THE IMPACT OF SPATIAL INTERPOLATION TECHNIQUES ON SPATIAL BASIS RISK FOR WEATHER INSURANCE: AN APPLICATION TO FORAGE CROPS

by
Daniel Turenne

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Asper School of Business
University of Manitoba
Winnipeg, Canada

Abstract

Weather index insurance has become a popular subject in agricultural risk management. Under these policies farmers receive payments if they experience adverse weather for their crops. Spatial basis risk is the risk that weather observed at stations does not correspond to the weather experienced by the farmer. The objective of this research is to determine to what extent spatial basis risk can be impacted by the interpolation technique used to estimate weather conditions. Using forage crops from Ontario, Canada, as an example, a temperature based insurance index is developed. Seven different interpolation methods are used to estimate indemnities for forage producers. Results show that the number of weather stations in the interpolation area has a larger impact on spatial basis risk than the choice of interpolation technique. For insurers wishing to implement this type of insurance, more focus should be placed on increasing the number of available weather stations.

Acknowledgments

I would like to thank both of my parents, Mike and Kathy, for all of the support they have given me over the past two years. From kind words of encouragement when I was feeling disheartened to hot meals when I was feeling tired and overwhelmed my parents have always done everything they could to give me all I needed.

I would like to thank my supervising professor Dr. Lysa Porth as well as committee members Dr. Milton Boyd, Dr. Barry Coyle, and Dr. Xuemiao Hao for all the time and effort that they have spent helping me complete my research. Their feedback and edits were invaluable to me throughout the writing process.

For providing me with all of the data for my analysis I would like to thank the risk management firm Agricorp. Without access to their data this research would not have been possible.

I would also like to thank the University of Manitoba and the Government of Manitoba for their extremely generous financial support through the Manitoba Graduate Scholarship. This scholarship has given me the freedom to devote my time and efforts to completing my research, and for that I am truly thankful.

Finally, I would like to thank all of the many contributors to the R-Sig-Geo mailing list. R-Sig-Geo is volunteer run and provides a valuable resource for people doing geostatistical analysis in R. Without the patience and advice of their many helpful members I would have never gotten this thesis off the ground.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Weather index insurance has become a popular subject in agricultural risk management with many papers devoted to its discussion (Heimfarth and Musshoff, 2011; Lin et al., 2015; Okhrin et al., 2013). Under these policies farmers receive payments if their crops experience adverse weather during the growing season. These policies offer many benefits, such as reduced administrative costs and decreased adverse selection and moral hazard. However, this type of insurance is particularly susceptible to the problem of spatial basis risk (Lin et al., 2015). Spatial basis risk occurs when the weather observed at weather stations does not match the weather experienced by the farmer's crops, causing improper indemnities to be paid to the farmer (Dick and Stoppa, 2011). However, spatial basis risk may be reduced through the use of averaging and spatial interpolation techniques such as inverse distance weighting and kriging. These techniques make it possible to incorporate multiple weather stations in the estimation process rather than using only the single closest station, potentially resulting in more accurate estimates and thereby reducing spatial basis risk.

The objective of this study is to determine if an insurer's choice of spatial interpolation technique can impact the amount of spatial basis risk in a weather based insurance model. To evaluate the performance of different spatial interpolation techniques, temperature based policies for forage crops in Ontario, Canada, are considered as an example. A weather insurance index is developed based on cooling degree days, a weather metric which represents the excess heat stress that the crops experience over the growing season. Seven different interpolation methods are applied to temperature data and estimated indemnities are calculated for forage producers across the province. By analyzing the correlation between the estimated indemnities and reported forage yields, the impact of the different interpolation techniques on spatial basis risk is quantified.

The results of this study will provide valuable insight for any insurer who wishes to develop these types of policies. By comparing the differences in spatial basis risk for various spatial interpolation techniques, it can be determined if there is any benefit to using more complex, computationally intensive methods such as spatio-temporal regression kriging over simpler methods such as inverse distance weighting. If the difference in spatial basis risk between these techniques is minimal, an insurer may prefer to avoid the added complexity of methods such as kriging in favor of simpler, easier to understand methods.

This study also contributes to the current body of literature by comparing the impact of spatial basis risk in regions with varying amounts of weather stations. The analysis is split into two regions, one with a large number of weather stations and one with a small number of weather stations. This distinction helps to determine if different approaches need to be considered for areas with varying amounts of weather stations.

In addition, this study examines whether temperature is a suitable variable for designing index-based insurance policies. In general, temperature mainly impacts forage quality while precipitation mainly impacts forage yields (Buxton, 1995), therefore it is likely that rainfall is a better proxy for forage yields than temperature. Temperature was chosen as the variable of interest in this study because the data sets used in the analysis have more extensive observations for temperature than for rainfall, which results in more information being available for spatial interpolation. In addition, past research has noted that relatively little is known about the basis risk for high temperature based products (Clarke et al., 2012) and this study provide valuable insight into the basis risk associated with these policies.

The remainder of this paper proceeds as follows. First, past research on the subjects of forage, basis risk, and spatial interpolation is reviewed. Second, the data used in the analysis is described. Third, the methodology for the analysis is outlined, including all of the spatial interpolation techniques and the design of the insurance index. Lastly, the results of the analysis are presented and the paper concludes with a discussion of the results and a summary.

Chapter 2: Background Review

2.1 Forage Insurance Challenges

This section consists of a brief background on the subject of forage crops, including their importance to the agricultural sector in Canada and the difficulties in designing insurance plans to protect them. Forage is defined as plant matter that is consumed by livestock for food and includes alfalfa, straw, hay, and other grasses (AAFC, 2016). In Canada the most popular forage products are compressed bales of hay and alfalfa with 19 million acres planted in 2007, grown mostly in the western provinces of Alberta and Saskatchewan. The production of forage materials is essential to Canadian agriculture, with 80% of cows for beef production and 60% of dairy cows being dependent on forage for feed (Porth and Tan, 2015). In addition Canadian hay and alfalfa are both valued around the world as high quality feed for dairy cows and other livestock, and in 2007 Canada exported over $194 million worth of processed forage products in the form of compressed hay bales, alfalfa pellets, and alfalfa cubes (AAFC, 2012b).

Canada exports forage all over the world, however, Japan and the United States are the largest single purchasers of Canadian forage exports, accounting for 67% and 25% of 2007 exports respectively (Yungblut and Jalbert, 2012). Given these statistics it is clear to see that the forage industry plays an important role in feeding livestock not just within Canada's borders, but all around the world as well.

Despite the importance of the forage industry in Canada and despite the fact that roughly 44% of Canada's farm lands are devoted to forage, participation rates in forage insurance programs have historically been much lower than for other forms of crop insurance. As of 2015 only 20% of all forage acres and 12% of all pasture acres in Canada were insured even though premiums were heavily subsidized (Porth and Tan, 2015). Ever since 1967 when forage became covered under AgriInsurance, there have been two main obstacles that have stood in the way of increasing participation rates. Since forage is generally consumed as necessary by livestock, many producers choose not to keep track of the amount of forage they produce. This makes it very difficult to make accurate estimates of forage yields which in turn makes it difficult to design an effective insurance plan that forage producers will want. In addition, forage is a commodity with a market price that can vary significantly over the course of a year, making the process of establishing actuarially fair premiums very challenging (AAFC, 2012a).

Faced with these low participation rates, the Canadian government has been pressured to help forage producers by providing disaster assistance. In the years since 2008 there have been four initiatives organized by the Canadian government in order to provide assistance to forage producers. The first two initiatives provided money to offset rising feed costs in times of drought while the other two initiatives covered feed and transportation costs during times of flooding (AAFC, 2012a). It is estimated that $148 million was distributed to forage producers through these four initiatives, demonstrating the necessity of adequate forage protection in Canada. Troubles in the forage industry can ripple throughout the entire livestock sector, and so it is vital that producers be offered affordable, effective insurance to protect themselves against downturns.

2.2 Basis Risk Review

This section includes a brief discussion about basis risk in general as well as a review of the existing literature regarding how spatial basis risk can be analyzed and quantified. Basis risk is defined as "...the difference between the loss experienced by the farmer and the payout triggered."(Dick and Stoppa, 2011, p. 22). Basis risk can result in a farmer receiving a payment without experiencing any losses or experiencing losses without receiving a payment. This definition of basis risk specifically refers to the loss experienced by the farmer, which is usually expressed as a percentage of historical average yield, however, because this information was not available for this analysis, yield was used as a proxy for losses under the assumption that lower yields correspond to higher losses and vice versa. In general, there are three different types of basis risk (Dick and Stoppa, 2011):

- **Product Basis Risk:** Occurs when there is no clear relationship between losses and the chosen weather index.
- **Temporal Basis Risk:** Occurs when insurance phases are not temporally aligned with the intended crop growth stage.
- **Spatial Basis Risk:** Occurs when there is local variation in the weather index within the area surrounding a weather station.

As an example of product basis risk, an insurance policy might be designed using temperature when in reality the amount of rainfall has a greater impact on crop losses. As an example of temporal basis risk, a farmer might receive no rain during the most important growth phases of their crops, yet still receive enough rain during the rest of the season to avoid triggering a payment. As an example of spatial basis risk, a weather station may record no rain when in fact the farmer's crops, located many kilometers away, receives enough rain to avoid any losses. While all of these situations can result in improper indemnities being paid to farmers, the focus of this research is on spatial basis risk, which occurs as a result of uncertainty in estimating the weather index at the farmer's property (Dick and Stoppa, 2011).

There are often a limited number of weather stations available to the insurer, and as a result indemnities must be calculated based on observations from weather stations that are not located in the farmer's fields. Therefore, there is a chance that the weather observed at the station may not match exactly with the weather experienced by the farmer, creating uncertainty when indemnities are calculated. Without placing weather stations in every farmer's field, a certain degree of spatial basis risk is unavoidable when dealing with index-based insurance. However, spatial basis risk may be reduced by using spatial interpolation techniques to create better estimates of weather conditions in the farmer's fields. Basis risk continues to be an important and challenging area of research, often limited by the amount of reliable data that is available (AAFC, 2012a; Lin et al., 2015).

Major (1999) was one of the first academic papers to propose a method for analyzing basis risk as it applies to catastrophe derivative contracts. These contracts are designed to payout if an index based on insured losses exceeds a certain threshold, and were first developed to help insurers hedge their positions. First, in order to simulate weather and loss data, a computer model was constructed based on the geographic features of the insured property and the location of the property underlying the index. A large number of simulations were then performed in order to calculate the sample correlation in between the index value and the losses from the insurer's book of business. A high degree of correlation, positive or negative, indicates that there is very little basis risk, meaning that the conditions measured by the index closely reflect the losses sustained by the insured. Conversely a correlation coefficient close to zero would indicate that the index gives very little information about the losses sustained by the insured. By comparing models Major (1999) was able to conclude that a model where indices are measured by Zip codes would help increase correlation when compared to a model where indices are measured by state.

Paulson and Hart (2006) approached the spatial basis risk problem by using a kriging technique with weather data from the state of Iowa in order to estimate the level of rainfall at unobserved locations. For comparison, a simple inverse distance weighted estimator was also used to provide estimates of rainfall. It was found that both methods produced nearly equivalent results despite the sophistication of the kriging model. In addition a Monte Carlo analysis was used to calculate premiums for precipitation based index insurance using both the kriging and inverse distance weighted models, and both methods were found to be nearly equivalent with respect to ratemaking. When they performed a historical analysis to determine how this index insurance would have performed in the past it was shown that it would have been successful in triggering payouts in areas of low precipitation. They note, however, that the loss areas for this type of insurance product would be geographically concentrated due to the nature of precipitation events. This implies that any such insurance policies would require a geographically large coverage area in order for the insurer to create a sufficient risk pool, meaning that these policies would be difficult for smaller, private companies to administer.

Norton et al. (2010) considered the inclusion of variables such as longitude, latitude, and elevation into their model when measuring spatial basis risk for both temperature and precipitation based insurance products. They concluded that for temperature events, differences in altitude have the most significant impact on basis risk while for precipitation events, distance in between the unobserved location and the weather station is the most significant factor. Interestingly the analysis also shows that other than these two relationships, there is little evidence to support any other relationships in between indemnities and other geographic variables. Their final conclusion was that the best possible approach to managing spatial basis risk might be to create insurance policies based on multiple weather stations. These policies would be relatively easy to price and would be easy to customize based on the perceived needs of the consumer. As such future research should focus on methods for calculating optimal weights to assign to each weather station.

2.3 Spatial Interpolation Review

The focus of the methodology of this study is on the use of several different spatial interpolation techniques. With this in mind this section is dedicated to a review of existing literature on the subject of spatial interpolation, in particular kriging. Kriging methods are the most statistically sophisticated of all of the methods in this analysis and so their background is covered here. The background of the simpler methods such as inverse distance weighting are covered in the methodology section. Kriging is a term that is often used synonymously with geostatistics and spatial interpolation. Originating in the 1950's, kriging was initially used to determine ore grades in South African mines (Krige, 1951). Since then the use of kriging has expanded into many different fields, including meteorology and agriculture.

Chien et al. (1997) used both ordinary kriging and cokriging to make predictions of soil properties in Taiwan. Of interest were the concentrations of iron, phosphorus, calcium and magnesium. The estimation of these properties is extremely important in managing agricultural fields, however, the sampling process can be time consuming and labor intensive. The results of this research were able to show that by using geostatistical techniques like kriging, the existing sampling density could potentially be reduced by up to half and still provide enough spatial information to make meaningful estimates of soil properties.

