estimator and the closer it is to zero (0), the better the estimator is. The RMSE is used to measure the accuracy of a model and the smaller the value, the better the model is. The ASE is the average standard deviation and should be close to the RMSE value. RMSSE is used to examine the variability of the estimations (under or overestimations) and should ideally be close to 1. If the RMSSE value is greater than 1, then the variability of the predictions is being underestimated, and vice versa (ESRI, 2011). By using these five metrics, the various kriging models can be confidently compared and contrasted against each other.

Table 4-2 Statistical Measures for Model Performance

Mean Squared Error	$MSqE = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{Z}(x_i) - Z(x_i)\right]^2$
Mean Standardized Error	$MStdE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{Z}(x_i) - Z(x_i)}{\hat{\sigma}^2(x_i)}\right]$
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Z}(x_i) - Z(x_i)\right]^2}$
Average Standardized Error	$ASE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^2(x_i)}$
Root Mean Squared Standardized Error	$RMSSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{Z}(x_i) - Z(x_i)}{\hat{\sigma}^2(x_i)}\right]^2}$

4.4 Enhanced Regression Kriging using Network Distances

The distance between any two points on a road network is bound by the roads connecting the two points. This network distance is not always the same as the Euclidean distance between them, especially if the points are separated by a wall, structure, or one-way streets (refer to Figure 1-2 from Chapter 1 that best illustrates this point). Therefore, the primary assumption that the Euclidean distance between any two points is the basis for the underlying covariance spatial structure may not be an accurate one for transportation engineering problems. But to confirm this hypothesis, a comparative analysis needs to be done between Euclidean distance results and Network distance results.

Overall, all the same steps are taken as described in the previous sections using the Euclidean distance with the addition of generating a semivariogram and kriging model using Network distances. The semivariogram model values between the two are recorded as are the interpolated model outputs and crossvalidation results. As per previously, the same crossvalidation metrics are calculated to compare the Network Distance models to the Euclidean distance models. To ensure appropriate comparisons, the same dataset is used in both cases.

The semivariogram and crossvalidation analysis using network distances was done via a combination of ArcMAP, python and R. First, an origin-destination (OD) matrix is required to obtain the network distances between all pairs of points in the study area. From ArcMAP, the OD matrix is generated by using the Network Analyst package built into the program. It is important to note that the distances between any two pairs of points is potentially unique. That is to say, the distance from A to B is not necessarily the same as B to A should there be constraints such as one-way streets or points on expressways with limited access ramps. Therefore the total number of OD pairs will always be n^2 where n is the number of points on the road network. This can quickly increase the computational requirements of the network trace as the number of OD pairs will increase quadratically for every additional point increasing the calculation times significantly. It is for this reason that this comparison is done on a sub-region and not for the entire state, and the region chosen is mostly a rural setting with a smaller density of road which will help keep the number of points down.

To generate the semivariance points, Python (Van Rossum & Drake, 2009), was used to do the binning and correlation calculation of the data before that information was passed to R to generate the semivariograms. The pseudo code for the Python process is as follows:

```
Python Pseudocode: variogram for R
data_file = file with the variates, and their point IDs
distance_file = csv file of all the OD distance pairs


define:
```

```
lag_size = (max distance from OD list) / (# of lags)

function: semivariance
   1) Calculate variance for every pair of points and their
      network distance
   2) Calculate bins limits from lag size
   3) Partition variance values into each bin and calculate
      the semivariance value for that lag bin

Return variogram csv file
```

The variogram table is then loaded into R for semivariogram model fitting. This was done using the gstat package for its vgm and vfit functions that will fit a variogram model to the sample variogram points. The pseudo code for this process is as follows:

```
R Pseudocode: Construct Semivariogram

roadfile = variogram csv from Python
for (model in list (Spherical, Gaussian, Exponential)){
     vfit = fit.variogram(roadfile, model)
     img = plot(roadfile,vfit)
     print(img) to jpeg file
     return csv with variogram values (nugget, sill, range)
}
```

The semivariogram results from R, are then used back in Python to do the model estimation and crossvalidation calculations. The crossvalidation results are recorded in a text file for easier record keeping and the estimates are outputted into a csv file. The estimates are then loaded back into ArcMAP to generate the surface map of the estimates. Once all the calculations are done, it is only a matter of putting the results side by side and comparing them.

The main issue of this method arises from the computation complexity just due to the sheer amount of OD pairs that need to be found and then used for modelling. This takes a lot of computing power and time to process thus making it potentially prohibitive to do as the smaller datasets being used is still over 1000 points and takes over 3 hours to compute just the crossvalidation. As stated above, a mostly rural zone would help in reducing the number of data points. But in this case, the road network was further simplified to reduce the number of data points even more as even its base condition still incurred an initial run time of over 60 hours. The most ideal location for this process was the Northcentral zone as it does not contain any major municipality, it is mostly a rural zone with some major state freeways and sections of interstate present.

4.5 Characterization of Underlying Spatial Structures

A key part of kriging is the assumption that the spatial structure of the data is the same throughout the entire region. By definition, it states that the intrinsic variance structure of the dataset is the same regardless of the translation throughout the area. This is the second order stationarity assumption (SOSA) and is required when making the interpolations for the whole study area. However, past studies utilizing kriging seldom went beyond a single municipality or county. Given the size of Iowa being near the size of their study area within Ontario, then a similar analysis should done to ensure that the best models are being obtained for use.

As stated in section 4.2, the underlying spatial structure is characterized by the semivariogram which can be used as a point of comparison for models encompassing various spatial ranges. The regional size would be the entire state of Iowa, and the zonal spaces will be the areas identified in Figure 3-1(b) and (c). The partitions were done in this fashion as it neatly divided the state into 4 quadrants along existing county lines. The Northcentral zone was partitioned for the reasons outlined in section 4.4. This will provide four zones (NE, NW, SE, and SW) with the same data density as the region and one zone (Northcentral) with a reduced data density, but using the same base dataset. The comparison values will be the five statistical measures used in the previous sections and will be based on the OK crossvalidation results and the RK crossvalidation results for each zone. The OK and RK results from Section 4.3 will be used to represent the region. The OK and RK models are developed in the same fashion as for the region.

By analyzing the semivariograms produced for the various spatial extents, it should reveal whether or not a single regional semivariogram is appropriate for the entire state of Iowa, or if zonal semivariograms must be generated in order to maximize the potential model benefits.

4.6 Summary

Presented here is the framework and methodologies for the core of this thesis to address the three objectives, which are to 1) develop and expand the hybrid geostatistical method known as regression kriging for use in modelling winter traffic safety problems; 2) enhance the estimates from this hybrid method by using network distances, and 3) characterize the underlying regional and zonal spatial structures.

The first objective saw the generation of a RK model by conducting a MLR analysis. This process will determine what regression model will be used to calculate the residuals for the RK estimations. As an important feature of geostatistics, especially kriging, the semivariogram was introduced along with its calculation, interpretation, and usage. With the spatial structure understood, a preliminary OK analysis was done to set a point for comparison for RK so see if it the estimation model is the better performer. Once RK has been determined to be the better performing method, the enhancement of its estimates by substituting network distances was explored next. This was done by implementing the same RK analysis method, but on a smaller more manageable study area, as done for the first objective, only the comparison being made is between RK with Euclidean distances and RK with Network Distances. The final objective looks to examine the underlying spatial structure and thus the SOSA by again, conducting a comparative analysis of the semivariogram, OK crossvalidation, and RK crossvalidation results only this time between the region, defined as the entire state of Iowa, and the zonal areas defined in Section 3.1.

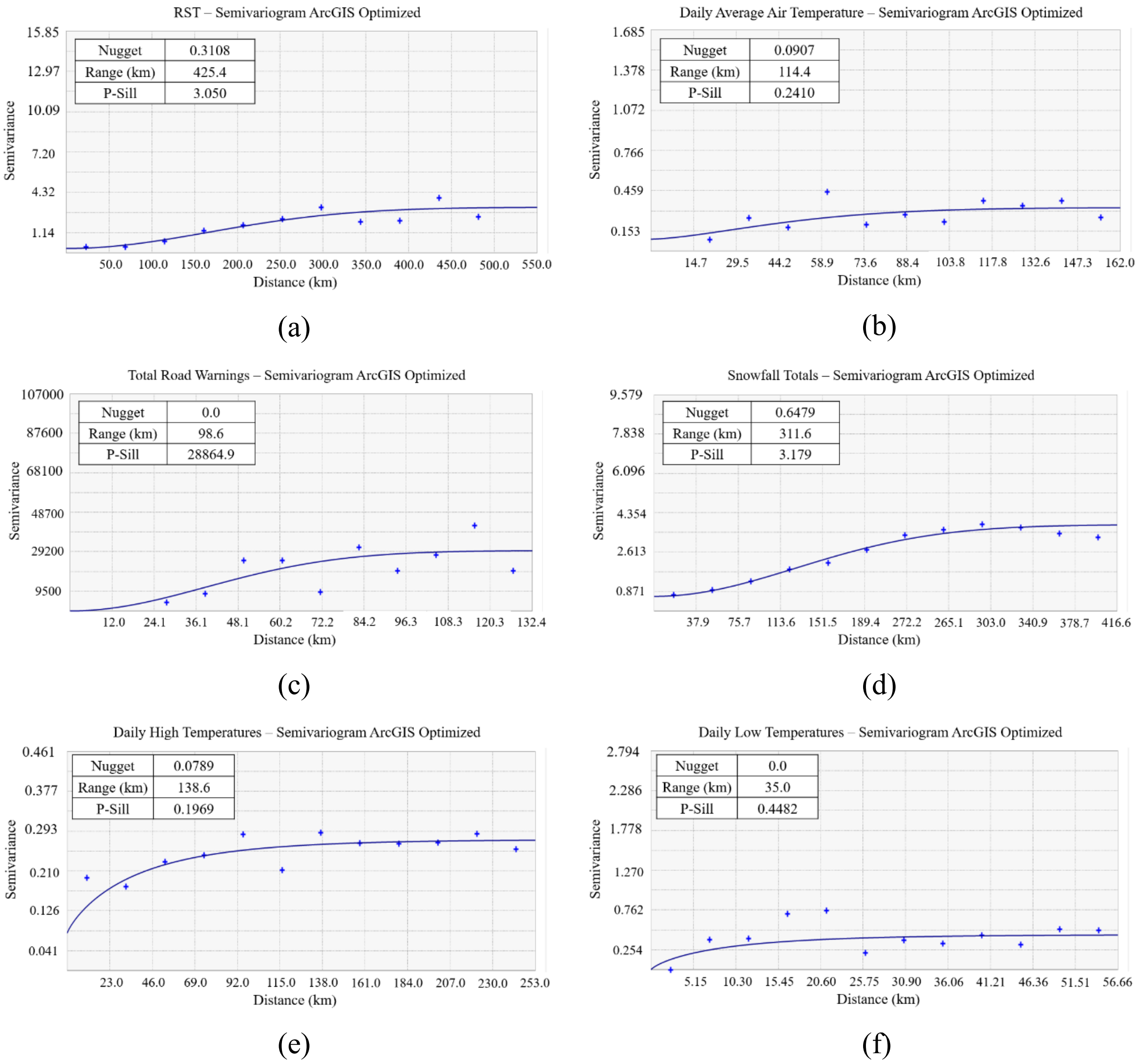
Chapter 5 RESULTS AND DISCUSSION

This chapter presents the results of the analysis and findings from the methodology outlined in Chapter 4. Section 5.1 details the interpolation of the road surface and environmental conditions to ensure every road segment has those values on them. Section 5.2 will discuss the multiple linear regression (MLR) work done and the results from it. Details about the coefficients, their magnitude, and sign will reveal details about the covariates and their relation to the variate. These model results will be used in the regression kriging (RK) process. Section 5.3 models the winter collision (WC) ratios with Ordinary Kriging (OK) to set a benchmark for comparison with RK. OK was chosen as the benchmark as through many studies, it has been proven to be an effective spatial estimation tool for transportation engineering problems. Section 5.4 progresses the analysis by generating RK models and their associated spatial structures and their performances analyzed. The results of the RK and OK model performances are then compared to each other. Section 5.5 goes through the results from applying Network distances in place of Euclidean distances and how the results from both compare to each other. Section 5.6 provides an in-depth examination into the second order stationarity assumption (SOSA) and its findings. Section 5.7 summaries this chapter with the work done and its findings.

