future transportation

MDPI

Article

Advances in Regression Kriging-Based Methods for Estimating Statewide Winter Weather Collisions: An Empirical Investigation

Andy H. Wong * and Tae J. Kwon

Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB T6G 2W2, Canada; tjkwon@ualberta.ca

* Correspondence: andyw@ualberta.ca; Tel.: +1-780-910-8221

Abstract: Winter conditions create hazardous roads that municipalities work hard to maintain to ensure the safety of the travelling public. Targeting their efforts with effective network screening will help transportation managers address these problems. In our recent efforts, regression kriging was found to be a viable and effective network screening methodology. However, the study was constrained by its limited spatial extent making the reported results less conclusive and transferrable. In addition, our previous work implemented what has long been adopted in most of conventional studies—the Euclidean distance; however, use of the road network distance would, intuitively, result in further improving kriging estimates, especially when dealing with transportation problems. Therefore, this study improves upon our previous efforts by developing a more advanced kriging model; namely, network regression kriging using the entire state of Iowa with the significantly expanded road network. The transferability of the developed models is also explored to investigate its generalization potential. The findings based on various statistical measures suggest that the enhanced kriging model vastly improved the estimation performance at the cost of greater computational complexity and run times. The study also suggests that regional semivariograms better represent the true nature of the local variances, though an overall model may still function adequately if higher fidelity is not required.

Keywords: regression kriging (RK); road network distances; network screening; geostatistics; second order stationarity assumption

Citation: Wong, A.H.; Kwon, T.J. Advances in Regression Kriging-Based Methods for Estimating Statewide Winter Weather Collisions: An Empirical Investigation. *Future Transp.* **2021**, *1*, x. https://doi.org/10.3390/futuretransp1030030

Academic Editor: Luigi dell'Olio

Received: 3 August 2021
Accepted: 24 September 2021
Published: 13 October 2021

1. Introduction and Background

Winter conditions (WC) create hazardous winter road conditions (WRC) that municipalities must contend with to ensure the safety of their road users. Snow and ice can buildup upon the road surface causing slippery conditions increasing the risk of collisions to the travelling public. In the United States, around 16% of traffic fatalities occur from collisions that were weather induced [1] while, in Canada, the Royal Canadian Mounted Police (RCMP) found that, in 2017, over 14,000 collisions occurred in December alone [2]. Norrman et al. (2000) was one of the first to quantify the relationship between road surface conditions and traffic safety, ultimately finding that between 50% and 70% of winter collisions are attributed to slippery road conditions [3].

Heqimi (2016) found, in their thesis, that, as snowfall totals increase, so do the frequencies of crashes on freeways [4]. Asano and Hirasawa (2003) determined that the majority of crashes in their region of Japan occurred between −5 and −3 °C, temperatures that favor freeze-thaw cycles that can form slippery conditions and even black ice [5]. Andersson (2010) found similar results noting that these temperatures are conducive to freezing rain events [6].

Municipalities have a duty of care to their citizens; thus, to make the roads safer, they undertake winter road maintenance (WRM) activities, such as plowing, salting, and abrasive dispersions. This process is time consuming, arduous, and expensive, and delays in response could result in unnecessary collisions. They strive to make efficient use of their resources, usually using subjective historical experiences of the planners and operators. Alternatively, targeting their efforts with effective network screening can help transportation managers address these problems objectively. Network screening is the first and most important step within the safety management cycle and is used to identify sites that should be the focus for potential assessment or treatments [7].

There are many established methods for network screening, such as the empirical Bayes method and the use of safety performance functions (SPF), but they become cumbersome to implement over a large spatial scale. These methods also suffer from site selection and categorization biases that can potentially influence the estimates [8]. As such, there has been a growing interest into Geostatistics as a possible replacement for these methods, especially for applications over a large spatial area. Of the various methods available, kriging has been found to be an extremely power predictor for transportation problems.

Thakali et al. (2015) showed how well ordinary kriging (OK) performed over the kernel density estimation (KDE) method [9]. Universal kriging (UK) was introduced as a method for estimating the annual average daily traffic in Texas [10] and to estimate the ridership on select New York subway lines [11]. Gu et al. (2018) demonstrated the effectiveness of regression kriging (RK) to estimate the winter road surface temperatures (RST) on Highway 16, in Alberta Canada [12]. The limitations of these studies lies in the road network expanse and time scale. These studies mainly focused their efforts on a single stretch of road or were limited to a single year's worth of data. This limits the conclusiveness of the results as there are inherent biases and missed trends that come with using a single year or a single stretch of highway.

In the quickly expanding field of machine learning (ML) and artificial intelligence (AI), it has been explored as a method that can make predictions and estimations without the need for assumptions and predefined relationships that are required in statistical methods. Various neural networks were implemented to model crash frequencies and severities [13–15]. Other methods used to model crash severities include the implementation of decision trees by Abellán et al. (2013) [16] and genetic algorithms by Das & Abdel-Aty (2010) [17].

However, all these methods are computationally laborious and intensive as they are often self-recursive in nature requiring many iterations before terminating. Another limiting characteristic of ML and AI is their black-box nature whereby the complete analytical process is not fully open to scrutiny or examination, and replication is often imprecise. As with the more established methods, these previous studies were also limited in spatial and temporal scales, and, specifically, these studies did not focus on determining winter collision behaviours.

Additionally, incidences or events that occur on the road network would have the distance between them measured not by the Euclidean distance, but by the distance via the road network. Previous studies have looked into using network distances for transportation problems, such as Selby and Kockelman's (2013) study where they estimated AADT on major highways throughout Texas using Universal Kriging (UK) [10]. Then, in the study by Zhang and Wang (2014), they also considered the use of network distance with kriging when studying ridership on New York's subway lines [11]. Both studies provided persuasive, but not conclusive, results that showed network distances can improve model performances. Further limitations of the two studies stem from using only one year's worth of data, and being limited to very select networks.

Our previous study sought to address the aforementioned limitations of previous studies by applying regression kriging to a larger scale, that being a quadrant of Iowa, and using five years' worth of data. Ultimately, it was shown that Regression Kriging (RK)

is a viable and effective network screening methodology, though it was still limited in scope by only focusing the method to that one quadrant of Iowa [18]. However, by increasing the spatial scope, the spatial transferability of the variance, known as the second order stationarity assumption (SOSA), must be checked, which was omitted. The premise for checking the SOSA is to see if the underlying spatial structure can be represented by a single semivariogram model, or if regional semivariograms are more appropriate [19]. It is unfortunate that, outside of textbooks, this assumption is rarely checked.

Another limitation of the initial study and existing literature on this very topic was bound by how geostatistics typically defines separation distances between points. Within kriging, the development of the semivariogram models uses the Euclidean distances to model how the semivariance changes as separation distances increase. However, for transportation problems, the true distance between points is bound by the road network, which is not always a straight line. Figure 1 best illustrates this concept with two points within a residential block in the City of Des Moines.

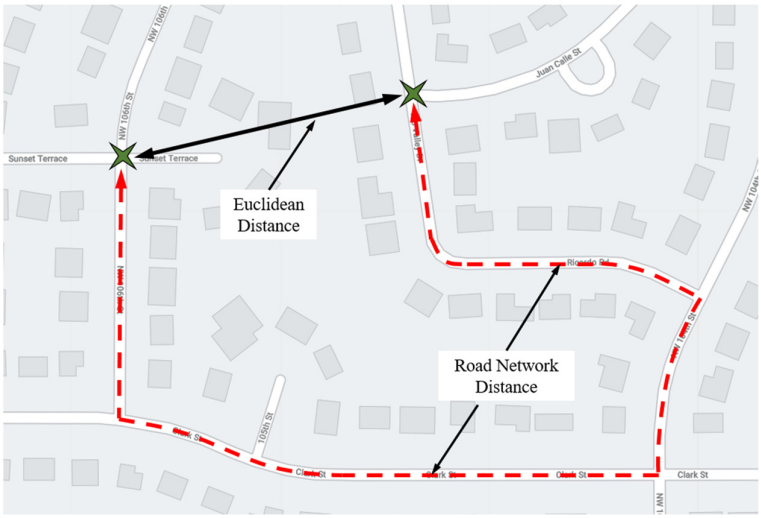

Figure 1. Euclidean distances versus road network distances.

