The impact of weather change on Nitrous Oxide Emission with Spatial Pattern Detection and Large Data Approximation

by

Francis Ohene Ofori

B.S., University of Ghana, 2009

M.S., University of Vermont, 2012

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

The correlations between agriculture, climate change, and greenhouse gas concentration are multiplex and manifold. Agriculture has been a focus due to its vital connection with climate and food supply. It could have substantial implications for the economy and agricultural management to study the detection of spatial pattern in regional climate change and the impact of weather change on greenhouse gases, specifically on nitrous oxide N_2O emission of state crops.

To capture the spatial pattern of significant regional climate change, a Process-based Geographical Algorithm Machine (PGAM) procedure is proposed by viewing the spatio-temporal data sets as a realizations of underlying random fields. Past and future climate scenarios of daily weather are simulated using multiple Global Circulation Models (GCMs). The simulation differences and consistency of precipitation, minimum, and maximum temperature in the state of Kansas produced by these climate models are assessed using the spatial Kolmogorov-Smirnov test. The climate change index described by a temporal distance metric from PGAM is used to study the adverse effect on N_2O emission connected with agricultural management practices based on a linear mixed-effect model.

This project further delves into the effect of weather change on N_2O emission using large data approximation technique; however, the size of the data set creates issues in estimation and prediction. This is because it involves the determinant and inversion of the $n \times n$ covariance matrix of the data process. Thus, an approximation technique for reducing the dimension of the covariance matrix is required. We theoretically and numerically investigate the conditions under which computational intensity and prediction accuracy are balanced by adopting a projection approach. An optimal rank selection method is proposed to achieve good efficiency in terms of Kullback-Leibler divergence and mean square prediction error

while reducing the computational cost. The accuracy and performance of the proposed method are evaluated via both simulation and spatial regression analysis of N_2O emission.

The impact of weather change on Nitrous Oxide Emission with Spatial Pattern Detection and Large Data Approximation

by

Francis Ohene Ofori

B.S., University of Ghana, 2009

M.S., University of Vermont, 2012

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Juan Du

Copyright

Abstract

The correlations between agriculture, climate change, and greenhouse gas concentration are multiplex and manifold. Agriculture has been a focus due to its vital connection with climate and food supply. It could have substantial implications for the economy and agricultural management to study the detection of spatial pattern in regional climate change and the impact of weather change on greenhouse gases, specifically on nitrous oxide N_2O emission of state crops.

To capture the spatial pattern of significant regional climate change, a Process-based Geographical Algorithm Machine (PGAM) procedure is proposed by viewing the spatio-temporal data sets as a realizations of underlying random fields. Past and future climate scenarios of daily weather are simulated using multiple Global Circulation Models (GCMs). The simulation differences and consistency of precipitation, minimum, and maximum temperature in the state of Kansas produced by these climate models are assessed using the spatial Kolmogorov-Smirnov test. The climate change index described by a temporal distance metric from PGAM is used to study the adverse effect on N_2O emission connected with agricultural management practices based on a linear mixed-effect model.

This project further delves into the effect of weather change on N_2O emission using large data approximation technique; however, the size of the data set creates issues in estimation and prediction. This is because it involves the determinant and inversion of the $n \times n$ covariance matrix of the data process. Thus, an approximation technique for reducing the dimension of the covariance matrix is required. We theoretically and numerically investigate the conditions under which computational intensity and prediction accuracy are balanced by adopting a projection approach. An optimal rank selection method is proposed to achieve good efficiency in terms of Kullback-Leibler divergence and mean square prediction error

while reducing the computational cost. The accuracy and performance of the proposed method are evaluated via both simulation and spatial regression analysis of N_2O emission.

Table of Contents

List of Figures

List of Tables

Acknowledgments

During my academic years at Kansas State University, I had Dr. Juan Du as my academic advisor from whom I learned everything I know about statistical research, specifically, spatial and large data analysis. After completing my qualifying exam, I have continuously benefited from long meetings with Dr. Du, who was always ready to help me understand the theoretical and computational aspect of the research. She was prepared to write down formulae and equations with me and to listen to me. I'm very grateful for her assistance, patience, and support. Additionally, I learned from Dr. Du, the joy of research. Every time we discussed new ideas and I was lost, she was always ready to help. I am very appreciative of and thankful to her for sharing her passion and for showing me that research is fun, challenging but doable, and I am proud and honored to be her student. I also enjoyed the advice and guidance from several other professors, who will have a lasting influence on my research direction and my life. Dr. Trevor Hefley did not hesitate to join my committee when I reached out to him after my qualifying exams; he has been an inspiration, and I am fortunate for his time and advice. I want to thank the outside chair Dr. Shawn Hutchinson and my committee members, Dr. James Neil, Dr. Trevor Hefley and Dr. Ignacio Ciampitti for assisting with their insightful thought on the research and suggestions. Many thanks go to Dr. James Dzikuni who was helpful during my work on the simulation studies. I am grateful to Dr. Bo Li and Dr. Banerjee for sharing their insights on regional climate models, approximation techniques and matrix projection. Their work inspired me to start working on the topic of this thesis. A very big thank you to all my classmates for hours of group studies and discussions. Thanks to Bonnie Messmer and Jo Blackburn for all they do in the department. Many thanks to my beloved family and friends who encouraged me during my stay in KSU. "All hard work brings a profit, but mere talk leads only to poverty."

Dedication

This thesis is dedicated to my sweet, charming and lovely wife, Claudia. Thank you for your love, support and encouragement during my studies. Yes, we made it together, Michochomicho. Thank you for the great moments we had. You are forever mine.

And to my mother, brothers and sisters-in-law.

Chapter 1

Spatial Pattern Detection of Regional Weather Change and the Effect on Nitrous Oxide Emission in Kansas Agro-Ecosystem

1.1 Introduction

The ability to study spatial pattern in climate models and the impact of weather change on nitrous oxide emission is one of several concerns in the agricultural industry. One of the many causes of weather change is industrial and human practices such as deforestation, urbanization, shifts in vegetation, burning fossil fuels, and other agricultural activities. These activities indeed increase the concentration of carbon dioxide and other greenhouse gases. Fluctuations in atmospheric formation represent changes in temperature, precipitation and amongst others, and alter marine and terrestrial ecosystems. The effects of weather change on agriculture include incidents of flooding, establishment of invasive species, modification to rangeland characteristics, and damage to livestock and crops from increased temperature to lower temperature.

Meanwhile, climate change, evidenced by changes in temperatures and precipitation patterns and magnitude, has an effect on nitrogen dynamics that impacts (N_2O) emissions. According to the Kansas Agriculture's Economic Impact Report (Updated October 2018), agriculture in the state of Kansas is an important contributor to the State's economic well-being. Agriculture accounts for over 42% of the total economy, and in 2012, there were almost 61,7333 farms in the state, producing crops and livestock. Given that Kansas is agriculture-based and currently 8^{th} in terms of oil and gas production, [39], it would be very important for the economy and agricultural management to study the spatial pattern of regional climate change and determine whether climate change has an adverse effect on greenhouse gases emission of state crops.

The fundamental tools to understand and project regional climate change due to increasing greenhouse gas concentrations are Global Circulation Models (GCMs) and a large body of observational and theoretical results [57]. GCMs are climate models that use Navier-Stokes equations and computer programs to simulate the earth's atmosphere. GCMs simulate contrasting changes in future climate [2] and are usually used for predicting weather, forecasting climate change, and understanding climate. However, numerous studies have examined the projection of global change and its impact on agricultural ecosystems. Evaluating these impact and mitigation strategies from agriculture has been focused on emission scenarios [31]. Its been established that integrating emission scenarios with ecosystem models improves the estimate of emission, impacts assessment, evaluates and identifies mitigation strategies [46]. In the past, impact assessment studies have focused on the effect of climate change on crop production [31]. Other studies have focused on the evaluation of climate change on greenhouse emission from the soil even though nitrous oxide emissions from agricultural soil have been projected to increase tremendously through 2030 as a result of expansion in crops and livestock [50].

Understanding the impact of weather/climate change on nitrous oxide emissions in agro-ecosystems is useful because agricultural soil management practices play a key role in nitrous oxide emissions. Moreover, Argoti (2013) established that the impact of weather/climate change on nitrous oxide emissions involves developing plausible future climate scenarios.

He further studied weather/climate change by using simulated climate data. However, the method (change factor methodology) used was a highly integrated index based on averaging weather information across space and over time. Recently, researchers have become interested in observing and understanding how climate change affects regional zones in the area under study. Regional climate studies assist in identifying regions that are most affected by the changes in climate. Furthermore, the implications of climate change on weather data can be divided into temporal and spatial patterns on regional scales [47, 49]. Understanding the concept of spatial and temporal patterns in temperature and precipitation therefore is important because they are used as mechanisms in ecologic, environmental, and hydrologic models [9].

There are several techniques to pattern detection, including the subjective eyeball technique, a kernel-based method that highlights differences on a surface (median smoothing techniques such as headbanging [27], artificial intelligence approaches (e.g., genetic algorithms and neural networks [45], and exploratory spatial data analysis [3]. Kernel-based method, genetic algorithms, neural network, exploratory spatial data analysis and many more rely on statistical methods for pattern detection. These techniques are used to determine whether an observed pattern of an event in one or more geographical regions occurs by chance alone. In our case, we borrow the idea of Geographical Algorithm Machine (GAM) which focuses on detection of intensity clusters rather than pattern detection. GAM depend on the overall pattern of an event over a large region, treating the data under study as one realization from a process and using a global process test to detect clusters. For example, the metric distance by [33] was used to assess spatial association and detection of spatial patterns via clustering. The idea of the metric distance was to show the global spatial pattern of the region under study rather than the micro-spatial pattern. Since the spatial pattern detection relies on the concept of location and distance, the inherent relationship between distance and similarity has been extended to space to represent similarity. Also, researchers such as [43, 20] have assessed spatial patterns using statistical measures such as local indicators of spatial association and local Moran's I in a global sense rather than as a local test.

In this study, we propose a statistical technique for detecting the spatial pattern that is in line with the Geographical Algorithm Machine (GAM) and distance test, [42, 33], but have a statistical technique for detection of spatial pattern at a local level. Most importantly, we delve into the process difference rather than the observational difference at each fixed location under study. We proposed a Process-based Geographical Algorithm Machine (PGAM) to detect spatial patterns in regions under study and used linear mixed models to determine the effect of weather change on nitrous oxide emission.

The remainder of the dissertation is organized as follows: Section 1.2 presents a description of weather variables and nitrous oxide emission data in the state of Kansas; Section 1.3, describes the periodic auto-regressive model, distance metric and Kolgomorov sample test; Section 1.4 explains the methodology and proposed Process-based Geographical Algorithm Machine (PGAM); Section 1.5 presents the results of our analysis using the proposed method; Finally, Section 1.6 presents our conclusion and discussions.

1.2 Kansas Weather and Nitrous Oxide Data

Kansas is known as an energy-producing and agriculture-based state and is located in the Great Plains region of the United States [40]. Corn, winter wheat, sorghum, and soybean, amongst others, are the main crops in Kansas. The state of Kansas covers an area of 213,096 km^2 (Institute for Policy and Social Research, 2011) and covers nine climate divisions (NCDC, 1994). The agricultural sector plays a pivotal role in the Kansas economy; in 2006, it accounted for $3.29 billion and about 3 percent of the state's GDP of about $114 billion. Correspondingly, the North American Regional Assessment Program (NARCAAP) supply dynamically downscaled GCMs output at a spatial resolution of 50 km. It also provides high resolution climate change simulations to investigate uncertainties in regional scale projections of future climate and generates scenarios for use [2].

Around 128 grid points that covered the state of Kansas were selected for this study. Different Regional Climate Models (RCMs) were developed via different Atmosphere - Ocean General Circulation Models (AOGCMs) to provide boundary conditions [38]. NARCCAP

was divided into two phases; phase 1, in which six RCMs use boundary conditions for National Center for Environmental Prediction (NCEP-DOE) Reanalysis II (R2) for a period of 25 years (1980-2004) and phase 2, in which the boundary conditions are provided by four AOGCMs for 30 years of current climate (1971-2000) and 30 years of future climate (2041 - 2070) for the Special report on emission scenarios (SRES) [2].

The datasets used for this analysis from phase 2 include a mixture of AOGCMs and RCMs (See Table 1.2) for description of the models used [2]. To be precise, six datasets, namely three from RCMs and three from AOGCMs, were used for this analysis. The three RCMs include; The Canadian Regional Climate Models (CRCM) from OURANOS/UQAM, the Regional Climate Model version 3 (RCM3) obtained from UC Santa Cruz, and the Weather Research and Forecasting model (WRFG) obtained from the Pacific Northwest National Lab. The three other AOGCMs include; The Community Climate System Model (CCSM) obtained from National Center for Atmospheric Research (NCAR), the third generation Coupled Global Climate Model (CGCM3) obtained from the Canadian Center for Climate Modeling, and the Geophysical Fluid Dynamics Laboratory (GFDL) [2]. Due to differences in coordinate systems, all models were re-gridded to an identical resolution using the linint2 Wrap from the NCL software NCL (2012), [2].

The dataset used for this analysis was accessed on September 2012 [38, 2]. The data was cleaned and imputed in cases with missing values using the local moving average method as it provides a full sample size, which can be advantageous for reducing bias and optimizing precision. In this study, data from (1971-1992) and (2041-2062) for current and future climate scenarios, respectively, were used. For each weather variable, there were two data set from each RCMs: the current and future climate scenarios. Three main weather variables were considered in this study: maximum and minimum temperature and precipitation. Maximum and minimum temperature were at monthly resolution Kevin (K), and precipitation was at 3 hr resolution ($kgm^{-2}s^{-1}$). Other variables used included regional simulated and observed nitrous oxide emissions (N_2O) from four crops (corn, sorghum, soybean and winter wheat), agricultural management practices (tillage no irrigation, no-tillage no irrigation, tillage irrigation, and no-tillage irrigation), crop-agricultural management practices (groups), locations

(longitude and latitude), years, and months.

Nitrous oxide emission data was obtained from the DeNitrification-DeComposition (DNDC) model, which was initially developed to model (N_2O) emission from cropped soils in the U.S, and since then, has been modified by many research groups and used in many countries and production systems [21, 2]. The DNDC model simulates biochemical and geochemical reactions common in agroecosystems, which are mainly carbon (C) and nitrogen (N) transport and transformation in plant-soil-climate systems [32]. The model consists of six submodels: thermal-hydraulic, aerobic decomposition, nitrification, denitrification, fermentation, and plant growth. The thermal-hydraulic submodel calculates soil temperature and moisture profiles based on soil physical properties, daily weather, and plant water use. The aerobic decomposition simulates production of soil organic matter driven by soil microbial respiration. The nitrification submodule calculates growth of nitrifiers and oxidation of ammonium to nitrate. The denitrification sub-model simulates denitrification and the production of nitric oxide, nitrous oxide, and dinitrogen at an hourly time step. The fermentation sub-model simulates methane production and oxidation under anaerobic conditions. Plant growth is modeled with the DNDC daily crop growth curve [21]. Detailed information about DNDC is in [32]. The parameters in DNDC used in this work are broadly classified into four categories: climate, soil, crop, and management [2].

1.3 Preliminary Results

1.3.1 Periodic Auto-Regressive Model

Evidence shows that some temporal weather processes exhibit stochastic trends and seasonal cycles. When these cycles exist in a temporal process, they do not act independently and hence differencing filters may not be ideal. Simple periodic auto-regressive models usually have the tendency to depict temporal weather data, and more specifically, the first order periodic auto-regressive model fits temperature data reasonably well. Also, the periodic auto-regressive model has a varying seasonal autoregressive parameter and a periodic differencing

filter that helps in understanding the mechanisms in the data. To formulate the process, let Y_t be an observed temporal process, a univariate representation of Periodic Autoregressive of order p, written for short as PAR(p) and given as

$$Y_t = \sum_{i=1}^{p} \phi_{i,s} Y_{t-i} + \epsilon_t, \tag{1.1}$$

where $\{s\}_{j=1}^{S}$ denotes the season for $\{t\}_{j=1}^{n}$, n is the total number of observations, $\{\phi_{i,s}\}_{i=1}^{p} \in \mathbb{R}$ are the autoregressive parameters for different seasons and ϵ_t is assumed to be a white noise process with constant variance σ^2. The assumption can be relaxed to account for the variability in the season, which is σ_s^2. Periodic autoregressive models have estimated parameters that change with the seasons. Moreover, there are several ways of estimating $\phi_{i,s}$; for example, an asymptotic efficient estimate of $\phi_{i,s}$ can be obtained by solving Yule - Walker equations [44]. Considering the normality assumption of ϵ_t and fixed starting values, and using the ordinary least squares estimation, the maximum likelihood estimates of the parameter $\phi_{i,s}$ can be obtained [17]. The order of a PAR (p) is obtained by using Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) in addition to a diagonistic test on the residual autocorrelation [18]. It is advisable to use the Lagrange Multiplier (LM) test to assess the periodicity of the autocorrelation in the residuals. Periodic moving average of order q is denoted as PMA(q) and is developed in a similar way. However, the identification of the order of a periodic model is not straightforward, and often, model selection criterion such as AIC or BIC are used to select a suitable model [16].

1.3.2 The Temporal Distance Metric

Two temporal processes are similar if the distance between them is small and dissmiliar if the distance is large. This is because processes from the same location or region have more similar traits than processes from different regions. To account for the distance between the processes at the same domain of study accounting for the seasonal structures, we consider the temporal distance metric. A temporal distance metric is a statistical test used to assess the

seasonal means, variances, and auto-correlation between two temporal processes of a spatio-temporal data simultaneously at a particular location [33]. Thus, it is used to determine whether two temporal processes have the same dynamics or not. In assessing that, let $\mathbf{X} = \{X_t\}$ and $\mathbf{Y} = \{Y_t\}$ be two temporal processes at the same location and $d(\mathbf{X}, \mathbf{Y})$ be the process difference rather than the locational difference; then the function d must satisfy the following properties [36]. For every $\mathbf{X}, \mathbf{Y}$;

(i) The symmetric property, $d(\mathbf{X}, \mathbf{Y}) = d(\mathbf{Y}, \mathbf{X})$.

(ii) The non-negative property, $d(\mathbf{X}, \mathbf{Y}) \geq 0$.

(iii) The identification marking property, $d(\mathbf{X}, \mathbf{X}) = 0$.

The temporal distance metric between two temporal processes at a location is given as:

$$d(\mathbf{X}, \mathbf{Y}) = \frac{1}{2N} \sum_{t=1}^{N} \left(\frac{X_t - \hat{X}_t - Y_t + \hat{Y}_t}{\nu_t^{1/2}} \right)^2, \tag{1.2}$$

where N represents the length of the two temporal processes, hat represents one-step-ahead prediction, and ν_t represents squared prediction error. This temporal distance metric tests the process difference of two temporal process at the same location and is computed under the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ has equal first two moments: $E[X_t] = E[Y_t]$ for all t and $cov(X_t, X_s) = cov(Y_t, Y_s)$ for all t, s. In conducting the test, the same model must be fitted for both $\mathbf{X}$ and $\mathbf{Y}$ at the same location.