Hudson and Wackernagel (1994) provides one of the earliest examples of using universal kriging as a method for interpolating temperatures. By using universal kriging with temperature data from Scotland, they were able to eliminate concerns of stationarity in the data, and concluded that performance was improved when compared to ordinary kriging. They note in their conclusion that further studies would benefit from incorporating data regarding elevation and proximity to large bodies of water.

Wu and Li (2013) used regression kriging to interpolate temperatures across the United States using data recorded at weather stations. The goal of this paper was to improve temperature prediction methods in response to increased demands caused by the greenhouse effect. They showed that by including elevation information in the kriging model, as well as cross products and squares of explanatory variables, temperature predictions became more accurate, and they concluded that regression kriging outperformed standard kriging algorithms employed by the popular ArcGIS software. In addition they showed that latitude and elevation were the most important of the explanatory variables included in the model.

Kilibarda et al. (2014) used spatio-temporal regression kriging with data from the MODIS satellite project to estimate daily temperatures across the entire globe. Temperature was modeled as a function of latitude and time of year. They were able to create predictions with an average error of ±3°C, however, it was found that spatio-temporal regression kriging did not significantly reduce the estimation error when compared to standard regression kriging. Despite the lack of improvement in the estimation error, they note that spatio-temporal techniques greatly reduce the complexity of parameter estimation, since only one model must be estimated for the entire season rather than estimating a new model for each day as would be required in a standard regression kriging model.

Chapter 3: Data

This section describes the data that was used in the analysis. The steps used to prepare the data for analysis are outlined and some basic geographic information about the Canadian province of Ontario is provided. Finally, the number of weather stations and forage farms in each region of Ontario is summarized and discussed. All data used in this analysis was provided by Agricorp, an agricultural risk management firm from Ontario.

3.1 Data Sets

Two separate data sets are used for the analysis of this paper. The first data set contains information collected from weather stations located throughout Ontario, Canada, from the years 1967 to 2004 in the months of April to August. Twenty-four of these weather stations have observations from 1967-2004 while the remaining stations only have observations from 1997-2004. Included in the weather data are the daily maximum and minimum temperatures in Fahrenheit as well as the longitude and latitude for each station. Many days also included measurements of rainfall, however there are more missing observations for rainfall than temperature.

Temperature was chosen as the variable of interest in this study because the data set has more temperature observations than rainfall observations, which results in more information being available for spatial interpolation. Since the goal of this study is to examine spatial basis risk, and not to develop the most optimal weather insurance index, it is preferable to use the weather variable with more complete and extensive observations. In addition, past research has concluded that more research needs to be done on basis risk in high temperature insurance policies (Clarke et al., 2012) and this study will help to address this gap in the literature.

The second data set contains information regarding Ontario forage producers. In addition to the latitude and longitude of each farm, the data set also contains the total acreage and the reported yield in tonnes per acre. This data was collected from 1981 to 2004, however, not all farms reported yields in each year. Forage crops can be harvested up to three times per year, however this data set provides information on the first harvest exclusively. The first harvest of forage is generally performed at the end of June or the beginning of July, therefore for this study the growing season is considered to be from April 1st to June 30th for a total of ninety-one days.

To prepare the data for analysis, the following steps were taken.

1. All temperatures were converted from Fahrenheit to Celsius $\rightarrow$ °C $= \frac{5}{9}(°F - 32)$.
2. Elevation data was added for each farm and weather station.
3. The distance to the Great Lakes was calculated for each farm and weather station.
4. Six weather stations had duplicate entries recorded at the same location which were removed.
5. One weather station was geographically isolated in the middle of Hudson Bay and was removed.
6. Five farms were geographically isolated and were excluded to restrict the analysis to within the borders of Ontario

All elevation data was obtained using the “elevation” function in the R package “rgbif”. Given a set of latitude and longitude coordinates this function retrieves elevation data using the Google Elevation API. For any weather stations located on the surface of the Great Lakes, the elevations were changed to accurately reflect the elevation on the surfaces of the lakes. Stations on Lake Superior were changed to 183 m, Lake Michigan and Lake Huron to 177 m, Lake Erie to 174 m, and Lake Ontario to 75 m (Herdendorf, 1982).

In addition to elevation data, the distance to the nearest large body of water was calculated using the ArcGIS software package. For the purposes of this research, "large body of water" includes all five Great Lakes along with the channels and locks between them. Distance to large bodies of water can have a significant impact on the temperatures of a region, and so it is important to include this variable in the analysis.

3.2 Region of Study - Ontario

At just over one million km^2, Ontario is Canada's second largest province and is bounded approximately by 42°N to 57°N latitude and 75°W to 95°W longitude. The climate of Ontario is characterized by cold, dry polar air coming from the Arctic during the winter and warm, moist air coming from the Gulf of Mexico during the summer. As expected, temperatures in Ontario increase from North to South, however, they can also be influenced by the presence of the Great Lakes as well as Hudson Bay (Baldwin et al., 2011). As an example of the impact these large bodies of water can have on temperatures, consider the example of Winnipeg, Manitoba and Cochrane, Ontario. Although these two locations have approximately the same latitude, Cochrane has about one thousand less growing degree days than Winnipeg (Baldwin et al., 2011). This is caused by Cochrane's proximity to Hudson Bay, a source of cold air from Canada's arctic. These types of patterns can be observed in other areas of the province as well, such as warm air coming from the South being cooled by Lake Superior, causing drops in temperature along the northern shore.

Ontario also contains several elevated highlands in the areas around Thunder Bay, Algonquin Park, and Sault Ste. Marie. It has been shown that there is a negative correlation in between elevation and temperature (Wu and Li, 2013), and as a result these highlands have fewer growing degree days than the areas surrounding them (Baldwin et al., 2011). Figure 3.1 shows a detailed map of the elevation of Ontario. Ontario is bordered in the

North by Hudson Bay, and so the elevation is close to sea level in these areas. Moving South, the elevation gradually increases to an average of around 175 m in the Great Lakes region and then slowly decreases to 75 m around Lake Ontario.

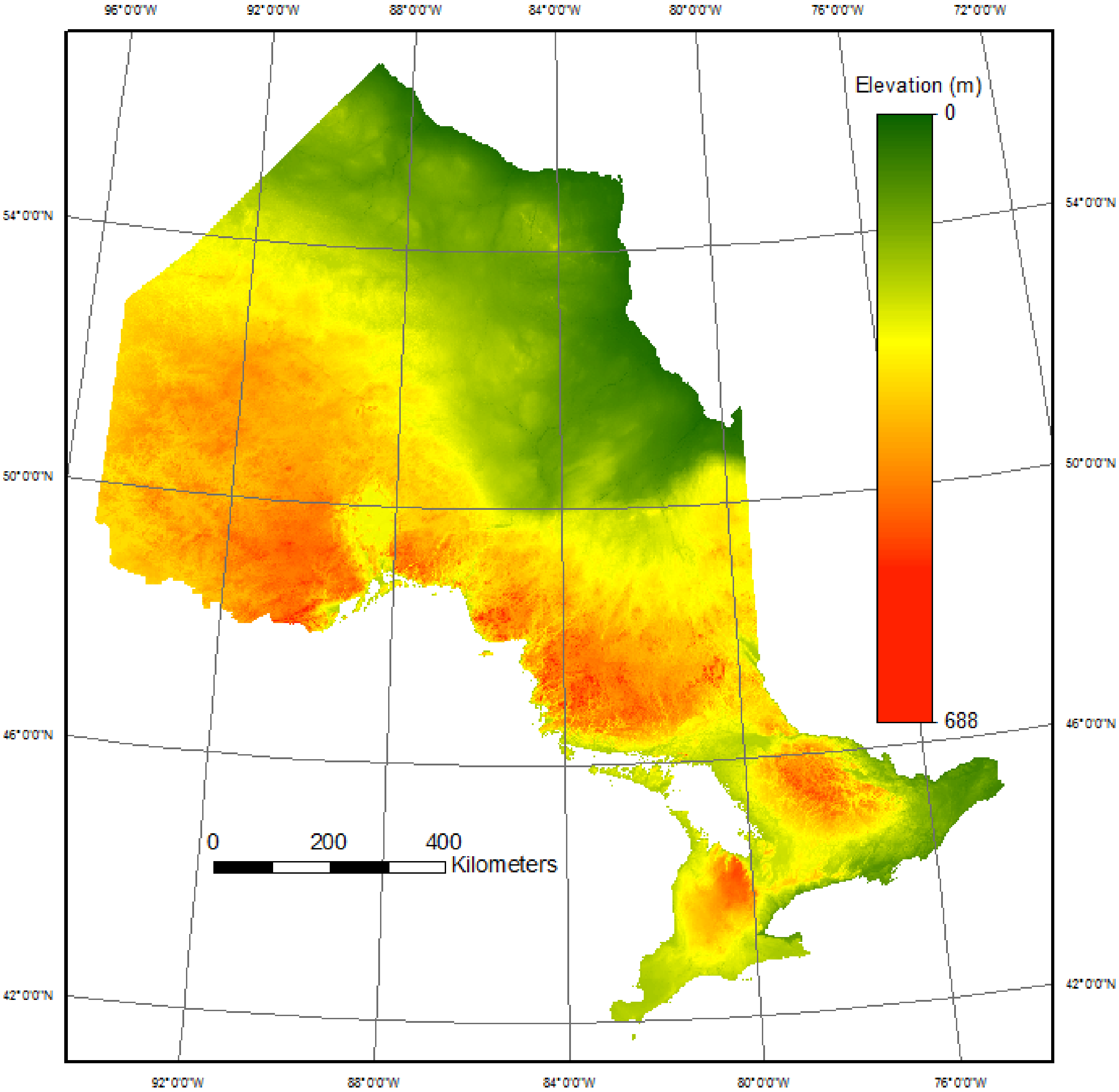
Due to a lack of long-term consistent weather data, it is difficult to analyze and identify any temporal trends in temperature. Studies by Environment Canada have shown that from the late 1800's onward a general warming trend has been present in Ontario, however it is unclear whether this is a result of short or long-term climate change (Baldwin et al., 2011). In more recent years, it has been observed from satellite data that subarctic regions in Ontario have had progressively earlier thaws, leading to a longer active growing season (Baldwin et al., 2011).

3.3 Distribution of Stations and Farms

Figure 3.2 shows a map of the agricultural divisions in Ontario while Table 3.1 provides statistics regarding the number of farms and weather stations in Ontario for each year. For the purposes of this analysis "southern Ontario" refers to the four southernmost agricultural divisions while "northern Ontario" refers to the northernmost agricultural division.

Figure 3.3 shows how the weather stations are distributed across the province, with the vast majority being located in southern Ontario. A similar pattern is visible in Figure 3.4, which shows the location of the long-term weather stations with data going back to 1967. In terms of spatial interpolation, southern Ontario represents a nearly optimal scenario where weather stations are very densely distributed throughout the entire prediction space. This dense distribution of weather observations is ideal for spatial interpolation, and it is expected that the interpolation methods will perform well in this region.

Figure 3.1: Elevation Map of Ontario



Note: This map was generated from the Ontario Provincial Digital Elevation Model - Version 3.0. The vertical grid lines represent longitude while the horizontal grid lines represent latitude.

Table 3.1: Summary of the Number of Weather Stations and Forage Farms in Each Region of Ontario

Region		1997	1998	1999	2000	2001	2002	2003	2004
All Ontario	Stations	189	210	212	205	197	179	170	154
	Farms	286	330	287	236	267	239	226	56
South Ontario	Stations	160	167	169	164	157	141	135	121
	Farms	221	264	224	187	210	188	180	26
North Ontario	Stations	29	43	43	41	40	38	35	33
	Farms	65	66	63	49	57	51	46	30

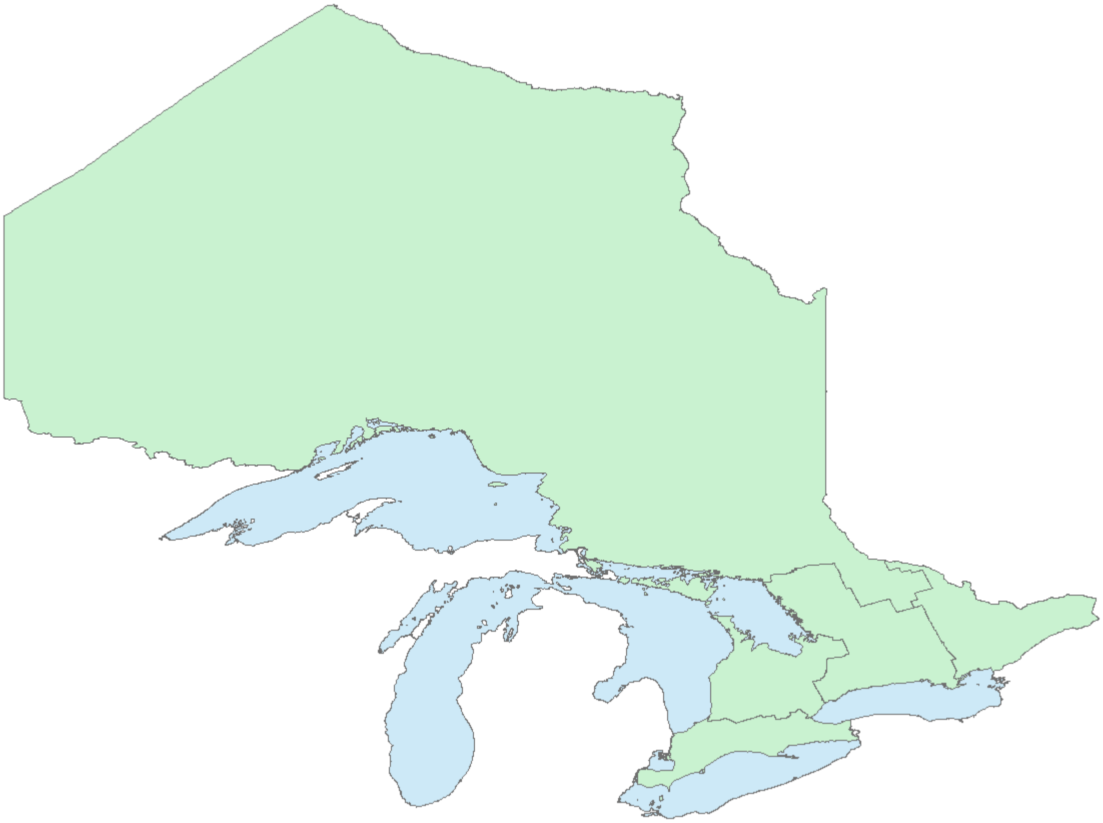
Note: This table shows the number of forage farms and weather stations that were used in each year of the analysis.