With considerations into the limitations of the previous studies and our past efforts, the primary purpose of this paper is to enhance and expand the use of Regression Kriging by incorporating network distances and to examine the underlying spatial structure to conduct a SOSA analysis to check if a single spatial model can sufficiently represent the whole region, or if individual models are more appropriate. This study will also be using a much larger and comprehensive data set that spans multiple years to provide more conclusive results.

The hypothesis being tested is whether or not network distance improves the outcome of regression kriging estimate values. The evaluation criteria are based on five (5) statistical measures namely the mean squared error, mean standard error, average standard error, root mean squared error, and the root mean squared standardized error. The paper is structured as follows: Section 2 explains the fundamentals of regression kriging; Section 3 describes the study area, the data, and subsequent pre-processing required; Section 4 outlines the methodology undertaken in great detail; Section 5 reports and discusses the results and findings; and Section 6 summarizes the findings, conclusions that can be drawn, shortcomings, and potential future research directions.

2. Regression Kriging Fundamentals

Regression kriging (RK) has gained notoriety for being a good estimator within geostatistics. Geostatistics is a broad term that incorporates many different numerical methods that are used to characterize spatial attributes [19] to analyze spatially or temporally autocorrelated data [20]. Autocorrelation is defined as data that is correlated with itself, usually based on physical or temporal separation measures (i.e., distance or time). As one of the renowned variants of kriging, RK has the ability to incorporate external covariates into the kriging analysis. It incorporates regression modelling to construct a regression function that models the local mean as a fitted regression model. This, in turn, improves the estimation variance results providing a higher level of confidence in the interpolated estimates. Figure 2 provides a visual example of how regression kriging functions.

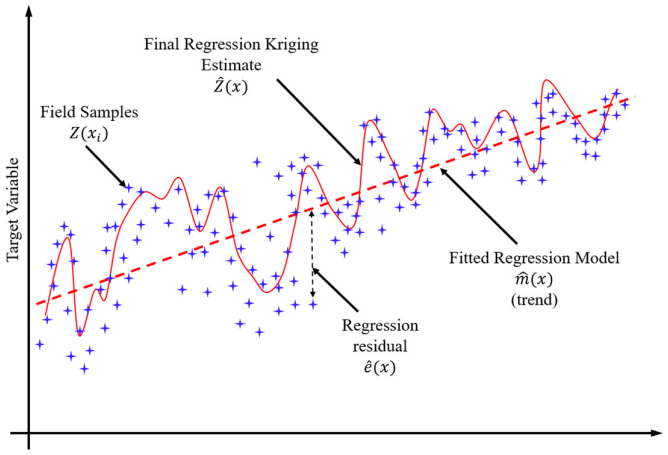

Figure 2. Graphical example of regression kriging.

At its core, the spatial structure of the data will dictate the outcome of the interpolation process. This structure is defined by the semivariogram plot whereby the known data points are plotted, analyzed, and used to find an optimal semivariogram function. The semivariogram is a plot that shows how the level of dissimilarity between pairs of points change as the separation distance between them increases. This change over space provides three important values, namely the nugget, sill, and range that are used to calculate the kriging weights for making an estimate at an unmeasured location. Figure 3 illustrates a typical semivariogram and its various components.

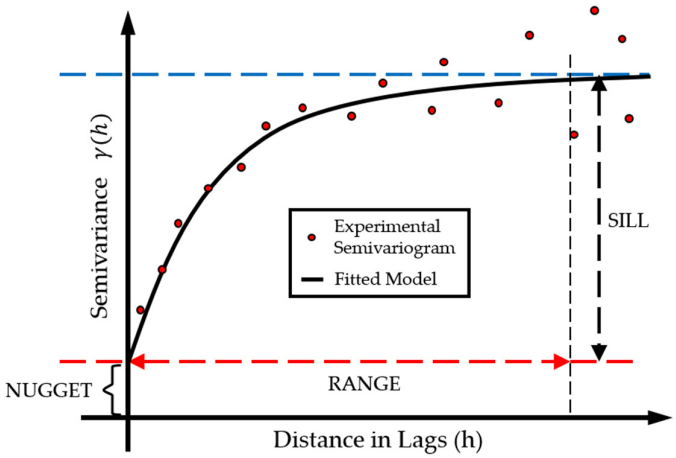

Figure 3. Typical semivariogram plot.

The nugget represents the localized measurement error and should be theoretically zero. However, no measurement is ever perfect due to human error, equipment accuracy, and imperfect recordings and is exhibited by the y-intercept in the plot. The sill is the maximal value of dissimilarity at which point the semivariance between pairs of points is no longer considered significant, and the range is the distance value at which this point is reached. The semivariogram can also be used to check the second order stationarity assumption (SOSA). If the variance or spatial structure is truly uniform, or at least very similar, throughout, then the semivariogram from region to region should be similar to each other. Thus, through visual and numerical comparisons, the SOSA can be checked.

The experimental semivariogram is calculated from the measured data and then a mathematical model is iteratively used to determine the curve of best fit. There have been many models that have been found for this process such as the cubic, power, sine hole, and pentaspherical; however, in practice, the most commonly used ones are the exponential, Gaussian, and spherical models as shown in Equations (1)–(3), respectively [19,21]. The semivariogram function selected is used within the weighting calculations that follows.

$$Exp(h) = C\left(1 - e^{-\frac{3h}{a}}\right) \tag{1}$$

$$Gau(h) = C\left(1 - e^{-3\left(\frac{h}{a}\right)^2}\right) \tag{2}$$

$$Sph(h) = \begin{Bmatrix} C\left(\frac{3h}{2a} - \frac{1}{2}\cdot\left(\frac{h}{a}\right)^3\right) & 0 \le |h| \le |a| \\ C & |a| \le |h| \end{Bmatrix} \tag{3}$$

Regression kriging follows the core tenants of kriging by using utilizing a deterministic component to reduce the uncertainty of the stochastic estimation. Mathematically, it takes the core form of Equation (4) with the estimator taking the form of Equation (5).

$$Z(x) = m(x) + \varepsilon(x) \tag{4}$$

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) + [1 - \sum_{i=1}^{n} \lambda_i]\mu \tag{5}$$

where $\hat{Z}(x_0)$ is the estimator at the unmeasured location x_0, x_i are values at measured locations, and λ_i are the weights for the kriging estimator that minimizes the variance of the estimator (estimation variance) and the mean squared error (MSqE). Within RK, the deterministic component is modified by the regression analysis. Therefore, when combining the MLR function generated via the regression analysis, it can be expanded into Equation (6).

$$\hat{z}(x_0) = \sum_{i=0}^{n} \widehat{\beta_i} \cdot q_i(x_0) + \sum_{i=0}^{n} \lambda_i(x_0) \cdot r(x_i) \tag{6}$$

where $\hat{\beta}_i$ are the model coefficients, $q_i(x_0)$ are the auxiliary variables, $\lambda_i(x_0)$ are the covariance weights, and $r(x_i)$ are the regression residuals. As noted, the goal is to minimize the estimation variance, which is represented by Equation (7).

$$\sigma^2(x_0) = 2\sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i, x_j) \tag{7}$$

where $\gamma(x_i, x_0)$ is the semivariogram function selected from the semivariogram analysis that was conducted earlier. From looking at Equations (5)–(7), it becomes clear that the weighting values play a pivotal role in the process. The weights are found by solving an

optimization problem with an objective function that is represented by a Lagrangian function as exemplified in Equation (8) [19].

$$L(\lambda_1, \lambda_2, \ldots, \lambda_n; \mu) = \sigma^2(x_0) + 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right) \quad (8)$$

With the Lagrangian function defined, then the weights will be the solution to the system of covariance equations as shown in Equation (9).

$$\left.\begin{aligned} \sum_{i=1}^{n} \lambda_i \mathrm{Cov}(x_i, x_1) + \mu = Cov(x_1, x_0) \\ \sum_{i=1}^{n} \lambda_i \mathrm{Cov}(x_i, x_2) + \mu = Cov(x_2, x_0) \\ \ldots \\ \sum_{i=1}^{n} \lambda_i \mathrm{Cov}(x_i, x_n) + \mu = Cov(x_n, x_0) \end{aligned}\right\} \quad (9)$$

Subject to

$$\sum_{i=1}^{n} \lambda_i = 1$$

Within Equation (9) is where the separation distance between points is used and, depending on the value, it can change the weightings significantly. This is also reflected in the semivariogram plot where the covariance is plotted against the separation distances between each pair. It is here where the use of network distances will affect how the semivariogram plot will change and, in turn, affect the values of the weights. Furthermore, network distances will affect the interpolation of estimates as this will also be reflected in the distances between the known and unknown locations. Altogether, with a more representative separation distance, the estimates should more accurate, the inherent measurement error (or nugget) should be reduced, and the estimation variances should improve.