5.1 Estimating Region-wide Environmental Variables

In order to correlate WC collisions to winter condition measurements, the point measurement nature of the environmental data needs to be spatially interpolated to cover the entire state and thus every road segment. Recall from chapters 2 and 3 that ordinary kriging (OK) was determined to be an effective estimator of environmental conditions (Eguía, et al., 2016; Kwon & Gu, 2017). Therefore, OK was used to calculate a meteorological surface map for each environmental covariate, which was then converted and averaged onto a 1 km x 1 km raster grid over the road network to maintain a high level of granularity by averaging onto the road segments. These environmental covariates were interpolated via ordinary kriging to build a surface map to provide data coverage for all roads in Iowa. This is then mapped onto the road segments along with the collision and road geometry data.

To ensure that the spatial interpolation is valid, there must be a proper underlying spatial structure modeled by the semivariogram that provides a nugget, range, and sill. Though the process has been covered in previous studies it still entails a significant amount of work as it requires spatial analysis, crossvalidation, mapping, and finally projection. The interpolation was done via ArcGIS's ordinary kriging interpolation tool as part of its Geostatistical Analyst toolbox. The semivariograms generated used the stable setting where it would self-determine the best fitting semivariogram function. Figure 5-1 shows the optimized semivariograms for each environmental covariate and their semivariogram values.



(a) (b) (c) (d) (e) (f)

Figure 5-1 Environmental and Road Surface Conditions Semivariograms (a) RST (b) Daily Average Air Temp (c) Total Road Warnings (d) Snowfall Totals (e) Daily High Temp (f) Daily Low Temp

The semivariogram plots all show that there is a spatial structure to the covariance over space, though some are stronger than others. The strongest ones are Figure 5-1(a), (d), and (e) where the semivariance points are closely aligned to the model. To check how reliable the models are, the crossvalidation results are considered. The geostatistics package in ArcMAP provides crossvalidation results with their modelling process. The five statistical measures covered in section 4.3 are used in this process as well. Table 5-1 below show the crossvalidation outcomes for each covariate.

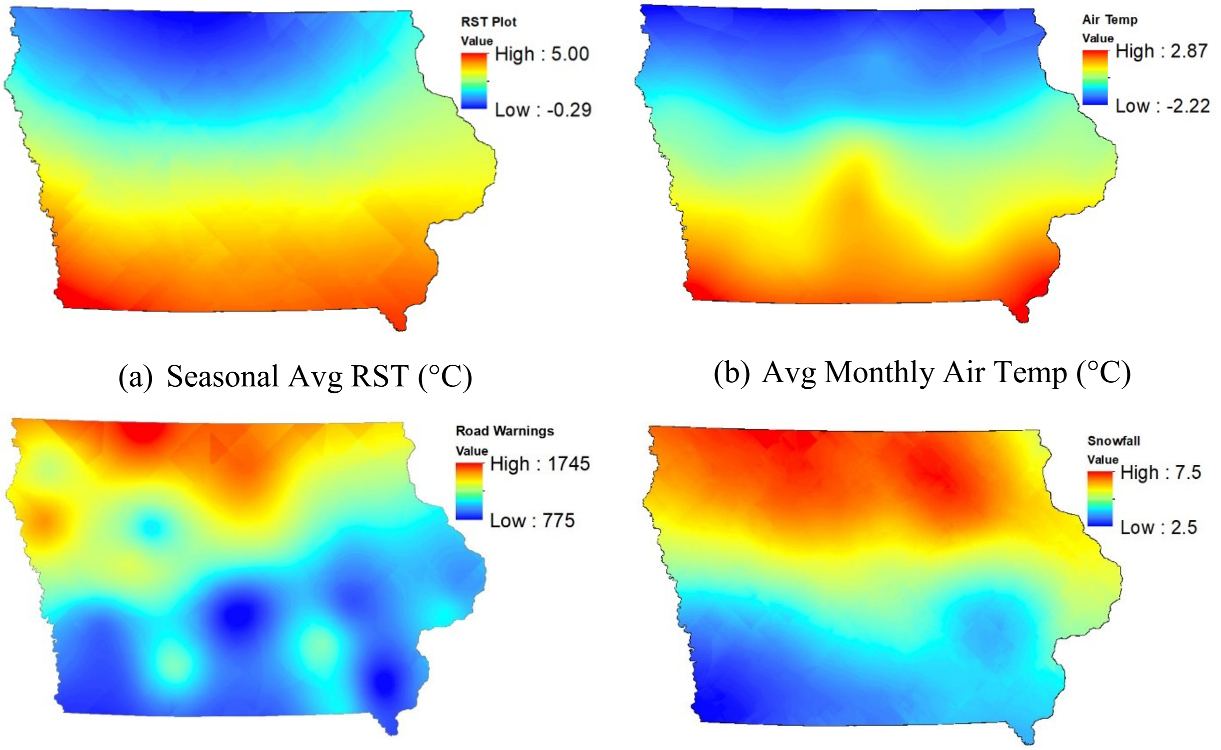
Table 5-1 Covariate Kriging Crossvalidation Statistics

	Road Surface Temp	Average Air Temp	Total Road Warnings	Snowfall Total	Daily High Temp	Daily Low Temp
MSqE	0.0001	0.0018	15.53	0	0	0.0004
MStdE	0.0008	-0.0596	-0.0231	-0.0046	0.0068	0.0159
ASE	0.6766	0.5369	151.3	0.8739	0.4568	0.6755
RMSE	0.5515	0.5367	153.5	0.8857	0.4481	0.5911
RMSSE	0.7884	1.001	1.002	1.019	0.9859	0.8610

The crossvalidation results show that the spatial structure for the total road warnings is highly irregular and not ideal. This could be due to skewness of the data counts or how the road warnings are sent out/determined. It could stem from the side of caution where a warning is issued when not required thus increasing the count of either Red, Orange, or Yellow warnings. However, serving as a surrogate to the RSI value, these covariates are kept and utilized. As for the other variables, the MSqE and MStdE are very small, close to zero, which is near ideal. The RMSE and ASE values also look to be acceptable as having a low RMSE value is ideal and an ASE value close to the RMSE value is desired. And finally, the RMSSE value indicates the level of over or under estimation as represented by how close it is to 1.0. For all the covariates, they are within a few points of 1.0 thus the model estimate variances are well accounted for. The crossvalidation and semivariogram results indicate that these covariate kriging models are acceptable.

Using the models chosen, an interpolation was done to generate a surface map of estimated values for each covariate. Figure 5-2 shows the various interpolated surface maps generated for the covariates being explored.

One clear takeaway from the environmental interpolations is that there is a clear divide between weather and environmental patterns from the north and south halves of the states. Another takeaway from these plots is the fact that there might be collinearity or multicollinearity between these factors when conducting the multiple linear regression (MLR) analysis. There may also be an issue with the Road Warnings as the crossvalidation results of the amalgamation of all the warnings indicate that the interpolation has a high degree of uncertainty in the estimates thus possibly inflating uncertainties or becoming non-significant covariates. The surface map for air temperature interpolation does not show a nice dividing line as with the other plots with a "spike-zone" near the middle. That spike-zone is directly over the capital city of Des Moines, Iowa and the high air temperature in this area may be attributed to the heat island affect that most major cities experience. The road warnings show a very blotchy pattern to the warning counts, however, there is an observable trend where the northern half gets more warnings than the southern half. The two areas of higher warnings in the southern half also coincide with the major interstate highways, while the lower counts are more rural roads, state highways, and arterial freeways. Overall, it still represents the network well.

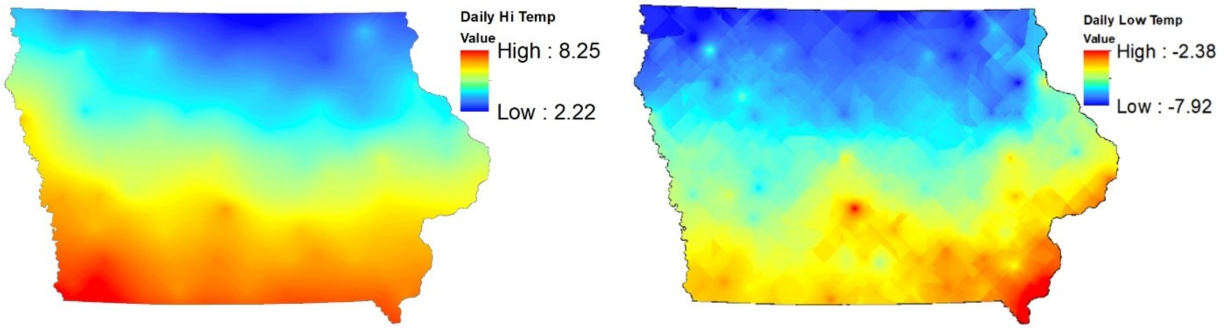


(a) Seasonal Avg RST (°C)

(b) Avg Monthly Air Temp (°C)

(c) Sum of Red, Orange, and Yellow Road Warning Messages

(d) Average seasonal snowfall totals (cm)



(e) Avg daily high temp (°C)

(f) Avg daily low temp (°C)

Figure 5-2 Spatially Interpolated Surface map of environmental and surface condition data

5.2 Deterministic Modelling of WC Ratios via MLR

Multiple Linear Regression (MLR) analysis was completed for the state of Iowa, and its sub regions as defined in Chapter 3. Using R, an open source statistical analysis package (R Core Team, 2020), a MLR was completed using a backwards removal process to determine the statistically relevant covariates. A confidence of 95% (α = 0.05) was used as the cutoff for statistical significance. The resulting statistically significant variables for each model need to be checked first for multicollinearity before the model can be considered statistically relevant. For the check, the VIFs of each significant variable within each model were calculated. Recall that any VIF value above 10 would signify that the associated variable shows multicollinearity with at least one other variable thus inflating its variance contribution. Table 5-2 below summarizes the statistically significant covariates and the values of their coefficients, and Table 5-3 summarizes their final VIF values as a check for multicollinearity. A discussion about the relevant variables, their signs, coefficient magnitudes, and intuitiveness follows afterwards.

During the regression analysis, RST was found to be a statistically significant variable in most of the regression models. However, as can be seen in the final models (Table 5-2), RST is only present in the overall regional model. This is because for the sub regions, RST had a large VIF value, much greater than 10, for those regression models indicating severe presence of multicollinearity. Once RST was removed, the models' VIF results were much better. Refer to Appendix A to see the VIF results for RST. As such, RST was not included in the models as the other variables were found to be more relevant. Perhaps it could be used in place of the other variables should they not be available. With the final models determined, the model details can now be discussed.