Conversely, northern Ontario represents a very poor scenario for performing spatial interpolation. Weather stations are grouped together in several small clusters that are separated by large distances, with many areas having no observations at all. Using spatial interpolation in an area like this is analogous to making predictions out of sample, and as such it is expected that the spatial interpolation methods will perform significantly worse in this region. In addition, northern Ontario has very few long-term weather stations, meaning that indemnities for farms are often calculated based on a weather station that is very far away.

In Figure 3.5 this trend continues with respect to forage farms. The vast majority of farms are located in the four southern agricultural divisions while very few are located in the North. By comparing Figure 3.3 and Figure 3.5 it is clear that the northern weather stations are located in roughly the same areas as the forage farms, however there are still large areas where there are no weather observations.

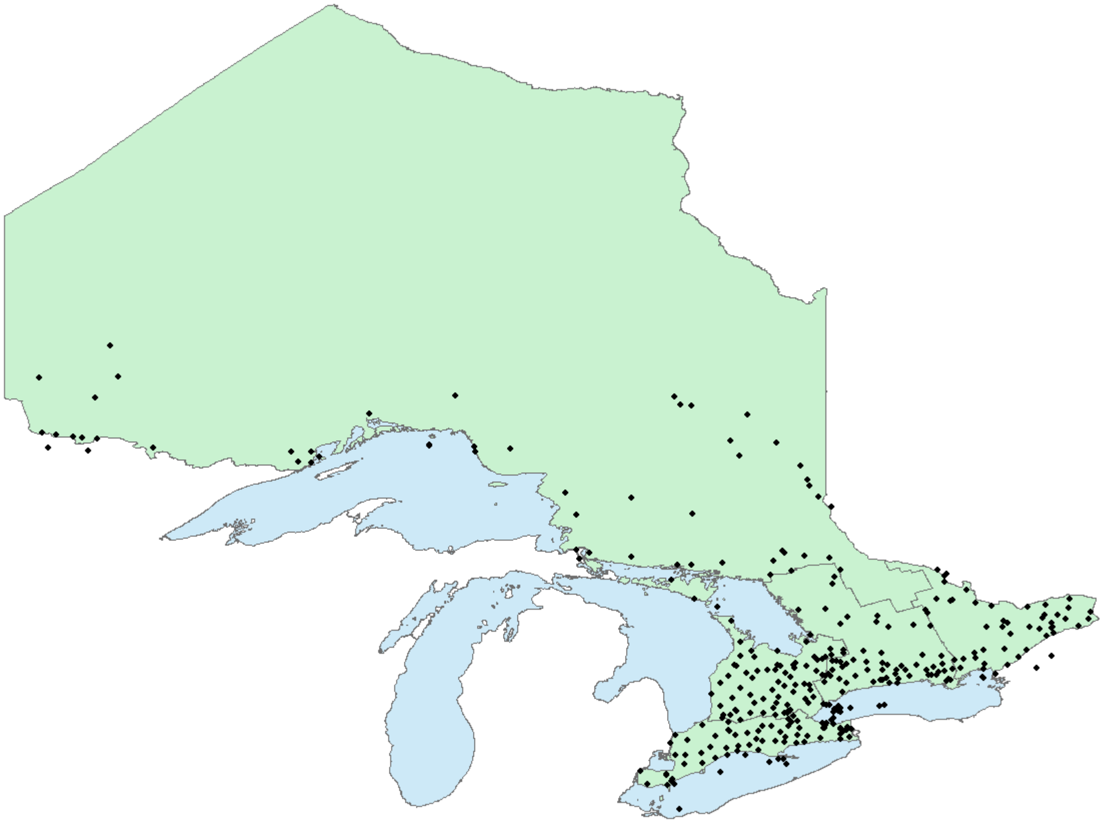
With all of this information in mind, the analysis of this paper is split into three sections. The first uses data from all of Ontario, the second uses data from southern Ontario only and the third uses data from northern Ontario only. By subsetting the data in this way, this study provides valuable insight as to how spatial basis risk is impacted by the number of weather stations available for performing spatial interpolation.

Figure 3.2: Map of Ontario's Agricultural Divisions



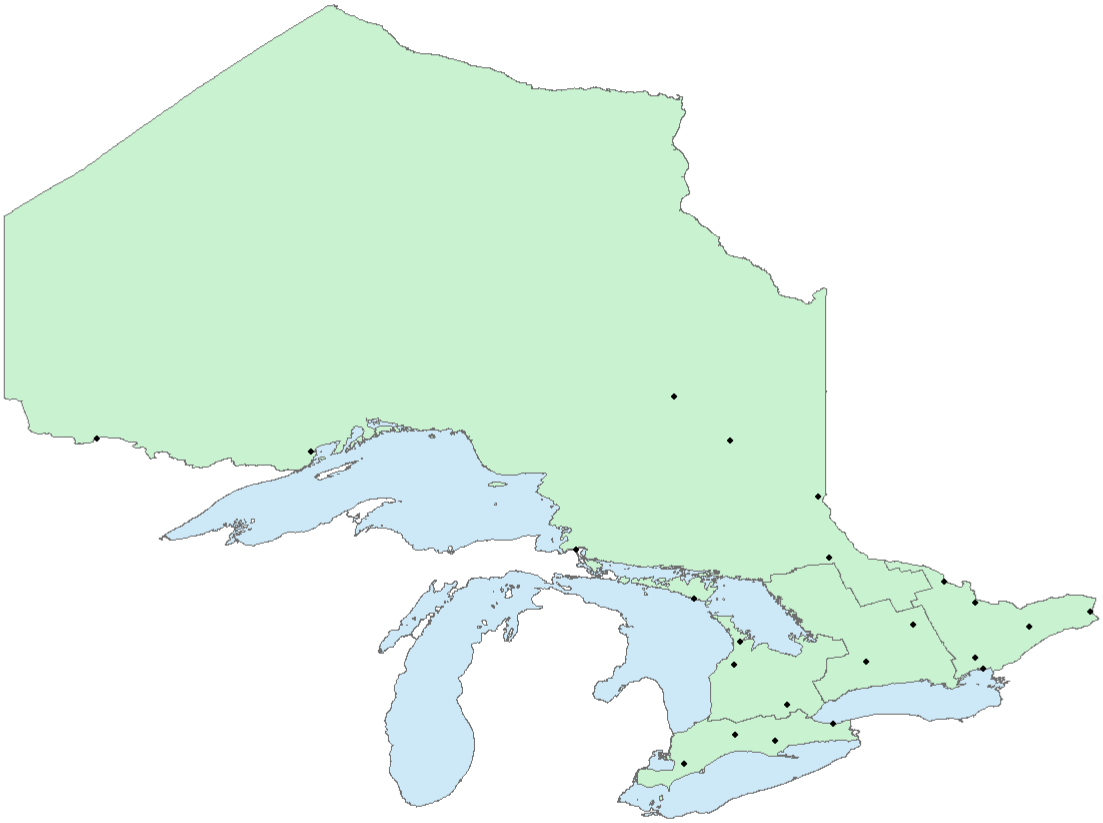
Note: These agricultural divisions are based on the 2011 census from Statistics Canada.

Figure 3.3: Map of Weather Stations in Ontario



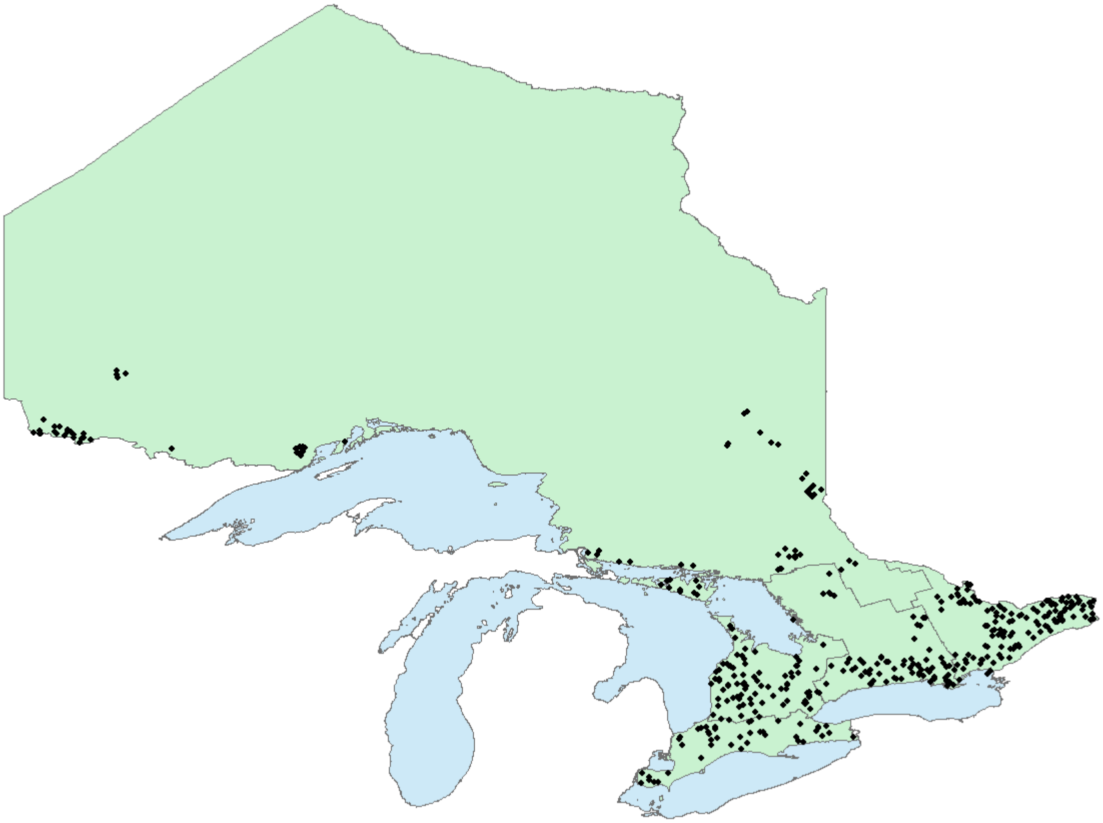
Note: This map shows the locations of all of the weather stations used to estimate temperature conditions at the forage farms. Not all stations were available in all years from 1997 to 2004.

Figure 3.4: Map of Weather Stations in Ontario With Long-Term Data



Note: This map shows the subset of weather stations which have long-term temperature data from 1967 to 2004. These stations are used to calculate the long-term average cooling degree days when calculating indemnities.

Figure 3.5: Map of Forage Farms in Ontario



Note: This map shows the locations of all of the forage farms in Ontario where temperature conditions were estimated. Not all farms reported yields in all years from 1997 to 2004.

Chapter 4: Methodology

In order to analyze and quantify any potential reduction in spatial basis risk, seven different spatial interpolation techniques are used to estimate the mean daily temperature at each of the forage producer's farms. These daily estimates are calculated for each farmer from April 1st to June 30th in the years 1997 to 2004. Using these temperature estimates, indemnities are calculated and spatial basis risk is analyzed by calculating the sample correlation between indemnities and forage yields. This process is repeated three times over: once for southern Ontario, once for northern Ontario, and once for all Ontario. This analysis gives insight into how spatial basis risk is impacted not only by the choice of interpolation technique, but also by the number of weather stations available for performing analysis.