1.3.3 The Kolmogorov-Smirnov (KS) test

Kolmogorov-Smirnov was developed in the 1930s by Andrei Nokolaevich Kolmogorov and Nikolai Vasilyevich Smirnov. In our settings, we will consider the two sample non-parametric test, which is used to compare distributions of statistical values in two datasets. The main purpose of performing this test is to determine whether or not data samples are realization from the same process. To proceed, suppose $Y_1, Y_2, ..., Y_m$ is the first data sample of size m with a cumulative distribution function $F(y)$, and $X_1, X_2, ..., X_n$ is a second sample of size n,

with cumulative distribution function $G(x)$. The hypothesis test under the null hypothesis is given as $H_0 : F = G, \quad H_1 : F \neq G$. Under H_0, the two sample KS test statistic is given as: $D_{mn} = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \sup_x |F_m(x) - G_n(x)|$ where F_m and G_n are empirical cumulative distribution functions. When H_0 is rejected, we conclude evidence that the two sample data come from the same distribution.

1.3.4 Linear Mixed Effect Model

The mixed effects model is a powerful statistical method for analyzing grouped data according to several classification factors. This model has been one of the important centerpieces of applied statistics in the agricultural and biological fields. Generally, a Linear Mixed Effects model (LME) is used to determine the relationship between the response and various classification factors for the observations. In this study, due to the structure of the data, an exploratory analysis and a determination of how both the fixed and random effect enters the model linearly require a linear mixed effect model. This model is formulated by following the idea of [30] as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is a column vector, the response variable; $\mathbf{X}$ is a matrix of independent variables; $\boldsymbol{\beta}$ is a column vector of the fixed-efffects regression coefficients; $\mathbf{Z}$ is the design matrix for the random effects; $\boldsymbol{\mu}$ is a vector of random effects and $\boldsymbol{\epsilon}$ is a column vector of random errors, which may or may not be spatially correlated. A complete description of this model can be found in the reference paper [30].

1.4 Methodology

Agricultural studies has been one of the main research areas in the world due to the important role it plays in climate change, food supply, and human activities. Since the state of Kansas is considered to be an agriculture-based area, it would be very important to study the spatial pattern of regional climate change and the impact of weather change on state crop

yield. Below, we will illustrate the steps in using our proposed Process-based Geographical Algorithm Machine, which targets micro-pattern detection based on significant differences in the underlying processes. The main procedure of the proposed method is presented in step 2 below.

1.4.1 Process-based Geographical Analysis Machine (PGAM)

Motivated by the idea of Geographical Analysis Machine (GAM) by [42], we propose Process-based Geographical Analysis Machine (PGAM), which can be used as an algorithm tool for detecting spatial pattern by viewing the dataset as a realization of the underlying processes. To effectively study the spatial pattern of regional climate change and the effect on Nitrous Oxide emission in an agro-ecosystem effectively, the steps below illustrate how to achieve the desired results.

Step 1. If the number of locations is not large, that is; if the data is distributed regularly at grid points with a small sample size (say; the sample size is less than 500). At each grid point, fit a suitable time series model. Numerous studies on time series with periodic and seasonal properties have been well researched and covered areas such as economics, climatology, signal processing, hydrology, electrical engineering, and genetics amongst others and researchers have applied periodic time-series models in these fields [34, 37]. Identification of such models is usually the painful part of the model building procedure. In the past, statistical techniques in periodic time series models have been attributed to Gladyshev [22]. However, Noakes (1985) suggested investigation into plots of periodic autocorrelation function as an ideal way to detect periodic and seasonal models [41]. Time series with periodic and seasonal properties can be modeled by using either periodic or seasonal processes. In general, for a temporal process $\epsilon_{nT+\nu}$, the univariate representation of periodic autoregressive of order p is written as:

$$\epsilon_{nT+\nu} = \sum_{i=1}^{p} \phi_{i,s}\epsilon_{nT+\nu-i} + a_{nT+\nu}, \tag{1.3}$$

where $s = 1, 2, ..., T$ is the season for time period $n = 1, 2, ..., N$, n is the total number of

periods, $\phi_{i,s} : i = 1, 2, ...p$ are the autogressive parameters for different seasons, and $a_{nT+\nu}$ is assumed to be a periodic white noise with mean 0 and constant variance σ^2. Periodic autoregressive models are often assumed to be stationary in the periodic sense [22]. In recent years, there has been much interest [26, 44] in periodic autoregressive model across many disciplines leading to estimation and testing of these models. Thus, researchers have reviewed and examined the fundamentals traits of periodic autoregressive models and the techniques for estimating parameters.

In this study, we concentrate on the first order periodic autoregressive model PAR (1) and all its traits because it describes the behavior of climatic weather data very well based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Our primary model for a temporal process $Y_{Tn+\nu}$ is given as

$$Y_{Tn+\nu} = \mu_{\nu}^{Y} + f(Tn + \nu) + \epsilon_{Tn+\nu}^{Y},$$

where μ_{ν}^{Y} is the seasonal intercept for the ν^{th} month, f is the trend function and $\epsilon_{Tn+\nu}^{Y}$ is the random error with mean 0 and variance σ_{ν}^{2}. In our case study, let $Y_{12n+\nu}^{c}$ and $Y_{12n+\nu}^{f}$, be the current and future processes respectively and model them using PAR (1) as;

$$Y_{12n+\nu}^{c} = \mu_{\nu}^{Y^c} + \alpha(12n + \nu) + \epsilon_{12n+\nu}^{Y^c},$$
$$Y_{12n+\nu}^{f} = \mu_{\nu}^{Y^f} + \alpha(12n + \nu) + \epsilon_{12n+\nu}^{Y^f},$$

where $n = 1, 2, ..., 22$, the number of years is assumed to follow a first - order periodic model with period 12 and $\nu = 1, 2, ..., 12$. The rest of the parameters are explained as described in the primary model.

Step 2. Perform the proposed Process-based Geographical Algorithm Machine (PGAM). To capture spatial pattern of significant weather change, we view the data sets (current and future spatio-temporal data) as realizations of the underlying process. In this procedure, we focus on the process difference rather than the observational difference. Motivated by the clustering detection procedure of GAM on the spatial aspect, we propose PGAM for weather

change pattern detection. This method is specifically employed by first creating a fine, dense grid mesh across the study region and building several circles of increasing diameters (say; 100m) at each mesh point. Secondly, at each circle, if the number of locations or the grid structure is small (say; less than 500 locations such that each circle contains enough sample locations say 30 to allow reasonable statistical test) otherwise scan each point performing temporal difference test. In general, a two sample spatial Kolmogorov Smirnov (KS) test is applied to two standardized data. The current data can be viewed as a realization of the first Gaussian random field indexed by both space and time while future data is another one determined by individual mean and covariance function structures (μ_1, C_1) and (μ_2, C_2) respectively. Those structures can be estimated by the likelihood method or another estimation method based on the underlying distribution of the data. We then test for the significant difference of the underlying structure by using the standardized samples $\tilde{Y_1} = \tilde{M}^{-1/2}(Y_1 - \tilde{m})$ and $\tilde{Y_2} = \tilde{M}^{-1/2}(Y_2 - \tilde{m})$ where $\tilde{m}$ and $\tilde{M}$ are estimated pooled mean and covariance matrix. Thirdly, a kernel estimation process [53] is applied to recognize the weather change spatial pattern and location scale. Thus, circles where the test shows significance will be flagged, and a kernel is passed over the study region of significant circles to detect areas with significant weather change.

Part I. In our study, due to the grid structure in the data (spatio-temporal structure), we propose the simplified Process-based Geographical Algorithm Machine (PGAM) for weather change spatial pattern detection. We first consider monthly precipitation, maximum and minimum temperature for two time-periods, current (1971 – 1992) and future (2040 – 2061), from each of the combination of the Climate Models (RCMs) in the state of Kansas. Each of the 128 grid points that cover Kansas has two temporal processes; now, we can build several circles of increasing diameter (100m) at each grid point. Secondly, we scan through each local circle by performing a temporal distance metric test d to assess the process difference between current and future data in this local area. The temporal spatial metric test embodies trend, covariates, seasonal mean and autocorrelation structures in the temporal process (for both current and future data).

Before computing the temporal spatial metric test, we first consider the computation of one-

step ahead prediction. Using the estimated parameters from PAR (1) in step 1 and following [33], let the superscript π, denote combined estimators from the process Y_t^c and Y_t^f; then the one step ahead prediction is given as:

$$\tilde{Y^c}_{12n+\nu} = \hat{m}^{\pi}_{12n+\nu} + \tilde{\phi}^{\pi}(Y^c_{12n+\nu-1} - \hat{m}^{\pi}_{12n+\nu-1}),$$
$$\tilde{Y^f}_{12n+\nu} = \hat{m}^{\pi}_{12n+\nu} + \tilde{\phi}^{\pi}(Y^f_{12n+\nu-1} - \hat{m}^{\pi}_{12n+\nu-1}),$$

where $\hat{m}^{\pi}_{12n+\nu} = \hat{\mu}^{\pi}_{12n+\nu} + \hat{\alpha^{\pi}}(12n+\nu), \hat{\alpha^{\pi}} = (\alpha^{\tilde{Y^c}} + \alpha^{\tilde{Y^c}})/2$ is the average of the regression estimator from the two-temporal process and

$$\hat{\mu}^{\pi}_{\nu} = \frac{\hat{\mu}^{Y^c}_{\nu} + \hat{\mu}^{Y^f}_{\nu}}{2}, \qquad \hat{\phi}^{\pi}_{\nu} = \frac{\hat{\phi}^{Y^c}_{\nu} + \hat{\phi}^{Y^f}_{\nu}}{2}, \qquad (\hat{\sigma}^2_{\nu})^{\pi} = \frac{(\hat{\sigma}^2_{\nu})^{Y^c} + (\hat{\sigma}^2_{\nu})^{Y^f}}{2},$$

are the arithmetic averages of the estimators from the two process. The mean squared prediction error from PAR (1) is given as $\nu_{12n+\nu} = (\hat{\sigma}^2_{\nu})^{\pi}$. The estimated parameters of $\hat{\sigma}^2, \nu$ and π in addition to $\tilde{Y_t}^c$ and $\tilde{Y_t}^f$ are used in the computation of the temporal distance metric as described below. This metric is performed by first considering the current $\{Y_t^c\}$ and future $\{Y_t^f\}$ data at each grid point (location) in Kansas.

Proposition 1.4.1 *Assume that* $\{Y_t^c\}$ *and* $\{Y_t^f\}$ *are a Gaussian process, if* $\{Y_t^c\}$ *and* $\{Y_t^f\}$ *have the same mean and autocovariance functions,*

$$d = \frac{1}{2N}\sum_{t=1}^{N}\left(\frac{Y_t{}^c - \tilde{Y_t}{}^c - Y_t{}^f + \tilde{Y_t}{}^f}{\nu_t^{1/2}}\right)^2 \sim \chi^2_N/N, \tag{1.4}$$

where $Y^c = \{Y_t^c\}$ and $Y^f = \{Y_t^f\}$ are the two temporal processes at each grid point (location), N denotes the length of the two-temporal process $(1, ..., N = 264)$, $\tilde{Y_t}{}^f$ and $\tilde{Y_t}{}^f$, denotes one-step-ahead prediction, and ν_t represents squared prediction error. This measure d produces the climate index for each of the regional climate models. The will be used as a bases for determining the spatial pattern. The distance metric must satisfy all properties of distances as described in [36]. At each time t, the distribution of the Gaussian process Y_t^f,

is approximately standard normal, and the hypothesis is given as;

H_0 : The Gaussian processes $\{Y_t^c\}$ and $\{Y_t^f\}$ are realization from the same process.

H_1 : The Gaussian processes $\{Y_t^c\}$ and $\{Y_t^f\}$ are not realization from the same process.

Under the null hypothesis, the two processes have the same dynamics and serves as a reference for one another. Also, the distance metric d is approximately distributed as χ^2_N/N, a chi square random variable with N degrees of freedom divided by N. When the null hypothesis is rejected we conclude there is statistical evidence that the two processes serve as a reference point for each other. From equation 1.4, $\hat{X}_t = E(X_t \mid 1, X_1, ..., \hat{X}_{t-1})$ is the best linear prediction of X_t from a constant and past observations $X_1, X_2, ..., X_{t-1}$. The squared prediction error, $E[(X_t - \hat{X}_t)^2] = \nu_t$ and $S_t = (X_t - \hat{X}_t)/\sqrt{\nu_t}$ is the scaled error for each t has zero white noise and a unit variance. This implies that when $t \neq s$, S_t and S_s are uncorrelated. $\hat{X}$ plays a critical role in calculating the exact distance between two temporal processes. It is based on finding the best linear predictor of $\hat{X}_t$ given past observations $X_1, X_2, ..., X_{t-1}$. See appendix 1.6.1 for how to deal with the computation of one-step ahead prediction $\hat{X}_t$ and proof of this proposition.

Part II We determine how different those climate data generated from the RCMs are in terms of their underlying process. This is achieved by performing a spatial two sample pairwise KS test on the temporal distance metric values (climate index). A spatial two sample Kolmogorov-Smirnov test determines whether two random fields observed in a spatial domain (Kansas) serves as a realization from the same process . In other words, to check the model consistency , the test is formulated by assuming a process $Y(x), x = (s, t) \in D \subset R^2 \times R$ to be the spatio-temporal random field where t is time and s is the spatial location. Let $Y_1(x)$ and $Y_2(x)$ be a random field defined in a spatial domain D and

$$Y_i(x) = \mu_i(x) + \epsilon_i(x), \tag{1.5}$$

where $\mu(x)$ represents the spatial mean of $Y(x)$ and $\epsilon(x)$; is a spatially correlated error with

zero mean and covariance matrix $C(.,.)$ and $\epsilon(x)$.
Let $\mu_i(.) = \Big(\mu_i(x_1), ..., \mu_i(x_n)\Big)_{i=1,2}$, $\quad Y_i = \Big(Y_i(x_1), ..., Y_i(x_n)\Big)_{i=1,2}$ and $C_i = \Big(c(x_i, x_j)\Big)_{1\leq i,j\leq n}$.
It's easier to see that the same traits of $Y_1(x)$ hold for $Y_2(x)$.
The objective is to assess whether the disparities between $Y_1(x)$ and $Y_2(x)$ are due to randomness of the particular observations and thus to assess whether $Y_1(x)$ and $Y_2(x)$ measure/share a common mean and correlation structure. The hypothesis is given as;

$$H_0 : \mu_1(x) = \mu_2(x) \text{ and } c_1(s,t) = c_2(s,t)$$

$$H_1 : \text{ Otherwise,}$$

for $x = (s,t) \in D \subset R^2 \times R$ and any h such that $x + h \in D \subset R^2 \times R$. Under the null hypothesis, for any observed processes $Y_1(x)$ and $Y_2(x)$ at identical sets of location, $\mu_1 = \mu_2$ and $c_1 = c_2$ and $y_1 = C^{-1/2}(y_1 - \mu)$ and $y_2 = C^{-1/2}(y_2 - \mu)$ yields two uncorrelated samples, noting μ and C are the pooled estimates. Both samples follow the central t_v distribution if P is n-variate t_v distribution with a degree of freedom ($v > 2$) and a normal distribution $N(0,1)$ if P is a n-variate normal. Under the null hypothesis, the spatial two sample KS test is obtained by accounting for the equality in y_1 and y_2. The mean μ and covariance C part of the random fields are usually unknown. The trend part of the field which is usually the parametric form of μ is used to estimate μ while C is obtained by fitting a parametric spatial model, such as spherical or an exponential or Matérn . Once the estimates of μ and C are obtained by maximum likelihood method or least square method, the estimation of $\tilde{y}_1$ and $\tilde{y}_2$ follows trivially. The KS test is then employed to $\tilde{y}_1$ and $\tilde{y}_2$ to determine whether they are realization from the same process.

1.4.2 The Effect of Climate Change on Nitrous Oxide Emission

A linear mixed effect model is fitted to determine the effect of climate change on N_2O emissions. The response is nitrous oxide emission, and the regressors are the set of climate indices for each weather variable; thus, maximum and minimum temperature and precipitation. Af-

ter considering a series of models, we chose the parsimonious model given as;

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + +\beta_3 x_{i1} x_{i2} + c_i + \epsilon_i,$$

where;

- μ is the overall mean.
- y_i is the nitrous oxide emission observed at i^{th} location in the state of Kansas.
- x_{i1} is the climate index for precipitation at the at i^{th} location in the state of Kansas.
- x_{i2} is the climate index for maximum temperature at the i^{th} location in the state of Kansas.
- $x_{i1}x_{i2}$ is the climate index for interaction between maximum temperature and precipitation at the i^{th} location in the state of Kansas.
- c_i are the random-effects regressor (group specific runs) and ϵ_i are spatially correlated error.

The following provides step by step directions on how to detect spatial pattern in a regional climate model and to determine its effect on nitrous oxide emission of state crop in Kansas.

1. Perform an exploratory data analysis.
2. At each grid point that covers Kansas, perform the Process-based Geographical Algorithm Machine (PGAM) procedure.
3. Perform the Spatial Kolgomorov sample test and fit a linear mixed effect model.

1.5 Data Analysis Results

According to step 1, we deliver details of the results for detecting the spatial pattern in Regional Climate Models (RCMs) and the effect of weather change on N_2O emissions on state crops. Figure 1.1 shows the time series plot for a randomly selected location in Kansas. This was used to initially examine patterns in weather variables (average precipitation, minimum and maximum temperature) across a period of 22 years. The time series plot in red and green represents maximum and minimum temperature, respectively, while the plot in blue represents precipitation. A regular repetition of patterns in temperature was observed as was a random fluctuation in precipitation, and so a periodic auto-regression model of order 1 was fitted for each weather variable.