Table 5-2 Regression Results for Iowa State and its sub regions

Coefficient Values	Iowa State	Northwest	Northeast	Southwest	Southeast	North Central
Number of Data Points	19591	3257	6284	2565	7504	1090
Adjusted R^2	0.0355	0.0190	0.0389	0.0390	0.0182	0.0403
Intercept	0.1182	-0.0839	-0.4897	-0.1691	0.0521	-0.5572
Number of Lanes	-0.0254	N/A	-0.0217	-0.0237	-0.0305	-0.0475
Speed Limit	0.0013	0.0015	0.0009	0.0020	0.0009	N/A
ln(AADT)	0.0165	N/A	0.0258	0.0220	0.0205	0.0470
RST	-0.0418	N/A	N/A	N/A	N/A	N/A
Avg. Air Temp	0.0397	N/A	N/A	N/A	-0.0300	N/A
Seasonal Snowfall Total	N/A	0.0558	0.0226	N/A	N/A	N/A
No. of Red Warnings	N/A	-0.0004	0.0001	0.0014	N/A	0.0002
No. of Orange Warnings	0.00001	N/A	0.0002	0.0003	N/A	0.0005
No. Of Yellow Warnings	0.0009	N/A	0.0040	-0.0074	0.0010	N/A

The results of MLR for all models show a very weak R^2 value indicating that regression alone is not a very good estimator with all of them being below 0.05. However, some information can still be gleaned from the results. In general variables with a positive correlation to WC ratios for all models including the speed limit, AADT, and orange stage warnings whereby indicating that if these values increase, so does the WC ratio. This makes intuitive sense as previous literature has shown that an increase in exposure (AADT) and speeds tend to lead to increased collisions in general, and under the hazards of winter conditions, they are more likely to occur (Andersson, 2010; El-Basyouny & Sayed, 2006; Usman, et al., 2012; Abdel-Aty & Radwan, 2000). It is worthwhile mentioning that the magnitude of their coefficients appears to be small as it needs to translate the covariates such as Speed Limits and Warning Counts from values that are well above 50 or 200, respectively, to values that fit within the WC ratio range between 0 and 1.

Table 5-3 VIF values for the MLR models

VIF Values	Iowa State	Northwest	Northeast	Southwest	Southeast	North Central
Number of Lanes	1.2061	N/A	1.2280	1.1146	1.2517	1.0255
Speed Limit	1.0901	1.0070	1.1451	1.0854	1.1034	N/A
ln(AADT)	1.2478	N/A	1.3203	1.1077	1.1526	1.3096
RST	2.1797	N/A	N/A	N/A	N/A	N/A
Avg. Air Temp	N/A	N/A	N/A	N/A	1.0149	N/A
Seasonal Snowfall Total	N/A	1.1602	3.7250	N/A	N/A	N/A
No. of Red Warnings	N/A	1.1610	1.9763	1.4599	N/A	1.0580
No. of Orange Warnings	2.2972	N/A	3.8264	1.2943	N/A	1.2320
No. Of Yellow Warnings	1.0972	N/A	1.8098	1.1603	1.0234	N/A

The number of lanes showed a negative correlation which indicates that as the number of lanes increase, the WC ratio decreases. This also makes sense as the increased space reduces the chance of a WC collision, or it could mean that more collisions occur during non-winter conditions as more lanes are indicative of roads that carry higher speeds and traffic volumes.

Another interesting outcome from this analysis shows that different variables are relevant in different regions. For example, snowfall totals seem significant for the north half of Iowa only. This makes sense as the northern half of Iowa tends to be colder than the south making it more susceptible to snowy weather events. Road warnings messages do not seem to follow any discernable pattern for significance. This may be a result of local trends in driver behavior reacting to these warnings, or how local maintenance crews react differently to the warnings thus possibly affecting the collision rates.

The less than ideal R^2 values indicate that MLR leaves a lot unaccounted for in its model. This means that there is room for improvement thus opening up the possibilities for geostatistics to be used to improve the WC ratio estimates for network screening.

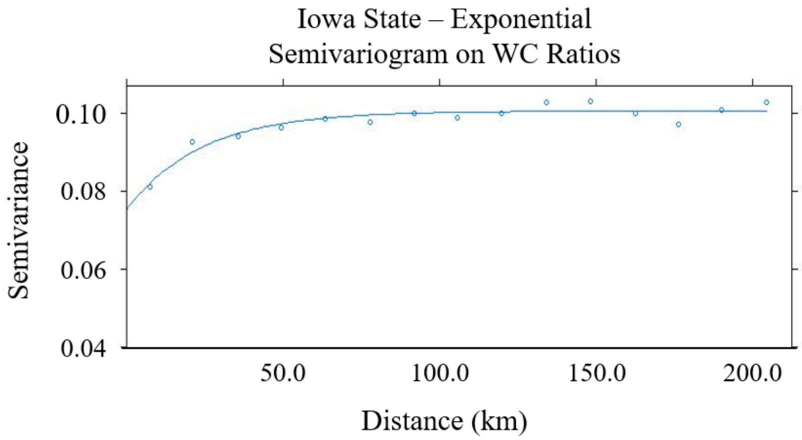
5.3 Stochastic Modelling with Ordinary Kriging

Prior to kriging interpolations, the spatial structure, as represented by the semivariogram, of the area must first be analyzed and formulated. This process was completed within R using the *gstat* and *sp* library packages. Three semivariogram models were initially chosen to be the basis for the spatial structure analysis and from these three, one will be used to conduct the kriging interpolation and its model performance assessed. As the variate of interest is the WC ratio, a baseline for comparison is done using Ordinary Kriging on WC ratios first. Here, the exponential, Gaussian, and spherical semivariograms were generated for the state of Iowa as a whole. Figure 5-3 shows the three semivariograms for the WC ratio while Table 5-4 summarizes the semivariogram values for each of the semivariograms generated.

Table 5-4 WC Ratio Semivariogram Variables and Crossvalidation Analysis results

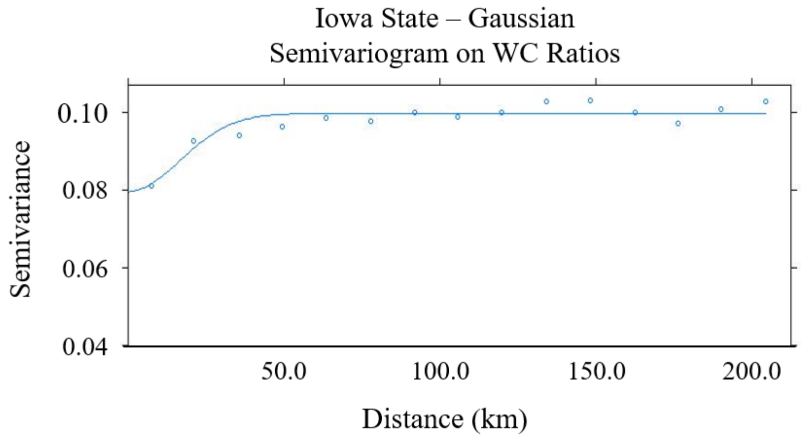
	Iowa State Ordinary Kriging		
Iowa State	Exponential Model	Gaussian Model	Spherical Model
Nugget	0.069	0.080	0.080
Range (km)	15.000	19.452	69.013
P-Sill	0.030	0.018	0.020
MSqE	0.098	0.098	0.100
MStdE	0.000	0.001	0.000
ASE	1.081	1.029	1.113
RMSE	0.313	0.313	0.316
RMSSE	0.612	0.436	0.752
No. of Pts		19591	

(a) Exponential Semivariogram

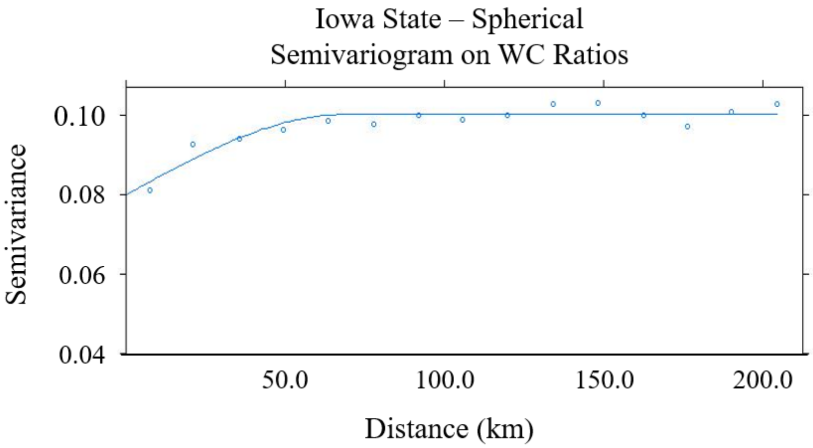


Iowa State – Exponential Semivariogram on WC Ratios

(b) Gaussian Semivariogram



Iowa State – Gaussian Semivariogram on WC Ratios



(c) Spherical Semivariogram

Figure 5-3 Semivariograms for WC Ratios for the State of Iowa

Looking at the crossvalidation results across the three models used, it appears that the three perform quite similarly. However, when looking at the RMSSE value, is apparent that the Gaussian model is the weakest of the three semivariogram models in accounting for the variance in the estimates. Going by the same statistical measure, then it appears that the Spherical model is the best performing one of the three. Here, it is good to also look at the semivariogram values from the model. As shown, the spherical model has the largest range, which means that its structure is able to account for more correlations before it becomes no longer significant. This would mean that this model is effective for a larger area. The tradeoff would be the larger nugget value implying that there may be more inherent measurement error that is not account for. Based on the results of Table 5-4, the spherical model is the best performing model of the three and thus has been used to create the surface map of estimations. Figure 5-4 shows the interpolated map of the WC ratio estimations using the OK model developed in this section.

From this map projection, it is obvious that there is a locational attribute to the WC ratio. Higher ratios tend to be on the northern half while the southern half have lower WC ratios. Another interesting pattern resulting from this mapping is that the urban centers all show a low to very low WC ratio as compared to rural roads. This highly suggests that the effects of winter conditions on collisions is greater outside of urban centers. One of the benefits of implementing any kriging estimate is that it comes with variance/error values, providing a measure of uncertainty to the estimates. Figure 5-5 is a plot of the estimation variance from the OK model estimation.