Apart from regression, all of the interpolation techniques featured in this study take the form of a weighted average of observations from the surrounding area. Each of these spatial interpolation techniques provides a different approach for calculating the weights associated with each station. These weighted average estimates are given by:

$$\hat{z}(s_0) = \sum_{i=1}^{N(h)} w_i \cdot z(s_i) \tag{4.1}$$

where s_0 represents the unobserved location (A farm with an insurance policy in this case), z represents the weather process being interpolated (mean daily temperature in this case), N(h) is the number of stations within distance h of the unobserved location, and w_i is the weight assigned to station i. Ideally, the weights should minimize the prediction error variance, which is given by (Li and Heap, 2008):

$$\mathrm{Var}\Big(\hat{z}(s_0) - z(s_0)\Big) = C(s_0, s_0) + \sum_{i=1}^{N(h)} \sum_{j=1}^{N(h)} w_i w_j C(s_i, s_j) - 2 \sum_{i=1}^{N(h)} w_i C(s_i, s_0) \qquad (4.2)$$

where $C(s_i, s_j) = \mathrm{Cov}[Z(s_i), Z(s_j)]$. For all the weighted average techniques h has been set to 200 km and the maximum number of stations to be used in estimations is ten. This means that if there are more than ten stations within 200 km of an unobserved location then only the closest ten stations are used, and if there are less than ten stations within 200 km then all the stations are used.

The maximum number of stations and the maximum search distance (h) can be difficult to determine. The most common method to objectively select these parameters is to perform cross-validation analysis (see Chapter 4.8) and to select the model with the smallest error, however this quickly becomes computationally prohibitive for any moderate to large sized problem. This combined with the fact that a model must be estimated for every day during the growing season makes this approach impractical. With this in mind the values of ten stations and 200 km were selected based on a heuristic examination of the number of weather stations available.

The methodology section of this paper proceeds by following a ten step process. The first seven steps consist of estimating mean daily temperature using each of the seven different interpolation methods and outlining the various procedures used to obtain the results. The seven interpolation methods are: regression, nearest neighbor, inverse distance weighting, regression-based inverse distance weighting, ordinary kriging, regression kriging, and spatio-temporal regression kriging. A brief description of each of these methods is given in Appendix A.

Step 8 is used to perform cross-validation and estimate the amount of variance caused by each spatial interpolation method. Step 9 involves designing an insurance index using cooling degree days for the purpose of estimating the indemnities that would have been paid to each farmer if they had purchased this type of insurance. Step 10 concludes the analysis by calculating the sample correlation between the estimated indemnities and reported forage yields in order to determine if any of the interpolation methods could potentially reduce spatial basis risk.

4.1 Step 1: Regression

The goal of step 1 is to develop daily linear regression models for estimating mean daily temperatures. This method is used as a benchmark to compare against the more sophisticated methods such as kriging. Regression is also the only one of the seven interpolation techniques in this analysis which is not a weighted average.

Regression is a well known and extensively documented estimation method that uses explanatory variables to make predictions. Assume that there are k explanatory variables that are observed at n locations, then X represents an n x k matrix of those variables. Let $\vec{Z}$ represent the vector of mean daily temperatures observed at n locations. The marginal effects of the explanatory variables can be estimated by:

$$\hat{\beta} = (X'X)^{-1}X'\vec{Z} \tag{4.3}$$

Once the marginal effects have been calculated, the regression estimate of the mean daily temperature is given by:

$$\hat{z}(s_0) = X_0\hat{\beta} \tag{4.4}$$

where X_0 is the vector of explanatory variables measured at s_0. For this analysis the explanatory variables that are used are: longitude, latitude, elevation, and distance to the

Great Lakes. The cross products of these variables are also included, making a total of ten explanatory variables. The regression estimator therefore takes the form of:

$$\begin{aligned}\hat{z}(s_0) = \widehat{\beta_0} + \widehat{\beta_1}\text{Long} + \widehat{\beta_2}\text{Lat} + \widehat{\beta_3}\text{Elev} + \widehat{\beta_4}\text{Dist} + \widehat{\beta_5}\text{Long}\cdot\text{Lat} + \widehat{\beta_6}\text{Long}\cdot\text{Elev} \\ + \widehat{\beta_7}\text{Long}\cdot\text{Dist} + \widehat{\beta_8}\text{Lat}\cdot\text{Elev} + \widehat{\beta_9}\text{Lat}\cdot\text{Dist} + \widehat{\beta_{10}}\text{Elev}\cdot\text{Dist}\end{aligned} \tag{4.5}$$

In order to determine which variables should be included in each linear model, stepwise regression is used. Stepwise regression is an algorithm which successively adds and removes variables into the regression model based on their significance levels. Once no more statistically significant variables can be added or insignificant variables removed, the algorithm terminates. More information on stepwise regression can be found in Mendenhall et al. (1996).

For this analysis, a regression model is estimated for every day in the growing season using stepwise regression. These models are then used in order to make predictions of the daily mean temperature for each farm. This process is also used in step 4, step 6 and step 7 to model the deterministic portions of the estimators. Regression has the advantage of being simple to implement and is also very well understood, however, the regression model does not incorporate any information about spatial autocorrelation. In addition, all weather stations in the sample are used to make predictions regardless of their proximity to the farm, which may result in a loss of accuracy when used in a large geographic area.

4.2 Step 2: Nearest Neighbor

The goal of step 2 is to implement an interpolation method which only considers information from one weather station. This is in order to determine how spatial basis risk is impacted when only one weather station is used for making estimations rather than using multiple weather stations as in step 3. This method is perhaps the most intuitive of all the interpolation techniques presented here. For each day in the growing season, the estimate of the mean daily temperature at the unobserved farm is equal to the mean daily temperature recorded at the closest weather station.

$$\hat{z}(s_0) = z(s_c) \tag{4.6}$$

where s_c represents the weather station that is closest to s_0 . The nearest neighbor method is a special case of Equation 4.1 where w_i=1 for the station closest to the unobserved farm. In order to find the weather station that is closest to each farm, the Haversine formula is used (see Appendix B).

This method is simple and easy to understand, however, it does not incorporate any additional information in terms of explanatory variables or a spatial autocorrelation structure. In addition, this method only considers information from a single weather station instead of making use of multiple stations in the surrounding area which may not result in the most accurate estimate possible.

4.3 Step 3: Inverse Distance Weighting (IDW)

The goal of step 3 is to begin incorporating information from multiple weather stations to determine if this provides any significant advantage over using only the closest weather station as in step 2. Inverse distance weighting is among the most simple of the spatial interpolation techniques and is the purest mathematical representation of Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236). The weight assigned to each station is simply the inverse of its distance to the unobserved location.

$$w_i = \frac{1/d_i^p}{\sum_{i=1}^{N(h)} 1/d_i^p} \tag{4.7}$$

where d_i is the distance from s_i to s_0, and p is the power-decay parameter. Higher values of p assign less weight to stations that are further away. In practice p=2 is a very common choice (Li and Heap, 2008) and this is the value used in this paper. The inverse distance weighting technique is applied for every day in the growing season. The weights for the closest ten stations (up to a maximum of 200 km away) are calculated using Equation 4.7.

The main advantage of inverse distance weighting is its simplicity, since these distances are very easily calculated using the longitude and latitude coordinates of the farmer's property and the model requires only one simple parameter. However, inverse distance weighting assumes that distance from nearby weather stations is the most important factor in calculating estimates, and this may not be the case (Paulson and Hart, 2006). Another disadvantage to inverse distance weighting is that the power-decay parameter p must be applied uniformly in all directions when in reality it may vary based on direction and distance.

4.4 Step 4: Regression-Based Inverse Distance Weighting

The goal of step 4 is to extend the inverse distance weighting model so that it becomes possible to incorporate variables other than distance in the estimation process, similar to step 1. Regression-based inverse distance weighting allows the model to incorporate additional information by combining inverse distance weighting and regression techniques. The estimator is split into two distinct terms: a stochastic term and a deterministic term.

$$\hat{z}(s_0) = \underbrace{\hat{m}(s_0)}_{\text{Deterministic}} + \underbrace{\hat{e}(s_0)}_{\text{Stochastic}} \tag{4.8}$$

The stochastic term represents the residual from a linear regression model and so is assumed to follow a N(0,1) distribution. The regression-based inverse distance weighting estimator is given by:

$$\hat{z}(s_0) = \underbrace{X_0\hat{\beta}}_{\text{Deterministic}} + \underbrace{\sum_{i=1}^{N(h)} w_i \cdot e_i}_{\text{Stochastic}} \tag{4.9}$$

where e_i is the residual calculated at station i and the w_i are calculated using Equation 4.7. For this interpolation method, the following steps are used for each day in the growing season:

1. Estimate a linear model using the process outlined in step 1.
2. Calculate the regression estimate for each farm.
3. Calculate the model residuals for each weather station.
4. Estimate the residuals at each farm using the inverse distance weighting technique from step 3.
5. The final temperature estimate is given by the sum of the regression estimate and the inverse distance weighting estimate.

In addition to incorporating explanatory variables, regression-based inverse distance weighting is nearly as easy to implement as inverse distance weighting. However, there is still no attempt to define a spatial autocorrelation structure and there is no way of knowing if the weights that are assigned to each station are optimal in terms of reducing prediction error variance.

4.5 Step 5: Ordinary Kriging

The goal of step 5 is to make temperature estimates more accurate by introducing a spatial autocorrelation function which describes the correlation between two points in space. This makes it possible to calculate station weights in a way which minimizes the prediction error variance. This is done by working with a type of spatial interpolation technique known as kriging. Ordinary kriging is an interpolation method which is subject to the following conditions:

1. The variable being interpolated has a constant, stationary mean over the entire interpolation area.
2. $\sum_1^{N(h)} w_i = 1$.
3. The w_i are selected such that the variance in Equation 4.2 is minimized.

This results in an estimator that is considered to be the best linear unbiased prediction model (BLUP) for spatial data (Li and Heap, 2008).

The minimization of the error variance is the defining characteristic of a kriging model. In order to calculate values for the covariance terms, the concept of semivariance needs to be introduced, often denoted by $\gamma(h)$. The semivariance is a function of distance and describes the dependency structure for the correlation between two locations. The semivariance is defined as:

$$\gamma(h) = \frac{1}{2}\mathbb{E}\Big[\big(Z(s_i) - Z(s_i + h)\big)^2\Big] \tag{4.10}$$

where Z(s_i) and Z($s_i + h$) are measured values of Z separated by a distance of h. From a set of spatial data, the semivariance is estimated as:

$$\hat{\gamma}(h) = \frac{1}{2M(h)} \sum_{i,j \in M(h)} \big[z(s_i) - z(s_j)\big]^2 \tag{4.11}$$

where M(h) is the number of pairs of sample points such that the distance between them is equal to h. A plot of $\hat{\gamma}(h)$ against h is known as a sample variogram, and provides information regarding how correlation varies with distance (Li and Heap, 2008).

A variogram can be defined using three model parameters: the nugget, the sill and the range. The nugget is a positive constant that represents $\gamma(h)$ for small values of h, the sill represents the asymptotic value of the variogram, and the range represents the distance required for the variogram to reach the sill for the first time. Using these three parameters the sample variogram can be fit to one of several predefined variogram functions. The three most popular are the spherical, exponential, and Gaussian models (Kuzyakova et al., 2001) :

$$\textbf{Spherical:}\quad \gamma(h) = \begin{cases} \lambda_0 + \lambda\left[\frac{3h}{2a} - \frac{h^3}{2a^3}\right] & \text{If } 0 < h \leq a \\ \lambda_0 + \lambda & \text{If } h > a \end{cases} \tag{4.12}$$

$$\textbf{Exponential:}\quad \gamma(h) = \lambda_0 + \lambda\left[1 - \exp\left(-\frac{h}{r}\right)\right] \tag{4.13}$$

$$\textbf{Gaussian:}\quad \gamma(h) = \lambda_0 + \lambda\left[1 - \exp\left(-\frac{h^2}{r^2}\right)\right] \tag{4.14}$$

where λ_0 is the nugget, λ is selected so that λ_0+λ is equal to the sill, and a is equal to the range. Because the exponential and Gaussian functions approach the sill asymptotically, they technically do not have a finite range. Therefore for these functions a represents the effective range. The effective range is defined as the distance required to reach 95% of the sill variance. For the exponential function the parameter r is selected so that a=3r and for the Gaussian $a = \sqrt{3}r$. These values are selected so that $\gamma(a) \approx \lambda_0 + 0.95\lambda$. See Appendix C for an example of fitting a spatial variogram.

In the case of ordinary kriging, once the variogram model has been defined the kriging weights for an unobserved location can be derived as (Hengl, 2009):

$$w_0 = C^{-1} \cdot c_0 \tag{4.15}$$

where C is the N(h) x N(h) covariance matrix and c_0 is the vector of the covariances measured at the unobserved location. In order to ensure that the kriging weights sum to one, an additional row and column of ones are added to the C matrix, making it $\big($N(h)+1$\big)$ x $\big($N(h)+1$\big)$. Equation 4.15 can be expanded as :

$$\begin{bmatrix} w_1 \\ \vdots \\ w_{N(h)} \\ \psi \end{bmatrix} = \begin{bmatrix} C(s_1, s_2) & \dots & C(s_1, s_{N(h)}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(s_{N(h)}, s_1) & \dots & C(s_{N(h)}, s_{N(h)}) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(s_0, s_1) \\ \vdots \\ C(s_0, s_{N(h)}) \\ 1 \end{bmatrix} \tag{4.16}$$

where ψ represents the Lagrange multiplier, a constant which is not used in any of the calculations for this paper.