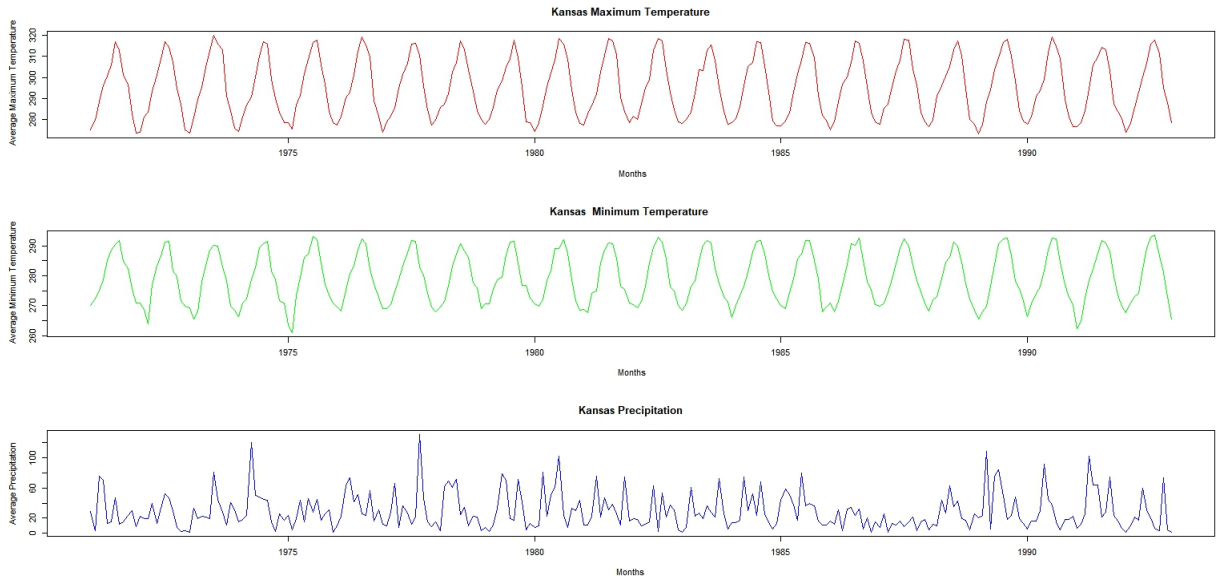

Figure 1.1: Weather data in kansas for a random year, red is Maximum Temperature, green is Minimum Temperature, and blue is Precipitation. Temperature is measured in (Kelvin) K and precipitation is measured $\text{kgm}^{-2}s^{-1}$.

To detect the spatial pattern in each regional climate model (RCMs), we used the proposed Process-based Geographical Algorithm Machine procedure. Following step 2, at a threshold of 0.05, the distance map was initially drawn for each weather variable in all RCMs to ascertain the spatial pattern in each of the models. Regions in red, yellow and white indicate areas where weather change was mostly, moderately and least effective. Fig-

ure 1.2 shows the distance map via kernel smoothing for precipitation for all climate models. According to model validation techniques, for example, comparing WRFGcc and WRFcg models, the northeastern part exhibited severe weather change. For the WRFGcc model, the southwestern part showed extreme weather while in the WRFGcg model, the northwestern part of Kansas was severe. Comparing the CRCMcg and RCM3gdfl models, we observed that the southeastern part of Kansas was mostly affected. In the results of the RCM3gdfl and CRCMcg models, weather change was moderately impacted in the central part of Kansas, but with the RCM3gdfl model, we noticed certain parts of the west and central were least affected by weather change. In RCM3cgcm and CRCMcc models, we observed that the southwestern part of Kansas was mostly affected.

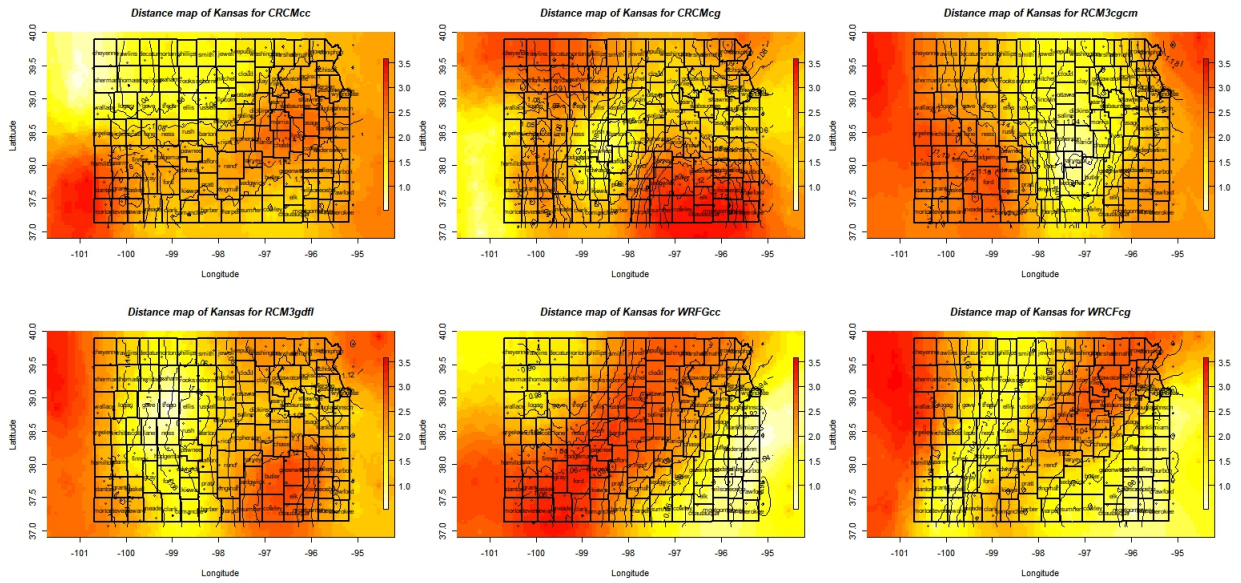

Figure 1.2: Initial Distance Map for Kansas precipitation

Based on our analysis, Figure 1.3 shows the distance map via kernel smoothing for maximum temperature. At a significance level of 0.05, we observed that the spatial pattern detected in all RCMs had some similarities and some differences. For example, WRCFcg and RCM3cgcm had similar traits; thus, the northeastern part of Kansas showed mild weather change while the southwestern part of Kansas showed extreme weather change. WRFGcc had identical characteristics as described above except that weather change exhibited mild weather change in the southeastern part of Kansas rather than the northeastern region. The

rest of the RCMs showed no similar places where the impact was the same.

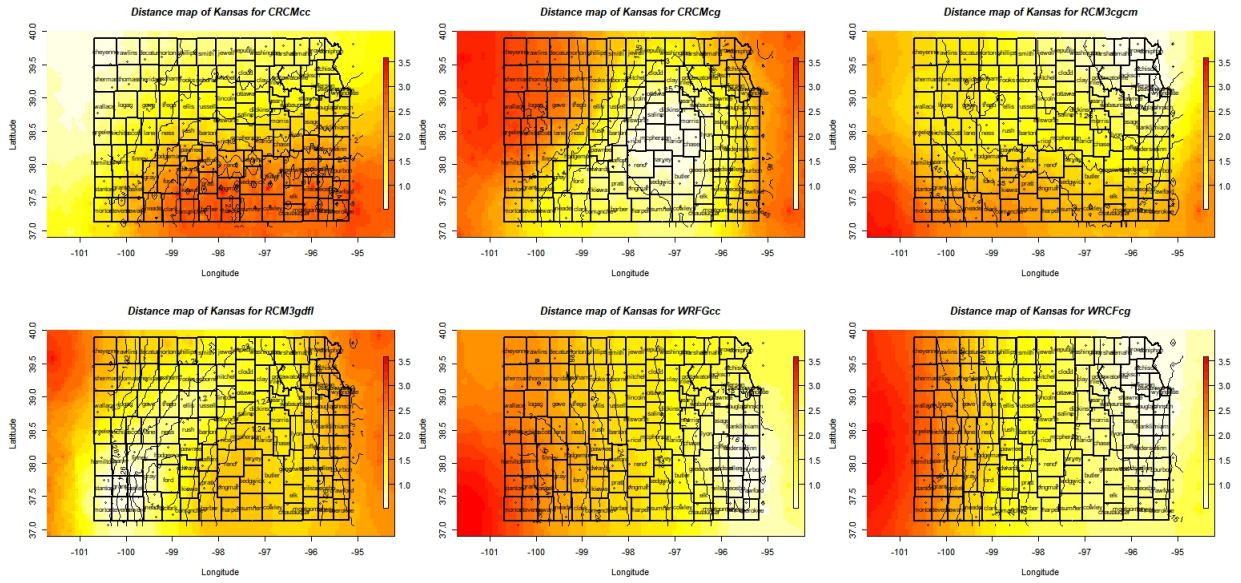

Figure 1.3: Initial Distance Map for Kansas maximum temperature

Figure 1.4 shows the distance map via kernel smoothing for minimum temperature. At a threshold of 0.05, we noticed that spatial pattern for some RCMs was somewhat similar and different at the same time across the state of Kansas. For example, WRFGcc and WRFGcg models showed similar traits; for instance, we observed that the entire region of Kansas was least affected by weather change except that the latter part of southwestern was mostly affected. For CRCMcc and RCM3cgcm, we observed that the northeastern part of Kansas was least affected by weather change. Considering just the CRCMcc model, more areas in some parts of the northern, central, and southern regions were least affected. For the RCM3cgcm model, weather change was moderately affected in some parts of the central and north of Kansas. The RCM3gdfl model showed the northern and northwestern parts of Kansas as having been affected the most.

Due to differences in RCMs, a two spatial Kolmogorov sample test was performed to check for model consistency. For precipitation there was enough evidence to support that possibly, most RCMs were a realization from the same process based on the p-values (p-value > 0.05). It was observed that, (CRCMcg and WRCFcc), (CRCMcc and RCM3gdfl), (RCM3cgcm and WRCFcc) models were significantly different from each other when consid-

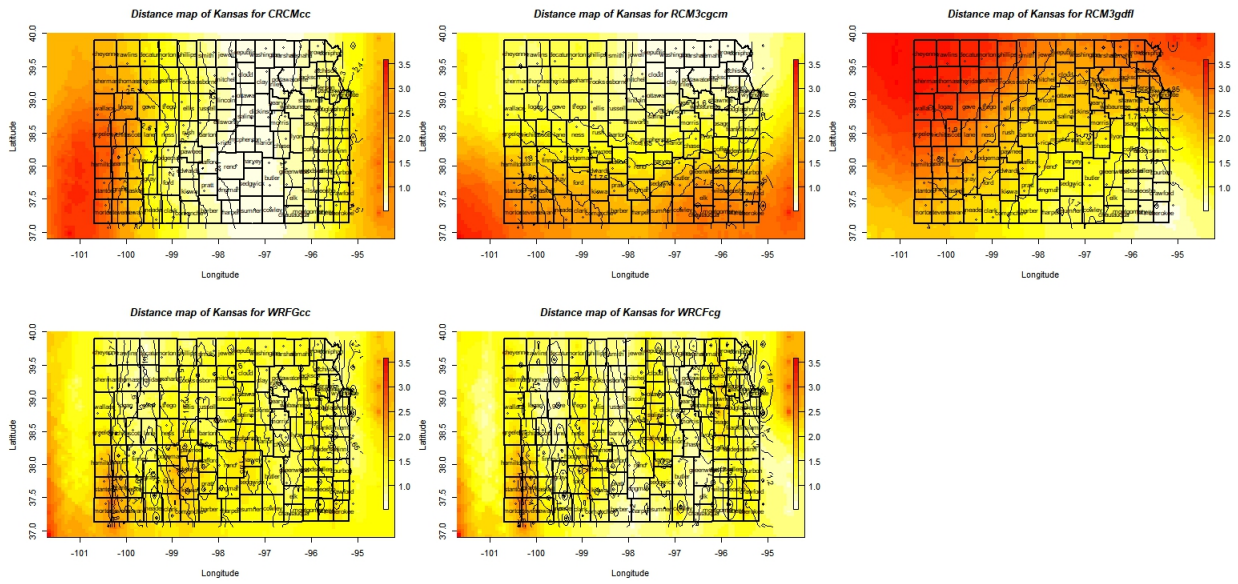

Figure 1.4: Initial Distance Map for minimum temperature

ering maximum temperature. In the case of minimum temperature, models (CRCMcg and RCM3gdfl), (CRCMcg and WRCFcc) were significantly different.

To draw the final distance map for the RCMs, for each weather variable, we used the average of climate index of RCM3cgcm and RCM3gdfl as a basis for our analysis. The results are in Figures 1.5 - 1.7. These distance maps exhibit a spatial pattern of high and low regions where climate change was affected. Figure 1.5 shows the distance map for maximum temperature, where lower weather change occurred in the northern and central part of Kansas. Thus, counties such as Washington, Marshall, Pottawatomie, Riley, and Nemaha were least affected. Greater weather change occurred in the southwestern part of Kansas, especially in Morton and Stanton counties, while the central part was moderately affected. For minimum temperature, we had similar results except that more counties experienced higher weather change than with maximum temperature. Also, greater weather change occurred in the southwestern part of Kansas specifically in Morton, Stanton, Grant, Hamilton, and Stevens counties. Less weather change happened in the north and east part of Kansas as illustrated in Figure 1.6. Counties such as Shawnee, Saline, Osage, Dickinson, Franklin, Douglas, Cloud, and Clay were least affected by weather change. Some parts of north, central, and southern Kansas were moderately affected, while the southwestern part of Kansas

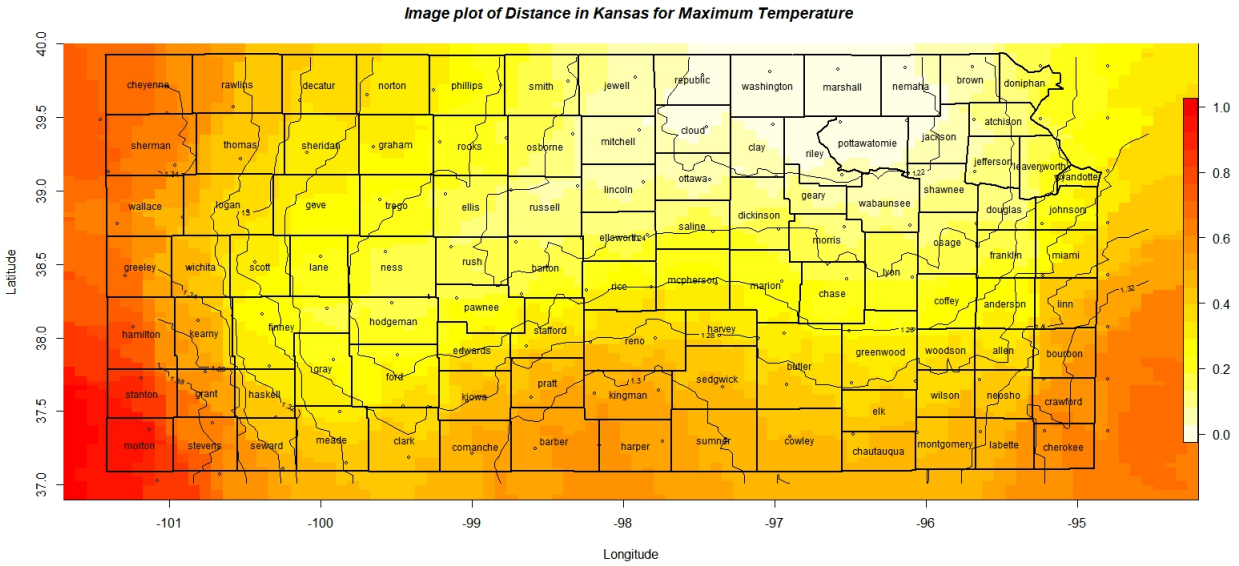

Figure 1.5: Distance map for maximum temperature by kernel estimation

was mostly affected by weather change. This includes counties such as Morton, Stanton, Hamilton, Stevens, and Grant, mainly.

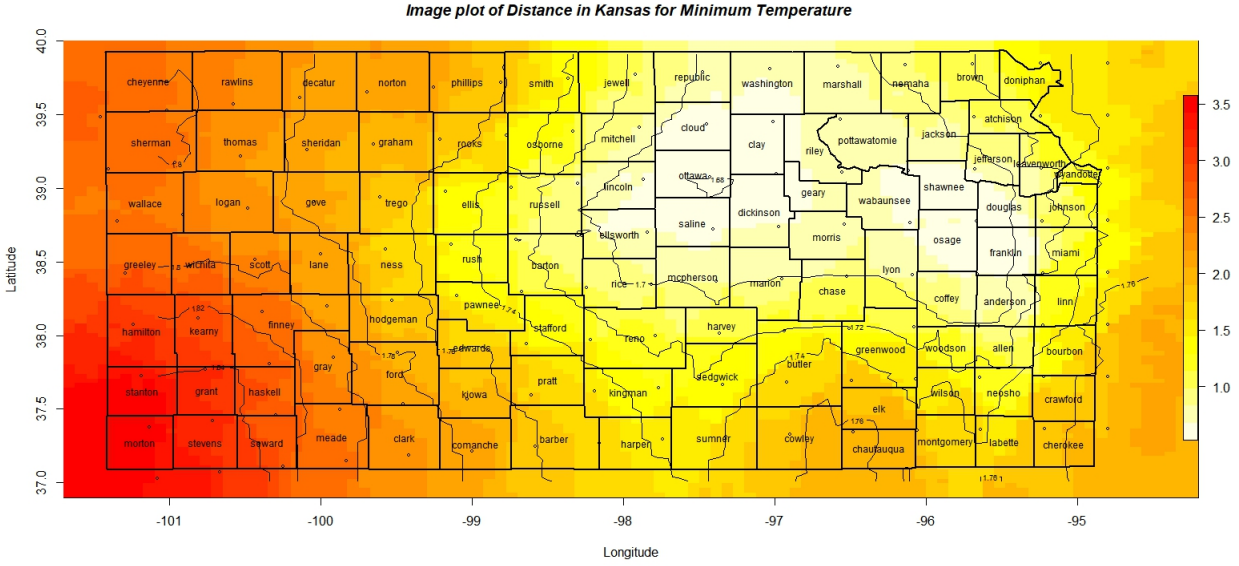

Figure 1.6: Distance map for minimum temperature by kernel estimation

Figure 1.7 shows the distance map for precipitation. We observed that the central part of Kansas experienced the least weather change while the northwestern part of Kansas was mostly affected by weather change. Counties such as Harvey, Saline, Lincoln, Mcpherson,

and Morton were least affected, while Cheyenne, Sherman, Wallace, and Greeley were mostly affected. Some parts of the eastern, northern, western, and southern Kansas were moderately affected by weather change.