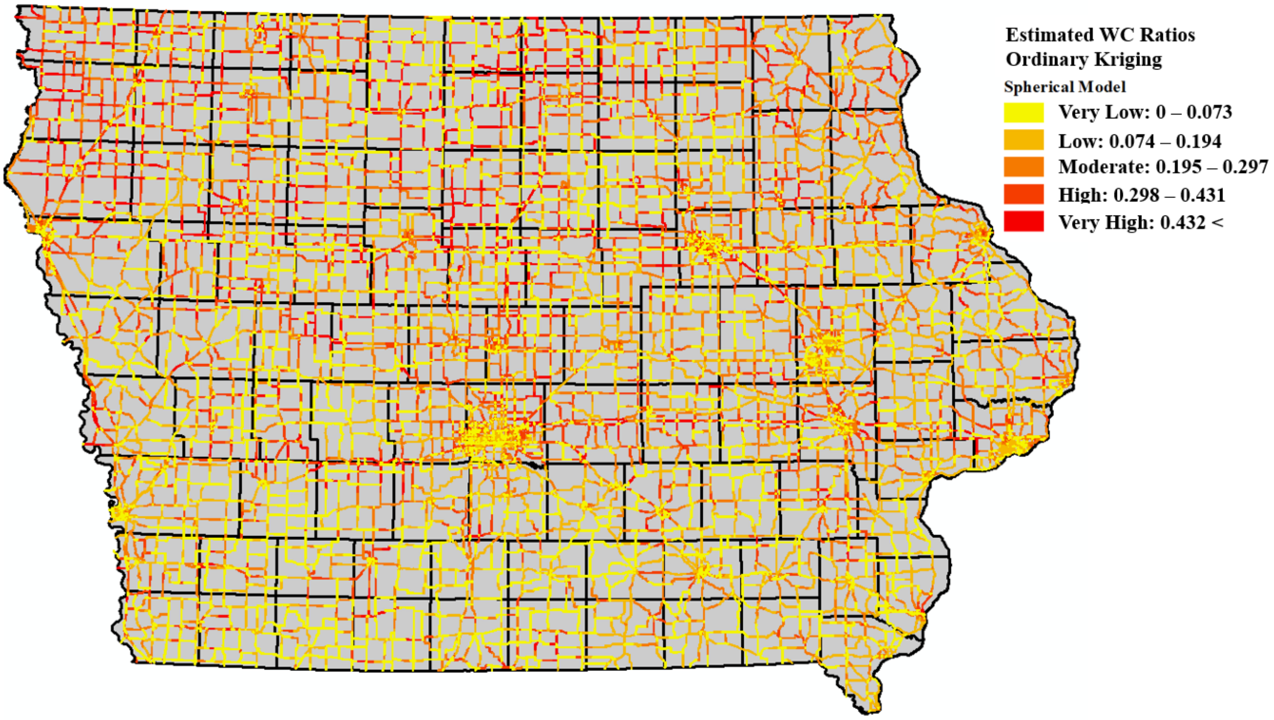


Figure 5-4 Map of OK Interpolated WC Ratios

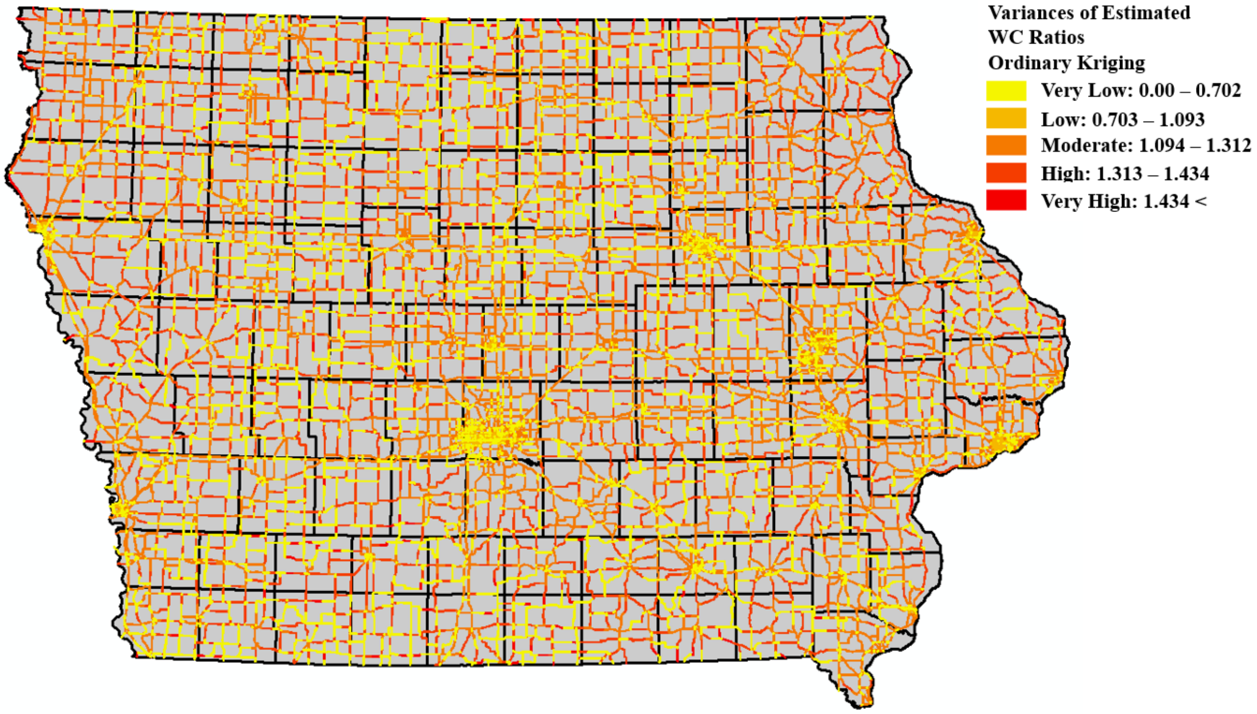


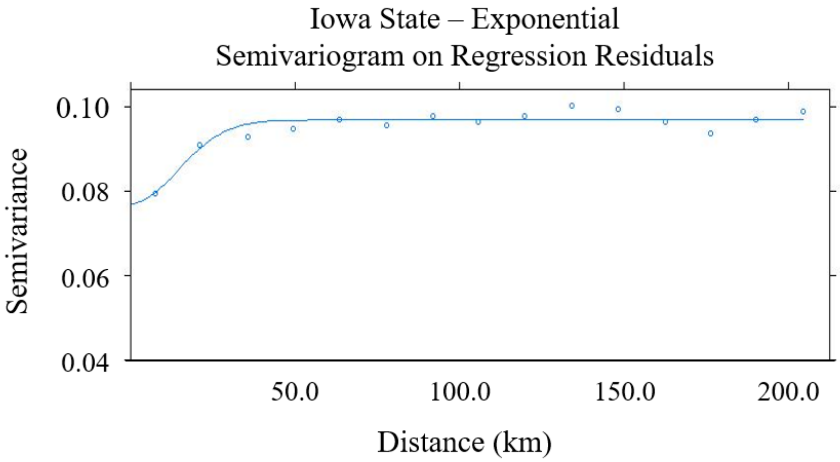
Figure 5-5 Map of the OK estimate variance

5.4 Stochastic Modelling with Regression Kriging

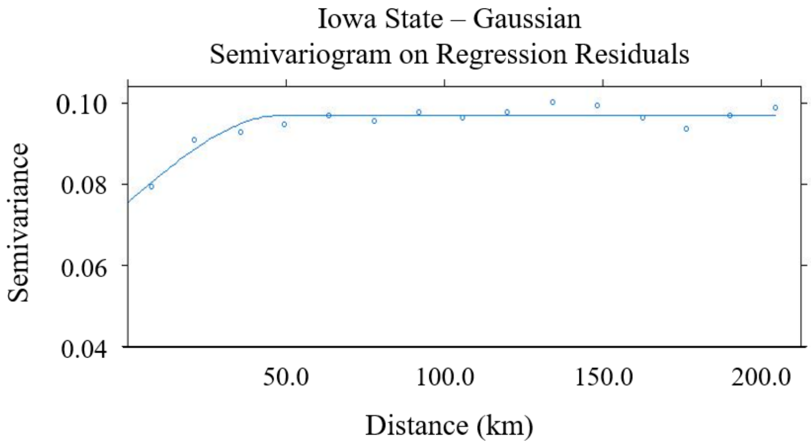
Using the regression models found earlier, the regression kriging process was conducted following the flowchart depicted in Figure 4-4. Following a similar process to OK, the residuals were calculated and then a semivariogram analysis was done upon the residuals. The residuals were then interpolated based off the underlying spatial structure and then the interpolated values they were added back into the MLR predictions to obtain the RK estimates.

For the state of Iowa, Figure 5-6 shows the semivariograms of the residuals for the three semivariogram models, namely Exponential, Gaussian, and Spherical. With the semivariogram values and crossvalidation results for RK obtained, they can be compared to the results from OK. The regression kriging results are presented in Table 5-5.

Exponential Semivariogram



Gaussian Semivariogram



Spherical Semivariogram

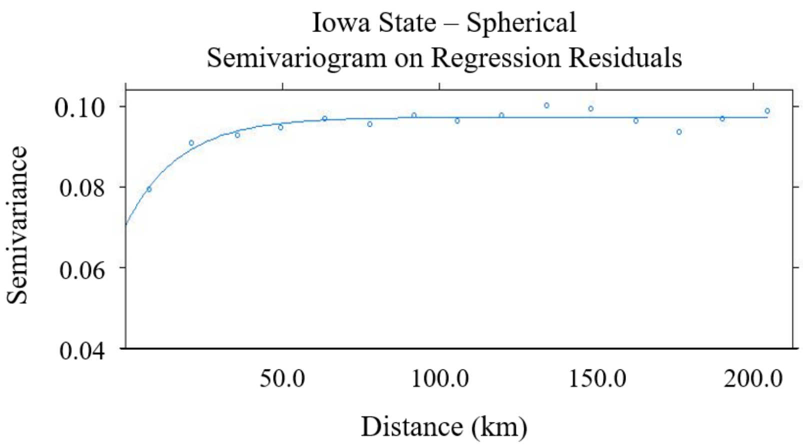


Figure 5-6 Semivariogram plots of Regression residuals for WC ratios

Table 5-5 Regression Kriging Semivariogram and Crossvalidation results for Iowa State

	Iowa State Regression Kriging		
Iowa State	Exponential Model	Gaussian Model	Spherical Model
Nugget	0.070	0.077	0.075
Range (km)	16.925	20.446	49.089
P-Sill	0.027	0.020	0.021
MSqE	0.098	0.098	0.100
MStdE	0.000	0.000	0.000
ASE	1.039	1.014	1.054
RMSE	0.312	0.312	0.315
RMSSE	0.659	0.489	0.795
No. of Pts		19591	

Within RK itself, the same conclusion can be made as was for OK and that the Spherical model performed the best. Again, it had the best RMSSE value and larger range while the other statistical metrics were nearly identical across all three models. But when contrasted with OK, the MStdE and ASE remained stable showing neither improvement nor deterioration in the model errors. The nugget values for the spherical models, the best of each set, shows that it got marginally smaller with RK. This means that the perceived measurement error is reduced suggesting that some measurement error has been accounted for. Furthermore, the RMSSE value in all cases saw an improvement by getting closer to the ideal value of 1. What this implies is that more of the variability in the kriging estimates is accounted for from RK than from OK thereby better capturing the change in spatial structure with the change in separation distance. Based on these crossvalidation results, it can be concluded that RK outperforms OK, albeit marginally. This marginality is made up for by the larger gain in insight into how additional covariates may be relevant in their influence over WC ratios, how they influence it, and where they are prevalent.

Using the spherical model, the residuals were interpolated and added back into the MLR estimated values to obtain the RK estimated values. Figure 5-7 is the surface map of the RK estimated values. From the plot results, it can be noted that the WC ratio values are lower than those found with OK. The grouping of higher WC ratios being in the northern half still remains, but is now more spotty, than uniform across the top half. To compare the variance between the two models, the variance plot will have the same values for the various levels as depicted in Figure 5-8 below.