3. Study Area and Data

The state of Iowa was chosen for its openly accessible and non-proprietary format datasets, relatively flat and consistent topography, distinct winter weather conditions, and their weather station network. Road, traffic, and collision data was sourced from the Iowa Department of Transportation's (DOT) Open Database that is made freely available online [22]. The environmental, road surface, and weather data was obtained through the Iowa Mesonet database as maintained by the Iowa State University [23] and is also freely available online.

The study area is the state of Iowa and its divided sub regions as shown in Figure 4 totaling six (6) zones. All of the zones will be used in this study to show the performance of regression kriging (RK) while simultaneously providing the necessary analyses for conducting the second order stationarity assumption (SOSA) or model transferability analysis. Using ArcMAP by ESRI, the working coordinate system used was NAD83 UTM 15N that projected the GIS data into metric values for measurements and calculations.

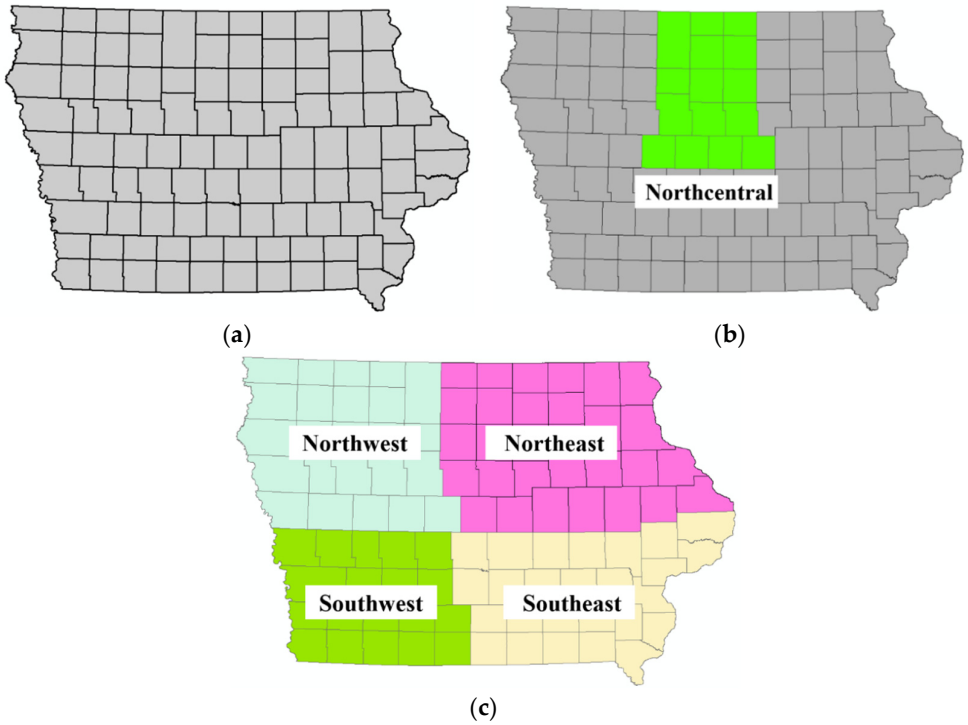

Figure 4. Iowa state study areas (**a**) Whole state, (**b**) Northcentral Zone, and (**c**) Quadrant Zones.

The study period spans five winter seasons, including the months of October to March. This longer time range is used to reduce the effects of outliers and the chance of biased data. This also will reduce the effects of the phenomena known as regression to the mean (RTM) whereby abnormally high or low values of random samples will trend to the mean over time [24]. As defined by the National Oceanic and Atmospheric Administration (NOAA) for the state of Iowa [25], the winter season are the months of December to February and shoulder, or transitional months, are October, November, and March, which also tend to experience winter events. The five most recent winter seasons from the data set encompasses the 2013/14 to 2017/18 seasons.

3.1. Meteorological and Road Conditions Data

The meteorological data used are measureable quantitative seasonal averages for the study area. Values such as snow fall amounts, air temperatures, road surface temperatures (RST), and road condition warnings provide a measure of winter variables that can be attributed to WC ratio results. There are many ways to obtain these values, but one of the more effective sources are from Road Weather Information Systems (RWIS) as they provide a near-instantaneous and continuous record of weather and road conditions at their location.

However, as point measurements they do not provide sufficient information to all roads and areas natively; thus, these values need to be interpolated to ensure complete statewide coverage. As found in previous transportation and environmental studies, the use of kriging is an effective and efficient method for spatially interpolating road surface and environmental data for widespread coverage [12,18,26,27]. Therefore, following their methods, ordinary kriging was used to interpolate these values to ensure statewide coverage for all road segments. Figure 5 shows the locations of the NWS COOP and RWIS stations.

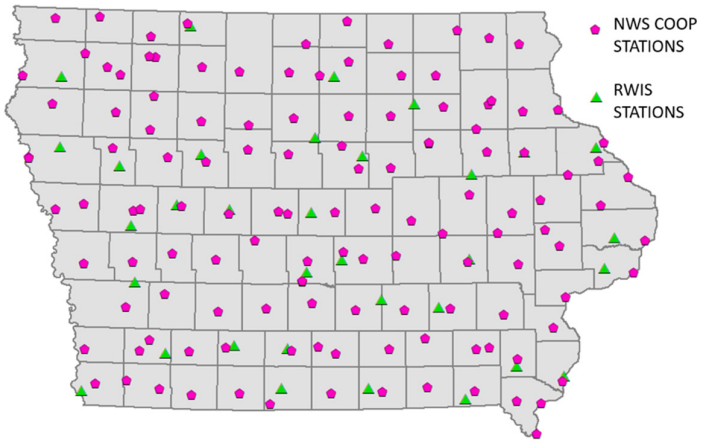

Figure 5. RWIS and NWS COOP station location map.

Following the original study, the covariates include the annual average daily traffic volumes (AADT), road surface temperatures (RST), seasonal snowfall averages, daily air temperatures (average, max, and min), and the road surface index surrogate of road warning messages (red, orange, and yellow classifications) [18]. Furthermore, additional road characteristics, such as the posted speed limits and the number of lanes, were also incorporated into the regression analysis. Table 1 provides the summary statistics for the covariates used.

Table 1. Summary statistics for the 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	

3.2. Road Network

The road network used in this study encompasses all major interstate highways, principal arterial freeways and expressways, minor arterial, and major collector roads. The quality of the road data is inconsistent, and the lengths of roads can vary greatly. To smooth out the data distribution, it is important that the roads are segmented into lengths that are no longer than 5.0 km. Road lengths may be shorter as intersections are natural break points and within urban areas, the distance between intersections are often less than 5.0 km.

The complete network connectivity needs to be checked to ensure that there are no gaps in the GIS road files that would inhibit a network trace when computing the network distance between any two points, an extremely time consuming, but necessary process. In total, over 46,000 km of road was used for this study. Figure 6 shows all the roads used in this study. Both the red and blue roads were used to conduct the expanded RK study and its SOSA analysis, while the road roads within the Northcentral region were used to conduct the Network to Euclidean distance comparison. The red colored roads are roads classified as federal functions 1 to 4, while the blue roads are classified as federal function 5.