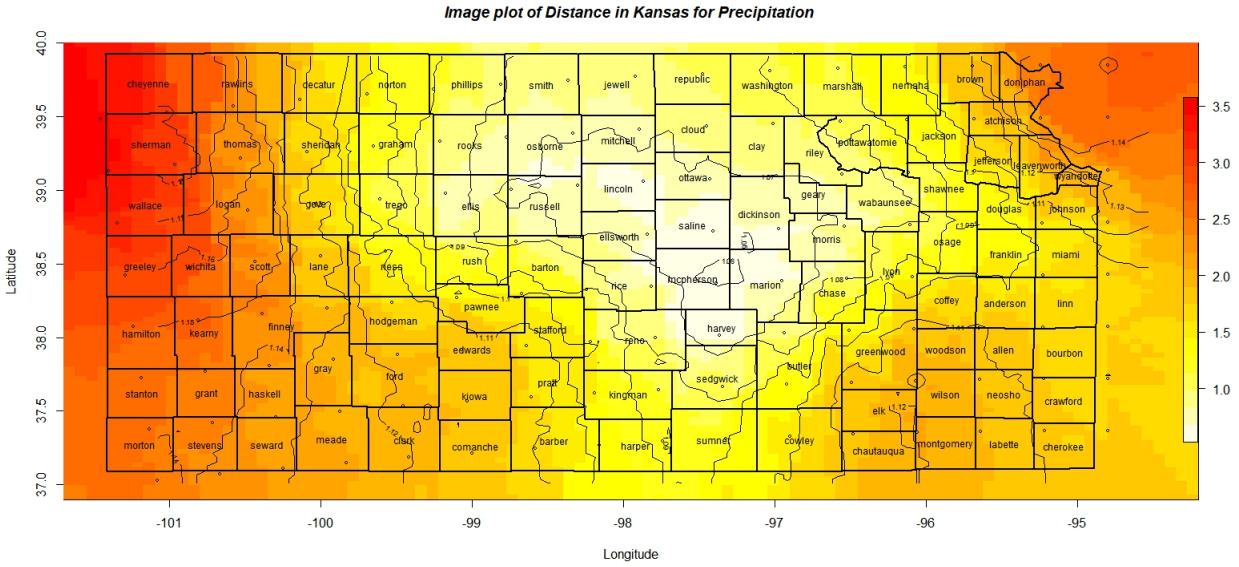

Figure 1.7: Distance map for precipitation by kernel estimation

Statistical analysis indicates that, in general, for maximum temperature, there was an increase in temperature from April through to October and a decrease in temperature from February through to March. This trend shows the effect maximum temperature played on weather change from April through to October and from February through to March. Greater weather change occurred from September through to November in the southwestern part of Kansas, specifically in Morton and Stanton counties, while least climate change occurred mostly in the northern part of Kansas from February through to March, specifically in Marshall and Washington counties. For minimum temperature, weather change decreased from January through to August while it increased from September through to December. Significantly greater weather change occurred in the southwestern part of Kansas, specifically in Morton, Grant, Stevens and Stanton counties from February through to April while least weather change occurred in the north and east part of Kansas, mostly in Ottawa, Clay and Shawnee counties. Lastly, for precipitation, climate change increased from January through to June and decreased from August through to December. Significantly greater weather

change occurred from May through to July in the northwestern part of Kansas, specifically in Cheyenne and Sherman counties, while least climate change occurred in the central part of Kansas from September through to December, specifically in Saline and Mcpherson counties.

Based on different climate change areas, a linear and nonlinear mixed effect model was fitted to assess its effect on N_2O emission with respect to the grouping variables. After comparing a pool of candidate models for all weather variables based on their AIC and BIC, we chose the most parsimonious linear mixed effects model. The results of the relationship between climate change and N_2O emission with respect to the main effects (maximum temperature and precipitation) and interaction term (maximum temperature $\times$ precipitation) are presented in Table 1.1.

There was a negative relationship between the joint effect of precipitation and maximum temperature on nitrous oxide emission. This was not surprising as a similar relationship was observed in the literature and is partially due to the application of the denitrification and to decomposition over the entire state of Kansas [13]. Thus, at higher levels of precipitation change, the effect of temperature change on nitrous oxide emission is low. The interaction term was a significant predictor in predicting nitrous oxide emission as (p-value = 0). Given that the interaction term has been accounted for, the impact of the individual weather variables on nitrous oxide emission was significant and positively related. In short, we conclude that the effect of temperature change on nitrous oxide emission is low at different levels of precipitation.

Table 1.1: Linear mixed effect model: Fixed effects estimates for maximum temperature, precipitation and its interaction

Fixed Effect	Est	Std. Err	Df	t - test	P-value	Rand Eff	Std. Err
Intercept	-15.3320	2.4016	31005	-6.3840	0	Group	0.4642
Max temp	19.4743	1.8520	31005	10.5151	0	Resid	0.8175
Precipitation	16.7239	2.1389	31005	7.8190	0	.	.
Max temp $\times$ Precipitation	-19.2955	1.6479	31005	-11.7094	0	.	.

1.6 Conclusion and Discussions

The primary motivation of the current study has been to contribute to the existing techniques used by climatologists and applied statisticians to detect spatial pattern in regional climate models and its effect on nitrous oxide emission. We showed that the Process-based Geographical Machine, motivated from GAM, can be used to detect the spatial pattern in regional climate models, by performing local scale process difference testing with spatial extension to the whole region. In doing so, the seasonal mean and autocovariance in the process was accounted for. We found that most RCMs were consistent with each other with respect to the weather variables.

We also presented the results of the spatial pattern using our proposed PGAM discussed in Section 1.4.1. The results indicate a different spatial pattern detected in different regions across the state of Kansas. At a significance level of 0.05, for maximum temperature, Morton and Stanton counties exhibit severe weather change, while Marshall and Washington counties were mild. For minimum temperature, Grant county was affected most while Ottawa, Clay, and Shawnee counties were least affected. Lastly, for precipitation, Cheyenne and Sherman counties were mostly affected by weather change, while Saline and Mcpherson counties were least affected.

There was a negative relationship between nitrous oxide emissions and the interaction term between precipitation and maximum temperature. Although nitrous oxide emission will likely increase due to changes in precipitation and increasing temperature, our studies showed that at each level of precipitation change, the effect of temperature change on the emission was low. This observed relationship was not surprising as De Vries in his paper [13] observed a negative relationship between the joint effect of precipitation and temperature on nitrous oxide emission in European forests. Our relationship was partially due to the application of denitrification and to decomposition (DNDC) over the entire state of Kansas and the correlations between the covariates [13, 35].

It is important to note that, for maximum temperature, areas with greater significant climate change, showed increased nitrous oxide. Also, areas where climate change was ex-

perienced least showed reduced nitrous oxide emission. In general, for precipitation, it is important to note that for areas where climate change was mostly affect, greater climate change led to an increase in nitrous oxide emission, and less climate change led to a decrease in nitrous oxide emission. Lastly, for minimum temperature, areas where weather change was experienced least showed a greater change in climate where nitrous oxide emission was experienced least.

Throughout this study, the main focus has been to examine the impact of weather change on nitrous oxide emissions and to detect spatial pattern. We found that performing a localized process difference test rather than a globalized process difference test at each grid point helps improve results for the spatial patterns detected. Also, we note that it is advisable to fit the same model for both temporal processes at each grid point. In our study, we used the periodic autoregressive model of order PAR(1), but if it is not adequate, then a higher order should be considered. For dense data such that the size of the sampled locations is large enough to allow reasonable statistical test for a given circle, we recommend using the two sample spatial Kolmogorov test to determine whether the two processes have the same dynamics or not. Next, we showed that the proposed based Geographical Algorithm Machine detects areas of interest across the study region. The spatial patterns in addition to the climate index obtained enabled us to conclude on regions where weather change was significantly pronounced. The spatial linear mixed model was used to determine the long-range effect of weather change on nitrous oxide emission.

For future work, one could improve on the model by taking care of the edge effect. One could also focus on areas where weather change was severe and mild and investigate reasons as to why there was different observations. For the effect of weather change on nitrous oxide emission, a sensitivity analysis could be done to determine the impact of each variable on nitrous oxide emission. One could also use a different linear model (for example, the spatial partial linear model) to assess the dynamic effect of the weather change on nitrous oxide emission based on the availability of data.

1.6.1 Chapter 1 Appendix

Table 1.2: Model Simulated in NARCCAP used on N_2O Regional Simulations

		AOGCM		
		CCSM	CGCM3	GFDL
RCM	CRCM	×	×	
	RCM3		×	×
	WRFG	×	×	

The six RCMs used are as follows; CRCMCCSM , CRCCMCGCM3, RCM3CGCM3, RCM3GFDL, WRFGCCSM and WRFGCGCM3. In the study, we adopted the short form of the abover as CRCMcc, CRCMcg RCM3cgcm, RCM3gfdl, WRFGcc and WRFGcg, respectively.

1.6.2 Prediction Error and One-Step-Ahead Prediction

Prediction error and one-step ahead prediction is an important tool that plays a significant role in the computation of $d(\mathbf{X}, \mathbf{Y})$. To proceed, consider m-steps ahead prediction and its associated error, thus $(X^n_{n+m} - X_{n+m})$. This is orthogonal to the prediction variables $1, X_1, ..., X_n$ and $X_0 = 0$. A one-step-ahead linear prediction is obtained when $m = 1$, and it's written as

$$X^n_{n+1} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \phi_{n3}X_{n-2} +, ..., +\phi_{nn}X_1, \tag{1.6}$$

which satisfies the prediction equations $E(X_{n+1} - X^n_{n+1}) = 0$ and $E[X_i(X_{n+1} - X^n_{n+1})] = 0, \text{for } i = 1, 2, ..., n$. Also $E(X_{n+1}X_i) = \sum_{j=1}^{n} \phi_{nj}E(X_{n+1-j}X_i)$ and $\sum_{j=1}^{n} \phi_{nj}\gamma(i-j) = \gamma(i)$ with $\Gamma_n\phi_n = \gamma_n$,

where $\Gamma_n = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{pmatrix}$, $\phi_n = \begin{pmatrix} \phi_{n1} \\ \phi_{n2} \\ \phi_{n3} \\ \vdots \\ \phi_n \end{pmatrix}$, $\gamma_n = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(3) \\ \vdots \\ \gamma(n) \end{pmatrix}$.

The mean square error associated with this one-step-ahead linear prediction is given as

$$\begin{aligned} P_{n+1}^n &= E(X_{n+1} - X_{n+1}^n)^2 \\ &= E[(X_{n+1} - X_{n+1}^n)(X_{n+1} - X_{n+1}^n)] \\ &= E(X_{n+1}(X_{n+1-X_{n+1}^n})) \\ &= \gamma(0) - E(\phi' X X_{n+1}) \\ &= \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n \\ &= \mathrm{Var}(X_{n+1}) - \mathrm{Cov}(X_{n+1}, X)\mathrm{Cov}X, X)^{-1}\mathrm{Cov}(X, X_{n+1}) \\ &= E(X_{n+1} - E(X_{n+1}))^2 - \mathrm{Cov}(X_{n+1}, X)\mathrm{Cov}(X, X)^{-1}\mathrm{Cov}(X, X_{n+1}), \end{aligned}$$

where $X = (X_n, X_{n-1}, X_{n-2}, ..., X_1)$. In general, the procedure for m-step ahead linear predicition is similar to the one-step ahead linear prediction. The formulation of m-step ahead linear prediction is given as

$$X_{n+m}^n = \phi_{n1}^{(m)} X_n + \phi_{n2}^{(m)} X_{n-1} + \phi_{n3}^{(m)} X_{n-2} +, ..., + \phi_{nn}^{(m)} X_1,$$

$$\Gamma_n \phi_n^{(m)} = \gamma_n^{(m)},$$

noting that $\mathrm{Cov}(X_{n+m}, X) = \phi_n^{(m)} = (\gamma(m), \gamma(m+1), ..., \gamma(m+n-1))$. For example, for an autoregressive of order AR(P), given as:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + a_t \quad \rightarrow \quad AR(P),$$

and one-step-ahead prediction can be calculated as

$$\begin{aligned} X_{n+1}^{n} &= P(X_{n+1}|X_1, \cdots, X_n) \\ &= P(\sum_{i=1}^{p} \phi_i X_{n+1-i} + a_{n+1}|X_1, \cdots, X_n) \\ &= \sum_{i=1}^{p} P(X_{n+1-i}|X_1, \cdots, X_n) \\ &= \sum_{i=1}^{p} \phi_i X_{n+1-i} \quad \text{for} \quad n \geq p. \end{aligned}$$

The Proposition 1.4.1 can be proved by noting that $Y_t^c - \tilde{Y}_t^c$ and $Y_t^f - \tilde{Y}_t^f$ are independently identically distributed normal random variables if Y^c and Y^f have the same process structures. One can also use the Durbin-Levinson Algorithm and the Innovations Algorithm to calculate the one-step-ahead prediction.

Chapter 2

Spatial Impact of Weather Change on Nitrous Emission with Large Data Approximation Analysis

By the beginning of 1993, the concentration of nitrous oxide (N_2O) emission in the atmosphere had increased from 257 *ppbv* in the pre-industrial era to about 311 *ppbv*. This is primarily credited to increasing anthropogenic emission via production and use of nitrogen fertilizers, increased biomass burning, and tropical land conversion from forest to agriculture amongst other reasons [48]. Increase in concentration of nitrous oxide emissions in the atmosphere is a contributing factor to global warming/climate change in part because the effects of climate change interact with precipitation patterns and magnitude so as to impact nitrogen dynamics, which in turn affects nitrous oxide emissions. These emissions have a higher capacity for absorbing radiation, and the net gain in energy causes warming that affects the earth's climate. Agriculture is one of the main areas that accounts for about 10 to 12% of global anthropogenic emissions of greenhouse gases of which approximately 60 to 80% are related to N_2O emission. The behavioral patterns of weather can be used to develop a model that can explain the effect of weather on nitrous oxide emissions. Accordingly, the data used for this study were obtained from 2022 locations across the state and form part of

the data used by [2]. Figure 2.1 shows 2022 locations presented in blue from which nitrous oxide emissions, precipitation, and maximum and minimum temperature were measured or obtained. Notice that these locations were fairly distributed across the state of Kansas.

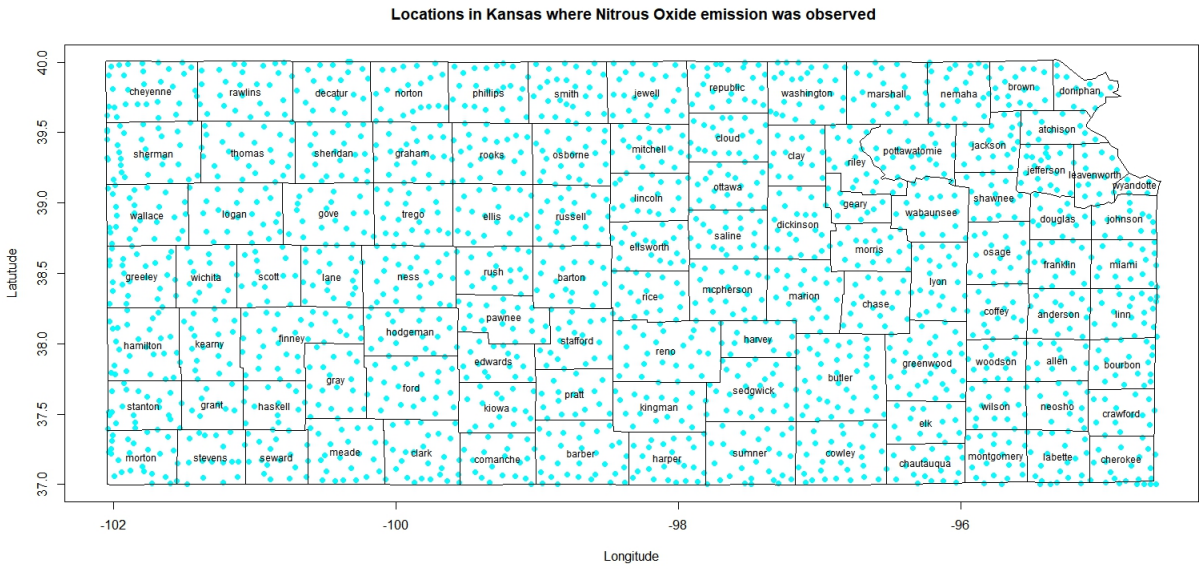

Figure 2.1: Locations where information was obtained in Kansas

Figure 2.2 - 2.3 illustrates the plot of daily maximum and minimum temperatures and precipitation at randomly chosen location in the state of Kansas. Temperature was measured in celsius and precipitation in millimeters. We observed a random fluctuation in weather across the year.

It is important to know that N_2O emission and weather data (precipitation, minimum temperature, and maximum temperature) were observed at all locations in the study region. According to Argoti, an increase and variation in temperature and precipitation, respectively, occurred across the climatic zone in the state of Kansas [2]. These changes contributed to a difference in nitrous oxide emissions. Due to the nature of the data, we considered a spatial linear model with a large data approximation technique. Often, a modified regression analysis procedure works where predictors and scalar response are linked at a location. Numerous methods have been suggested for dealing with data of this kind. Consequently, due to the nature of the data, one of the goals of this paper is to evaluate the effect of actual weather measurements on nitrous oxide emission accounting for spatial components. The changes

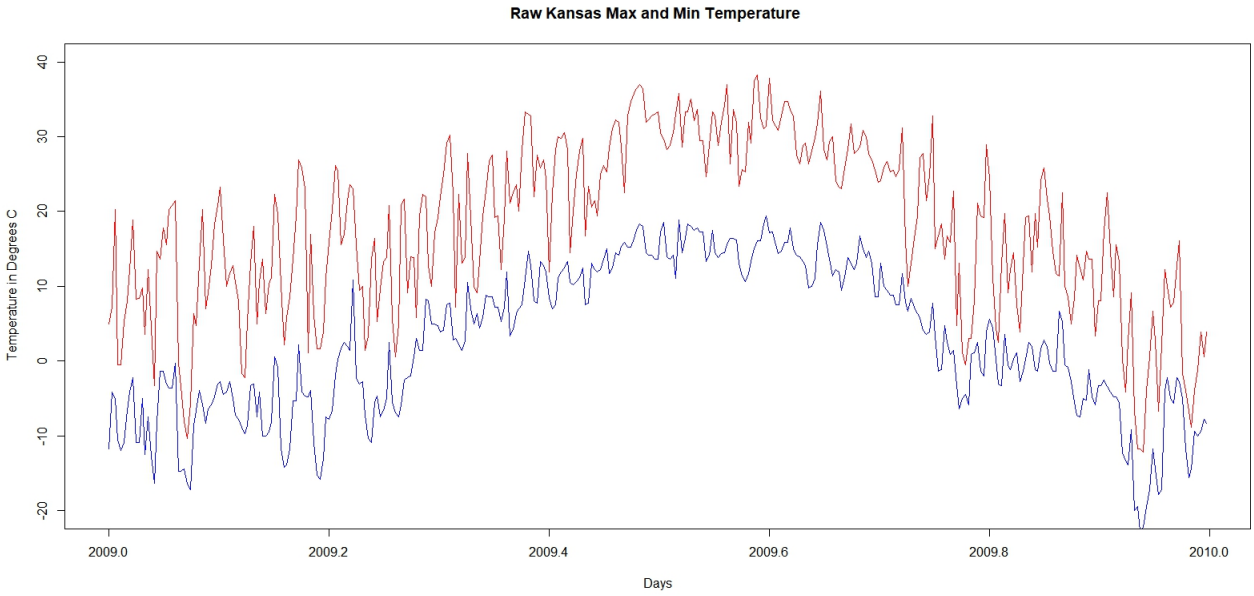

Figure 2.2: Random location where maximum and minimum temperature was measured: red and blue color shows maximum and minimum temperature respectively

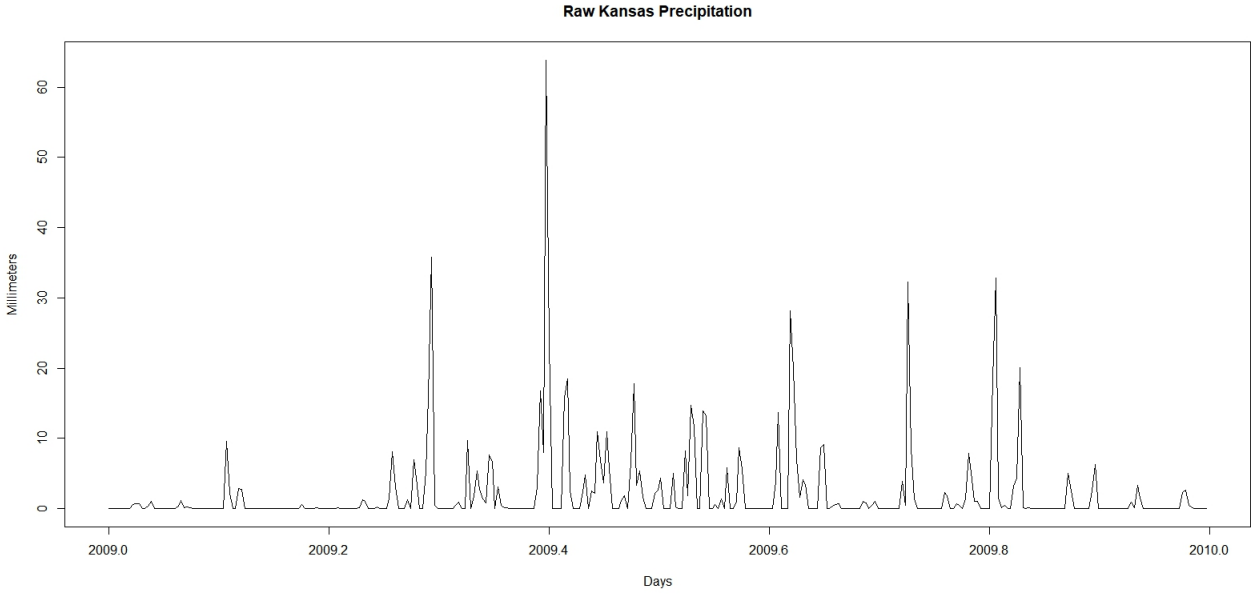

Figure 2.3: Random location where precipitation was measured

in nitrous oxide emission was highly variable and it is interesting to know how they are associated with changes in climate and other agricultural management practices [2].