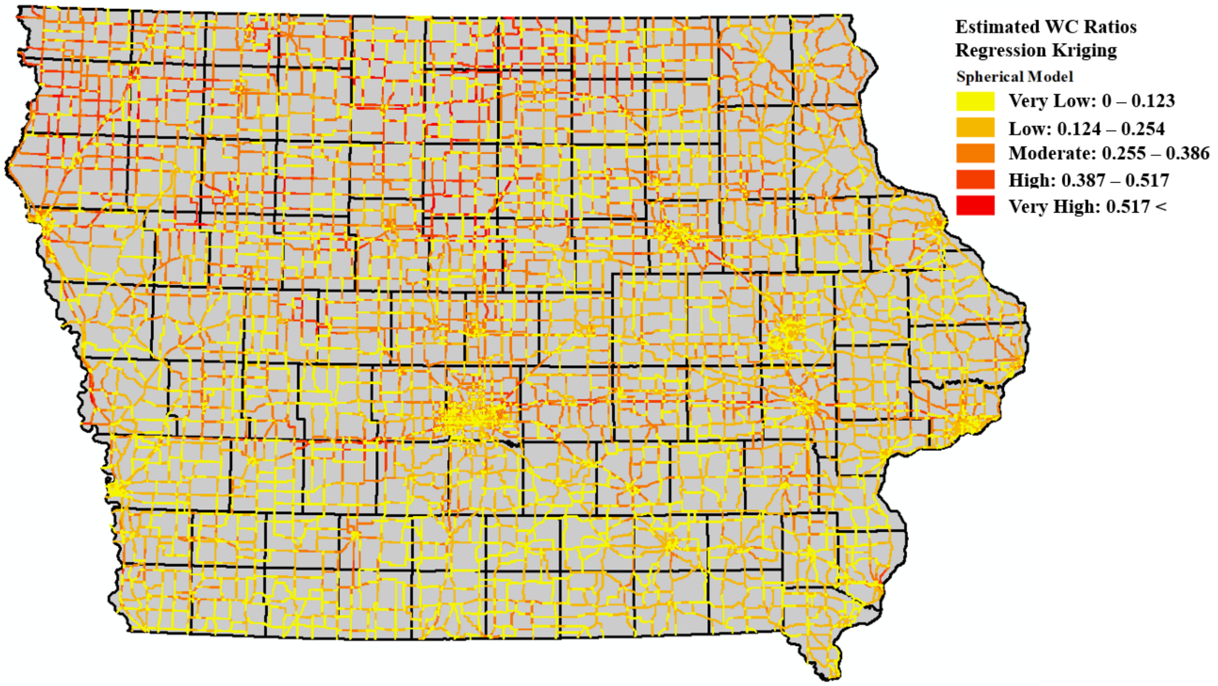


Figure 5-7 Map of RK Interpolated WC ratios

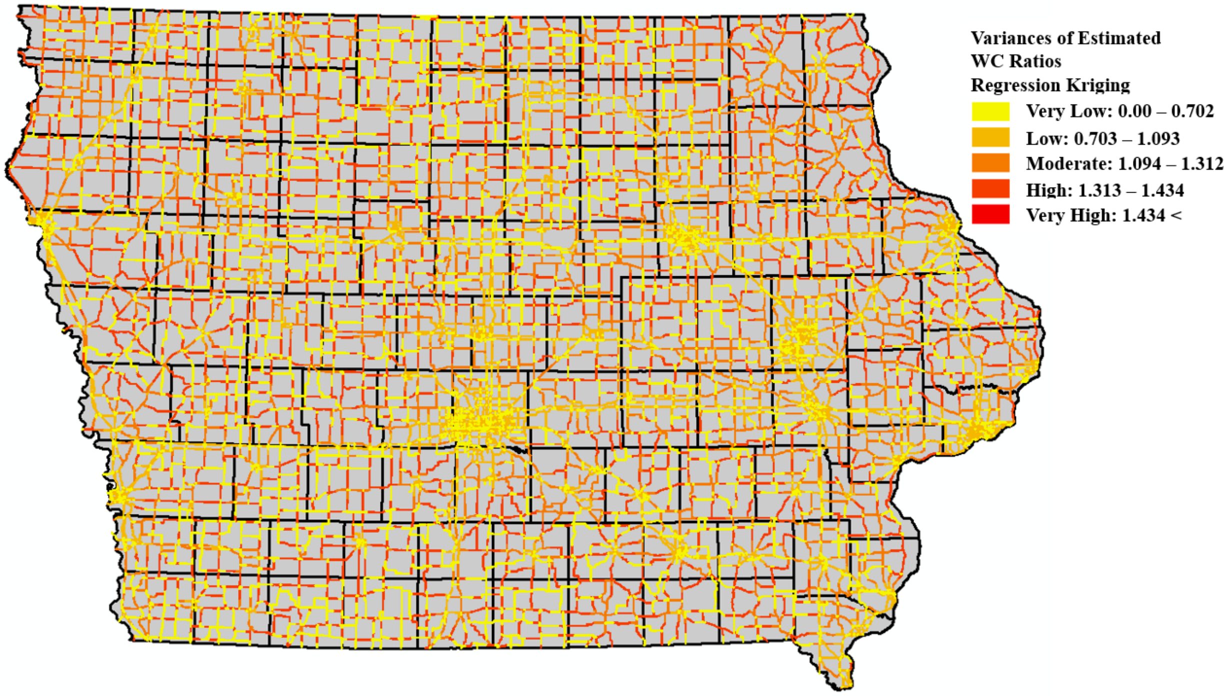


Figure 5-8 Map of the RK estimation variance

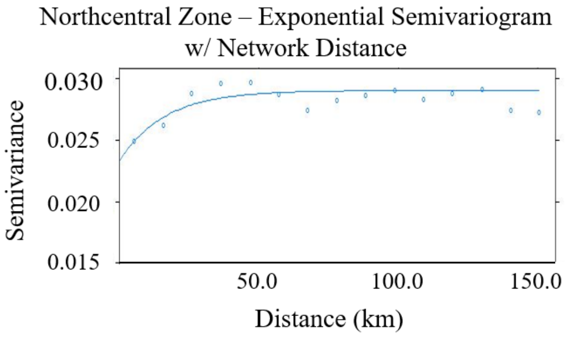
In visually comparing the two estimation variance maps, it is clear that OK has a greater number of higher variances over that of RK as depicted by the greater density of red. This further confirms that even though RK only provides marginally more accurate estimates, the variance or error associated with those estimates have been reduced.

5.5 Enhanced Regression Kriging using Network Distances

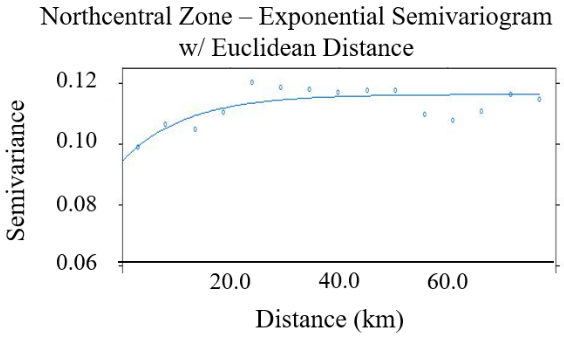
In an attempt to enhance the RK estimates, network distances were substituted for Euclidean distances. As discussed thoroughly in the previous section, this distance measure is more intuitive for the intended analysis as the true separation distance between two points on a road network is bound by the network itself. To find out whether such a hypothesis is true, RK was done for the Northcentral zone in Iowa using both Euclidean and Network distances and then are compared to each other using the five (5) metrics. Figure 5-9 below shows the semivariograms generated using Euclidean and network distances for the Northcentral zone, and a side-by-side comparison of the semivariogram results is shown in Table 5-6 below.

Table 5-6 Side-by-Side Results of RK using Euclidean Distances and Network Distances

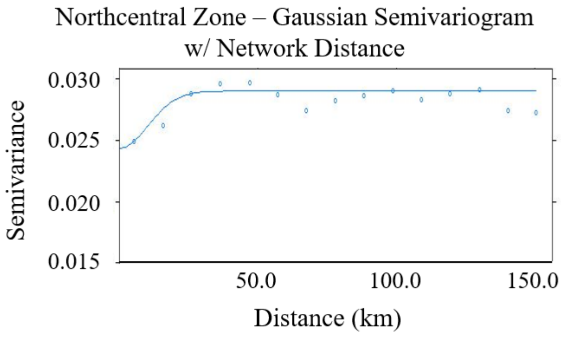
North Central Region	Exponential		Gaussian		Spherical	
	Euclidean Regression Kriging	Network Regression Kriging	Euclidean Regression Kriging	Network Regression Kriging	Euclidean Regression Kriging	Network Regression Kriging
Nugget	0.094	0.023	0.098	0.024	0.097	0.024
Range (km)	11.801	16.412	11.584	14.716	30.300	40.300
P-Sill	0.022	0.006	0.016	0.005	0.020	0.005
MSqE	0.120	0.116	0.119	0.116	0.123	0.116
MStdE	0.000	0.000	0.001	0.000	-0.001	0.000
ASE	1.111	0.160	1.051	0.160	1.152	0.160
RMSE	0.347	0.341	0.345	0.340	0.350	0.340
RMSSE	0.544	1.001	0.442	1.001	0.652	1.001



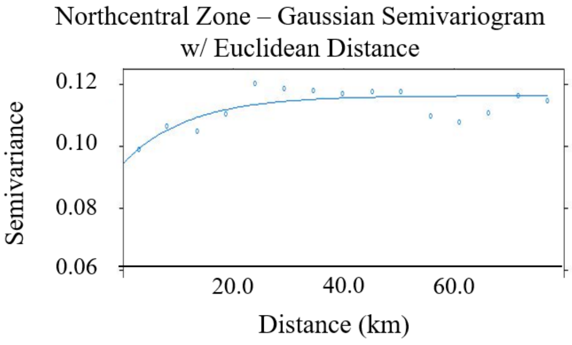
(a)



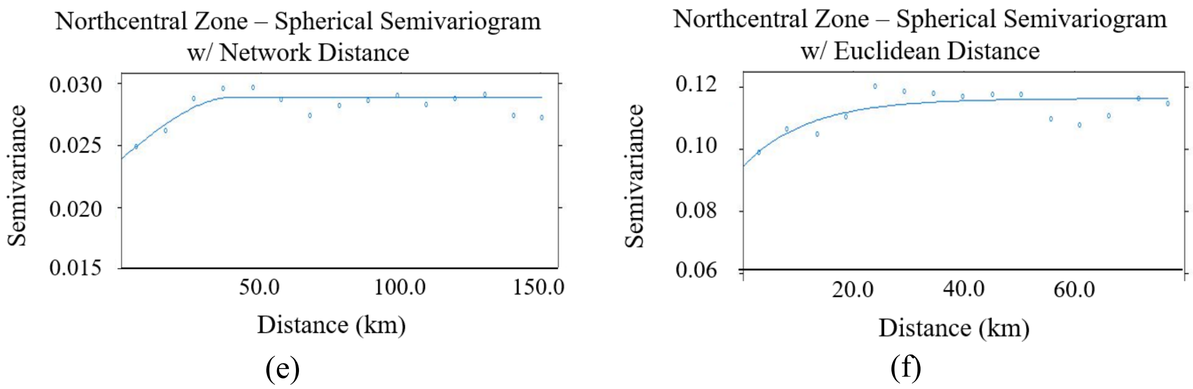
(b)



(c)



(d)



(e) (f)

Figure 5-9 Northcentral Euclidean and Network Distance Semivariograms

Looking at the resulting crossvalidation statistics from Table 5-6, it is clear that network distances will result in better model estimate results by all measures. The nugget across all models is reduced thus indicating less inherent measurement errors are present. The increase in the range means that the effectiveness of the spatial model is increased encompassing a greater spatial reach before it is no longer effective. The MSqE and RMSE have lower values indicating a reduction in the overall value of the errors, and with RMSSE near the ideal value of 1.0, the variability of the model estimates is almost perfectly accounted for. And finally, the ASE values are now lower than, but much closer to, the RMSE values indicating that the model now overestimates the outcomes, but not to the extent that it had underestimated it. By all accounts, this shows that Network distances work well for a large rural region.