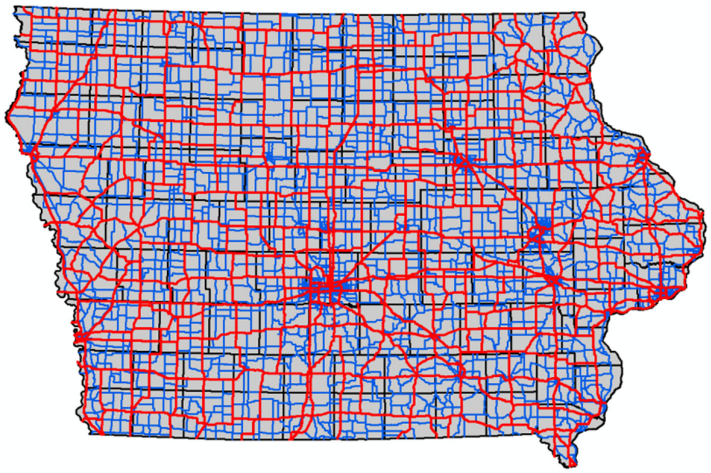

Figure 6. The study area road network.

For the portion of the study that seeks to enhance RK by utilizing network distances over Euclidean distances, the study area will be limited to the Northcentral zone as shown in Figure 4c. Additionally, the road used for this investigation was limited to the red roads from Figure 6 within this area. This was done to accommodate the limitations in computing power of the workstation. This is because network distance kriging requires the shortest distance between every pairing of data points and running an origin-destination (OD) algorithm for all points can become computationally intensive as the OD matrix of distances will grow quadratically as more points are added. Therefore, to reduce the data size for this process, the study area was reduced to this zone only.

3.3. Collision Data

To properly relate winter collisions to the various environmental and road covariates, the dependent variable that will be used is the Winter Collision (WC) ratio and is the ratio of collisions that occurred under winter conditions to all collisions that have occurred on the road segment. This relative collision valuation follows Khan et al. (2008) study and allows for a relationship between weather/condition elements that may influence collisions be developed [28]. As before, a winter collision is defined as a collision that had a snowing/snowy, icy, or slushy road or environmental conditions reported at the time of the collision. Collisions are also random events that are independent of each other; thus, it is important to have a sufficiently large sample size to minimize biases and outlier events.

The dataset obtained from the Iowa DOT source mentioned above included all collisions that occurred from January 2008 to June 2018. This was then truncated so that only collisions that occurred within the timeframes and on the selected roads were included. Figure 7 shows all the collisions used in this study totalled 111,699 reported collision events. Table 2 provides a summary of the collisions used in the study as provided by the Iowa DOT open database.

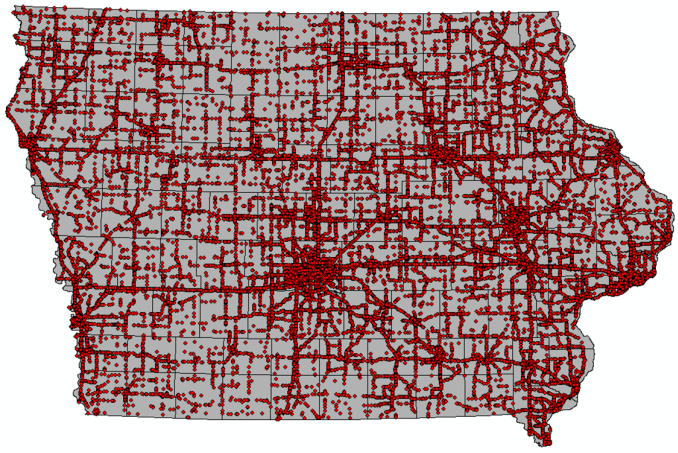

Figure 7. Reported collisions on the road network.

Table 2. Seasonal collision statistics.

Seasonal Collision Statistics	2013–2014 Season	2014–2015 Season	2015–2016 Season	2016–2017 Season	2017–2018 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	25,767	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

Intuitively, the majority of collisions shown in Figure 7 are mostly centered on major urban centers and major roads. This trend on location suggests that these locations may benefit the most from WRM activities. However, given the size of the network that needs to be serviced, any improvement in the planning stages will result in a higher level of service to their citizens.

It is important to discuss the significant drops in WC ratios for the 2015–2016 and 2016–2017 winter seasons. NOAA records for Iowa show that these particular winter seasons experienced temperatures that were close to 4.0 °C (6.0 °F) above normal overall, which also resulted in the snowfall totals being below normal by up to 35.5 cm (14 inches) in 2015–2016 and 20 cm (8 inches) in 2016–2017 [29,30]. Given the propensity of winter collisions being heavily influenced by winter conditions, the milder winters naturally resulted in an overall lower Winter Collision Ratio.

4. Methodology

This section outlines the methodology employed for this study. With the vast amount of data used, it was necessary to utilize software, such as ESRI's ArcGIS as our Graphical Information Systems (GIS) platform [31] and Microsoft Excel for tables and descriptive statistics [32]. For the calculations undertaken, R and its gstat package [33,34], and Python were primarily used [35]. The overall workflow for this study is outlined in Figure 8.

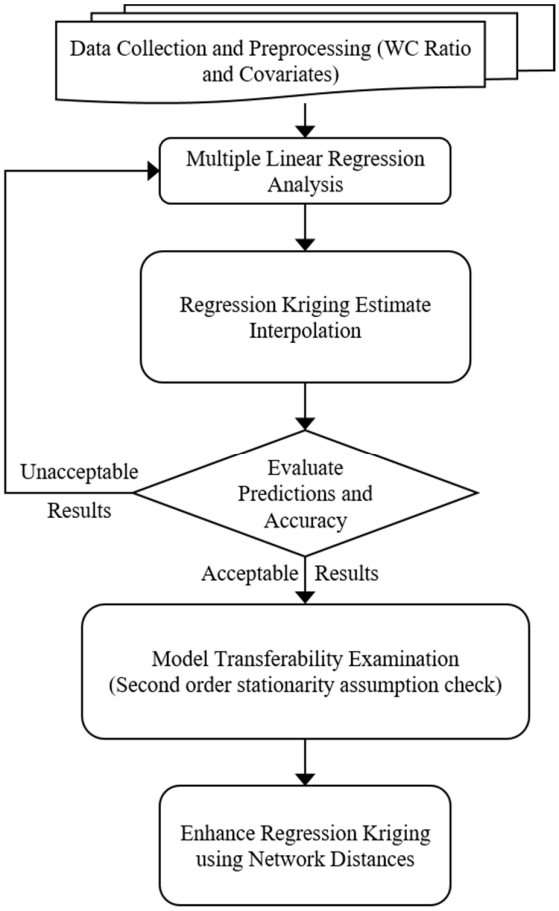

Figure 8. Proposed schematic of the overall workflow.

In general, the statewide data processing and regression analysis is first done to obtain the residuals. Then, the regression kriging estimates are calculated and evaluated. If the results are accurate and reliable, and then a second order stationarity assumption check is conducted to check model transferability. Finally, the enhancement of RK is explored using network distances.

4.1. Data Requirements

To ensure accuracy and to reduce biases, the quality of the datasets first needs to be ensured. As such, erroneous or missing data points were scrutinized for completeness and outlier frequency. For example, within the Iowa collision records dataset, reported collisions with no known coordinates were given a default location outside of the state to the southwest as a placeholder and were subsequently removed as they were not on a road or within the state. Meteorological and environmental data relies heavily on the collection station to be functioning properly and continuously.

As such, stations that were not operational or had missing data that exceeded 30% of the dataset for that station were subsequently omitted from the analysis to maintain a high level of data quality while also ensuring sufficient quantity for analysis and interpolation.

4.2. Spatial Interpolation via Regression Kriging (RK)

The overall methodological RK workflow is summarized in Figure 9 below. Detailed descriptions of each step will follow throughout this section.