In this paper, another main focus is on analyzing large spatial data. Our goal is to study statistical properties in large weather data and determine their effect on nitrous oxide

emission accounting for spatial dependence in the residual term. In doing so, we theoretically investigate techniques to deal with the issue of inversion of the covariance matrix of large data. Given its breadth and utility in research, large data is useful in a variety of research areas such as computer science, statistics, climatology, hydrology, agronomy, and government amongst others. Additionally, large data serves as a source of productivity, competition, and creativity [8]. However, analyzing large data sets of size n statistically comes with its own challenges, one being the inversion of the $n \times n$ covariance matrix, which requires $O(n^3)$ operations and is computationally complex and costly. Another problematic task often encountered is the loss of information, which tends to affect the accuracy of results, such as estimation and prediction. Nevertheless, the computation of the inverse of the covariance matrix is the centerpiece for determining the Best Linear Unbiased Prediction (BLUP) and maximum likelihood estimation. Kriging (BLUP) as well as other likelihood methods of estimation are special spatial tools that have become well liked in the world of statistics, geography, geology, environmental science, amongst others. In spatial statistics, Kriging involves the inversion of the covariance matrix of order $n \times n$, which sometimes is not easily solvable especially when dealing with a large dataset. With this computational difficulty at hand, namely the inversion of the covariance matrix and the need to estimate parameters using the likelihood function followed by prediction, one of the main objectives of this chapter is to find a way of reducing computation intensity by approximation through a projection approach.

There are abundant techniques in literature dedicated to computation reduction dealing with large spatial data sets [58, 19, 15, 4, 55]. We mainly focus on two main procedures; the first procedure employs the fixed/reduced rank approximation, and the second utilizes the sparse approximation technique. The first procedure relies on traits of the reduced rank approximation of the underlying process. Fixed reduced rank approximation has been successful in accounting for large scale structure of spatial process but has neglected means to account for small scale structure in the data accurately [55]. Examples of methods under the first procedure are the predictive process, low splines, basis functions, and kernel convolution amongst others [5, 28, 25, 58]. The second procedure searches a sparse approximation matrix

to the covariance function and uses the sparse matrix techniques to obtain computational efficiency. This procedure has been proven useful in capturing the short-range dependence structure of a spatial process but fails to account for long-scale structure in a spatial process [15]. In addition to sparse approximation, covariance tapering falls in line with the sparse approximation method and has recently been used to develop the sparse covariance matrices. Notably, a good approximation to the original covariance function is not readily obtained when a tapered covariance function with a smaller taper range is used because this promotes a huge bias in spatial prediction and parameter estimation. Therefore, several procedures and algorithms have been developed for the method to handle spatial prediction and parameter estimation [19]. For example, to deal with the shortfalls of the reduced rank and sparse approximation, Sang and Huang (2012) proposed the full-scale approximation [51]. Her procedure minimizes computation inefficiencies in large data sets thereby preventing the drawbacks in reduced rank and sparse approximation methods. The full-scale approximation procedure combines ideas in reduced rank and sparse approximation to successfully obtain a good approximation to the original covariance matrix [51]. Next, Banerjee (2012) proposed the linear projection method that approximates the original covariance matrix [4]. He does this by linearly projecting all data points onto a lower-dimensional subspace thereby yielding a near-optimal rank performance at a cheaper computational cost [4]. Thus, the linear projection and the reduced rank approximation form a centerpiece for our studies.

In these studies, one of the main goals is to find the near-optimal rank for which we can achieve a desired accuracy. This method contributes to the usual norm of knot-based methods, which is easier to implement and has theoretical justification. In the past, emphasis has been on randomly selecting different target ranks for the approximation covariance matrix without any theoretical justification. Now, we propose an optimal rank approximation coupled with a projection method via reduced rank approximation to mitigate the computational burden. The mitigation is shown to be proper by approximating the covariance matrix with a theoretical investigation of optimal conditions under which we can achieve good accuracy in terms of Kullback-Leibler divergence and mean squared prediction error.

The paper is divided into 2 parts: the first part deals with the theoretical techniques

used for large datasets while the second part deals with the application of the spatial linear model with large data approximation using the nitrous oxide emission and weather data. The remainder of the paper is structured as follows: Section 2.1 discusses some previous techniques on covariance approximations; Section 2.2 gives the proposed method; Section 2.3 presents a simulation study for the approximation under different goals; Section 2.4 illustrates our method via real data; Section 2.5 presents conclusion and discussions based on the simulation studies and real data analysis and Section 2.6 presents appendix, proofs and results.

2.1 Preliminary Results on Covariance Approximation

There are several techniques to computationally and effectively approximate a large covariance matrix of size n. In this section, we will give some brief underlying methods from the literature for dealing with the problems encountered when inverting the covariance matrix of large data. First, Cressie and Johannesson (2008) considered a fixed rank approach for large datasets. They considered a flexible family of non-stationary covariance functions that were established by using a set of basis functions that was fixed in number and led to a spatial prediction method called fixed rank kriging. Fixed rank kriging is a spatial prediction or kriging within a group of non-stationary covariance function. They proposed techniques based on reducing the weighted Frobenius norm to produce the best estimators of the covariance function parameters [12]. These are then substituted into the fixed rank equations (see section 2.3 - 2.4 for the formula for the fixed rank equation). Next, Banerjee (2008) suggested instead, a predictive process model for spatial and spatio-temporal data. He suggested every spatial process generates a predictive process model that projects process realizations of a spatial process to lower dimensional subspace, which tends to reduce computational burden [5]. Furthermore, researchers such as [19, 29] proposed a technique to approximate the covariance function by a compactly supported function using tapering or simply introducing zeros into the covariance matrix to make it sparse. This makes the best linear unbiased prediction at an unobserved location with a sparse covariance matrix

easily solvable using the sparse matrix algorithms. Du and Zhang (2009) also studied how covariance tapering can be used to overcome the numerical challenges when the sample size is extremely large [14]. They investigated how tapering affects the asymptotic efficiency of the maximum likelihood estimator for the microergodic parameter in the Matérn covariance function by establishing a fixed-domain asymptotic distribution of the exact estimator and that of the tapered estimator. Meanwhile, Sang and Huang (2012) proposed full-scale covariance approximation combining reduced rank covariance approximation and sparse covariance approximation [51]. They suggested that the full-scale covariance technique offers a high propensity for approximating the covariance matrix at small and large spatial settings. It also has the tendency to allow efficient computation of the maximum likelihood, spatial prediction, and Bayesian inference. Lastly, Banerjee proposed a means of dealing with the inversion of the large covariance matrix by linearly projecting all the data points from a higher dimensional space to a lower dimensional subspace [4]. He used the idea of the projection approximation method and reduced-rank matrix approximation to illustrate the superiority of his method via a theoretical perspective.

These methods represent various approaches to approximate the covariance matrix when extremely large sample size is considered. In this study, we contribute to the collection of techniques addressing the inversion of a large covariance matrix, thereby reducing the computational cost by effectively approximating the covariance matrix with a theoretical investigation of optimal conditions. Importantly, we want to theoretically find the near-optimal rank for which we can achieve the desired accuracy. We will focus on the projection method via reduced rank approximation since almost all techniques can be put under a common umbrella. Let's start by first assuming that $Y(\mathbf{s})$ is the response variable at location $\mathbf{s} \in D \subseteq R^d$ with $p \times 1$ vector of regressors $\mathbf{x}(\mathbf{s})$ at location $\mathbf{s}$ then

$$Y(s) = \mathbf{x}^{\mathbf{T}}(\mathbf{s})\boldsymbol{\beta} + \boldsymbol{\varepsilon}(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients while $\boldsymbol{\varepsilon}$ is the demeaned spatial process, which is normally distributed with zero mean and a nonzero variance term. Rewrite Equation 2.1 to account

for the smooth scale spatial process $w(\boldsymbol{s})$ and the random error process $\epsilon(\boldsymbol{s})$. Write $\boldsymbol{\varepsilon}(\mathbf{s}) = w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$ then Equation 2.1 is given as

$$Y(\boldsymbol{s}) = \mathbf{x}^{\mathbf{T}}(\mathbf{s})\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{2.2}$$

where $w(\boldsymbol{s})$ is assumed to be a Gaussian process with zero mean and a covariance function $c(\mathbf{s}, \boldsymbol{s'})$; thus $w(\boldsymbol{s}) \sim GP\{0, \boldsymbol{c(.,.)}\}$. Usually we assume that $\boldsymbol{c(.,.)}$ is a constant process variance with a covariance function defined as $\boldsymbol{c(s, s')} = \sigma^2\rho\boldsymbol{(s, s'; \theta)}$ where $\boldsymbol{\theta}$ is a vector of correlation parameters, and $\rho(.,.;\boldsymbol{\theta})$ is the correlation function. The error process $\epsilon(\boldsymbol{s})$ is assumed to be normally distributed with zero mean and a variance ς^2, so-called nugget effect [11], at every location $\boldsymbol{s}$. The smooth scale spatial process $\boldsymbol{w(s)}$ often models the spatial relationship while the nugget effect determines the spatial pattern in unobserved covariates or models the measurement error [10].

Let $\mathbb{X} = (\mathbf{x(s_1)}, \mathbf{x(s_2)}, ..., \mathbf{x(s_n)})^T$ and $\mathbf{Y} = (Y(s_1), Y(s_2), ..., Y(s_n))^T$, keeping in mind that the focus is estimating the unknown parameters $\Psi = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \varsigma^2)$. Estimating the unknown parameters involves the application of the maximum likelihood. With the same description as in above, let $\boldsymbol{C} = [\boldsymbol{c(s_i, s_j)}]_{\boldsymbol{i=1:n, j=1:n}}$ be the covariance matrix of the smooth scale spatial process at each location $\boldsymbol{s_i}$. The log-likelihood function is defined as

$$\begin{aligned} lnL(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \varsigma^2) = & - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log[\det\boldsymbol{\{C(\theta, \sigma^2)} + \varsigma^2\mathbf{I_n}\}] & (2.3) \\ & - \frac{1}{2}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta})^T\boldsymbol{\{C(\theta, \sigma^2) + \varsigma^2 I_n\}}^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}), & (2.4) \end{aligned}$$

noting that the covariance matrix of $\mathbf{Y}$ is given as $\boldsymbol{C} + \varsigma^2\mathbf{I_n}$. Estimating the unknown parameters Ψ involves the inversion of the $n \times n$ covariance matrix $\boldsymbol{C} + \varsigma^2\boldsymbol{I_n}$, which is computionally costly and requires $O(n^3)$. This is one of the problems researchers encounter estimating the unknown parameters and spatial prediction at an unobserved location. The next section 2.1.1 - 2.2.2 presents modified ideas to address this problem.

2.1.1 Predictive Process via Reduced Rank Approximation

The reduced rank approximation method has been widely used successfully to approximate the covariance matrix for large data. It does so by creating a fixed approximated process $\mathbf{w}(\mathbf{s}*)$ that resides in a lower finite subspace from the smooth scale spatial process $\mathbf{w}(.)$ in Equation 2.2. Computational efficiency thereby is achieved by using a reduced rank approximation. [5] successfully developed reduced rank approximation using spatial interpolation, while [59] did likewise using Karhunen-Loeve expansion. Due to the context of this study, we proceed by following the ideas referenced in [5]. For now, we assume the mean part is zero and write Equation 2.2 as

$$Y \sim N_n(0, C + \varsigma^2 I_n). \tag{2.5}$$

Consider the set of knots $\{x(s_i^\star)\} : (i = 1, ..., m)$, and let $\mathbf{C}_{\star\star} = (c(x(s_i^\star), x(s_j^\star)))_{m\times m}$, and $\omega_\star = \{w(s_1^\star), w(s_2^\star), ..., w(s_n^\star)\}^T$. Using the predictive Gaussian process method, the best linear unbiased predictor (BLUP) at a new location is given as $w^{ppm}(\cdot) = E\{w(\cdot)|w_\star\}$ with a covariance function $c'_{s,\star}C^{-1}_{\star\star}c_{\star s'}$ where $c_{s,\star} = \{c(s, s_i^\star)\}^T : i = 1, 2, ..., m$ and $c_{\star,s'} = \{c(s_i^\star, s')\}^T : i = 1, 2, ..., m$. Plugging the (BLUP) into Equation 2.5 yields $Y \sim N_n(0, C_{X\star}C^{-1}_{\star\star}C_{\star X} + \varsigma^2 I_n)$, where $C_{\star X}$ is the covariance matrix between Y and $\omega_\star$. With these traits, anytime $s \notin \{x(s_i^\star)\}^T, i = 1, 2, ..., m$ the variance is underestimated since for each $s \in S$, $E(var\{w(s)|w_\star\}) = var\{w(s)\} - var\{w^{ppm}(s)\}$. The problem is rectified by introducing an independent process $\epsilon_2(s) \sim N(0, c(s, s') - c^{ppm}(s, s'))$ where c^{ppm} is described as above and then the rectified Gaussian process $w^{rppm}(s)$ has a covariance function

$$c^{rppm}(s, s') = c^{ppm}(s, s') + \varkappa(s, s')\{c(s, s') - c^{ppm}(s, s')\}, \tag{2.6}$$

where

$$\varkappa(s, s') = \begin{cases} 1 & \text{if } s = s' \\ 0 & \text{otherwise.} \end{cases}$$

2.1.2 Linear Projection Method

As generalization to the predictive process via reduced rank, [4] proposed a linear projection method where he defined a new kriging; thus, $w^{lpm}(\cdot) = E\{w(\cdot)|\Theta w_X\}$, $w_X = (w(s_1), w(s,2), ..., w(s,n))$ instead of w^{ppm} as described in 2.1.1, where Θ is an $m \times n$ matrix. Using the linear projection method, the improved marginal form is

$$Y \sim N_n(0, C^{lpm} + \varsigma^2 I_n), \tag{2.7}$$

where $C^{lpm} = (\Theta C)^T(\Theta C \Theta^T)^{-1}\Theta C$, which is induced by applying a Gaussian process with covariance function $c^{lpm}(s, s') = (\Theta c_{s,w})^T(\Theta C \Theta^T)^{-1}\Theta c_{w,s'}$, where
$c_{w,s'} = \{c(s_1, s'), c(s_2, s'), \ldots, c(s_n, s')\}^T$ and $c_{s,w} = \{c(s, s_1), c(s, s_2), \ldots, c(s, s_n)\}^T$ and $s, s' \in X$. The variance is underestimated and is rectified by introducing a nugget effect term so that a modified linear approximation is defined as $w^{mlpm}(.)$ with covariance function $c^{mlpm}(s, s') = c^{lpm}(s, s') + \varkappa(s, s')\{c(s, s') - c^{lpm}(s, s')\}$ for any $s, s' \in S$.