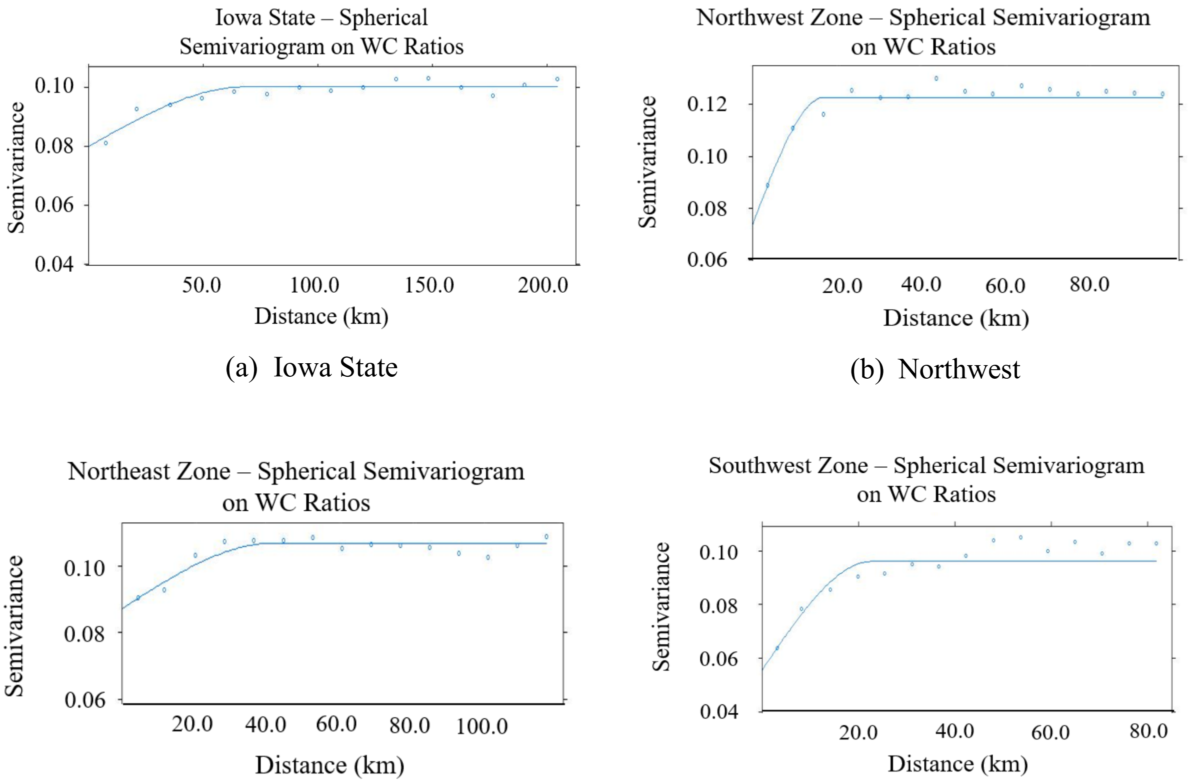
However, this comes at the expense of computational time as it took over 2 hours to run the models for this single region compared to a minute or two using Euclidean distances. Recall that this region underwent significant data filtering to reduce the amount of data points by about half as detailed in Chapter 3. A previous run was done for another zone with about 3400 data points and that analysis took well over 60 hours of computer run time. This goes to show that network distances require a lot more computational work than using Euclidean distances.

5.6 Zonal Characterization of Spatial Structures

For the final objective of this thesis, the regional and zonal spatial structures are characterized using their semivariograms and are then compared to each other to check if the SOSA is met. Ideally, the spatial structure for any zone should closely match the structure for the region as a whole. A first point of comparison would be the semivariogram values: the nugget, range, and sill. As determined in section 5.3 and 5.4, the spherical model tends to be the better performing model of the three, thus only the spherical model is considered in this section. For this section, only Euclidean distances are used. The spherical semivariogram models for each quadrant of the state were generated and their values tabulated. Following Kwon, et al (2019) OK first used to explore this case. Figure 5-10 are the spherical semivariograms for each of the zones and Table 5-7 summarizes the semivariogram values from each of those plots.

Table 5-7 Semivariogram values for the Iowa region and its sub-regions

Value	Iowa	Northwest	Northeast	Southwest	Southeast	Northcentral
Nugget	0.08	0.073	0.087	0.055	0.064	0.098
Range (km)	69.013	17.041	41.827	22.877	25.082	37.442
P-Sill	0.02	0.049	0.02	0.041	0.026	0.024


Iowa State – Spherical Semivariogram on WC Ratios

Northwest Zone – Spherical Semivariogram on WC Ratios

(a) Iowa State

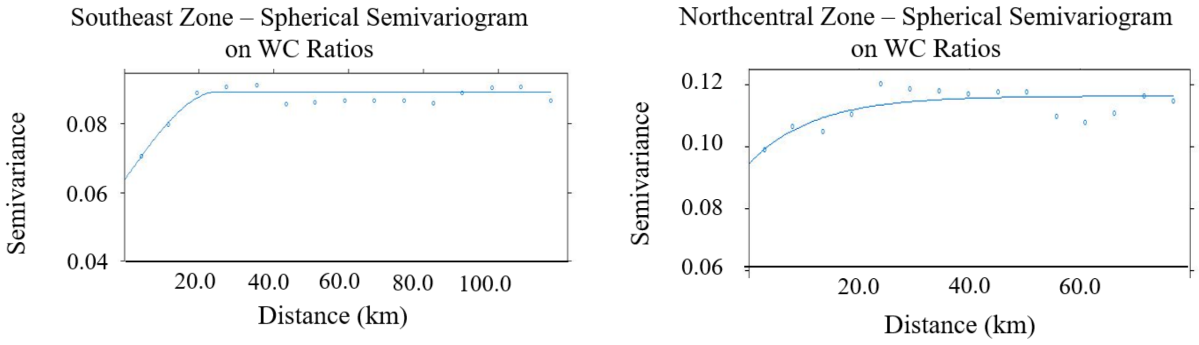
(b) Northwest

Northeast Zone – Spherical Semivariogram on WC Ratios

Southwest Zone – Spherical Semivariogram on WC Ratios

(c) Northeast

(d) Southwest



(e) Southeast

(f) Northcentral

Figure 5-10 Modelling of Spherical Semivariograms using WC Ratios

It is apparent that a single region-wide semivariogram does not result in a spatial structure that is similar to any of the zonal semivariograms. This is visually apparent within the semivariogram plots as shown in Figure 5-10 and by their semivariogram values shown in Table 5-7. The shape of the semivariogram plot is a visual representation of how it changes over distance and it can be seen that no two plots are similar in shape. When looking to the values, it becomes clear that the differences between Iowa and its zones are significant. This implies that the WC ratio is highly sensitive to the spatial structure and possibly confounding factors as well. It is possible that unique zonal characteristics are lost when aggregated with the whole and the loss of those details greatly affects the spatial structure (semivariogram) and the resulting OK model generated. To further confirm this suspicion, each zone had their model crossvalidated and their statistical measures recorded in Table 5-8 below.

Table 5-8 Spherical model OK crossvalidation results

	Iowa State	Northwest	Northeast	Southwest	Southeast	North Central
P-Sill	0.020	0.049	0.020	0.041	0.026	0.024
MSqE	0.100	0.127	0.106	0.100	0.083	0.124
MStdE	0.000	0.000	0.000	0.000	0.000	-0.001
ASE	1.113	1.147	1.115	1.136	1.090	1.152
RMSE	0.316	0.356	0.326	0.316	0.289	0.352
RMSSE	0.752	0.648	0.764	0.206	0.028	0.600

It becomes clear from the crossvalidation results that a region wide spatial structure may not be appropriate for some of the zones as the zonal semivariogram results are quite different to that of the region. Therefore, there is a high case for suggesting the use of zonal semivariograms. The next step was to confirm if this result would follow with the use of RK. Following the method done for OK, the same results for RK were found and tabulated in Table 5-9 below.

Table 5-9 RK semivariogram model values and crossvalidation results

Values	Iowa	Northwest	Northeast	Southwest	Southeast	North Central
Nugget	0.075	0.074	0.083	0.054	0.062	0.096
Range (km)	49.089	18.772	36.663	21.04	25.105	30.272
P-Sill	0.021	0.048	0.021	0.037	0.025	0.02
MSqE	0.1	0.127	0.106	0.099	0.083	0.123
MStdE	0	0	0	0	0	-0.001
ASE	1.054	1.147	1.115	1.136	1.09	1.152
RMSE	0.315	0.356	0.325	0.315	0.287	0.35
RMSSE	0.795	0.685	0.83	0.356	0.032	0.652

As Table 5-9 clearly illustrates, the same result is found with RK as was with OK. Therefore, this is a clear indication that the SOSA is not met and that zonal spatial structures are required to have the best possible outcome.

Having the crossvalidation results on hand for the zones for both OK and RK also provides a secondary result for this section. By the same metrics used to determine that RK was the better estimator over OK, the same can be said for each of the zone here for the same reasons. The RMSSE values are better and the nugget values are smaller. This further shows how RK provides marginally better estimates, but substantially has better estimation variance results.

5.7 Summary

The results of the study have been able to address the three objects set forth in Chapter 1. To conduct the analysis, spatial interpolation via ordinary kriging (OK) was done on the road surface and environmental conditions to ensure that all road segments had the appropriate information. This task also shows how effective of a tool that OK is for handling these types of variates, similar to the results of earlier studies exploring OK.

Once the weather and road condition information was interpolated then the first objective was studied utilizing the whole state of Iowa and modelling the WC ratios using ordinary kriging (OK) and its more sophisticated variant, Regression Kriging (RK). By all measures, RK outperformed both MLR and OK in their estimates proving how strong of an estimator tool it can be.

With RK being found to be an excellent estimator, the second objective was to enhance RK using network distances. When RK was utilized with network distances, the results were further enhanced as the model was able to better account for the variability of the estimates as noted by the improved RMSSE, smaller nugget, and greater range values.

The final objective was to conduct an examination of the underlying spatial structure both regionally and zonally. This final task serves to determine if a single semivariogram is sufficient enough to represent the entire region, or if the region is too varied with its road network conditions and thus zonal semivariograms are better suited. When taking into account all the semivariograms and models generated and the results therein, it was found that WC ratios are quite sensitive to the underlying spatial structure. This would imply that for modelling WC ratios with kriging, the SOSA is not met and that a regional semivariogram should be replaced by zonal semivariograms as they better capture any unique spatial structures for those sub regions.

Chapter 6 CONCLUSION AND FUTURE WORKS

Geostatistics is a burgeoning field within the transportation engineering profession as recent studies in other disciplines have shown it to be a good estimator. More importantly, the use of kriging has started to gain in popularity given its propensity to provide some of the best geostatistical model estimates. Several transportation engineering studies have made use of kriging, however, those studies were limited in the spatial and temporal scope of their study area and were limited to the simpler kriging variants of simple kriging or ordinary kriging.