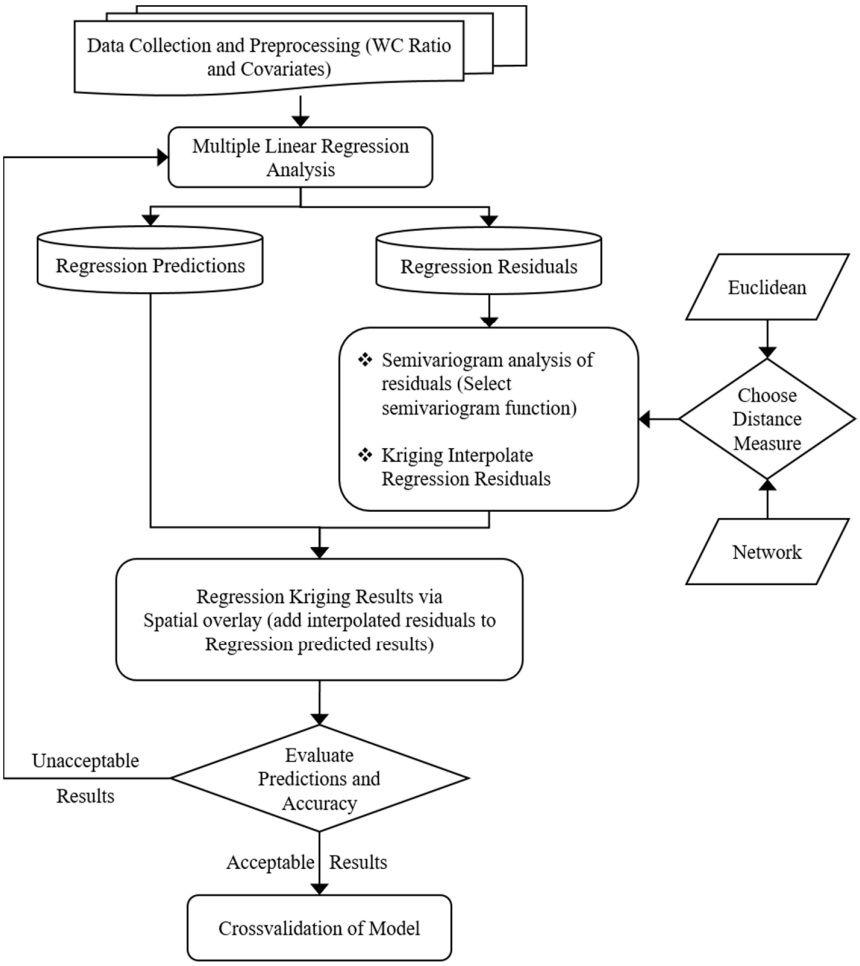

Figure 9. Simplified map of the Regression Kriging method.

The majority of the results are obtained following the RK process detailed in our previous study [18] but now expanded to the state and regional study areas and with higher data densities. All relevant collisions were categorized as a winter collision or not and then mapped to the appropriate road segments. The environmental and road surface conditions were spatially interpolated using ordinary kriging into a 500 m × 500 m raster grid, which was then used to project their average values onto the overlapping road segments. Since kriging works on point-based data points, all road segments were reduced to their midpoints for regression analysis and kriging interpolation.

A linear regression analysis is first done to determine which covariates are significant for the region or not. Statistically relevant covariates are determined by their p-value at the 95% confidence interval ($\alpha = 0.05$). In the regression process, it is important to ensure that no multicollinearity is present between covariates; thus, a Variance Inflation Factor (VIF) analysis is completed during the regression analysis.

A covariate with a VIF value greater than 10 shows high collinearity with at least one other covariate and, thus, is flagged for removal from the model [36]. With the regression models generated, the residuals are calculated, and the semivariogram plot is constructed and analyzed for each region and situations (e.g., network vs. Euclidean distances). The semivariogram values are then used to iteratively solve the systems of covariance equations providing the weighting values to make an estimate for each unmeasured location using the surrounding measured locations. Once all the estimates have been calculated, the model's performance is evaluated using cross validation.

Cross validation is a typical method used to evaluate the accuracy of the method and model being employed. For this study, a Leave-One-Out (LOO) cross validation method way employed where iteratively, each known data point was estimated as an unknown location by leaving it out of the dataset, generating the kriging model with the remaining data, and then estimating the value at that location [37]. By this process, a complete list of estimates was created, which will provide measures of estimation accuracy and reliability. The comparison statistics used, their formulation, ideal values, and interpretations are shown in Table 3 below.

Table 3. Statistical measures for model performance.

Name	Formulation	Ideal Value
Mean Squared Error	$MSqE = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{Z}(x_i) - Z(x_i)\right]^2$	Close to 0
Mean Standardized Error	$MStdE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{Z}(x_i) - Z(x_i)}{\hat{\sigma}^2(x_i)}\right]$	Close to 0
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Z}(x_i) - Z(x_i)\right]^2}$	The smaller the value, the better the model
Average Standardized Error	$ASE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^2(x_i)}$	Close to RMSE
Root Mean Squared Standardized Error	$RMSSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{Z}(x_i) - Z(x_i)}{\hat{\sigma}^2(x_i)}\right]^2}$	Close to 1

The mean squared error (MSqE) and mean standard error (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 root mean squared error (RMSE) is used to measure the accuracy of a model, and the smaller the value, the better the model is. The average standard error (ASE) is the average standard deviation and should be close to the RMSE value. The root mean squared standardized error (RMSSE) is used to examine the variability of the estimations (under or overestimations) and should ideally be close to 1. If the RMSSE is greater than 1, the variability of the predictions are underestimated, and vice versa [31]. By these five metrics, the various kriging models can be confidently compared against each other.

Typically, for smaller regions, this is where the process ends. However, when working with such a large scale the SOSA, or transferability of the variance, needs to be examined to ascertain if a single semivariogram model adequately applies to the whole region. Two comparative methods will be used to check the SOSA. The simplest and most basic method would be comparing the semivariogram results for each region. First, a visual inspection of the semivariogram plot is conducted as intuitively, if the assumption holds and the variance structure is consistent throughout, then the shape, scale, and behaviour should be very similar from region to region.

Likewise, the three core values from each semivariogram, namely the nugget, sill, and range, can also be used as a comparative metric. Furthermore, the five statistical measures will also be used to compare model performance for both the SOSA check and the enhancement of RK using network distances.

5. Results and Discussion

5.1. Development of A Statewide Regression Kriging Model

The regression analysis across the various regions resulted in some foreshadowing for the remaining results. Table 4 shows the results from backward stepwise selection regression and VIF analysis. The results are quite apparent that the relevant covariates are not consistent throughout the state and sub regions.

Table 4. Regression coefficients for each region.

Coefficient Values	Iowa State	Northwest	Northeast	Southwest	Southeast	North Central
Number of Data Points	19,591	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	na	−0.0217	−0.0237	−0.0305	−0.0475
Speed Limit	0.0013	0.0015	0.0009	0.0020	0.0009	na
ln(AADT)	0.0165	na	0.0258	0.0220	0.0205	0.0470
RST	−0.0418	na	na	na	na	na
Avg. Air Temp	0.0397	na	na	na	−0.0300	na
Seasonal Snowfall Total	na	0.0558	0.0226	na	na	na
No. of Red Warnings	na	−0.0004	0.0001	0.0014	na	0.0002
No. of Orange Warnings	0.00001	na	0.0002	0.0003	na	0.0005
No. of Yellow Warnings	0.0009	na	0.0040	−0.0074	0.0010	na

The resulting low R^2 values from the MLR analysis shows that these variables alone cannot be used to estimate WC ratios. Instead the minor relationships that were found can be used within kriging as a method for detrending. Though each model has a very weak R^2 value, some information can still be obtained. Variables with a positive coefficients show a positive effect on WC ratios, and vice versa. In this case, the speed limit, ln(AADT), and orange stage warnings will increase the WC ratio, which follows the findings of previous collision studies investigating traffic behaviour [6,38–40].