2.2 Optimal Rank Determination for Projection Method

Given the goal of reduced rank approximation, the spatial process $w(s)$ can be approximated by another process w that lies in lower dimensional space based on Karhunen–Loève theorem. For a zero mean stochastic process w, the Karhunen–Loève expansion of the spatial process is given as $w(\cdot) = \sum_{i=1}^{\infty} \varrho_i(\sqrt{\lambda_i})e_i(s)$, where e_i, and λ_i are the eigenfunctions and eigenvalues of the covariance function $c(.,.)$, respectively and ϱ_i are uncorrelated random variables with mean zero and a unit variance. Often, the key terms of the spatial process $w(s_i)$ capture important properties of the process, and the remaining terms are eliminated from the Karhumen - Loeve expansion [1]. This leads to the truncated Karhumen - Loeve expansion, which produces the optimal m - term approximation. The truncated Karhumen - Loeve expansion for the spatial process $w^{plp}(s)$ is given as

$$w^{plp}(s) = \sum_{i=1}^{m} \varrho_i(\sqrt{\lambda_i})e_i(s),$$

which can be expressed in linear projection, $E(\omega|R_m^T w_X)$, where R_m^T equals to the $m \times n$ matrix of eigenvectors corresponding to the m largest eigenvalues of C, e_i, and λ_i are the eigenfunctions and eigenvalues of the covariance function $c(.,.)$, and ϱ_i are uncorrelated random variables with mean zero and a unit variance. The covariance of the processs $w^{plp}(\cdot)$ is $C(s,s') = (R_m^T c_{s,w})^T (R_m^T C R_m)^{-1} (R_m^T c_{w,s'})$. The covariance of the process produces the optimal m- term approximation to the covariance function of the original spatial process $w(s)$. With this setting, the modified marginal form is given as

$$Y \sim N_n(0, C^{plp} + \varsigma^2 I_n), \tag{2.8}$$

which is induced by applying a Gaussian process with a covariance function

$$C^{plp} = \begin{pmatrix} R_m^T & c_\star \end{pmatrix}^T \left(R_m^T \begin{pmatrix} R_m & R_{(n-m)} \end{pmatrix} \begin{pmatrix} D_{mm} & 0 \\ 0 & D_{(n-m)(n-m)} \end{pmatrix} \begin{pmatrix} R'_m \\ R'_{(n-m)} \end{pmatrix} R_m \right)^{-1} \times \begin{pmatrix} R_m & c_\star \end{pmatrix},$$

where $c_\star = \{Cov(s, s_i)\}^T \ i, j = 1, 2, ..n$ and $C = \{Cov(s_i, s_j)\}^T$. The underestimated variance is improved by introducing a nugget effect as described in section 2.1.1, and then the covariance function of the proposed revised linear projection is given as

$$C^{prlp} = c_\star R_m D_{mm}^{-1} R_m^T C + \varkappa(s, s')\{c(s, s') - C^{plp}(s, s')\},$$

where

$$\varkappa(s, s') = \begin{cases} 1 & \text{if } s = s' \\ 0 & \text{otherwise.} \end{cases}$$

2.2.1 Reduced-Rank Matrix Approximation and Linear Projection Construction

We adapt a technique for calculating near-optimal approximation by applying the idea of linear projection and reduced rank approximation. For simplicity, we consider an $n \times n$ positive definite covariance matrix C of real values. Let $||\cdot||_2$ be the spectral norm and $||\cdot||_F$ be the Frobenius norm of C defined as the largest singular value and the square root of the sum of the absolute squares of its element, respectively. Using singular value decomposition, C can be written as $C = RDR^T$ where R is the eigenvectors and D is the $n \times n$ diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_n$ arranged in descending order of magnitude. Split C into block matrices as

$$C = \begin{pmatrix} R_m & R_{(n-m)} \end{pmatrix} \begin{pmatrix} D_{mm} & 0 \\ 0 & D_{(n-m)(n-m)} \end{pmatrix} \begin{pmatrix} R_m^T \\ R_{(n-m)}^T \end{pmatrix}. \tag{2.9}$$

If the rank m is known, by the Eckart-Young theorem [56], the best rank m approximation to C in spectral norm and Frobenius norm is given by $C_m = R_m D_{mm} R_m^T$. Using our settings, the original covariance matrix C is replaced by $C^{plp} = (R_m C)^T (R_m^T C R_m)^{-1} R_m^T C$. Then, Equation 2.8 is modified into $Y \sim N(0, C^{plp} + \varsigma^2 I_n)$. Aiming to control loss of information with reduced computation, we propose a technique that can address this problem of inversion of C by finding a near-optimal rank that yields good accuracy given a desired tolerance level. However, finding the near optimal rank requires a spectral decomposition of the covariance matrix C such that the $n \times n$ eigenvector matrix (which is orthogonal) spans the column space of C noting that $C = RR^T C$. For a simple error computation, we often consider the column space approximator $R_m R_m^T C = C_m$. The optimal rank m spectral decomposition corresponds to the best rank-m column space of this estimator, so it's enough to find a favorable column space approximator [4]. At any fixed target error, we only need to compute an approximate spectral decomposition and search for a projection matrix R_m^T for the column space approximation. Also with unknown target rank m, we propose Theorem 2.2.1 that finds the optimal rank m such that the computational intensity and prediction accuracy are well balanced. A modified Eigenvalue Decomposition via Nyström Method by [52] is used to help implement the task. In the same framework, one can adapt the modified Adaptive Randomized Range Finder Algorithm by [24].

Theorem 2.2.1. *Let $P = N_n(\mu_x, \sum = C_{xx} + \sigma^2 I)$ and $Q = N_n(\mu_x, \sum^{plp} = C_{xx}^{plp} + \sigma^2 I)$ be the marginal distributions of the original response vector y and a projection approximation with rank $m*$ respectively. If $m* = min\{m; \sum_{i=1}^{m} \lambda_i^2 \geq \sum_{ij}^{n} C_{ij}^2 - \Delta^2\}$ with $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_n$ then*

$$D_{KL}(P||Q) \leq \frac{n\Delta}{\sigma^2},$$

where $D_{KL}(\cdot, \cdot)$ represents the Kullback - Leibler divergence between probability densities, C is the true covariance, C^{plp} is the approximated covariance matrix, and λ_i and C_{ij} are the eigenvalues and entries of the covariance matrix, respectively, with $\sigma^2, \Delta > 0$.

The proof of theorem 2.2.1 can be found in the Appendix 2.6.3. Theorem 2.2.1 is used in our simulation study under different covariance functions to assess the accuracy of the

approximation. Algorithm 1 provides a fast way to calculate the eigenvectors and eigenvalues starting from the largest to a specific rank of interest. It aids in reducing computational intensity without changing much of the result accuracy [4].

Algorithm 1. *Given a positive definite matrix $C \in \mathbb{R}^{n\times n}$ and a random generated Johnson - Lindernstrauss matrix Ω of order $n \times r$, find the projection matrix of order $m \times n$ which approximates the column space, and compute the approximate spectral decomposition through Nyström approximation with the projection matrix.*

See Appendix 2.6.2 for details on Algorithm 1.

2.2.2 Approximation conditions based on mean squared prediction error

We will follow some important traits in chapters 2 and 3 of [54] that are in line with our results. In our study, we propose an optimal condition in terms of kriging for approximating the linear spatial predictor, which tends to be accurate with computational efficiency.

Assume $Y(\mathbf{s})$ be a Gaussian process with mean 0 and a covariance function $c(.,.,\eta,\sigma^2)$. In geology and geographical statistics, the main focus is to predict $\hat{Y}(s*)$ at any given location $s*$ in the spatial domain D. This prediction relies on the concept of minimum mean squared error [54]. At any unobserved location $\mathbf{s}*$, the best linear unbiased prediction is given by

$$\hat{Y}(\mathbf{s}*) = \mathbf{c_*^T}\sum\nolimits^{-\mathbf{1}}\mathbf{Y},$$

where $\mathbf{Y} = (Y(\mathbf{s_1}), Y(\mathbf{s_2}), ..., Y(\mathbf{s_n}))^T$, is the observation at $\mathbf{s_i}$ $i = 1, 2, ..., n$ locations, $\sum = C + \sigma^2\mathbf{I_n}$ and $\mathbf{c}_*$ is the covariance between all locations in the spatial domain D and the unobserved location $s*$; thus, $c_* = (c(s_1, s_*), \ldots, c(s_n, s_*))^T$. In order to make any kind of prediction, one needs to first evaluate $\sum^{-1}$ (as in Equation 2.3) since it is used when estimating the parameters of the model. We use the idea of reduced rank approximation to approximate $\sum$ and call it misspecified covariance, thus $\widetilde{\sum}$, without changing the trait of the true covariance matrix. Under reduced rank approximation, the best linear unbiased

predictor is given as

$$\hat{Y}(s*) = c_*^T \widetilde{\sum}^{-1} Y.$$

Inverting the misspecified covariance matrix $\widetilde{\sum}$ has more significant advantage computationally than inverting the true covariance matrix $\sum$. We will employ the Woodbury matrix identity to evaluate it. The mean squared prediction error is used to assess the accuracy of the best linear unbiased predictor (BLUP).

If indeed the BLUP is calculated under the reduced rank approximation technique or the misspecified covariance matrix, then the mean squared prediction error (MSPE) is given as

$$\begin{aligned} MSPE(s*, \tilde{C}) &= E(\hat{Y}(s*) - Y(s))^2 && (2.10) \\ &= C(s*, s*) - 2c_*^T \widetilde{\sum}^{-1} c_* + c_*^T \widetilde{\sum}^{-1} \sum \widetilde{\sum}^{-1} c_*. && (2.11) \end{aligned}$$

If the BLUP is calculated under the true covariance matrix $\sum$, then the mean squared prediction error (MSPE) is given as

$$\begin{aligned} MSPE(s*, \tilde{C}) &= E(\hat{Y}(s*) - Y(s))^2 && (2.12) \\ &= C(s*, s*) - c_*^T \sum^{-1} c_*. && (2.13) \end{aligned}$$

where $C(s*, s*)$ is the variance for the unobserved location $s*$. See Appendix 2.6.4 for details on how to obtain Equations 2.11 and 2.13. It is essential to know the difference between $\sum$ and $\widetilde{\sum}$ used in equations 2.11 and 2.13. In this study, our goal is to apply Theorem 2.2.1 - 2.2.2 for optimal rank selection and computational time saving. This is done through the measure of accuracies such as mean squared prediction error and Kullback-Leibler divergence. Consequently, we propose optimal rank conditions in terms of MSPE for a finite sample by using the following theorem.

Theorem 2.2.2. *Let* $P = N_n(\mu_x, \sum = C_{xx} + \sigma^2 I)$ *and* $Q = N_n(\mu_x, \sum^{plp} = C_{xx}^{plp} + \sigma^2 I)$ *be the marginal distributions of the response under the true model and a projection approximation with rank* $m*$*, respectively. If* $m* = min\{m; \sum_{i=1}^{m} \lambda_i^2 \geq \sum_{ij}^{n} C_{ij}^2 - \Delta^2\}$ *with* $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_n$ *then*

$$\frac{1}{\|c_* c_*^T\|_F} \left| MSPE_t(s*) - MSPE_m(s*) \right| \leq \Delta\sqrt{n}, \tag{2.14}$$

where $MSPE_t(s*)$ *and* $MSPE_m(s*)$ are the mean squared prediction error under the true and the projected matrix at a fixed location $s*$ in the spatial domain, respectively.

Proposition 2.2.2 *In general, for large sample approximation, we have that for fixed sample size* n *and varying* m*,*

$$\frac{MSPE_m(s*)}{MSPE_t(s*)} \xrightarrow{m \to n} 1.$$

We observed the following tendencies by using simulation studies under some mild assumptions on the covariance structure. Thus, for fixed target rank m, and a varying sample size n and vice-versa, we have the following;

$$\frac{MSPE_m(s*)}{MSPE_t(s*)} \xrightarrow{n \to \infty} 1. \tag{2.15}$$

The proof of Theorem 2.2.2 and Proposition 2.2.2 is in Appendix 2.37 and 2.6.6 of this chapter.

2.3 Simulation Study

We investigate the impact of the approximation techniques for large data in this section. The simulation studies are divided into three main parts.

- We investigate how Algorithm 1 can facilitate the application of Theorems 2.2.1 and 2.2.2. That is, one does not need to calculate the entire eigenvalues and eigenvectors of a large matrix of size n but rather select a reasonable rank and its associated eigenvalues and eigenvectors to achieve the same goal.

- We analyze the asymptotic trends of Theorems 2.2.1 and 2.2.2 based on Proposition 2.2.2 and Equation 2.15. This is done with the help of a randomized singular value decomposition package in R. Here, we will consider two scenarios: In scenario 1, we fix sample size n and vary target rank(s) m; and in scenario 2, we vary the sample size n and fix the target rank(s) m. The idea is to illustrate how the misspecified covariance mimics the true covariance matrix under certain conditions.

- We apply Theorem 2.2.1 and 2.2.2 for optimal rank selection and check for computational time-saving in terms of mean squared prediction error and Kullback - Leibler divergence. Here, without knowledge of the target rank, we seek to find the optimal rank via our theorem and establish that the misspecified covariance converges to the true covariance matrix.

In the first simulation, the Matérn covariance function with smoothness parameters $\nu = 0.5$ and 1.5 is used as a basis for comparison. The covariance function $C(x, y)$ is evaluated over a uniform grid of n locations in $[0, 1]$ and an $n \times n$ covariance matrix C is obtained. We apply Algorithm 1 to assess the computational efficiency in obtaining the desired results. In the simulation, the target rank m is obtained by using $2\%, 4\%$ and 6% of the sample size. Using a range and nugget parameter of 1, a Relative Structured Variability (RSV) of 50% and 90% is set for the Matérn covariance function at each ν, and the results are in Table 2.1 - 2.2 for an RSV of 50%. Tables 2.1 and 2.2 show that in general, given the sample size, as the target rank m increases, the difference between the RMSE of the eigenvalues and eigenvalues shrinks. For fixed sample size, the bigger the target rank, the more time is saved in the calculation. Comparing the difference in time spent using the entire eigenvalues and eigenvectors of the covariance matrix versus using a fraction of the eigenvalues and eigenvectors, we notice computationally we save time. The results, in general, are comparable, and this approach has the advantage of saving time and cost. The case with an RSV of 90% and $\nu = 0.5$ and $\nu = 1.5$ are presented in Appendix 2.6.7 because of similar results.

In the second simulation, under scenario 1, data are generated from the Matérn covariance

Table 2.1: Root mean square results for Matérn covariance function with $\nu = 0.5$ and RSV of 50%

Sample Size	Target rank	RMSE.Ev[1]	RMSE.Evc [2]	Time Partial (Original)[3]
100	2	0.000010	0.000005	0.08 (0.10)
	4	0.000156	0.000683	0.07 (0.10)
	6	0.001662	0.002665	0.06 (0.10)
500	10	0.000242	0.000329	0.20 (0.25)
	20	0.028245	0.009012	0.16 (0.25)
	30	0.075755	0.019947	0.15 (0.25)
1000	20	0.008921	0.002066	1.47 (1.68)
	40	0.092111	0.012875	1.34 (1.68)
	60	0.124733	0.016219	1.20 (1.68)
2000	40	0.044063	0.005486	10.77 (12.08)
	80	0.140236	0.011863	9.66 (12.08)
	120	0.1401755	0.013770	8.51 (12.08)
3000	60	0.067577	0.006160	34.93 (38.83)
	120	0.157688	0.010663	30.94 (38.83)
	180	0.138747	0.011984	26.97 (38.83)
4000	80	1038022	0.006530	100.17 (110.53)
	160	0.163492	0.009620	90.06 (110.53)
	240	0.137901	0.010706	78.39 (110.53)
5000	100	0.117288	0.006543	175.66 (195.04)
	200	0.164963	0.008738	156.31 (195.04)
	300	0.134672	0.009810	135.57 (195.04)

[1] RMSE.Ev is the RMSE for the difference in Eigenvalues.

[2] RMSE.Evc is the RMSE for the difference in Eigenvectors.

[3] Time Partial (Original) is time spent on the calculation when considering partial and full eigenvalues respectively.

Table 2.2: Root mean square results for Matérn covariance function with $\nu = 1.5$ and RSV of 50%

Sample Size	Target rank	RMSE.Ev[1]	RMSE.Evc [2]	Time Partial (Original)[3]
100	2	0.000000	0.000043	0.08(0.10)
	4	0.054755	0.027955	0.06 (0.10)
	6	0.093846	0.050158	0.04 (0.10)
500	10	0.110120	0.0197877	0.18 (0.23)
	20	0.076285	0.030079	0.22 (0.23)
	30	0.056612	0.032489	0.21 (0.23)
1000	20	0.103734	0.020028	1.47 (1.67)
	40	0.0619708	0.023333	1.28 (1.67)
	60	0.044948	0.024378	1.11 (1.67)
2000	40	0.086734	0.016021	11.89 (13.22)
	80	0.050673	0.017455	10.61 (13.22)
	120	0.036289	0.018042	9.34 (13.22)
3000	60	0.080711	0.013734	38.89 (42.62)
	120	0.046410	0.014537	34.72 (42.62)
	180	0.032778	0.014866	30.62 (42.62)
4000	80	0.074597	0.012094	84.34 (93.75)
	160	0.042724	0.012730	75.17 (93.75)
	240	0.029965	0.012975	66.48 (93.75)
5000	100	0.069080	0.011027	181.94 (199.61)
	200	0.039267	0.011493	158.33 (199.61)
	300	0.027548	0.011671	137.33 (199.61)

[1] RMSE.Ev is the RMSE for the difference in Eigenvalues.

[2] RMSE.Evc is the RMSE for the difference in Eigenvectors.

[3] Time Partial (Original) is time spent on the calculation when considering partial and full eigenvalues respectively.

function based on different parameters with sample sizes 100, 500, 1000, 2000, 3000, 4000 and 5000. Consider the Matérn covariance function, evaluate it over a uniform grid of n locations in $[0, 1]$, and consider the resulting $n \times n$ matrix C. The main idea is to illustrate how the misspecified covariance performs under Theorems 2.2.1 and 2.2.2 based on Proposition 2.2.2 and Equation 2.15. In the first scenario, it is important to note that the range parameter and the RSV play a major role in the outcome of the results. For consistency, we fix the range and nugget parameter at 1 and vary the RSV. The Matérn covariance matrix is evaluated with $\nu = 0.5$, RSV of 50%, and target ranks $m = 50, 60, 70, 80$ and 90. For Matérn covariance matrix with $\nu = 1.5$, we considered an RSV of 50%, and target rank $m = 13, 15, 17, 20$ and 25.

Figures 2.4 and 2.5 show results in using the randomized singular value decomposition package (rsvd) in r, Theorems 2.2.1 and 2.2.2 based on Proposition 2.2.2 and Equation 2.15 for Matérn covariance matrix with an RSV of 50% at $\nu = 0.5$ and 1.5, respectively. From Figure 2.4, we can see a decreasing trend as m increases; at $m = 60$, the graph seems to be leveling off. We also see that the closer m is to n, the better the estimate and closer the lines (graph) are to the horizontal axis or to the line $y = 1$. Next, Figure 2.5 shows the results for Matérn covariance function at $\nu = 1.5$. Here, we notice that the rate of decrease is much faster and steeper, and at a rank of $m = 15$, the graph levels off. Table 2.3 provides more detailed information of the actual measurements showing that as m increases, the MSPE for the misspecified covariance converges to the true covariance matrix.