This thesis first attempted to show the viability of using regression kriging, a hybrid form of geostatistics, as a tool for use in winter collision analysis. To improve upon generated kriging estimates, network distances were used in place of Euclidean distances, a consideration that has not been extensively explored in previous studies. Likewise, no existing studies looked into characterizing the underlying spatial structure even when considering unique zonal spatial features and properties. This thesis, therefore, set out to fill in those gaps in current literature to evaluate the feasibility and applicability of kriging for tackling challenging transportation problems, one being a network screening analysis as done herein. The next being the enhancement of estimates using the more intuitive network distances over Euclidean distances. And finally, the examination and verification of the second order stationarity assumption (SOSA) to ensure the best models are being constructed.

6.1 Research Findings

The Winter Collision ratio (WC ratio) is a value that was used as it easily provides a way to relatively compare how collisions due to winter conditions differ between road segments. This was treated as the dependent variable in MLR and in Regression Kriging (RK) while AADT, lane numbers, speed, RST, air temperatures, road surface condition warnings, snowfall totals, and daily high and low temperatures were used as covariates. The datasets used in this study come from five winter seasons from 2013 to 2018 within the state of Iowa for an expansive spatial and temporal study area. Following the methodology outlined, the key findings from this study are as follows:

- Using the state of Iowa, WC ratios were estimated by using MLR, OK, and RK and their results statistically measured. As a result, RK was found to be the better estimator over MLR or OK as determined by the five statistical metrics used. This shows that RK is indeed a viable hybrid method for modelling winter collisions when there are covariates available to use. During this process, it was found that OK estimates also outperformed MLR, thus if no covariates are available, OK can still be used.

- When substituting network distances in place of Euclidean distances, the model performance only resulted in marginally better estimates, but it better accounts for the variability of the estimates substantially. This improvement in variability description comes at the cost of computational complexity and time. As such, the cost of increased computing resources for such marginal benefits needs to be considered.
- In the characterization of the underlying spatial structure, it was found that an overarching regional semivariogram does not perform as well as zonal-semivariograms that better capture the localized spatial structure. Furthermore, it was found that many urban centers did not have a spatial structure present based off the available data thus implying that kriging may not be useful in densely packed zones.

6.2 Research Contributions

The primary contribution of this thesis was the development and benchmarking of the hybrid method of regression kriging for used in winter collision analysis. Through this thesis, a methodological framework was developed to take in covariates and utilize them to make better winter collision estimates. The results show that RK is a better estimator over MLR and OK by all metrics used in this study and ultimately why it should be considered a power modelling tool.

The second contribution made by this thesis was the enhancement of RK by using Network distances in place of Euclidean distances. This thesis showed that RK estimates are marginally improved by using network distances, but the real gain came in the form of the variability of the estimates being better captured as shown by the significant improvement in the RMSSE values. This means that the uncertainty of the estimates is greatly reduced. This shows what the real benefit is to using network distances over Euclidean distances.

The final contribution from this thesis is the characterization of the underlying spatial structure for the study area. Based off the many semivariogram analysis completed, it is clear that WC ratios are sensitive to the underlying spatial structure thereby suggesting that a singular region wide spatial structure cannot be used for the whole study area. Rather, zonal semivariograms need to be developed in order to attain the best outcome from the models. Therefore, this would suggest that the stationarity assumption required for applying a singular semivariogram to any form of kriging for the region does not hold.

Traditional methods such as the safety performance function or Empirical Bayes method require an extensive amount of high quality data in order to conduct a good network screening. Few road authorities have the means and/or ability to collect, maintain, and utilize an extensive amount of data that can be used for good network screening. Due to limited finances or infrastructure, many road authorities around the world have limited collision records, weak record keeping practices, or unreliable reporting. Kriging makes use of the spatial correlation of limited data that is not often accounted for in other methods. Regression kriging takes in external covariate that are often easier to record and can use them to improve their collision modelling. Best of all, the analysis and calculations are all open to the user and not done behind a black box such as neural networks or artificial intelligence computing. This provides a more transparent, evidence based analysis that can be audited and relied on by public officials. Overall, as shown here, RK can effectively model WC ratio hotspots over a large spatial area and produce a hotspot map that authorities can use to make more informed decisions in regards to their WRM programs.

6.3 Future Research Directions

This study is without its limitations and could be expanded upon in future studies. Some recommendations are as follows:

- Include the use of additional weather stations such as those used by the airport authorities. This will increase the weather data point density that will improve the spatial modelling of some of the weather covariates. Weather stations for air travel also have specific sensors for air traffic use and can potentially add additional covariates that were not available from the other stations, such as visibility distances, wind speeds, wind direction, gust speeds, and sunlight intensity.
- Including the use of maintenance activities as covariates to the regression analysis to see if that has an influence on the collision outcomes. As was covered in Chapter 2, WRM can affect the level of risk and occurrences of collisions on the roads during and after a weather

event. Including maintenance times, activities, or service level as covariates can affect the outcome the regression modelling, potentially improving the detrending ability within RK.

- This study only took into account one study area so a repeat of this using another state or even country altogether would provide additional support for the outcomes found here. As with many studies and as a key part of the scientific process, a sample or study of one can set an example and a benchmark. But it is only by conducting the same experiment with different data can the methodology and theory be upheld and verified.
  Though this was limited to winter collisions only, it could possibly be expanded to account for other severe environmental conditions such as fog, heavy rainstorms, etc. Winter conditions are not the only weather condition that has been known to increase the risk of collisions. Foggy weather reduces visibility and ice fog has been known to generate extremely slippery conditions with the formation of ice on roads. Heavy rains can cause slippery conditions as well but can also cause flooding of roads or cause vehicles to hydroplane if water cannot drain from the roads fast enough.

References

Abdel-Aty, M. A. & Radwan, A. E., 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention,* 32(5), pp. 633-642.

American Association of State Highway and Transportation Officials (AASHTO), 2010. *Highway Safety Manual Vol. 1.* 1st ed. Washington D.C: AASHTO.

Andersson, A. K., 2010. *Winter Road Conditions and Traffic Accidents in Sweden and UK-Present and Future Climate Scenarios.* Gothenburg: Department of Earth Sciences; Institutionen för geovetenskaper.

Andersson, A. K. & Champman, L., 2011. The impact of climate change on winter road maintenance and traffic accidents in West Midlands, UK. *Accident Analysis and Prevention 43(1),* pp. 284-289.

Andrey, J. C. & Mills, B. E., 2003. *Collisions, Casualties, and Costs: Weathering the elements on Canadian roads.* 1 ed. London, ON: Institute for Catastrophic Loss Reduction.

Asano, M. & Hirasawa, M., 2003. Characteristics of traffic accidents in cold, snowy Hokkaido, Japan. *Proceedings of the Eastern Asia Society for Transportation Studies. Vol. 4,* pp. 1426-1434.

Atkinson, P. M. & Lloyd, C. D., 1998. Mapping Precipitation in Switzerland with Ordinary and Indicator Kriging. *Journal of Geographic Information and Decision Analysis 2.1-2,* 2(2), pp. 72-86.

Black, W. R. & Thomas, I., 1998. Accidents on Belgium's Motorways: A Network Autocorrelation Analysis. *Journal of Transport Geography, Vol. 6, No. 1,* pp. 23-31.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C., 1994. *Time Series Analysis: Forecasting and Control.* 3rd ed. New Jersy: Prentice-Hall.

Christensen, R., 1991. *Linear Models for Multivariate, Time Series, and Spatial Data.* New York: Springer Science & Business Media.

Cressie, N., 1990. The Origins of Kriging. *Mathematical Geology, Vol. 22, No. 3,* pp. 239-252.

De Pauw, E. et al., 2014. *The Magnitude of The Regression to the Mean Effect In Traffic Crashes.* Vienna, International Co-operation on Theories and Concepts in Traffic Safety (ICTCT).

Eguía, P. et al., 2016. Weather datasets generated using kriging techniques to calibrate building thermal simulations with TRNSYS. *Journal of Building Engineering, 7,* pp. 78-91.

Einax, J. & Soldt, U., 1999. Geostatistical and multivariate statistical methods for the assessment of polluted soils—merits and limitations. *Chemometrics and Intelligent Laboratory Systems 46,* pp. 79-91.

Eisenberg, D. & Warner, K. E., 2005. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. *American Journal of Public Health,* pp. 120-124.

El-Basyouny, K. & Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models.. *Transportation Research Record,* 1950(1), pp. 9-16.

ESRI, 2011. *ArcGIS Desktop: Release 10.* Redlands: Environmental Systems Research Institute.

Fu, L. & Perchanok, M. S., 2006. Effects of winter weather and maintenance treatments on highway safety. *TRB 2006 Annual Meeting CD-ROM,* p. Paper No. 06–0728.

Fu, L., Thakali, L., Kwon, T. J. & Usman, T., 2017. A risk-based approach to winter road surface condition classification.. *Canadian Journal of Civil Engineering, 44(3),* pp. 182-191.

Gräler, B., Pebesma, E. & Heuvelink, G., 2016. Spatio-Temporal Interpolation using gstat. *The R Journal,* 8(1), pp. 204-218.

Gu, L., Kwon, T. J. & Qiu, T. Z., 2019. A Geostatistical Approach to Winter Road Surface Condition Estimation using Mobile RWIS Data. *Canadian Journal of Civil Engineering,* 46(6), pp. 511-521.

Harirforoush, H. & Bellalite, L., 2019. A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. *Accident Analysis & Prevention 130,* pp. 62-74.

Hengl, T., 2009. *A Practical Guide to Geostatistical Mapping.* Amsterdam: Office for Official Publications of the European Communities.

Hengl, T., Heuvelink, G. B. & Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma 120.1-2,* pp. 75-93.

Heqimi, G., 2016. *Using spatial interpolation to determine impacts of snowfall on traffic crashes.* East Lansing: Michigan State University.

Heqimi, G., Kay, J. J. & Gates, T. J., 2017. *Using Spatial Interpolation to Determine Effects of Snowfall on Traffic Crashes: A Case Study of Interstate-94 in Southwest Michigan..* Washington, DC, Masters Thesis, Michigan State University.

Iowa Department of Transportation, 2020. *Iowa DOT Open Data.* [Online] Available at: https://public-iowadot.opendata.arcgis.com/

Iowa State University, 2020. *Iowa Environmental Mesonet.* [Online] Available at: https://mesonet.agron.iastate.edu/

Islam, M. T., El-Basyouny, K., Ibrahim, S. E. & Sayed, T., 2016. Before-After Safety Evaluation Using Full Bayesian Macroscopic Multivariate and Spatial Models. *Transportation Research Record: Journal of the Transportation Research Board, No. 2601,* pp. 128-137.

Khan, G., Qin, X. & Noyce, D. A., 2008. Spatial Analysis of Weather Crash Patterns in Wisconsin. *Journal of Transportation Engineering 134(5),* pp. 191-202.

Krige, D. G., 1981. *Lognormal-de Wijsian geostatistics for ore evaluation.* Johannesburg: South African Institute of mining and metallurgy.

Kutner, M. H., Nachtsheim, C. J. & Neter, J., 2003. *Applied Linear Regression Models.* 4 ed. New York: McGraw-Hill/Irwin.