Likewise, an increase in the number of lanes will reduce the WC ratios as more space tends to reduce collisions, reflected by the negative sign. The magnitudes of the coefficients are small as they need to translate the values to be in-between 0 and 1, but the signs of them are quite intuitive as discussed. With the regression models in hand, the semivariogram can be generated to model the spatial trend resulting in Figure 10. With the semivariogram, a semivariogram function was then chosen, in this case, the spherical function, to determine the weighting values by solving the system of covariance equations for each unmeasured location point. This culminates in Figure 11, the final hotspot heat map generated via the Regression Kriging process.

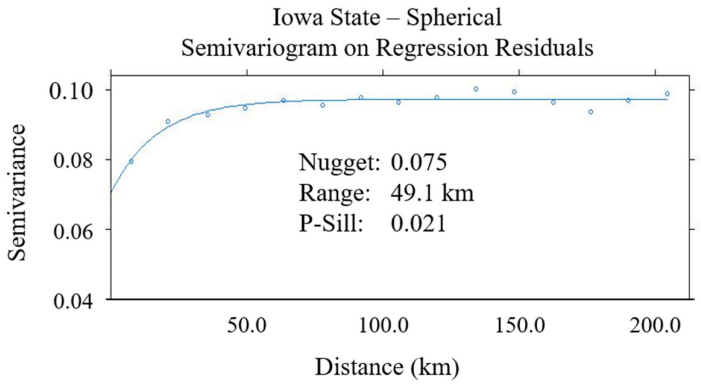

Figure 10. Semivariogram for the state of Iowa RK residuals.

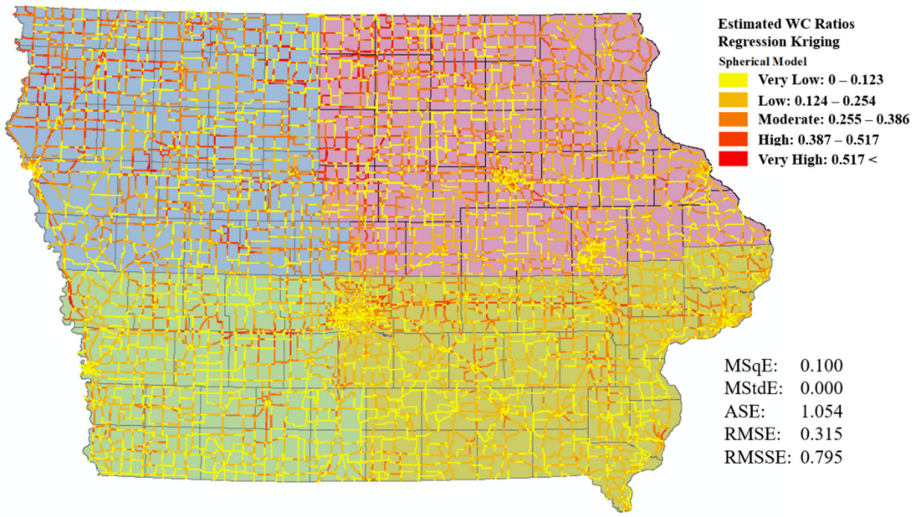

Figure 11. Regression Kriging interpolated WC ratio hotspot map.

The resulting semivariogram model, the heat map of the WC ratio distribution, and the values of the statistical measures all coincide with our preceding smaller scaled study that only focused on the Northeast quadrant of Iowa. In that previous study, RK was shown to be an effective prediction tool over a larger spatial extent, thus, showing how powerful a tool RK can be [18].

5.2. Validation of Model Transferability (Second Order Stationarity Assumption)

The results of the statewide estimates are strong; however, having a singular model represent an entire state must be checked. A good first step to checking the SOSA would be to visually compare the semivariograms between the regions. Following the same procedure used to generate the overall statewide estimates, the semivariograms, estimates, and statistical measures were iteratively generated. If the variance is transferrable throughout, the semivariogram and its values should be similar to each other. Of the three common functional semivariograms, the spherical model is used, as it was found in Wong (2021) to be the overall better performing function of the three [41]. Figure 12 shows all the semivariograms for each zone using the spherical model.

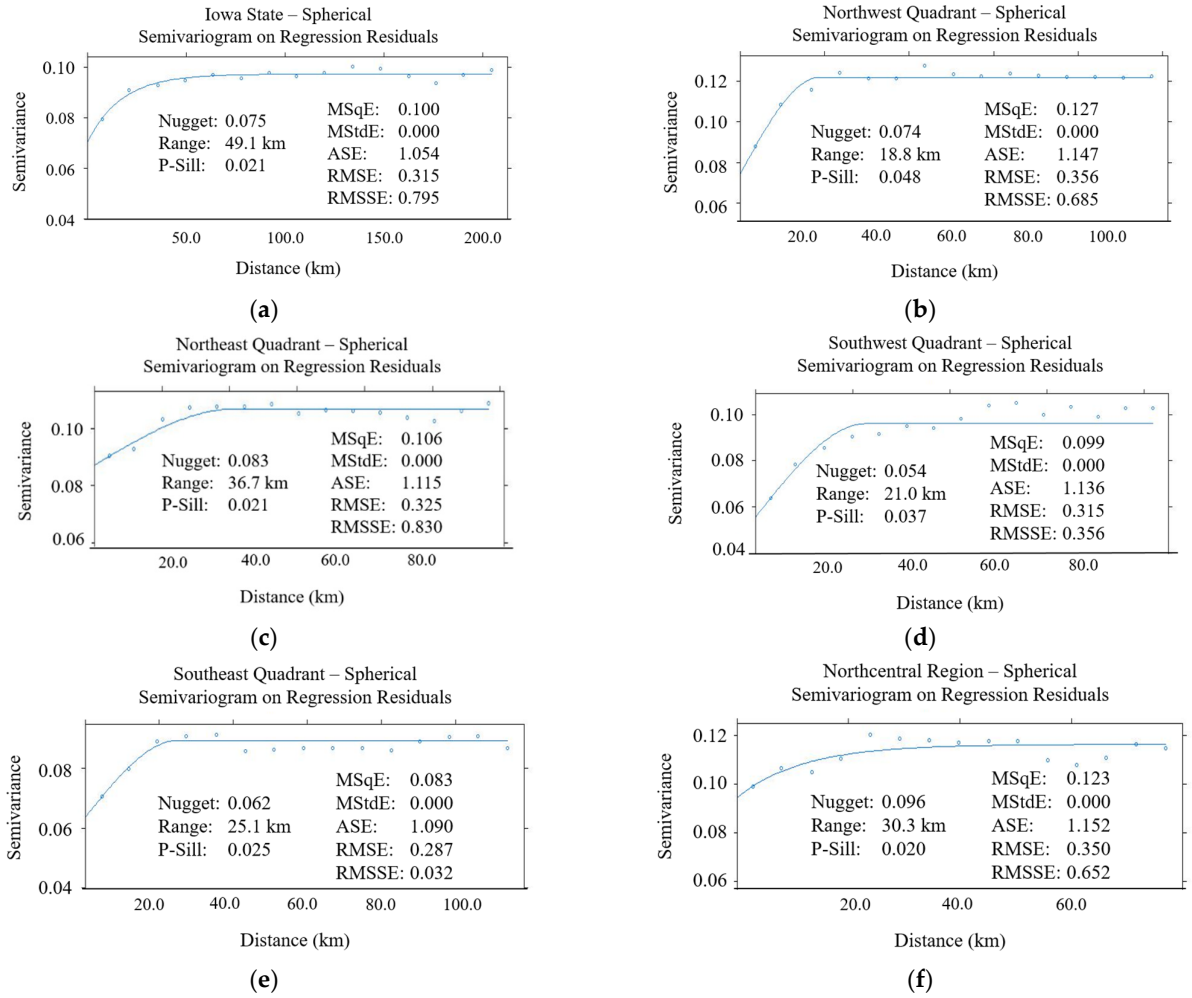

Figure 12. Spherical model semivariograms for (**a**) Iowa State, (**b**) Northwest Iowa, (**c**) Northeast Iowa, (**d**) Southwest Iowa, (**e**) Southeast Iowa, and (**f**) Northcentral Iowa.