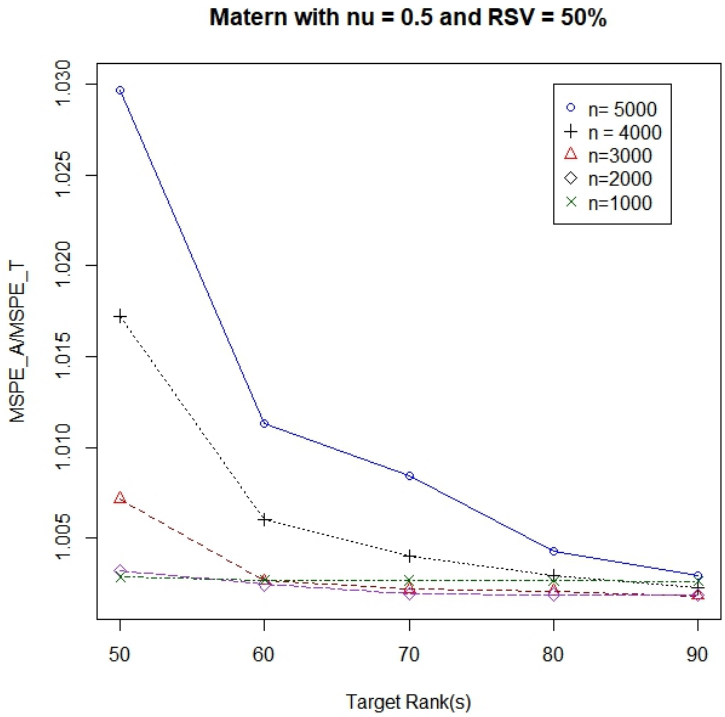

Figure 2.4: Resulting graph of Matérn at an RSV of 50% and $\nu = 0.5$

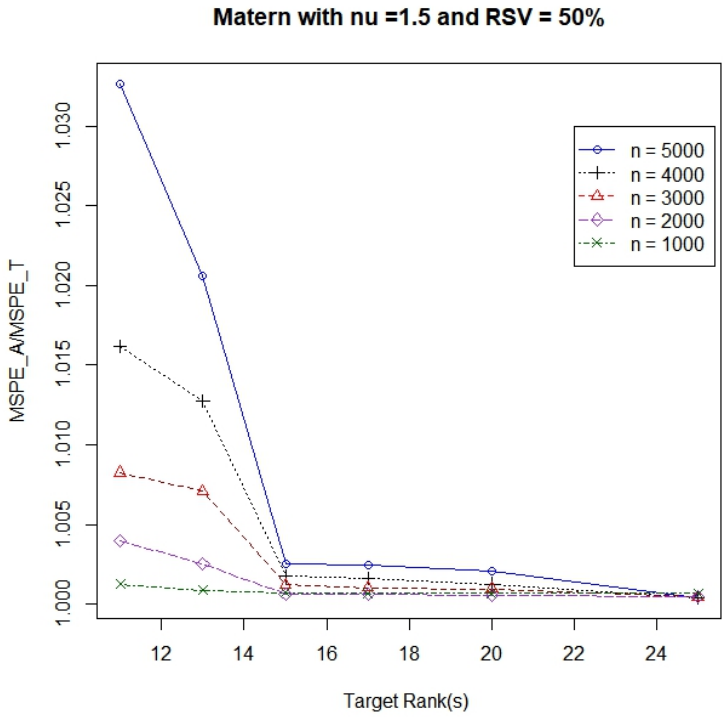

Figure 2.5: Resulting graph of Matérn at an RSV of 50% and $\nu = 1.5$

Table 2.3: Results of ratio of MSPEs at fixed N and varying M with an RSV of 50%

Sample Size	Rank ($\nu = 1.5$)	M.Ratio[1]	Rank ($\nu = 0.5$)	E.Ratio[2]
1000	50	1.00288	13	1.00086
	60	1.00271	15	1.00072
	70	1.00268	17	1.00068
	80	1.00267	20	1.00067
	90	1.00264	25	1.00066
2000	50	1.00323	13	1.00251
	60	1.00245	15	1.00061
	70	1.00197	17	1.00060
	80	1.00191	20	1.00054
	90	1.00188	25	1.00049
3000	50	1.00719	13	1.00707
	60	1.00265	15	1.00120
	70	1.00219	17	1.00103
	80	1.00209	20	1.00095
	90	1.00184	25	1.00043
4000	50	1.01724	13	1.01276
	60	1.00605	15	1.00173
	70	1.00400	17	1.00163
	80	1.00293	20	1.00126
	90	1.00230	25	1.00042
5000	50	1.02968	13	1.02062
	60	1.01134	15	1.00255
	70	1.00844	17	1.00248
	80	1.00428	20	1.00209
	90	1.00292	25	1.00041

[1] M.Ratio is ratio between the MPSEs for Matérn with $\nu = 1.5$.
[2] E.Ratio is ratio between the MPSEs for Matérn with $\nu = 0.5$.

In scenario 2, the sample size n is allowed to vary keeping the target rank m fixed. Here, emphasis is on how RSV affects the ratio between the MSPEs when it either increases or decreases. Figure 2.6 and Table 2.4 show the results of Matérn at $\nu = 1.5$ with an RSV of 30% and 50%. Figure 2.6 shows that the graph increases and levels off at some point. When the rank is small, the graph ill behaves, and when it's bigger, it stabilizes. Also, the smaller the RSV, the smoother the graph, and its distributions converge at some point. Table 2.4 gives detailed information of the measurements showing that smaller rank performs better in terms of the approximation. The case with $\nu = 0.5$ is presented in Appendix 2.6.8 because of similar results.

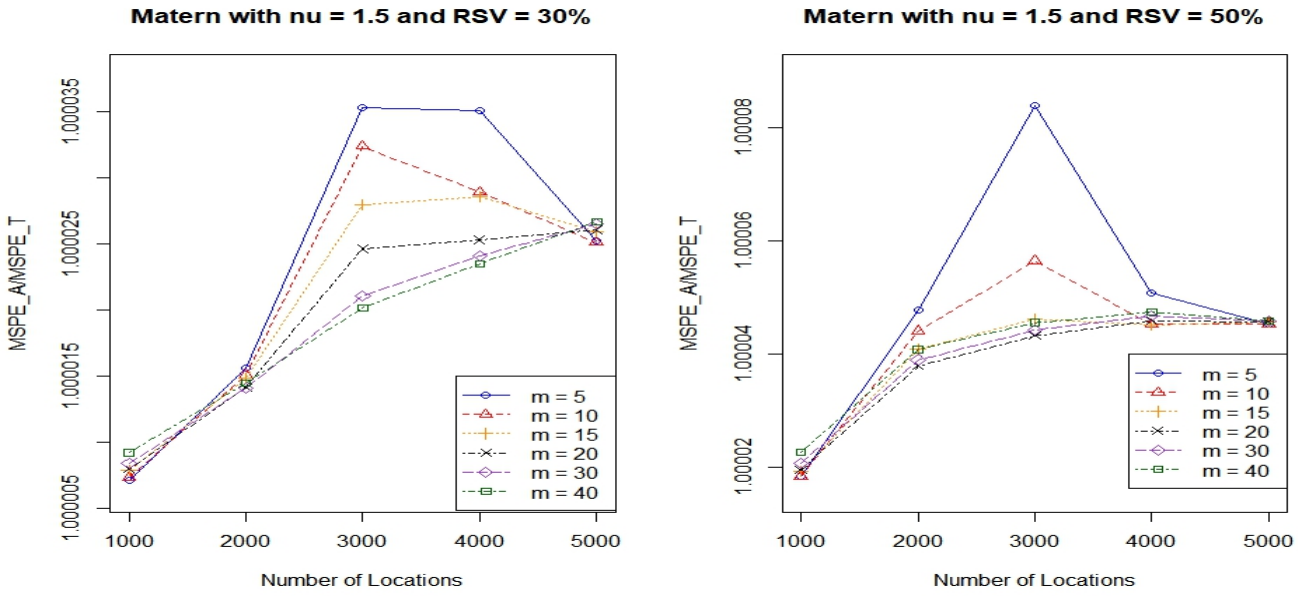

Figure 2.6: Results for Matérn covariance function for fixed M and varying N

In the third simulation, we illustrate how Theorem 2.2.1 and 2.2.2 can be used to assess the accuracy of the covariance approximation under the Matérn covariance function. To be precise, mean squared prediction error (MSPE) and the Kullback-Liebler (KL) divergence between distributions via covariance approximation techniques are used to achieve the approximation technique. Here, with unknown optimal rank, we use Theorems 2.2.1

Table 2.4: Results for ratio of MSPE for fixed M varying N for Matérn with $\nu = 0.5$, an RSV of 50% and 30% respectively

Target Rank	Sample size	Ratio RSV of 50	Ratio at RSV of 30
5	1000	1.0000183	1.0000071
	2000	1.0000478	1.0000156
	3000	1.0000840	1.0000353
	4000	1.0000507	1.0000350
	5000	1.0000453	1.0000252
10	1000	1.0000183	1.0000074
	2000	1.0000441	1.0000151
	3000	1.0000565	1.0000324
	4000	1.0000452	1.0000289
	5000	1.0000453	1.0000251
15	1000	1.0000192	1.0000079
	2000	1.0000408	1.0000149
	3000	1.0000462	1.0000280
	4000	1.0000451	1.0000286
	5000	1.0000456	1.0000259
20	1000	1.0000195	1.000008
	2000	1.0000379	1.0000141
	3000	1.0000432	1.0000246
	4000	1.0000459	1.0000253
	5000	1.0000457	1.0000260
30	1000	1.0000206	1.0000084
	2000	1.0000390	1.0000141
	3000	1.0000442	1.0000211
	4000	1.0000466	1.0000241
	5000	1.0000458	1.0000265
40	1000	1.0000225	1.0000092
	2000	1.0000408	1.0000145
	3000	1.0000455	1.0000202
	4000	1.0000474	1.0000235
	5000	1.0000458	1.0000267

and 2.2.2 with a reasonable tolerance level to reduce computational burden without losing most of the results accuracy. We start by randomly generating data from simulated samples from the Matérn covariance function that is constructed over a uniform grid of n locations in $[0, 1]$. The true covariance matrix C is obtained, and Theorem 2.2.1 and 2.2.2 is used to find the optimal rank m for the misspecified covariance matrix. As a golden rule, it is useful to specify an appropriate tolerance level Δ to achieve a reasonable optimal rank m for which the computation of the MSPE is well defined. In general, let $\Delta = \alpha F$ where F is the first eigenvalue of the true covariance matrix C and $\alpha > 0$. The misspecified covariance matrix is achieved by using the proposed theorems and rsvd package in R. For comparison, data is generated from the Matérn covariance function with $\nu = 0.5, 1, 1.5$, and an RSV of $50\%, 70\%$ and 90%. The range parameter and nugget parameter are held constant while the smoothness parameter is allowed to vary. We do this to avoid singularity problems in the covariance matrix. For instance, in our simulation, we used the following: range parameter =1, nugget parameter =1, $\nu = 0.5, 1, 1.5$, and $\phi = 1, 2.3, 9$. The following tolerance level $\Delta = 0.001F,\ 0.01F,\ 0.05F$ and $0.1F$ is used to obtain the optimal rank m. Notably, Δ plays an important role in getting a reasonable optimal rank m, which in turn reduces computional cost and time. Theorem 2.2.1 and 2.2.2 in addition to Algoritm 1 are used to find m; then the MPSEs under the true and misspecified covariance matrix are calculated. Finally, the ratio between the MSPEs is computed. Tables 2.5 - 2.6 provide results for Matérn covariance function with $\nu = 0.5$. We observe that as Δ decreases, the rank increases with an increase in the sample size. It is important to note that as the tolerance level decreases so does the divergence between the two distributions, and the ratio between the MSPEs tends to be good. In other words, the MSPE under the misspecified covariance converges to the true covariance with a smaller divergence between the two distributions. Comparing the times to compute the MSPEs, we observe that it takes longer to obtain the MSPE under the true covariance function than in the misspecified covariance matrix. This was expected as the computation of the MSPE under the true covariance involves inverting the true covariance matrix of size n using the normal approach while that of the misspecified covariance involves using the Sherman - Morrison formula that computationally saves time. Generally, as the

sample size increases, the misspecified covariance mimics the traits of the true covariance matrix. We also observe that, as the relative structured variability decreases, the ratio between the MPSEs gets better. Thus, the lower the RSV, the better the estimates. With a smoothing parameter $\nu = 1$, the results for Matérn covariance function are presented in Tables 2.7 - 2.8. The result shows that as the tolerance level decreases, the ratio yields good estimates. The time taken to calculate the MSPE under the misspecified covariance matrix is less than that of the true covariance. We also note that, across RSV, the smaller the RSV, the better the estimate, so one doesn't need a huge optimal rank to get a good estimate. Ultimately, the MSPE under the misspecified covariance matrix approaches that of the true covariance matrix with a smaller tolerance level. Because of similar conclusions, the results for Matérn with a smoothness parameter of 1.5 and an RSV of $50\%, 70\%$ and 90% Matérn with $(\nu = 0.5, 1)$ at an RSV of 90% are left in Appendix 2.6.9.

Table 2.5: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 0.5$ and RSV 50%

		Tolerance Level = Δ		
Sample		$\Delta_1 = 0.01F^3$	$\Delta_2 = 0.05F^3$	$\Delta_3 = 0.1F^3$
100	Ratio	1.00003	1.002268	1.013865
	Rank	18	8	5
	Time.True[1]	0.015	0.015	0.015
	Time.Miss [2]	0.08	0.07	0.05
500	Ratio	1.000237	1.005284	1.024885
	Rank	36	16	11
	Time.True[1]	0.23	0.23	0.23
	Time.Miss [2]	0.19	0.17	0.16
1000	Ratio	1.000265	1.007934	1.037813
	Rank	48	22	15
	Time.True[1]	1.48	1.48	1.48
	Time.Miss [2]	0.89	0.43	0.46
2000	Ratio	1.000533	1.025697	1.048749
	Rank	66	29	21
	Time.True[1]	12.62	12.62	12.62
	Time.Miss [2]	3.92	2.06	1.56
3000	Ratio	1.00243	1.016728	1.071978
	Rank	82	36	26
	Time.True[1]	40.88	40.88	40.88
	Time.Miss [2]	10.35	5.22	4.08
4000	Ratio	1.000907	1.0238	1.139153
	Rank	96	42	30
	Time.True [1]	94.75	94.75	94.75
	Time.Miss [2]	21.32	10.76	8.25
5000	Ratio	1.000696	1.033283	1.095292
	Rank	110	47	33
	Time.True [1]	184.12	184.12	184.12
	Time.Miss [2]	37.75	17.64	13.20

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.6: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 0.5$ and RSV 70%

		Tolerance Level = Δ		
Sample		$\Delta_1 = 0.01F^3$	$\Delta_2 = 0.05F^3$	$\Delta_3 = 0.1F^3$
100	Ratio	1.000097	1.009734	1.039421
	Rank	27	12	8
	Time.True [1]	0.011	0.011	0.011
	Time.Miss [2]	0.009	0.005	0.004
500	Ratio	1.000331	1.014306	1.086852
	Rank	56	24	17
	Time.True[1]	0.24	0.24	0.24
	Time.Miss [2]	0.28	0.14	0.09
1000	Ratio	1.000413	1.027845	1.077359
	Rank	76	32	23
	Time.True [1]	1.38	1.38	1.38
	Time.Miss[2]	1.34	0.64	0.48
2000	Ratio	1.002059	1.04823	1.265167
	Rank	102	44	31
	Time.True[1]	12.91	12.91	12.91
	Time.Miss [2]	6.00	2.77	2.05
3000	Ratio	1.005773	1.056763	1.180285
	Rank	105	55	39
	Time.True[1]	41.28	41.28	41.28
	Time.Miss [2]	13.05	7.39	5.77
4000	Ratio	1.001419	1.072739	1.294263
	Rank	152	64	45
	Time.True [1]	94.99	94.99	94.99
	Time.Miss [2]	33.61	15.03	11.08
5000	Ratio	1.001142	1.077658	1.379489
	Rank	174	73	51
	Time.True [1]	184.19	184.19	184.19
	Time.Miss [2]	59.06	25.07	17.89

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.7: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1$ and RSV 50%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000008	1.009884	1.019341	1.019341
	Rank	8	4	2	2
	Time.True [1]	0.012	0.012	0.012	0.012
	Time.Miss[2]	0.007	0.007	0.005	0.003
500	Ratio	1.000004	1.000737	1.001442	1.634426
	Rank	14	7	5	4
	Time.True[1]	0.18	0.18	0.18	0.18
	Time.Miss [2]	0.09	0.16	0.14	0.07
1000	Ratio	1.000027	1.004345	1.038748	1.038748
	Rank	17	9	5	5
	Time.True[1]	1.31	1.31	1.31	1.31
	Time.Miss [2]	1.52	0.23	0.19	0.25
2000	Ratio	1.000058	1.002477	1.025798	1.060442
	Rank	21	11	7	5
	Time.True[1]	12.02	12.02	12.02	12.02
	Time.Miss [2]	1.61	1.19	1.06	1.06
3000	Ratio	1.000032	1.006814	1.065251	1.070792
	Rank	24	12	7	6
	Time.True[1]	39.78	39.78	39.78	39.78
	Time.Miss [2]	3.55	2.92	2.64	2.51
4000	Ratio	1.000068	1.012287	1.121105	1.123732
	Rank	26	13	8	7
	Time.True[1]	100.00	100.00	100.00	100.00
	Time.Miss[2]	7.37	6.23	5.92	5.93
5000	Ratio	1.000122	1.003287	1.196391	1.201403
	Rank	28	14	9	7
	Time.True [1]	196.33	196.33	196.33	196.33
	Time.Miss[2]	12.20	9.78	8.93	8.52

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.8: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1$ and RSV 70%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000095	1.004326	1.198917	1.299862
	Rank	10	5	3	2
	Time.True[1]	0.011	0.011	0.011	0.011
	Time.Miss[2]	0.008	0.005	0.005	0.04
500	Ratio	1.000094	1.011849	1.023423	1.023423
	Rank	18	9	5	5
	Time.True[1]	0.15	0.15	0.15	0.15
	Time.Miss[2]	0.11	0.10	0.10	0.07
1000	Ratio	1.000156	1.013001	1.079322	1.51769
	Rank	22	11	7	5
	Time.True[1]	1.11	1.11	1.11	1.11
	Time.Miss[2]	0.75	0.70	0.65	0.61
2000	Ratio	1.000231	1.004719	1.36305	1.375799
	Rank	27	14	9	7
	Time.True [1]	9.92	9.92	9.92	9.92
	Time.Miss[2]	3.20	3.03	2.83	2.86
3000	Ratio	1.000488	1.011855	1.879564	1.884828
	Rank	31	16	10	8
	Time.True[1]	34.19	34.19	34.19	34.19
	Time.Miss [2]	10.02	9.30	8.86	8.78
4000	Ratio	1.000283	1.020313	2.626606	2.626636
	Rank	34	17	10	9
	Time.True[1]	79.43	79.43	79.43	79.43
	Time.Miss [2]	14.93	13.45	13.01	13.05
5000	Ratio	1.000363	1.034078	1.356085	3.603772
	Rank	36	18	11	9
	Time.True [1]	152.38	152.38	152.38	152.38
	Time.Miss [2]	25.98	23.56	22.75	22.18

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

2.4 Kansas Data on Nitrous Oxide Emission and Weather Variables

The use of the proposed spatial linear model with a large data approximation technique is illustrated with nitrous oxide emission and actual weather data from the state of Kansas. The research aims to study the effect of weather change on nitrous oxide emission in the state of Kansas using a modified version of linear projection and reduced rank approximation. In this study, observed nitrous oxide emission was obtained from 2022 random locations across the state of Kansas for the year 2009. Daily temperature and precipitation were recorded from each of the locations, and the average yearly nitrous oxide emission and weather data were used.

In this section, we focus on a large data approximation technique to assess the spatial impact of actual weather measurement rather than effects of long-range weather change on nitrous oxide emission. To demonstrate the utility of the technique, we consider information on nitrous oxide emission and climate data in all 2022 locations in Kansas. The difficulty here comes with the inversion of the covariance matrix of size 2022×2022. The goal is to assess the effect of actual weather measurement on nitrous oxide emission and to perform spatial prediction by applying Theorems 2.2.1 and 2.2.2 for optimal rank selection and computational time-saving in terms of KL and MSPE. To achieve this, we proceeded by using the maximum likelihood (ML) methods of estimation. Under ML, there are two scenarios, the traditional method and the reduced rank approximation method. We first divided the data into training and testing sets. The training set is usually used to estimate the unknown parameters from the model and to calculate the prediction under consideration, while the testing set is used to evaluate the predictions. In our study, we considered two different sets of training and testing sets by using the sample function in r for our data analysis. The sample function does that by taking a sample of a specified size from the data without replacement [6]. We divided the data into 99% to 1% and 80% to 20% of the sample size. Therefore, the first setting had randomly selected 2001 training data points and 21 testing

sets, while the second setting had 1617 points for training and 405 for testing.