Kwon, T. J. & Gu, L., 2017. Modelling of winter road surface temperature (RST)—A GIS-based approach. *2017 4th International Conference on Transportation Information and Safety (ICTIS),* pp. 551-556.

Kwon, T. J., Muresan, M., Fu, L. & Usman, T., 2019. Development of Zonal-Specific Semivariograms for a Strategic RWIS Network Optimization: Case Study. *Journal of Infrastructure Systems,* 25(2).

Levine, N., Kim, K. E. & Nitz, L. H., 1995. Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accident Analysis and Prevention, Vol. 27, No. 5,* pp. 663-674.

Lindner, A., Pitombo, C. S., Rocha, S. S. & Quintanilha, J. A., 2016. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. *Geo-spatial Information Science 19.4,* 19(4), pp. 245-254.

Manepalli, U. R. & Bham, G. H., 2011. Crash Prediction: Evaluation of Empirical Bayes and Kriging Methods. *In Proc., 3rd International Conference on Road Safety and SImulation,* pp. 1-13.

Microsoft, 2016. *Excel 2016,* Redmond: Microsoft.

Minnesota Department of Transportation, 2020. *SCAN Glossary.* [Online] Available at: http://rwis.dot.state.mn.us/scanweb/SWFrame.asp?Pageid=Home&Units=&Groupid=&DisplayClass=Java&SenType=All
[Accessed 25 July 2020].

Montgomery, D. C., Peck, E. A. & Vining, G. G., 2001. *Introduction to Linear Regression Analysis.* 3 ed. New York: John Wiley & Sons, Inc.

Morgan, A. & Mannering, F. L., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. *Accident Analysis and Prevention,* 43(5), pp. 1852 - 1863.

Nicholson, A., 1999. Analysis of spatial distributions of accidents. *Safety Science 31,* pp. 71-91.

NOAA, 2020. *Iowa Climate Normals Map.* [Online]
Available at: https://www.weather.gov/dmx/climatenormals

Norrman, J., Eriksson, M. & Lindqvist, S., 2000. Relationships between road slipperiness, traffic accident risk and winter road maintenance activity. *Climate Research,* 15(3), pp. 185-193.

Olea, R. A., 1999. *Geostatistics for Engineers and Earth Scientists.* New York: Springer Science + Business Media.

Olea, R. A., 2006. A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment 20.5,* pp. 307-318.

Oliver, M. A. & Webster, R., 2015. *Basic steps in geostatistics: the variogram and kriging.* New York, NY: Springer International Publishing.

Pebesma, E., 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences,* Volume 30, pp. 683-691.

Peden, M. et al., 2004. *World report on road traffic injury prevention.* Geneva: World Health Organization.

Pei, X., Wong, S. C. & Sze, N.-N., 2012. The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention,* Volume 48, pp. 464-471.

Peng, G., Bing, W., Guangpo, G. & Guangcan, Z., 2013. Spatial distribution of soil organic carbon and total nitrogen based on GIS and geostatistics in a small watershed in a hilly area of northern China. *PloS one, 8(12),* p. e83592.

Pennelly, C., Reuter, G. W. & Tjandra, S., 2018. Effects of Weather on Traffic Collisions in Edmonton, Canada. *Atmosphere-Ocean, 56(5),* pp. 362-371.

R Core Team, 2020. *A language and environment for statistical computing,* Vienna: R Foundation for Statistical Computing.

Retting, R. A., Ferguson, S. A. & Hakkert, A. S., 2003. Effects of red light cameras on violations and crashes: a review of the international literature. *Traffic injury prevention,* 4(1), pp. 17-23.

Royal Canadian Mounted Police, 2019. *Just the Facts - Winter Driving.* [Online] Available at: https://www.rcmp-grc.gc.ca/en/gazette/just-the-facts-winter-driving [Accessed 01 04 2021].

Selby, B. & Kockelman, K. M., 2013. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography,* May, Volume 29, pp. 24-32.

Srinivasan, R., Gross, F. B., Lan, B. & Bahar, G. B., 2016. *Reliability of Safety Management Methods: Network Screening,* s.l.: United States Federal Highway Administration Office of Safety.

Thakali, L., Kwon, T. J. & Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation, 23(2),* pp. 93-106.

Tingvall, C. & Haworth, N., 1999. *Vision Zero-An ethical approach to safety and mobility,* Brisbane: Queensland University of Technology.

U.S. Census Bureau, 2021. *Iowa QuickFacts.* [Online] Available at:
https://www.census.gov/quickfacts/IA?
[Accessed 02 May 2021].

US DOT Federal Highway Administration, 2020. *How Do Weather Events Impact Roads.*
[Online] Available at: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm
[Accessed 25 July 2020].

Usman, T., Fu, L. & Miranda-Moreno, L. F., 2010. Quantifying safety benefit of winter road maintenance: Accident frequency modeling. *Accident Analysis & Prevention,* 42(6), pp. 1878-1887.

Usman, T., Fu, L. & Miranda-Moreno, L. F., 2012. A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level. *Accident Analysis & Prevention,* Volume 48, pp. 368-378.

Van Rossum, G. & Drake, F. L., 2009. *Python 3 Reference Manual,* Scotts Valley: CreateSpace.

World Health Organization, 2018. *Global Status Report On Road Safety 2018,* s.l.: World Health Organization.

Zhang, D. & Wang, X. C., 2014. Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC. *Journal of Transport Geography,* December, Volume 41, pp. 107-115.

Zhu, Q. & Lin, H. S., 2010. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes.. *Pedosphere, 20(5),* pp. 594-606.

Ziakopoulos, A. & Yannis, G., 2020. A Review of Spatial Approaches in Road Safety. *Accident Analysis & Prevention 135,* pp. 105323:1-30.

APPENDIX A – Regression Models

Table A-0-1 Backwards Elimination Stepwise Regression Results prior to VIF elimination

	Iowa State	Northwest	Northeast	Southwest	Southeast	North Central
Number of Data Points	19591	3257	6284	2565	7504	1090
Intercept	0.1480	-0.0839	0.4097	-0.5641	0.5631	1.3460
Number of Lanes	-0.0260	N/A	-0.0203	-0.0294	-0.0299	-0.0467
Speed Limit	0.0012	0.0015	0.0008	0.0019	0.0008	N/A
lnAADT	0.0167	N/A	0.0240	0.0244	0.0215	0.0521
RST	-0.0575	N/A	-0.2172	-0.3172	N/A	-0.2563
Avg Air Temp	0.0397	N/A	0.1270	0.1721	0.0328	N/A
Avg Daily Hi Temp	-0.0141	N/A	N/A	0.1534	-0.0825	N/A
Avg Daily Low Temp	N/A	N/A	-0.0343	0.0605	N/A	N/A
Snowfall	N/A	0.0558	-0.0226	-0.0711	-0.0347	-0.2467
No. of Red Warnings	N/A	-0.0004	-0.0004	0.0023	N/A	N/A
No. of Orange Warnings	0.0001	N/A	0.0002	0.0010	0.0001	0.0004
No. Of Yellow Warnings	0.0011	N/A	-0.0052	-0.0063	0.0017	N/A

Table A-0-2 VIF values for the initial Regression analysis

Number of Lanes	1.2106	N/A	1.2417	1.1370	1.2547	1.0262
Speed Limit	1.1154	1.0070	1.2038	1.1027	1.1417	N/A
lnAADT	1.3281	N/A	1.4402	1.2596	1.3594	1.2787
RST	17.0624	N/A	45.7930	68.1959	N/A	42.5372
Avg Air Temp	22.6663	N/A	19.9496	32.0681	5.9549	N/A
Avg Daily Hi Temp	19.9911	N/A	N/A	28.6329	6.8761	N/A
Avg Daily Low Temp	N/A	N/A	9.7026	6.0729	N/A	N/A
Snowfall	N/A	1.1602	8.2108	19.4435	2.2690	42.6072
No. of Red Warnings	N/A	1.1610	6.4068	1.8429	N/A	N/A
No. of Orange Warnings	4.3531	N/A	7.6093	11.5085	3.0722	1.9015
No. Of Yellow Warnings	2.1290	N/A	5.8279	4.0733	2.5777	N/A

APPENDIX B – Exponential and Gaussian Ordinary and Regression Kriging Comparisons

Table B-0-1 Exponential model OK and RK crossvalidation results

	Iowa State		Northwest		Northeast	
Exponential Model	**OK**	**RK**	**OK**	**RK**	**OK**	**RK**
No. of Pts	19591	19591	3257	3257	6284	6284
Nugget	0.069	0.070	0.062	0.064	0.086	0.081
Range (km)	15.000	16.925	6.608	7.041	20.414	16.250
P-Sill	0.030	0.027	0.062	0.060	0.022	0.024
MSqE	0.098	0.098	0.123	0.123	0.104	0.104
MStdE	0.000	0.000	0.000	0.000	0.000	0.000
ASE	1.081	1.039	1.104	1.104	1.083	1.083
RMSE	0.313	0.312	0.351	0.351	0.323	0.322
RMSSE	0.612	0.659	0.525	0.568	0.690	0.767

	Iowa State		Northwest		Northeast	
Exponential Model	**OK**	**RK**	**OK**	**RK**	**OK**	**RK**
No. of Pts	2565	2565	7504	7504	1090	1090
Nugget	0.051	0.050	0.058	0.057	0.097	0.094
Range (km)	10.548	9.566	9.585	9.607	18.673	11.801
P-Sill	0.048	0.044	0.032	0.030	0.028	0.022
MSqE	0.097	0.096	0.083	0.082	0.121	0.120
MStdE	0.000	0.000	0.000	0.000	0.000	0.000
ASE	1.096	1.096	1.066	1.066	1.111	1.111
RMSE	0.311	0.310	0.287	0.286	0.348	0.347
RMSSE	0.177	0.334	0.028	0.019	0.492	0.544

Table B-0-2 Gaussian model OK and RK crossvalidation results

	Iowa State		Northwest		Northeast	
Gaussian Model	**OK**	**RK**	**OK**	**RK**	**OK**	**RK**
No. of Pts	19591	19591	3257	3257	6284	6284
Nugget	0.080	0.077	0.089	0.088	0.089	0.087
Range (km)	19.452	20.446	11.050	11.055	18.091	19.679
P-Sill	0.018	0.020	0.034	0.034	0.019	0.018
MSqE	0.098	0.098	0.124	0.124	0.104	0.104
MStdE	0.001	0.000	0.001	0.000	-0.001	-0.001
ASE	1.029	1.014	1.039	1.039	1.029	1.029
RMSE	0.313	0.312	0.352	0.352	0.323	0.322
RMSSE	0.436	0.489	0.372	0.424	0.510	0.585
	Southwest		**Southeast**		**North Central**	
Gaussian Model	**OK**	**RK**	**OK**	**RK**	**OK**	**RK**
No. of Pts	2565	2565	7504	7504	1090	1090
Nugget	0.060	0.059	0.070	0.068	0.101	0.098
Range (km)	9.857	9.418	14.749	15.092	17.243	11.584
P-Sill	0.034	0.031	0.019	0.019	0.022	0.016
MSqE	0.096	0.095	0.083	0.082	0.120	0.119
MStdE	0.001	0.001	0.001	0.001	0.000	0.001
ASE	1.037	1.037	1.023	1.023	1.051	1.051
RMSE	0.310	0.308	0.288	0.286	0.346	0.345
RMSSE	0.309	0.432	0.043	0.058	0.413	0.442