From Figure 12, the shape of the semivariogram and the scales of the axis vary considerably from region to region. This strongly implies that the second order stationarity assumption does not hold when applying a single model to the whole state providing an indicator that further investigation is required. To do this, the semivariogram values and results from the cross validation analysis were calculated to further show how each region differs from the state significantly.

The numerical values shown within each plot in Figure 12 provides a clearer look into the differences that range between the six regions. The differences are the greatest between the northern and southern regions of the state but are smaller between each northern region and southern regions. From the statistical measures, the models for the overall state and the northern regions performs much better than in the southern regions suggesting that a single regional model may not accurately represent all regions appropriately showing that the second order stationarity assumption (SOSA) does not hold. However, given the performance of the overall state model, it can still be used as a more generalized model such as the Figure 11 estimation map.

The differences between regions could be a result of population density differences between the north and south, where the north is mostly rural while the southern half has major cities, such as Des Moines, Cedar Rapids, and Davenport. There could also be influential localized factors or features that are either not captured or averaged out in the

statewide model, but become prominent on a smaller scale, be it for better or worse. With that in mind, given the performance of the southern region models, the overall state model would be a good substitute. These results also provide further support to the findings from the previous study showing that regression kriging performs well as an estimator based on the results of the statistical measures.

5.3. Network Regression Kriging Using Road Distances

Knowing that regional models perform well, an examination into the enhancement of RK using network distances is focused on the northcentral region. Transportation events occur on the road network and the distance between any two events are bound by the network. Therefore, to improve upon the estimates, the distance values to be used in the kriging process should intuitively be the network distance. Here, RK was completed using both Euclidean and network distances and then compared.

For comparisons between using different distance measures, all three of the common semivariogram models were used along with their semivariograms. Cross validation was done for each model, and the five statistical measures were calculated. Figure 13a,b are the hotspot plots for both Euclidean and Network distances, respectively, along with the results of the cross validation and significant semivariogram values.

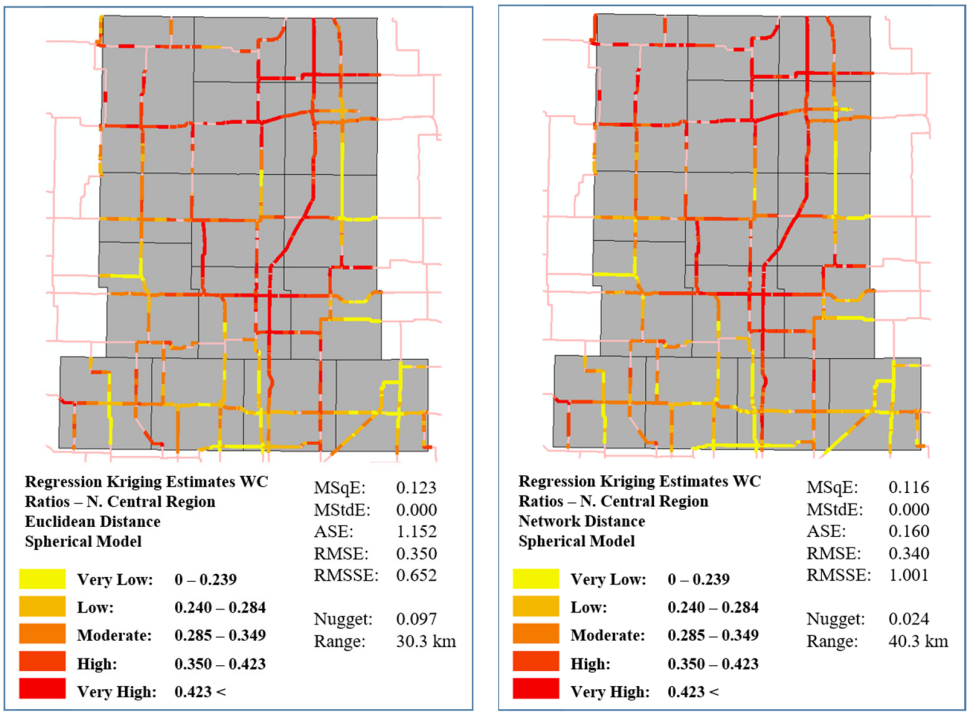

Figure 13. Regression Kriging using (**a**) Euclidean Distances and (**b**) Network Distances.

The resulting plots and values from Figure 13 show that network distances improve the model performance, estimates, and more importantly the estimation variances. The reduction in nugget values indicates that there is less inherent measurement error present within the model, and the increased ranges means the spatial effectiveness of the model is larger before it becomes no longer effective. The MSqE and RMSE have lower values showing a reduction in the errors, and the RMSSE improved to near 1.0, suggesting that it now accounts for more of the variability of the model estimates.

Finally, the ASE values are lower than, but much closer to, the RMSE values meaning the model now overestimates the outcomes, but not to the extent that it underestimated

it. By all accounts, network distances improve the RK model performance at the cost of computational complexity and run times as even this simplified analysis required 3 h of run time.

These results fall in line with the studies done by Selby and Kockelman (2013) and by Zhang and Wang (2014) when they applied network distances to their transportation problems. As mentioned in the introduction, their studies were very limited in size, scope, and timeline; thus, their results were suggestive but not conclusive. With this much larger spatial and temporal scale, this result is now considered conclusive.

6. Conclusions

The overall intent of this study was to examine the applicability of Regression Kriging on a statewide level and the resulting model's transferability and stability throughout the region via the second order stationarity assumption check. Furthermore, the study then looks into the possibility of improving RK estimates by intuitively using network distances for this transportation engineering problem. The results of this research will give regulating authorities and maintenance bodies an additional analytical and prediction tool to assess the state of their winter infrastructure and to improve maintenance operations. This research also provides a more conclusive result in support of the use of network distances in kriging for transportation problems. A summary of the findings from this study are as follows:

- The regression analysis conducted for the six regions of the study area showed that not all covariates have the same effects within each region. Despite this, the results do support previous findings of factors that are connected to higher collision rates such as higher speed limits, less number of lanes, greater traffic volumes, and deteriorating road conditions. This shows how covariate selection itself is an important step worthy of its own project scope before applying it to regression kriging as it lays an important foundation for RK to build upon to further increase the estimation accuracy.
- The performance of regression kriging at a much larger scale with increased data quantity and density was found to be very robust based on the five statistical measures used. However, we found that the second order stationarity assumption did not hold, as the semivariogram and cross validation results for each of the six regions differed substantially. This also showed how the urban/rural setting of the region can greatly affect the model's performance whereby rural road networks benefit from this process the most. However, an overall model is still adequate should higher fidelity not be required or if certain regions have insufficient data quality or quantity. This demonstrates how powerful a tool that RK can be for winter collision modelling.
- Finally, RK was enhanced by using road network distances over Euclidean distances. By the semivariogram value results and the five statistical measures, it was clear that RK with network distances outperformed its Euclidean distance counterpart. Applied over a large spatial scale, over a much larger and more complex connected road network, this study provides conclusive evidence that network distances can improve kriging estimation performance.

This study does come with some limitations and assumptions that can be expanded upon in future studies. Some limitations with suggestions for further research are as follows:

- One major assumption made is that the placement of the RWIS stations is optimal and substantially affects the outcome of weather induced collisions. The true effectiveness of RWIS and its warning system may provide insight into their effectiveness in reducing collisions, and its numerical valuation may be incorporated as a covariate.