The ordinary kriging estimation proceeds as follows for every day of the growing season:

1. Estimate the variogram function using the procedure in Appendix C.

2. Use Equation 4.15 to calculate the kriging weights for the closest ten stations (up to a maximum distance of 200 km away) for each farm.

While ordinary kriging is a relatively straight forward process, the entire method relies on an assumption that may not be applicable in practice: stationarity. Ordinary kriging assumes that all observations are identically and independently distributed with a stationary mean, an assumption that is not reasonable for the applications of this paper.

Given the geographic size of Ontario, as well as the known correlations in between elevation/latitude and temperature (Wu and Li, 2013), the idea of a stationary mean across the entire sampling area seems unlikely. It is for this reason that the ordinary kriging model is now extended to the regression kriging model. Cressie (1988) provides an in depth discussion on the assumptions underlying ordinary kriging, including independence and stationarity.

4.6 Step 6: Regression Kriging

The goal of step 6 is to extend the ordinary kriging model so that it becomes possible to incorporate variables other than distance in the estimation process. This is done by combining linear regression methods with ordinary kriging. Regression kriging splits the estimator into two distinct terms, a deterministic term and a stochastic term. This is similar to the motivation behind step 4, and in fact these two steps differ only in how the stochastic portion of the estimator is calculated. The regression kriging estimator is given by (Hengl, 2009):

$$\hat{z}(s_0) = \underbrace{\hat{m}(s_0)}_{\text{Deterministic}} + \underbrace{\hat{e}(s_0)}_{\text{Stochastic}} \tag{4.17}$$

Once again the stochastic term is assumed to follow a N(0,1) distribution. Combining Equation 4.4 and the kriging weights from Equation 4.15 the regression kriging estimate is given by:

$$\hat{z}(s_0) = \underbrace{X_0\hat{\beta}}_{\text{Deterministic}} + \underbrace{C^{-1}c_0 \cdot \vec{e}}_{\text{Stochastic}} \tag{4.18}$$

where $\vec{e}$ is the vector of residuals from the stations in the surrounding area. The regression kriging method can be summarized in the following steps for every day of the growing season:

1. Estimate a linear model using the process outlined in step 1.
2. Calculate the regression estimate for each farm.
3. Calculate the model residuals for each weather station.
4. Estimate the residuals at each farm using ordinary kriging from step 5.
5. The final temperature estimate is given by summing the regression estimate and the ordinary kriging estimate.

In addition to allowing the incorporation of explanatory variables, the regression kriging model also addresses the issue of stationarity that is present in the ordinary kriging model. Because the residuals from a linear regression model are used, the assumption of mean stationarity is more realistic in the regression kriging model than in the ordinary kriging model.

4.7 Step 7: Spatio-Temporal Regression Kriging

The goal of step 7 is to extend the regression kriging model to include the dimension of time as well as space. This is done not only to include more information in the estimation process, but also to simplify the variogram fitting process by only requiring one variogram to be fitted for each year. Given that kriging was first developed for use in the fields of mining and geology, many applications of kriging are not concerned with the dimension of time. Unlike meteorological events such as rainfall and temperature, mineral deposits do not change drastically over short time periods, and so there is no need to analyze the impact of time on interpolations. However in the last several years, geostatistics has been applied to many fields of study other than geology such as agriculture, meteorology, and epidemiology. As a result there has been an increased interest in methods of interpolation that can account for autocorrelation across time as well as space.

The rationale behind spatio-temporal regression kriging is that for certain processes (such as weather events), there exists not only a spatial autocorrelation structure, but also a temporal autocorrelation structure which impacts observations that are taken at the same location but separated through time (Hengl, 2009). The spatio-temporal regression kriging estimator can be expressed as:

$$\hat{z}(s_0,t) = \underbrace{\hat{m}(s_0,t)}_{\text{Deterministic}} + \underbrace{\hat{e}(s_0,t)}_{\text{Stochastic}} \tag{4.19}$$

The first step of the spatio-temporal regression kriging model is identical to the regression kriging model outlined previously, where the deterministic part of the model is calculated using regression. Where the model differs is in the estimation of the stochastic portion. As with ordinary kriging there are many forms that the spatio-temporal variogram can take, however, the analysis of this paper makes use of the sum-metric variogram model which allows the variogram to be broken into three distinct parts (Kilibarda et al., 2014).

$$\gamma(h,u) = \frac{1}{2}\mathbb{E}[\mathrm{e}(s,t) - \mathrm{e}(s+h,t+u)]^2 = \gamma_S(h) + \gamma_T(u) + \gamma_{ST}\Big(\sqrt{h^2 + (\alpha \cdot u)^2}\Big) \tag{4.20}$$

where e(s,t) represents the zero-mean stochastic residual, γ_S, γ_T, and γ_{ST} represent the spatial, temporal, and joint variograms respectively and α represents the spatio-temporal anisotropy ratio. The anisotropy ratio takes a value between zero and one, and serves to convert the temporal units (u) into spatial units (h). Thus, the spatio-temporal variogram is a function of ten parameters: three sill parameters, three range parameters, three nugget parameters, and the spatio-temporal anisotropy ratio. See Appendix D for an example of fitting a spatio-temporal variogram.

The spatio-temporal regression kriging method can be summarized in the following steps for every year from 1997 to 2004:

1. Estimate daily linear regression models using the techniques from step 1.
2. Calculate the model residuals for each weather station for every day in the growing season.
3. Estimate the spatio-temporal variogram using the methods in Appendix D.
4. Use the spatio-temporal variogram to estimate the residuals at each farm for every day in the growing season
5. The final temperature estimate is given by summing the regression estimate and the spatio-temporal regression kriging estimate.

In addition to allowing the incorporation of temporal autocorrelation, spatio-temporal regression kriging is also advantageous because it only requires one variogram model to be fitted in each year. Under the regression kriging model one variogram model must be estimated for every day in the growing season. This simplified variogram model means that it is much easier to estimate the variogram parameters visually rather than relying on automated variogram fitting using rules of thumb to make initial estimates of variogram parameters.

4.8 Step 8: Cross-Validation

The goal of step 8 is to calculate the variance for each one of the spatial interpolations techniques using cross-validation. For each of the spatial interpolation methods, cross-validation is performed using leave-one-out analysis. Leave-one-out cross-validation is an iterative process that removes the stations one by one from the sample, and then uses the model to predict mean temperature at the station that was removed. The end result is that every weather station has a predicted mean temperature and an observed mean temperature for each day in the growing season. Once the predicted temperature value is calculated at each of the stations, the root mean square error is calculated as (Hengl, 2009):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{z}(s_i) - z(s_i)\right)^2}{n}} \tag{4.21}$$

where $\hat{z}(s_i)$ and $z(s_i)$ are the predicted and observed values at station i, and n is the number of observations. Techniques that give a lower value for RMSE have less variance in their estimates and vice versa. The cross-validation was performed individually for each year in the analysis as well as with all data from all years.

4.9 Step 9: Cooling Degree Days (CDD) Insurance Index

The goal of step 9 is to develop a temperature based insurance index using the concept of cooling degree days (CDD) in order to calculate indemnities based on the estimated temperatures from steps 1 through 7. CDD are primarily used in the heating and ventilation industry, where they are used to quantify the amount of energy required to cool a building (Büyükalaca et al., 2001). For example, if 18°C is used as a base temperature (approximately room temperature) and the daily mean temperature is 25°C, then the cooling degrees for the day is 25°C-18°C= 7°C and the building requires enough energy to lower the average temperature by 7°C. In the context of this analysis, cooling degree days is meant to capture the total heat stress experienced by the crops over the entire growing season instead of the amount of cooling required. Taking this into consideration, “Heating Degree Days” might be a more appropriate name for the index, however, in order to remain consistent with the existing literature, the term cooling degree days is used.

For any given year, the estimated cooling degree days index for an unobserved farm is calculated as:

$$\widehat{\mathrm{CDD}_0} = \sum_{i=1}^{n} \mathrm{Max}(0, \hat{z}_i(s_0) - 20) \tag{4.22}$$

where n is the number of days in the growing season, ninety-one in this case. For cool season species of forage (the species predominantly grown in Canada), the optimal temperature for growth is 20°C. Research has shown that for every 1 degree increase in temperature, the forage digestibility is decreased by 0.3-0.7 percentage units (Buxton, 1995). With this in mind the CDD index was designed to measure the cumulative heat stress experienced by the forage crops throughout the growing season.

In order to calculate indemnities, the CDD index is compared to the long-term average CDD measured at the weather station closest to the farm.

$$I_0 = \mathrm{Max}(0, \widehat{\mathrm{CDD}_0} - \overline{\mathrm{CDD}}_c) \tag{4.23}$$

Figure 3.4 shows the location of all weather stations with long-term observations dating back to 1967, and these are the stations used to calculate $\overline{\mathrm{CDD}}_c$. This process is repeated for every farm with recorded yields during the year.

4.10 Step 10: Correlation Analysis

The goal of step 10 is to determine whether spatial basis risk can be significantly reduced by any of the interpolation techniques. In order to assess the performance of each spatial interpolation technique, the sample correlation in between forage yields and indemnities is calculated.

$$\rho_{xy} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} \tag{4.24}$$

where σ represents the standard deviation. The definition of basis risk (see Chapter 2.2) refers to the loss experienced by the farmer, however this information was not available in the data sets provided. Forage yields were used as a proxy for losses under the assumption that higher yields represent lower losses and vice versa. If the hypothesis that increased CDD results in lower forage yields is true, then it is expected that the correlation in between CDD and yields will be negative.

Following the logic of Major (1999), the level of correlation is used to assess the amount of basis risk in each model. A higher (more negative) correlation value indicates less basis risk and vice versa. Given that the same weather index was used and that the index was calculated over the same time period, any difference in the amount of basis risk between the different methods can be attributed to spatial basis risk, and not temporal or product basis risk. Therefore if any of the methods produce correlation results which are drastically lower (less negative) than the others, it implies that the method is more susceptible to spatial basis risk.

Chapter 5: Results

5.1 Regression Variable Results

Table 5.1 shows the percentage of daily regression models which include each explanatory variable at the 5% significance level. The table is broken down by month in order to determine whether certain variables have a greater impact during different parts of the growing season. All four of the main variables (longitude, latitude, elevation and distance to the Great Lakes) are statistically significant in the vast majority of the models across all three regions. The one exception to this result is elevation in northern Ontario, which is included in 75% of the models as opposed to 95% in the other regions. Figure 3.1 shows a map of the elevation in Ontario, and this can explain why elevation is less significant in northern Ontario.

When Figure 3.1 is compared to Figure 3.5, most of the farms in northern Ontario are located to the West of 80° W and South of 50° N, an area which is fairly homogeneous in elevation when compared to southern Ontario. The elevation in southern Ontario varies greatly from the lowlands on the shores of Lake Ontario to the highlands on the Southeast shore of Lake Huron, and this is likely the reason that elevation is more statistically significant in this region. The larger differences in elevation creates more variance in daily mean temperature and so the variable is more significant. This trend is also visible in the cross products long · elev and elev · dist for the same reasons.

The cross product long · lat is more significant when creating models for all Ontario than for southern or northern Ontario alone. This is likely caused by the size of the interpolation area. The range of latitude and longitude is larger when considering all of Ontario, resulting in the interaction between longitude and latitude becoming more significant. For the remainder of the explanatory variables, there is little difference in between the different regions and even in between the different months. For any practitioner seeking to model temperatures, these results suggest that all ten of the explanatory variables featured here would be suitable candidates for regression analysis. In addition, it does not appear necessary to consider different regression variables for different months when such a short time period (3 months) is being considered.

5.2 Cross-Validation Results

The results of the cross-validation analysis from step 8 are summarized in Table 5.2. The first interesting result from this table is that the nearest neighbor method has the highest RMSE in all of the regions. If nearest neighbor is compared to inverse distance weighting the impact of using multiple weather stations becomes clear. When inverse distance weighting is used the RMSE is significantly reduced, and this is caused by the additional information that is used in the estimation process. This shows that insurers should avoid using only the closest station when estimating weather conditions since these estimates are prone to have higher variance than estimates which incorporate multiple weather stations.

For northern Ontario the regression method had the lowest RMSE not only overall but also for each year individually. Regression is the only one of the seven interpolation methods which does not take the form of a weighted average from surrounding stations and (unlike the kriging methods) does not make any attempt to define a spatial autocorrelation structure. The fact that regression has the lowest RMSE shows that great care needs to be taken when applying spatial interpolation techniques to areas with a small number of weather stations. By attempting to use less-than-ideal data to define spatial autocorrelation structures the model ends up having more variability than if simple regression had been used.