- The study did not consider the effects of maintenance operations that could skew the collision frequencies being recorded. Incorporating maintenance characteristics, such as plowing schedules or chemical use, may further improve the regression portion of the analysis.
- This study used Iowa for its relatively uniform terrain characteristics, which may limit the results to more mountainous or hilly regions. Repeating this study, but in a different state or country altogether with drastically different geography, will further develop the process and also show if it can be applied universally or if regional adjustments are required.
- The datasets used are subject to human error and biases, especially for data that are recorded manually. Fortunately, environmental data is mostly automated now; however, collision and near-miss reports are not. The development and utilization of automated monitoring systems for collisions and near misses will reduce errors and biases while also providing the added variable of near misses.
- Expanding the weather source dataset and its quantity and quality may improve upon the environmental aspects of the modelling process. Additional covariates, such as dew point temperatures, visibility, or solar factors, may be considered.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: A.H.W. and T.J.K.; data collection: A.H.W.; analysis and interpretation of results: A.H.W. and T.J.K.; draft manuscript preparation: A.H.W. and T.J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant: RGPIN-2017-06448

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Collision, road, and boundary data is freely available from Iowa DOT Open Data at https://data.iowadot.gov/, accessed on 20 November 2019. Weather and environmental data is freely available from the Iowa State University Mesnoet Database at https://mesonet.agron.iastate.edu/, accessed on 15 November 2019.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. US DOT Federal Highway Administration. How Do Weather Events Impact Roads. 2020. Available online: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm (accessed on 25 July 2020).
2. Royal Canadian Mounted Police. Just the Facts—Winter Driving. 2019. Available online: https://www.rcmp-grc.gc.ca/en/gazette/just-the-facts-winter-driving (accessed on 1 April 2021).
3. Norrman, J.; Eriksson, M.; Lindqvist, S. Relationships between road slipperiness, traffic accident risk and winter road maintenance activity. *Clim. Res.* **2000**, *15*, 185–193.
4. Heqimi, G. Using Spatial Interpolation to Determine Impacts of Snowfall on Traffic Crashes. Master's Thesis, Michigan State University, East Lansing, MI, USA, 2016.
5. Asano, M.; Hirasawa, M. Characteristics of traffic accidents in cold, snowy Hokkaido, Japan. In *Proceedings of the Eastern Asia Society for Transportation Studies*; Eastern Asia Society for Transportation Studies: Tokyo, Japan, 2003; Volume 4, pp. 1426–1434.
6. Andersson, A.K. *Winter Road Conditions and Traffic Accidents in Sweden and UK-Present and Future Climate Scenarios*; Department of Earth Sciences, Institutionen för Geovetenskaper: Gothenburg, Sweden, 2010.
7. American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*, 1st ed.; AASHTO: Washington, DC, USA, 2010; Volume 1.
8. Reyad, P.; Sacchi, E.; Ibrahim, S.; Sayed, T. Traffic conflict–based before–after study with use of comparison groups and the empirical Bayes method. *Transp. Res. Rec.* **2017**, *2659*, 15–24.
9. Thakali, L.; Kwon, T.J.; Fu, L. Identification of crash hotspots using kernel density estimation and kriging methods: A comparison. *J. Mod. Transp.* **2015**, *23*, 93–106.
10. Selby, B.; Kockelman, K. *Spatial Prediction of AADT in Unmeasured Locations by Universal Kriging;* Transportation Research Record: Washington, DC, USA, 2011.

11. Zhang, D.; Wang, X.C. Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC. *J. Transp. Geogr.* **2014**, *41*, 107–115.
12. Gu, L.; Kwon, T.J.; Qiu, T.Z. A Geostatistical Approach to Winter Road Surface Condition Estimation using Mobile RWIS Data. *Can. J. Civ. Eng.* **2019**, *46*, 511–521.
13. Zeng, Q.; Huang, H.; Pei, X.; Wong, S.C.; Gao, M. Rule extraction from an optimized neural network for traffic crash frequency modeling. *Accid. Anal. Prev.* **2016**, *97*, 87–95.
14. Abdelwahab, H.T.; Abdel-Aty, M.A. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transp. Res. Rec.* **2001**, *1746*, 6–13.
15. Chang, L.Y. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Saf. Sci.* **2005**, *43*, 541–557.
16. Abellán, J.; López, G.; De OñA, J. Analysis of traffic accident severity using decision rules via decision trees. *Expert Syst. Appl.* **2013**, *40*, 6047–6054.
17. Das, A.; Abdel-Aty, M. A genetic programming approach to explore the crash severity on multi-lane roads. *Accid. Anal. Prev.* **2010**, *42*, 548–557.
18. Wong, A.H.; Kwon, T.J. Development and Evaluation of Geostatistical Methods for Estimating Weather-Related Collisions—A Large Scale Case Study. *Transp. Res. Rec.* **2021**, 1–13, doi:10.1177/03611981211020008.
19. Olea, R.A. *Geostatistics for Engineers and Earth Scientists*; Springer Science & Business Media: New York, NY, USA, 1999.
20. Einax, J.; Soldt, U. Geostatistical and multivariate statistical methods for the assessment of polluted soils—merits and limitations. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 79–91.
21. Cressie, N. The Origins of Kriging. *Math. Geol.* **1990**, *22*, 239–252.
22. Iowa Department of Transportation. Iowa DOT Open Data. Available online: https://public-iowadot.opendata.arcgis.com/ (accessed on 31 May 2020) .
23. Iowa State University. Iowa Environmental Mesonet. Available online: https://mesonet.agron.iastate.edu/ (accessed on 6 January 2020).
24. De Pauw, E.; Daniels, S.; Brijs, T.; Elke, H.; Geert, W. *The Magnitude of The Regression to the Mean Effect In Traffic Crashes*; International Co-operation on Theories and Concepts in Traffic Safety: Vienna, Austria, 2014.
25. NOAA. Iowa Climate Normals Map. 16 May 2020. Available online: https://www.weather.gov/dmx/climatenormals (accessed on 16 May 2020).
26. Eguía, P.; Granada, E.; Alonso, J.M.; Arce, E.; Saavedra, A. Weather datasets generated using kriging techniques to calibrate building thermal simulations with TRNSYS. *J. Build. Eng.* **2016**, *7*, 78–91.
27. Atkinson, P.M.; Lloyd, C.D. Mapping Precipitation in Switzerland with Ordinary and Indicator Kriging. *J. Geogr. Inf. Decis. Anal.* **1998**, *2*, 72–86.
28. Khan, G.; Qin, X.; Noyce, D.A. Spatial Analysis of Weather Crash Patterns in Wisconsin. *J. Transp. Eng.* **2008**, *134*, 191–202.
29. NOAA, Climate Reports: February & Winter (Dec–Feb) 2015–2016. 2016. Available online: https://www.weather.gov/dvn/Climate_Monthly_02_2016 (accessed on 30 August 2021).
30. NOAA, Climate Reports: February & Winter (Dec–Feb) 2016–2017. 2017. Available online: https://www.weather.gov/dvn/Climate_Monthly_02_2017 (accessed on 30 August 2021).
31. ESRI. *ArcGIS Desktop: Release 10*; Environmental Systems Research Institute: Redlands, CA, USA, 2011.
32. Microsoft. *Excel 2016*; Microsoft: Redmond, WA, USA, 2016.
33. R Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
34. Pebesma, E. Multivariable geostatistics in S: The gstat package. *Comput. Geosci.* **2004**, *30*, 683–691.
35. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
36. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 3rd ed.; John Wiley & Sons Inc.: New York, NY, USA, 2001.
37. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; Springer International Publishing: New York, NY, USA, 2015.
38. Abdel-Aty, M.A.; Radwan, A.E. Modeling traffic accident occurrence and involvement. *Accid. Anal. Prev.* **2000**, *32*, 633–642.
39. El-Basyouny, K.; Sayed, T. Comparison of two negative binomial regression techniques in developing accident prediction models. *Transp. Res. Rec.* **2006**, *1950*, 9–16.
40. Usman, T.; Fu, L.; Miranda-Moreno, L.F. A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level. *Accid. Anal. Prev.* **2012**, *48*, 368–378.
41. Wong, A.H. Advances in Kriging-Based Modelling Approaches of Winter Weather Vehicular Collisions—A Region-Wide Geostatistical Investigation. Master's Thesis, University of Alberta, Edmonton, AB, Canada, 9 June 2021.