Table 5.1: Percentage of Daily Regression Models Where Each Explanatory Variable is Statistically Significant

Region	Month	long	lat	elev	dist	long·lat	long·elev	long·dist	lat·elev	lat·dist	elev·dist
All Ontario	April	99.58	100.00	99.17	97.08	79.17	55.83	65.83	43.75	71.25	44.58
	May	99.60	100.00	95.97	99.60	76.21	51.61	64.11	50.00	59.68	38.31
	June	99.58	100.00	97.50	99.58	70.83	53.33	65.42	51.25	64.17	44.17
	Total	99.59	100.00	97.53	98.76	75.41	53.57	65.11	48.35	64.97	42.31
South Ontario	April	96.67	99.17	97.08	95.00	54.17	45.83	50.83	36.25	51.67	49.17
	May	97.98	98.39	96.37	96.77	63.31	41.94	58.87	46.37	43.95	63.71
	June	97.08	100.00	97.50	93.33	61.67	44.17	58.33	46.25	43.75	63.75
	Total	97.25	99.18	96.98	95.05	59.75	43.96	56.04	42.99	46.43	58.93
North Ontario	April	98.75	98.75	78.33	93.75	59.17	36.67	57.08	36.67	53.75	26.25
	May	97.18	90.73	81.45	92.34	56.85	33.06	56.85	37.50	41.53	31.05
	June	97.08	96.25	75.83	92.08	58.33	26.25	59.17	47.50	44.58	29.17
	Total	97.66	95.19	78.57	92.72	58.10	32.01	57.69	40.52	46.57	28.85

Note: This table shows the percentage of daily regression models which include each of the explanatory variables used in the analysis. The percentages were calculated using the regression models from April 1st to June 30th in the years from 1997 to 2004. The total percentages were calculated using all daily models across all months. The table is divided into the three regions of Ontario: all Ontario, South Ontario, and North Ontario. Variables were selected using stepwise regression. All variables are significant at the 5% level.

In southern Ontario, it is interesting to note that regression kriging has a very high RMSE when compared to the other methods (except nearest neighbor). To gain some insight into why this occurs, the regression kriging approach should be compared to the regression-based inverse distance weighting. These methods both produce temperature estimates by summing a regression estimator and a stochastic estimator that is used to estimate model residuals. The only way in which these two methods differ is the estimation of the model residuals. Regression kriging uses a variogram function to determine a spatial autocorrelation structure while regression-based inverse distance weighting uses a (nearly) non-parametric equation to calculate station weights. With this in mind it is interesting that the overall RMSE for regression kriging is so much larger than it is for regression-based inverse distance weighting.

When comparing the regression kriging RMSE for each year in southern Ontario, 1999 is the only year in which regression kriging has a larger RMSE than regression-based inverse distance weighting. For all of the regions in Ontario, 1999 is a year with significantly higher RMSE than all of the others. The regression kriging model has exaggerated this increased variability enough to significantly impact the overall RMSE. It was expected that the regression kriging model would have less variance than the ordinary kriging model, therefore these results are somewhat unexpected.

The regression kriging model requires that a variogram be estimated for every day in the three month (91 day) growing season. With this many models an automated process for variogram fitting is required in order to perform analysis efficiently. In order for optimization algorithms to fit the variogram parameters, initial estimates must be provided and in order to automate this process, some “rules of thumb” must be used (see Appendix C). Most literature on the subject of fitting variograms suggests that the process be done by hand after thoroughly examining the data to ensure that appropriate parameter values are used (Olea, 2006).

Fitting the variogram function is the most crucial step in any kriging method, therefore the fact that the variograms are fitted using an automated procedure is the most likely cause for this increased RMSE. This conclusion is supported by the RMSE results from the spatio-temporal regression kriging method, which produced the lowest RMSE of 1.231. Due to the fact that only one variogram must be modeled for each year, the spatio-temporal variograms were all inspected visually and initial estimates for parameters were estimated by hand (see Appendix D). This results in better fits for the variogram functions when compared to standard regression kriging and a lower RMSE in the year with the most variability.

5.3 Correlation Analysis Results

The results of the correlation analysis from step 10 are summarized in Table 5.3. Temperature estimates from April 1 to June 30th were used for these calculations. Initial impressions from this table are that temperature based policies administered in southern Ontario have the highest (most negative) correlations, followed by all Ontario. Meanwhile, policies administered in northern Ontario have correlation values that are closest to zero. This is consistent with expected results based on the fact that southern Ontario has a larger number of weather stations. With few exceptions the results for each interpolation method are mostly consistent in south Ontario and all Ontario while the results from northern Ontario appear more varied.

Similar to the RMSE analysis, the regression kriging method for southern Ontario produces results that are distinctly different from the other interpolation methods. Regression kriging produces a correlation value of -0.032 while the other methods have correlations closer to -0.15. This indicates that the regression kriging model is more susceptible to spatial basis risk when applied in southern Ontario. This is contrary to the expected results, which assumed that regression kriging would outperform ordinary kriging. The

Table 5.2: Root Mean Square Error (RMSE) for Each Interpolation Technique

Region	Method	1997	1998	1999	2000	2001	2002	2003	2004	Total
All Ontario	Nearest Neighbor	1.016	1.576	2.994	1.365	1.334	1.306	1.365	1.494	1.678
	Regression	1.140	1.347	2.199	1.289	1.212	1.182	1.277	1.311	1.420
	IDW	0.838	1.330	2.406	1.127	1.087	1.085	1.133	1.213	1.371
	Regression IDW	0.868	1.328	2.401	1.138	1.082	1.068	1.130	1.209	1.371
	Ordinary Kriging	0.804	1.262	2.276	1.091	1.063	1.043	1.111	1.169	1.312
	Regression Kriging	0.898	1.282	2.268	1.108	1.072	1.036	1.117	1.179	1.324
	ST Regression Kriging	0.889	1.271	2.218	1.139	1.147	1.112	1.175	1.210	1.337
South Ontario	Nearest Neighbor	1.035	1.570	2.821	1.282	1.227	1.232	1.272	1.440	1.594
	Regression	1.040	1.227	1.988	1.092	1.057	1.009	1.092	1.169	1.258
	IDW	0.833	1.310	2.264	1.055	1.012	1.018	1.062	1.195	1.302
	Regression IDW	0.825	1.304	2.232	1.061	1.014	1.007	1.065	1.184	1.292
	Ordinary Kriging	0.796	1.242	2.130	1.016	0.994	0.968	1.033	1.139	1.239
	Regression Kriging	0.810	1.242	3.146	1.031	0.999	0.973	1.046	1.148	1.511
	ST Regression Kriging	0.862	1.225	2.035	1.046	1.014	0.977	1.059	1.156	1.231
North Ontario	Nearest Neighbor	0.905	1.609	3.402	1.678	1.690	1.550	1.672	1.687	1.939
	Regression	0.830	1.093	2.360	1.192	1.103	1.114	1.124	1.042	1.338
	IDW	0.904	1.419	2.902	1.428	1.365	1.334	1.410	1.329	1.646
	Regression IDW	0.968	1.344	2.779	1.344	1.257	1.209	1.278	1.200	1.551
	Ordinary Kriging	1.061	1.332	2.720	1.352	1.298	1.271	1.331	1.257	1.564
	Regression Kriging	0.939	1.228	2.573	1.296	1.217	1.158	1.231	1.141	1.461
	ST Regression Kriging	0.934	1.212	2.429	1.305	1.182	1.158	1.211	1.133	1.417

Note: This table shows the root mean square error (RMSE) for each interpolation method in each year from 1997 to 2004. These values were calculated using temperature estimates from April 1 to June 30th with each different interpolation method. The total RMSE was calculated using all estimates across all years. The table is divided for the three regions of Ontario: all Ontario, South Ontario, and North Ontario.

increased variability discussed in the RMSE analysis has translated into a lower degree of correlation in the year 1999, where regression kriging gave a correlation of -0.032 while the other interpolation methods (except nearest neighbor) gave correlation values as low as -0.072. Norton et al. (2010) found that changes in elevation had the most significant impact on spatial basis risk for temperature based policies, and southern Ontario is the region with the largest variance in elevation (see Chapter 5.1). The changes in elevation in this region combined with the difficulties in the variogram fitting process described earlier create a situation which may result in increased spatial basis risk.

The regression method also shows significantly increased spatial basis risk when used for all Ontario. Regression gives a correlation value of -0.065 while the other methods have correlation value closer to -0.090. Regression does not take the form of a weighted average and does not take into account the distance in between weather stations and farms when making predictions. When using regression for all Ontario, this means that weather stations in northern Ontario are used to make predictions at farms in southern Ontario and vice versa. Conversely, the kriging and inverse distance weighting methods only use observations in the surrounding area up to a maximum of 200 km away. This shows that linear regression models lose some of their predictive power when they are applied to a geographic area that is too large, resulting in increased spatial basis risk.

When compared to southern Ontario, the insurance policies created in northern Ontario are much more sensitive to the type of interpolation method used. The correlation values range from -0.009 to 0.076, with many of the interpolation methods producing different results. The methods which produce the most negative correlations are the regression-based methods, although these correlations are close to zero. Nearest neighbor, inverse distance weighting, and ordinary kriging have positive correlation values significantly different from zero, which is contrary to the assumption that increased CDD should result in decreased yields. This is likely caused by the fact that these methods do not consider any information except the proximity of the stations to the farm, resulting in inaccurate predictions in an area with very few weather stations. These results suggest that temperature based policies in northern Ontario are very susceptible to basis risk in general, with most of the correlation values being very close to zero. In addition, the difference in correlation between the regression-based methods and the other methods shows that it is important to include explanatory variables when the amount of weather stations available is limited, since this can help to reduce the amount of spatial basis risk and these methods have the lowest RMSE in Table 5.2.

Taking these results into consideration, the best approach to reducing spatial basis risk is to increase the amount of weather stations that are available for performing spatial interpolation, rather than focusing on the type of spatial interpolation method that is used. In an area like southern Ontario where there are many weather stations the differences between spatial interpolation methods is minimal, with the exception of regression kriging, indicating that the choice of interpolation technique has only a limited impact on the amount of spatial basis risk in the insurance model. Meanwhile in an area like northern Ontario with a small number of stations, the impact of the spatial interpolation method becomes more pronounced with different methods producing significantly different results.

It also should be noted that although the policies in southern Ontario have the least amount of basis risk, the correlation values are still quite low for this index. A correlation value of -0.15 is not strong enough to design a profitable and effective insurance index, and this indicates that temperature based contracts are highly susceptible to product basis risk (see Chapter 2.2) when used to insure forage crops. This is as expected, since it is known that temperature impacts forage quality rather than forage yields (Buxton, 1995). These results show that temperature based policies such as these should be combined

Table 5.3: Correlations Between Estimated Indemnities and Reported Forage Yields for Each Interpolation Technique

Region	Method	1997	1998	1999	2000	2001	2002	2003	2004	Total
All Ontario	Nearest Neighbor	-0.163	0.141	-0.044	0.026	-0.096	0.106	-0.104	0.067	-0.093
	Regression	-0.126	0.224	-0.026	0.006	-0.015	0.137	-0.166	0.088	-0.065
	IDW	-0.166	0.179	-0.074	0.017	-0.051	0.115	-0.139	-0.012	-0.093
	Regression IDW	-0.171	0.156	-0.077	0.014	-0.075	0.098	-0.154	-0.012	-0.094
	Ordinary Kriging	-0.161	0.215	-0.073	0.025	-0.002	0.138	-0.114	NA	-0.086
	Regression Kriging	-0.148	0.185	-0.080	0.023	-0.057	0.107	-0.143	NA	-0.092
	ST Regression Kriging	-0.180	0.160	-0.064	0.024	-0.067	0.100	-0.140	0.194	-0.091
South Ontario	Nearest Neighbor	-0.238	0.069	-0.014	0.011	-0.131	0.086	-0.216	-0.051	-0.154
	Regression	-0.300	0.107	-0.048	-0.001	-0.050	0.119	-0.228	-0.330	-0.146
	IDW	-0.245	0.112	-0.054	0.005	-0.071	0.110	-0.226	-0.112	-0.151
	Regression IDW	-0.269	0.072	-0.062	0.011	-0.108	0.096	-0.208	-0.104	-0.154
	Ordinary Kriging	-0.227	0.134	-0.072	0.014	-0.037	0.137	-0.209	NA	-0.150
	Regression Kriging	-0.261	0.081	-0.032	0.018	-0.089	0.108	-0.202	-0.152	-0.032
	ST Regression Kriging	-0.259	0.051	-0.060	0.001	-0.088	0.119	-0.218	NA	-0.152
North Ontario	Nearest Neighbor	0.018	-0.015	-0.032	-0.160	0.186	0.078	0.381	NA	0.091
	Regression	-0.099	-0.165	-0.092	NA	0.023	-0.331	-0.015	NA	-0.009
	IDW	0.038	0.056	-0.036	NA	0.183	-0.153	0.267	NA	0.069
	Regression IDW	-0.099	-0.180	-0.080	NA	-0.035	-0.308	0.116	NA	0.003
	Ordinary Kriging	0.143	0.049	-0.009	NA	0.234	-0.219	0.185	NA	0.076
	Regression Kriging	-0.073	-0.165	-0.070	NA	-0.001	-0.384	0.171	NA	-0.001
	ST Regression Kriging	-0.136	-0.169	-0.060	NA	0.036	-0.047	0.260	NA	0.022

Note: This table shows the sample correlation in between the estimated indemnities and reported forage yields for each interpolation method in the years from 1997 to 2004. Calculations were done using temperature estimates from April 1st to June 30th. Total correlation was calculated using all estimates across all years. The table is divided for the three regions of Ontario: all Ontario, South Ontario, and North Ontario. Entries with NA correspond to years with no indemnities, making correlation calculations impossible.

with policies based on other variables such as rainfall in order to improve correlation and reduce product basis risk.

5.4 Summary of Results

There are several interesting results that can be taken from the analysis of this study. From the regression analysis it is clear that longitude, latitude, elevation, distance to the Great Lakes, and their cross-products are all significant variables when considering mean daily temperatures. Furthermore, there is no need to consider different explanatory variables when estimating temperatures in different months. In the RMSE analysis it is shown that the variance of temperature estimates can be significantly reduced by including multiple weather stations in the estimation process rather than considering only the closest weather station. The RMSE analysis also shows that problems with variance can arise when automated variogram fitting procedures are applied. Using "rules of thumb" to estimate variogram parameters can result in increased variance which in turn can cause increased spatial basis risk.

The correlation analysis shows that the number of weather stations available in the interpolation area can have a significant impact on the amount of spatial basis risk. For northern Ontario where there are very few weather stations, the different interpolation methods produce more varied results, indicating that policies in this area are susceptible to spatial basis risk. Meanwhile in southern Ontario where there is a large number of weather stations, most of the interpolation methods produce very consistent results with only minor differences in between them, indicating that policies in this region are less susceptible to spatial basis risk. In addition, the correlation analysis shows that the cooling degree days index is not strongly correlated with forage yields, indicating that these types of policies should be combined with policies based on other weather variables like rainfall in order to reduce overall basis risk.

As weather based insurance products become more popular and more prominent in the marketplace, these results are important for any insurer who wants to design these types of policies. The results imply that insurers should spend more of their time and effort on increasing the amount of weather stations that are available in the area of interest, rather than being preoccupied with the interpolation method that is used to make estimates. Concerning the choice interpolation techniques, an insurer should choose a method which uses multiple weather stations from the surrounding area rather than only the closest station. If the number of weather stations available in the area is a concern, regression-based methods should be implemented in order to incorporate as much information as possible and reduce the prediction variance. Insurer's should also take great care if using methods like kriging which require parameter fitting, since these methods can result in higher variance due to poor fits of the variogram function.

Chapter 6: Summary

Weather index insurance has become a popular subject in agricultural risk management with many papers devoted to its discussion (Heimfarth and Musshoff, 2011; Lin et al., 2015; Okhrin et al., 2013). Under these policies farmers receive payments if they experience adverse weather for their crops during the growing season. These policies offer many benefits, such as reduced administrative costs and decreased adverse selection and moral hazard. However, this type of insurance is particularly susceptible to the problem of spatial basis risk (Lin et al., 2015). Spatial basis risk occurs when the weather observed at weather stations does not match the weather experienced by the farmer's crops, causing improper indemnities to be paid to the farmer (Dick and Stoppa, 2011). However, spatial basis risk may be reduced through the use of averaging and spatial interpolation techniques such as inverse distance weighting and kriging. These techniques make it possible to incorporate multiple surrounding weather stations in the estimation process rather than using only the single closest station, potentially resulting in more accurate estimations and thereby reducing spatial basis risk.

The objective of this study was to determine if an insurer's choice of spatial interpolation technique can impact the amount of spatial basis risk in a weather based insurance model. To evaluate the performance of different spatial interpolation techniques, temperature based policies for forage crops in Ontario, Canada, were considered as an example. A weather insurance index was developed based on cooling degree days, a weather metric which represents the excess heat stress that the crops experienced over the growing season. Seven different interpolation methods were applied to temperature data and estimated indemnities were calculated for forage producers across the province. By analyzing the correlation between the estimated indemnities and reported forage yields, the impact of the different interpolation techniques on spatial basis risk was quantified.

The analysis of this study was repeated three times over. The first analysis used data from all of Ontario, the second used data from southern Ontario only and the third used data from northern Ontario only. This subsetting of the data was done in order to compare the impact of spatial basis risk in areas with many weather stations (southern Ontario) to areas with few weather stations (northern Ontario). The results of this analysis show that the impact of spatial interpolation techniques on spatial basis risk is determined mainly by the number of weather stations that are available for analysis. When there are many stations available, the insurer's choice of interpolation technique has only a limited impact on the amount of spatial basis risk in the index insurance model. In this situation, the technique which was the most susceptible to spatial basis risk was regression kriging, an issue likely caused by the fact that the daily variograms were estimated using an automated fitting process. When there are few weather stations available the difference in between the methods becomes more pronounced, with regression-based interpolation methods showing the least amount of spatial basis risk.

The correlation analysis also showed that the cooling degree days index is not well suited for designing weather based index insurance. Even in the area with the most weather stations (southern Ontario) the correlation values never exceeded -0.154, and this is not a strong enough correlation to design effective index insurance. This indicates that the cooling degree day policies designed here are highly susceptible to product basis risk (see Chapter 2.2). Future attempts at designing temperature based insurance policies should focus on trying to include information regarding other variables such as rainfall in order to reduce basis risk in general.

It was also shown that the variance of the temperature estimates can be significantly reduced by using interpolation methods which incorporate information from multiple weather stations into the estimate, rather than relying only on the closest available station. Even the simplest of these methods (inverse distance weighting) showed a significantly improved root mean square error over the nearest neighbor method. Given the simplicity and ease of implementation of inverse distance weighting, there is little (if any) reason why a practitioner should choose to consider only the single closest weather station when making estimates of weather conditions.

Taking these results into consideration, an insurer may prefer to forgo the additional steps and complexities of a kriging model in favor of an easier to implement method such as inverse distance weighting. If a practitioner is determined to apply more complex methods to their analysis, spatio-temporal regression kriging is a better option than either ordinary kriging or regression kriging since only one variogram model needs to be fitted in each year rather than one variogram model for each day. This makes it easier to fit these models visually, avoiding the complications that can arise from the automated variogram fitting process. The results of this study are primarily of interest to insurance firms who are in the process of designing weather based insurance policies. For these practitioners it is important to know whether the additional effort and computational power required by methods like kriging produce any tangible results in terms of reducing spatial basis risk.

Future research should examine the impact of using less than ten weather stations in the spatial interpolation methods. It was shown that increasing the number of stations can reduce prediction variance, however, it may be the case that fewer than ten stations are required for this to occur. It is desirable for the index insurance to use fewer stations in order to keep the policy as simple as possible, making this topic an important one for consideration. In addition, future research should examine the possibility of combining temperature based policies with those based on rainfall. The temperature based policies were shown to be highly susceptible to product basis risk and this risk must be reduced in order to implement these policies practically and effectively.

References

AAFC (2012a). Evaluation of the agriinsurance, private sector risk management partnerships and wildlife compensation programs. Technical report, Agriculture and Agri-Food Canada. Available online at: http://www.agr.gc.ca/eng/about-us/offices-and-locations/office-of-audit-and-evaluation/audit-and-evaluation-reports/agriculture-and-agri-food-canada-evaluation-reports/evaluation-of-the-agriinsurance-private-sector-risk-management-partnerships-and-wildlife-compensation-programs/?id=1367338599421 (Accessed May 5, 2016).

AAFC (2012b). Forage statistics. Technical report, Agriculture and Agri-Food Canada. Available online at: http://www.agr.gc.ca/eng/industry-markets-and-trade/statistics-and-market-information/by-product-sector/crops/pulses-and-special-crops-canadian-industry/forage/forage-statistics/?id=1174494927045 (Accessed May 5, 2016).

AAFC (2016). Forage. Technical report, Agriculture and Agri-Food Canada. Available online at: http://www.agr.gc.ca/eng/industry-markets-and-trade/statistics-and-market-information/by-product-sector/crops/pulses-and-special-crops-canadian-industry/forage/?id=1174594338500 (Accessed May 5, 2016).

Baldwin, D. J., Desloges, J. R., and Band, L. E. (2011). *Ecology of a managed terrestrial landscape: patterns and processes of forest landscapes in Ontario*, chapter 2: Physical Geography of Ontario, pages 12–29. UBC Press.

Buxton, D. (1995). Growing quality forages under variable environmental conditions. In Kenelly, J., editor, *Proceedings of the Western Canadian Dairy Seminar*, Edmonton, AB, Canada: University of Alberta. Available online at: http://www.wcds.ca/proc/1995/wcd95123.htm (Accessed January 12, 2016).

Büyükalaca, O., Bulut, H., and Yılmaz, T. (2001). Analysis of variable-base heating and cooling degree-days for turkey. *Applied Energy*, 69(4):269–283.

Chien, Y.-J., Lee, D.-Y., Guo, H.-Y., and Houng, K.-H. (1997). Geostatistical analysis of soil properties of mid-west taiwan soils. *Soil Science*, 162(4):291–298.

Clarke, D. J., Clarke, D., Mahul, O., Rao, K. N., and Verma, N. (2012). Weather based crop insurance in india. *World Bank Policy Research Working Paper*, (5985).

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4):405–421.

Dick, W. and Stoppa, A. (2011). Weather index-based insurance in agricultural development: A technical guide. Technical report, International Fund for Agricultural Development (IFAD).

Heimfarth, L. E. and Musshoff, O. (2011). Weather index-based insurances for farmers in the north china plain: An analysis of risk reduction potential and basis risk. *Agricultural Finance Review*, 71(2):218–239.

Hengl, T. (2009). *A practical guide to geostatistical mapping*, volume 52. Office for Official Publications of the European Communities, Luxembourg, 2 edition.

Herdendorf, C. E. (1982). Large lakes of the world. *Journal of Great Lakes Research*, 8(3):379–412.

Hudson, G. and Wackernagel, H. (1994). Mapping temperature using kriging with external drift: theory and an example from scotland. *International journal of Climatology*, 14(1):77–91.

Jiang, W. and Li, J. (2014). The effects of spatial reference systems on the predictive accuracy of spatial interpolation methods. Technical Report GeoCat 76314, Geoscience Australia, Canberra. Record 2014/01.

Kilibarda, M., Hengl, T., Heuvelink, G., Gräler, B., Pebesma, E., Perčec Tadić, M., and Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, 119(5):2294–2313.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of Chemical, Metallurgical, and Mining Society of South Africa*, 52(6):119–139.

Kuzyakova, I., Romanenkov, V., and Kuzyakov, Y. V. (2001). Geostatistics in soil agrochemical studies. *Eurasian Soil Science*, 34(9):1011–1017.

Li, J. and Heap, A. D. (2008). A review of spatial interpolation methods for environmental scientists. Technical Report GeoCat 68229, Geoscience Australia. Record 2008/23, 137 pp.

Lin, J., Boyd, M., Pai, J., Porth, L., Zhang, Q., and Wang, K. (2015). Factors affecting farmersâĂŹ willingness to purchase weather index insurance in the hainan province of china. *Agricultural Finance Review*, 75(1):103–113.

Major, J. (1999). Index hedge performance: insurer market penetration and basis risk. In Froot, K. A., editor, *The Financing of Catastrophe Risk*, pages 391–432. University of Chicago Press.

Mendenhall, W., Sincich, T., and Boudreau, N. S. (1996). *A second course in statistics: regression analysis*, volume 5, chapter 4.11: Stepwise Regression, pages 242–251. Prentice Hall.

Norton, M., Osgood, D., and Turvey, C. G. (2010). Weather index insurance and the pricing of spatial basis risk. In *the Annual Meeting of the American Agricultural Economics Association*, Denver, CO.

Okhrin, O., Odening, M., and Xu, W. (2013). Systemic weather risk and crop insurance: the case of china. *Journal of Risk and Insurance*, 80(2):351–372.

Olea, R. A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, 20(5):307–318.

Paulson, N. D. and Hart, C. E. (2006). A spatial approach to addressing weather derivative basis risk: A drought insurance example. In *the Annual Meeting of the American Agricultural Economics Association*, Long Beach, CA.

Porth, L. and Tan, K. S. (2015). Agricultural insurance - more room to grow? *The Actuary Magazine*, 12(2):34.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240.

Veness, C. (2011). Calculate distance and bearing between two latitude/longitude points using haversine formula in javascript. Availble online at: http://www.movable-type.co.uk/scripts/latlong.html (Accessed March 4, 2016).

Wu, T. and Li, Y. (2013). Spatial interpolation of temperature in the united states using residual kriging. *Applied Geography*, 44:112–120.

Yungblut, D. and Jalbert, J. (2012). Assessing the potential impact of Roundup Ready® alfalfa on Canada's forage industry. Technical report, Canadian Forage & Grassland Association and Saskatchewan Forage Council.

Appendix A - Summary of Spatial Interpolation Techniques

Table A.1: Summary of the Spatial Interpolation Techniques Used to Estimate Temperature

Interpolation Technique	Description	Advantages	Disadvantages
Nearest Neighbor	For each day, the temperature estimate is equal to the observed temperature at the weather station closest to the farm.	Simple to implement and only requires the coordinates of the farms and stations to calculate distances.	Only the closest station is used in the estimation process, regardless of its proximity to the farm.
Regression	Daily linear regression models are used to make temperature estimates. Latitude, longitude, elevation, and distance to the Great Lakes are the explanatory variables, along with their cross products.	Simple to implement, well documented, and well understood. Incorporates information other than distance in the form of explanatory variables.	All weather stations are used in the estimation process regardless of their proximity to the farm and no attempt is made to define a spatial autocorrelation structure.
Inverse Distance Weighting	Weights for multiple stations around the farm are calculated based on the inverse of the distance from the station to the farm. These weights are used to estimate temperature conditions at the forage farms.	Multiple stations are used rather than only using the closest station. The model only requires one easy to fit parameter and only the coordinates of the stations and farms are required to calculate distances.	No explanatory variables are incorporated in the estimation process. Because no spatial autocorrelation structure has been defined, the weights assigned to each station may not be those which minimize the prediction error variance.

Continued on next page

Continued from previous page

Interpolation Technique	Description	Advantages	Disadvantages
Regression-Based Inverse Distance Weighting	Daily regression models are used to make temperature estimates and calculate model residuals at weather stations. Inverse distance weighting is used to estimate the model residuals and the final estimate is the sum of these two estimates.	Compared to inverse distance weighting, more information is included in the model in the form of explanatory variables. This method is also relatively simple to implement.	Because no spatial autocorrelation structure has been defined, there is no way of knowing whether the weights assigned to each station during the inverse distance weighting are minimizing the prediction error variance.
Ordinary Kriging	Every day, the temperature data is used to estimate a variogram function. This variogram function is used to determine the station weights which minimize the prediction error variance.	Once the variogram has been defined, the kriging formulas are straightforward. Because of the variogram function, the weights that are assigned to each station are optimal in terms of reducing prediction error variance.	Ordinary kriging assumes that the variable being estimated has a stationary mean over the entire interpolation area. This method also does not incorporate explanatory variables in the estimation process.

Continued on next page

Continued from previous page

Interpolation Technique	Description	Advantages	Disadvantages
Regression Kriging	Daily regression models are used to make temperature estimates and calculate model residuals at weather stations. Ordinary kriging is used to estimate the model residuals and the final estimate is the sum of the regression and kriging estimates.	Compared to ordinary kriging, more information is included in the model in the form of explanatory variables. Because ordinary kriging is being applied to residuals from regression models, concerns of stationarity are largely eliminated.	The variogram function must be estimated for every day during the growing season, meaning that automated variogram fitting methods must be used in order to perform analysis efficiently.
Spatio-Temporal Regression Kriging	Each year, daily regression models are used to make temperature estimates and calculate model residuals at weather stations. A spatio temporal variogram function is estimated from the data and used to make estimates of the model residuals for each day. The final estimate is the sum of the regression and kriging estimates.	Only one variogram model must be estimated for each year, rather than one for each day. This makes it easier to fit the variogram function visually to ensure that appropriate parameter values are selected.	Spatio-temporal regression kriging is more computationally intensive then other interpolation methods. In addition, the software and algorithms for performing this type of kriging are still being actively developed.

Appendix B - The Haversine Formula

When calculating the distance between two points on a flat two dimensional surface, Euclidean distance is used:

$$d = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

where Δx and Δy represent the the differences in x and y coordinates respectively. However the Earth is roughly a sphere, and so Euclidean distance is not an accurate representation of the distance in between two points on the Earth's surface. For these calculations a formula called the Haversine formula is used. This formula assumes that longitude and latitude have been converted to radians (Veness, 2011):

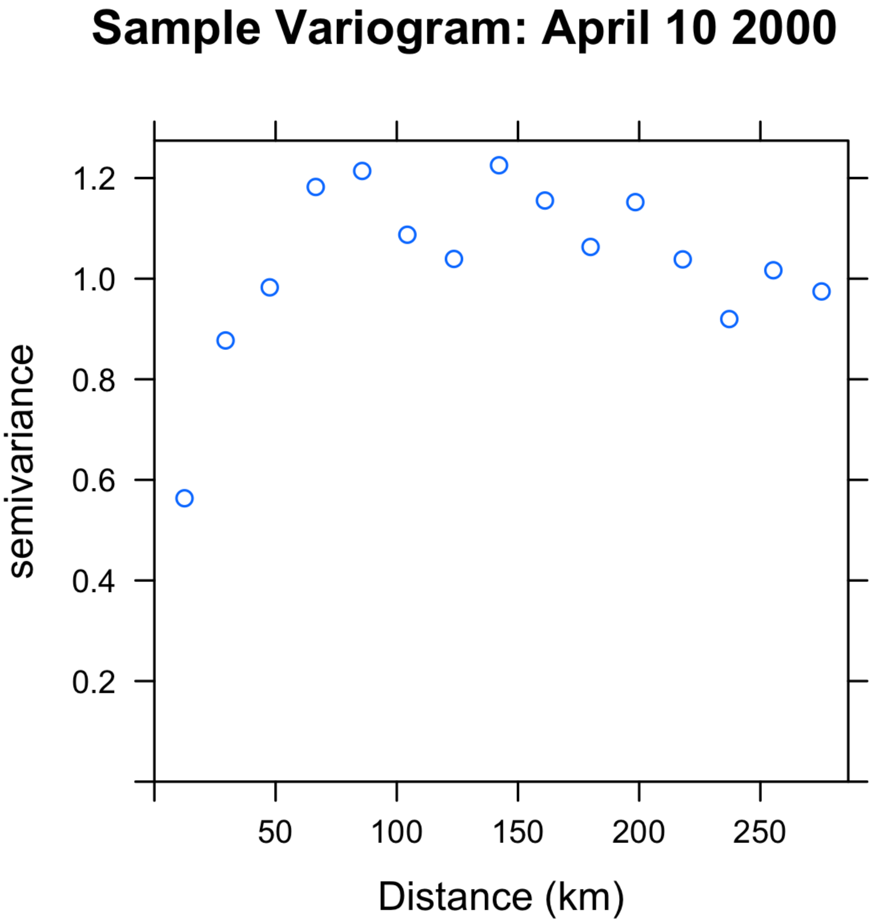
$$a = \sin^2\left(\frac{\Delta \mathrm{Lat}}{2}\right) + \cos(\mathrm{lat}_1) \cdot \cos(\mathrm{lat}_2) \cdot \sin^2\left(\frac{\Delta \mathrm{Long}}{2}\right)$$

$$c = 2 \cdot \mathrm{atan2}(\sqrt{a}, \sqrt{1-a})$$

$$d = R \cdot c$$

Here ΔLat and ΔLong represent the change in latitude and longitude respectively, lat_1 and lat_2 represent each of the latitude coordinates, and R represents the radius of the Earth in the desired unit of measurement. The Haversine formula is suitable for distance calculations over short distances and produces distance measurements with an average error of 0.3% (Veness, 2011).

Appendix C - Spatial Variogram Fitting



The graph above shows the sample variogram for the regression model residuals in southern Ontario on April 10th, 2000. The variogram is a function of three parameters: The nugget, the sill, and the range. These parameters can be visually identified from the sample variogram using the following reasoning. The nugget is defined as the value of the variogram for small distances, and theoretically it should be equal to 0. However due to measurements errors this is often not the case. From the sample variogram, the nugget is equal to approximately 0.5. The sill is the value of the plateau that the variogram eventually reaches as h increases. From the sample variogram, the sill is equal to approximately 1.2. Finally, the range is defined as the distance required for the variogram to reach the sill for the first time. From the sample variogram the range is equal to approximately 75 km. Using these values as initial estimates, the "fit.variogram" function from the gstat package in R is used to fit the three most popular variogram models (exponential, spherical, and Gaussian) and the model with the lowest sum of squared errors (SSE) is selected.

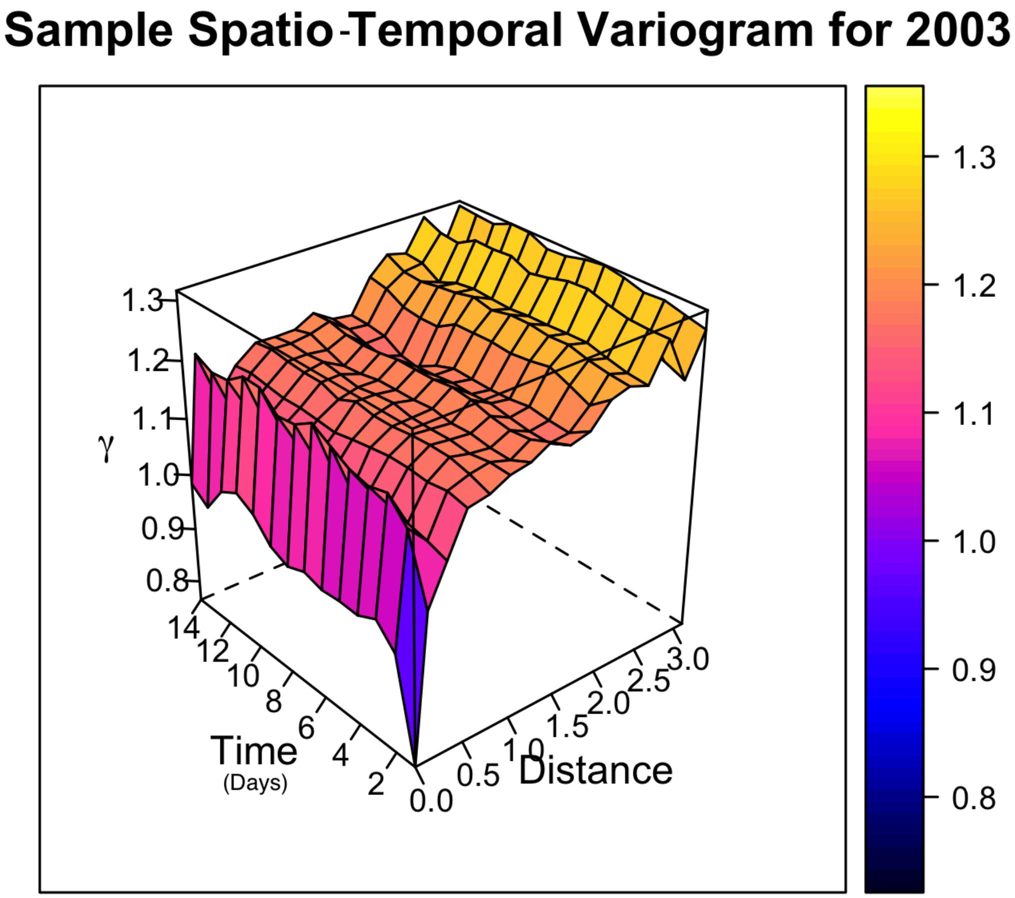
This procedure works well in theory, however, the analysis of this study covers eight years of data with each year having a ninety-one day growing season. This combined with the fact that the analysis is repeated three times for the different regions (southern Ontario, northern Ontario, and all Ontario) means that a total of $91 \cdot 8 \cdot 3 = 2184$ variogram models must be fitted for the analysis of this study. It is not feasible to fit all of these models visually, and so an automated fitting procedure must be implemented. In order to provide initial estimates for the variogram parameters, the method of Hengl (2009) is used:

1. Nugget=0
2. The sill is set to the variance of the data set → Sill=Var($\vec{Z}$)
3. The range is set to 100 km[1]

Using these initial estimates, the "fit.variogram" function fits the sample variogram to each of the three variogram functions(exponential, spherical, and Gaussian) and the function with the lowest SSE is selected.

[1] Hengl (2009) recommends the range be set to one quarter the diagonal of the interpolation area, however, due to the size of the interpolation area in this analysis this value was modified to 100 km. This results in a more accurate estimate of the range.

Appendix D - Spatio-Temporal Variogram Fitting



The graph above shows the sample spatio-temporal variogram for southern Ontario from the year 2003. The first difference in between this variogram and the purely spatial variogram is the units used for distance. The algorithms and software packages for performing spatio-temporal kriging in R are still actively being developed and as such there are certain features that have not been implemented yet. One of these features is the use of coordinates that are given in the longitude/latitude format, and so a Cartesian projection of the data must be used. Therefore, the distances in this variogram are given in the form of Euclidean distances assuming that one degree of latitude is equal in distance to one degree of longitude. It is a known fact that there is no method for projecting data from a sphere to a flat surface without distorting some of the distances, and indeed this is visible in Figure 3.1 where it can be seen that one degree of latitude covers more distance than one degree of longitude. However, research has shown that transforming projections from a spherical surface to a planar surface has a negligible impact on the spatial interpolation process (Jiang and Li, 2014).

Since only one model must be fit for each year, there are only twenty-four spatio temporal variograms that must be fitted for this analysis (eight years times three regions) and so this task is much less daunting than the process of fitting 2184 spatial variograms. As mentioned in Chapter 4.7, the spatio temporal variogram is a function of three distinct variograms with a total of ten parameters.

To fit the spatial variogram, consider the shape of the graph when time is 0 and follow the procedure outlined in Appendix C. The spatial nugget is estimated as 0, the spatial sill as 1.2 and the spatial range as 1.0. The temporal variogram is fit in a similar fashion, except by examining the shape of the variogram when distance is 0. This gives a temporal nugget of 0, a temporal sill of 1.05 and a temporal range of approximately 4 days. The spatio-temporal anisotropy ratio is estimated as the ratio of the spatial range to the temporal range which gives a value of 1.0/4=0.25.

The joint variogram is estimated by considering the shape of the variogram as both the distance and time increase. The joint nugget is estimated as 0 and the joint sill as 1.3. To estimate the joint range, the spatial and temporal ranges are considered. Using the anisotropy ratio the temporal range can be converted to spatial units, and the joint range is equal to the distance to the point with the spatial and temporal ranges as coordinates. Therefore, the joint range can be estimated as the diagonal of a square with the spatial range as the length of its sides and the estimate is given by $\sqrt{2} \cdot 1.0 = 1.41$.

Once the parameters have been estimated, the variograms are each fitted to the three most popular variogram functions (exponential, spherical, and Gaussian) and the model with the lowest SSE is selected. Since there are three marginal variograms to be fitted and three possible variogram functions, there are 3^3=27 possible combinations of models.