As an intermediate step, we fit an appropriate linear model and assessed the empirical variogram based on the residuals obtained in the linear model. Initial weighted least squares (WLS) estimates were obtained and used to find the estimates of the maximum likelihood method. Under the traditional method, we obtained the estimates using the traditional maximum likelihood method (that is, using the nlm function in R). The estimated parameters, the time recorded to compute the estimation and prediction, and the results are presented in Tables 2.9 - 2.10. Under the approximated method, the misspecified covariance matrix is obtained by applying the proposed theorems. The parameters for precipitation (β_1), minimum temperature (β_2), maximum temperature (β_3), the variance parameters; partial sill σ^2, nugget effect τ^2 and the time taken to calculate estimation and prediction were recorded, and the results are in Table 2.9 - 2.10.

Table 2.9: Results using Maximum Likelihood Method based on the first setting

Assignment	Traditional Method	Approximation Method
Time	618.83	105.63
MSPE	0.203796	0.360353
σ^2	0.207481	0.208350
τ^2	0.172158	0.172008
β_0	-4.959915	-5.404774
β_1 95% CI	2.216246 (1.996303, 2.436189)	2.584739 (2.372892, 2.796585)
β_2 95% CI	0.373021 (0.348652, 0.397390)	0.441984 (0.418654, 0.465315)
β_3 95% CI	1.136830 (1.102189, 1.171472)	1.090725 (1.057733, 1.123716)

From Tables 2.9 - 2.10, we observe that irrespective of the setting used, the approximation method computationally outperforms the traditional method in terms of the time to estimate and predict; meanwhile, the estimated parameters β, σ^2 and τ^2 are comparable. At a significance level of 0.05, we observe that precipitation, maximum and minimum temperature had a significant positive impact on nitrous oxide emission in the state of Kansas. Subsequently, the confidence interval for the estimated β's is presented, so we can determine the effect of weather change on nitrous oxide emission.

From Tables 2.9 - 2.10, regardless of the method and setting you used, we conclude, we are

Table 2.10: Results using Maximum Likelihood Method based on the second setting

Assignment	Traditional Method	Approximation Method
Time	350.76	92.21
MSPE	0.214929	0.245329
σ^2	0.199010	0.198933
τ^2	0.074549	0.159319
β_0	-4.911568	-6.326268
β_1 95% CI	3.122546 (2.911928, 3.333165)	4.067708 (3.825054, 4.310334)
β_2 95% CI	0.238031 (0.215640, 0.260423)	0.632190 (0.604901, 0.659479)
β_3 95% CI	0.717481 (0.684285, 0.750678)	0.626414 (0.588514, 0.664315)

95% certain that the true parameter of each of the weather variables lies in the interval and that those variables had a significant positive impact on nitrous oxide emission. This implies that precipitation and temperature have an effect on nitrous oxide emission [7, 23]. Using p-values, the results show, precipitation is most impactful, while minimum temperature as the least impact

2.5 Conclusion and Discussions

This study proposes a new method for calculating the optimal rank of the covariance matrix given the tolerance level Δ. The method serves as a tool for approximating the true covariance for spatial modeling through the Kullback-Leibler divergence and mean squared prediction error when considering a large data of size n. Randomized singular value decomposition and Theorems 2.2.1 and 2.2.2 uses the idea of linear projections to project data from a higher dimensional space into a lower dimensional space to reduce computational cost without sacrificing much results accuracy. This research is useful because in literature, researchers make an educated guess about the optimal rank without any theoretical justification. We have suggested a procedure of obtaining an optimal rank m with theoretical justification which is computational efficient. In the simulation studies, we have shown how to obtain the optimal rank for a fixed accuracy level so we can reduce computational burden. We have indicated that, the larger the range, the better the approximation method.

Thus, the misspecified covariance matrix mimics the traits of the true covariance matrix or converges to the true covariance matrix. A lower relative structured variability yields better results, and as the divergence between the two distributions decreases, the faster the convergence, and it yields good results. Also for large sample approximation, we showed that for a fixed sample size n and different values of the rank m, the ratio between the MSPEs under the true covariance and misspecify covariance tends to one. In the data analysis, we used our proposed method to assess the accuracy of the estimates in using the maximum likelihood method. We demonstrated with a tolerance level of 0.05F that, the approximated method under the maximum likelihood outperforms the traditional method. We showed that the proposed method is computationally accurate, efficient, and saves more time than the traditional method. We have established that the lower the training set, the more accurate the results will be in terms of the MSPE. Finally, we established that precipitation and temperature have a positive effect on nitrous oxide emission in the state of Kansas [7, 23].

An improvement and extensions can be studied in the future. First, we can further investigate the impact of each weather variable on nitrous oxide emission by doing a sensitivity analysis. Second, the study focused only on one year's worth of data in 2022 locations in Kansas. If more information is useful, it would be advisable to explore large scale influence of weather change over several years on nitrous oxide emission. Also, we only studied nitrous oxide emission in this work; for interest in various greenhouse gases emitted in Kansas, a multivariate spatial linear function can be considered, taking advantage of the theorems proposed in Chapter 2. Third, it would be beneficial to explore the dynamic effect of weather change on nitrous oxide emission. If the data is well defined, a spatial partial functional linear model with large data approximation would also be an important work to consider in the future.

2.6 Chapter 2 Appendix

2.6.1 Adapative Randomized Range Finder Algorithm

We follow the following steps in order to achieve an $m \times n$ projectiion matrix R_m^T.

1. Initialize $j = 0$ and $R^{(0)} = [\,]$, the $0 \times n$ empty matrix.
2. Draw a standard normal vectors $\omega^{(1)}, \omega^{(2)}, ..., \omega^{(r)}$ of length n.
3. For $i = 1, 2, ..., r$, compute $\kappa^{(i)} = C\omega^{(i)}$.
4. Is $\max_{i=1,2,...,r}(\parallel \kappa^{(j+i)} \parallel) < \epsilon/(10\sqrt{2/\pi})$? If yes, go to Step 11. If no, go to Step 5.
5. Recompute $j = j + 1, \kappa^{(j)} = [I - R^{(j-1)}{R^{(j-1)}}^T]\kappa^{(j)}$ and $r^{(j)} = \kappa^{(j)} / \parallel \kappa^{(j)} \parallel$.
6. Set $R^T(j) = [R^{(j-1)}{R^{(j-1)}}^T]^T r^{(j)}$.
7. Draw a standard normal vector $\omega^{(j+r)}$ of length n.
8. Compute $\kappa^{(j+r)} = [I - R^{(j)}{R^{(j)}}^T]C\kappa^{(j+r)}$.
9. For $i = (j+1), ..., (j+r-1)$, recompute $\kappa^{(i)} = \kappa^{(i)} - r^{(j)}\langle r^{(j)}, \kappa^{(i)}\rangle \kappa^{(j)}$.
10. Go back to the target error check in Step 4.
11. If $j = 0$, output $R^T = \{\parallel \kappa^{(1)} \parallel^{-1} \kappa^{(1)}\}^T$; else output $R^T = {R^{(j)}}^T$.

2.6.2 Modified Eigenvalue Decomposition via Nyström Method

1. Form the matrix product $C\Omega$.
2. Compute R_m^T, the left factor of the rank m spectral projection of the small matrix $C\Omega$.
3. Form $C_1 = R_m^T C R_m$.
4. Perform a Cholesky factorization of $C_1 = BB^T$
5. Calculate the Nyström factor $C_2 = CR_m^T(B^T)^{-1}$.

6. Compute a singular value decomposition for $C_2 = UDV^T$.

7. Draw a standard normal vector $\Omega^{(j+r)}$ of length n.

8. Calculate the approximate spectral decomposition for $C \approx C_{ms} = UD^2U^T$.

2.6.3 Proof Theorem 2.2.1

The explicit form of $\| C_{xx} - C_{xx}^{lp} \|_F$ is given by first writing the original covariance matrix as :

$$C_{xx} = \begin{pmatrix} R_m & R_{(n-m)} \end{pmatrix} \begin{pmatrix} D_{mm} & 0 \\ 0 & D_{(n-m)(n-m)} \end{pmatrix} \begin{pmatrix} R'_m \\ R'_{(n-m)} \end{pmatrix}, \tag{2.16}$$

where $D_{KL}(\cdot,\cdot)$ represents the Kullback - Leibler divergence between probability densities.

$$\begin{aligned} C_{xx}^{lp} &= R_m D_{mm} R'_m \\ &= \begin{pmatrix} \vec{e}_1(\cdot), ..., \vec{e}_m(\cdot) \end{pmatrix} D_{mm} \begin{pmatrix} \vec{e}_1(\cdot), ..., \vec{e}_m(\cdot) \end{pmatrix}'. \end{aligned} \tag{2.17}$$

Thus $C_{xx} = \sum_{i=1}^n \lambda_i \vec{e}_i \vec{e}'_i$, $C_{xx}^{lp} = \sum_{i=1}^m \lambda_i \vec{e}_i \vec{e}'_i$ implies that $||C_{xx}||_F = \sum_{i=1}^n \lambda_i^2$ and $||C_{xx}^{lp}||_F = \sum_{i=1}^m \lambda_i^2$.

$$\begin{aligned} ||C_{xx} - C_{xx}^{lp}||_F &= || \sum_{i=m+1}^{n} \lambda_i \vec{e}_i \vec{e}'_i ||_F && (2.18) \\ &= ||R_{n-m} D_{(n-m)(n-m)} R'_{n-m}||_F && (2.19) \\ &= \sqrt{tr\left(R_{n-m} D_{(n-m)(n-m)} R'_{n-m} R_{n-m} D_{(n-m)(n-m)} R'_{n-m}\right)} && (2.20) \\ &= \sqrt{tr\left(R_{n-m} D^2_{(n-m)(n-m)} R'_{n-m}\right)} && (2.21) \\ &= \sqrt{\sum_{i=m+1}^{n} \lambda_i^2} \leq \Delta && (2.22) \end{aligned}$$

In short $\forall \Delta > 0$,

$$||C_{xx} - C_{xx}^{lp}||_F^2 = ||C_{xx}||_F^2 - \sum_{i=1}^{m} \lambda_i^2 \tag{2.23}$$

$$= \sum_{i:j=1}^{n} C_{ij}^2 - \sum_{i=1}^{m} \lambda_i^2 < \Delta^2, \tag{2.24}$$

hence; $m = min\{m; \sum_{i=1}^{m} \lambda_i^2 \geq \sum_{ij}^{n} C_{ij}^2 - \Delta^2\}$. The Kullback - Leibler divergence between two multivariate normal distributions of dimension n; thus $N_F = MVN(\mu_F, \sum = C_{xx} + \sigma^2 I)$ and $N_R = MVN(\mu_R, \widetilde{\sum} = C_{xx}^{plp} + \sigma^2 I)$ is given as

$$D_{KL}(N_F \parallel N_R) = \frac{1}{2}\left[tr\left(\widetilde{\sum}^{-1}\sum\right) + (\mu_F - \mu_R)^T \widetilde{\sum}^{-1}(\mu_F - \mu_R) - n + \log\det\left(\widetilde{\sum}^{-1}\sum\right)\right].$$

Under our settings, consider the multivariate normal distributions
$N_F = MVN(0, \sum = C_{xx} + \sigma^2 I)$ and $N_R = MVN(0, \widetilde{\sum} = C_{xx}^{plp} + \sigma^2 I)$ then $\parallel \sum - \widetilde{\sum} \parallel_F = \parallel C_{xx} - C_{xx}^{plp} \parallel_F \leq \Delta$, and also $D_{KL}(N_F \parallel N_R) = \frac{1}{2}\left[tr\left(\widetilde{\sum}^{-1}\sum\right) - n + \log det\left(\widetilde{\sum}^{-1}\sum\right)\right]$. The theorem can be proved by breaking the expression for the Kullback - Leibler divergence into two parts. Consider first;

$$
\begin{aligned}
tr\left(\widetilde{\sum}^{-1}\sum\right) - n &= tr\left[\left(\widetilde{\sum}^{-1}\sum - I\right)\right] \\
&= tr\left[\widetilde{\sum}^{-1}\left(\sum - \widetilde{\sum}\right)\right] \\
&= tr\left[\left(C_{xx}^{plp} + \sigma^2 I\right)^{-1}\left(\sum - \widetilde{\sum}\right)\right] \\
&= tr\left[\left(C_{xx}^{plp} + \sigma^2 I\right)^{-1}\left(R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right)\right] \\
&= tr\left[\left(R_m D_m R_m^T + \sigma^2 I\right)^{-1}\left(R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right)\right] \\
&= tr\left[\left(\sigma^{-2}I - \sigma^{-2}IR_m\left(D_m^{-1} + R_m^T\sigma^{-2}IR_m\right)^{-1}R_m^T\right)\right. \\
&\quad \left.\times\left(R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right)\right] \\
&= tr\left[\left(\sigma^{-2}I - \sigma^{-2}IR_m\left(D_m^{-1} + \sigma^{-2}I\right)R_m^T\sigma^{-2}\right) \times \left(R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right)\right] \\
&= \sigma^{-2}tr\left[R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T - \sigma^{-2}R_m(D_m^{-1} + \sigma^{-2})R_m^T\right. \\
&\times \left.\left(R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right)\right] \\
&= \sigma^{-2}tr\left[R_{n-m}D_{(n-m)(n-m)}R_{n-m}^T\right] \\
&= \sigma^{-2}tr\left[D_{(n-m)(n-m)}R_{n-m}^T R_{n-m}\right] \\
&= \sigma^{-2}tr\left[D_{(n-m)(n-m)}\right] \\
&= \sigma^{-2}\sum_{i=m+1}^{n}\lambda_i \quad \text{By Cauchy Schwarz inequality} \\
&\leq \sigma^{-2}\sqrt{\sum_{i=m+1}^{n}1\sum_{i=m+1}^{n}\lambda_i^2} \quad \text{Noting that } \sqrt{\sum_{i=m+1}^{n}\lambda_i^2} \leq \Delta \\
&= \frac{\sqrt{n}\Delta}{\sigma^2},
\end{aligned}
$$

thus

$$
tr\left(\widetilde{\sum}^{-1}\sum\right) - n = \frac{\sqrt{n}\Delta}{\sigma^2}. \tag{2.25}
$$

Secondly, we consider bounding of $-\log\left(\det\left[\widetilde{\sum}^{-1}\sum\right]\right)$. First claim that

$$-\log\left(\det\left[\widetilde{\sum}^{-1}\sum\right]\right) \leq \frac{n\Delta}{\sigma^2-\Delta}. \tag{2.26}$$

By definition $\det\left[\widetilde{\sum}^{-1}\sum\right] = \frac{\det\sum}{\det\widetilde{\sum}} = \frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^F+\sigma^2 I)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^R+\sigma^2 I)} = \frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)}$ where $\tilde{\lambda}_i^F$ and $\tilde{\lambda}_i^R$ are the eigenvalues of the covariance matrix $\sum$ and $\widetilde{\sum}$ respectively, then $-\log\left(\det\left[\widetilde{\sum}^{-1}\sum\right]\right) = -\log\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)} = \log\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)}$. Since $\sum$ and $\widetilde{\sum}$ are symmetric and by Hoffman - Weilandt inequality (Bhatia, 1997), there exists a permutation π of the indices $1,2,\ldots,n$ such that $\sum_{i=1}^{n}\left|\tilde{\lambda}_{\pi_{(i)}}^R - \tilde{\lambda}_i^F\right|^2 \leq \| \sum - \widetilde{\sum} \|_F^2 \leq \Delta^2$. For each i,

$$\left(\tilde{\lambda}_{\pi_{(i)}}^R - \tilde{\lambda}_i^F\right)^2 \leq \Delta^2 \tag{2.27}$$

$$(\tilde{\lambda}_i^F)^2\left(\frac{\tilde{\lambda}_{\pi_{(i)}}^R}{\tilde{\lambda}_i^F}-1\right)^2 \leq \Delta^2 \tag{2.28}$$

$$\left(\frac{\tilde{\lambda}_{\pi_{(i)}}^R}{\tilde{\lambda}_i^F}-1\right)^2 \leq \frac{\Delta^2}{(\tilde{\lambda}_i^F)^2} \leq \frac{\Delta^2}{(\sigma^2)^2} = \frac{\Delta^2}{\sigma^4}, \tag{2.29}$$

noting that $(\tilde{\lambda}_i^F)^2 \leq \sigma^4$. With the same permputation π, $\frac{\tilde{\lambda}_{\pi_{(i)}}^R}{\tilde{\lambda}_i^F} \in \left[1-\frac{\Delta}{\sigma^2}, 1+\frac{\Delta}{\sigma^2}\right]$. Using the idea that

$$-\log\left(\det\left[\widetilde{\sum}^{-1}\sum\right]\right) = -\log\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)} = \log\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)},$$

we have that $\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)} \leq (1+\frac{\Delta}{\sigma^2})^n$ impying that $\log\frac{\prod_{i=1}^{n}(\tilde{\lambda}_i^R)}{\prod_{i=1}^{n}(\tilde{\lambda}_i^F)} \leq n\log(1+\frac{\Delta}{\sigma^2})$ thus by using the fact that $\frac{x}{x+1} \leq \log(1+x) \leq x$ then

$$-\log\left(\det\left[\widetilde{\sum}^{-1}\sum\right]\right) \leq \frac{n\Delta}{\sigma^2}. \tag{2.30}$$

Combining equation 2.25 and 2.30 , we obtain

$$\begin{aligned} 2D_{KL}\Big(N_F \parallel N_R\Big) &= \frac{\sqrt{n}\Delta}{\sigma^2} + \frac{n\Delta}{\sigma^2} \\ &\leq \frac{2n\Delta}{\sigma^2} \\ &= \frac{2n\Delta}{\sigma^2}. \end{aligned}$$

Thus the Kulllback-leibler diverence is given as

$$D_{KL}(N_F \parallel N_R) = \frac{n\Delta}{\sigma^2}.$$

2.6.4 Proof of MSPEs under true and misspecified covariance matrix

Assume $Y(\mathbf{s})$ be a Gaussian process with mean 0 and a covariance function $c(.,.,\eta,\sigma^2)$. In geology and geographical statistics, the main focus is to predict $\hat{Y}(s*)$ at any given location $s*$ in the spatial domain D. This prediction relies on the concept of minimum mean squared error [54]. At any unobserved location $\mathbf{s}*$, the best linear unbiased prediction is given by

$$\hat{Y}(\mathbf{s}*) = \mathbf{c}_*^{\mathbf{T}} \sum\nolimits^{-\mathbf{1}} \mathbf{Y},$$

where $\mathbf{Y} = (Y(\mathbf{s_1}), Y(\mathbf{s_2}), ..., Y(\mathbf{s_n}))^T$, is the observation at $\mathbf{s_i}$ $i = 1, 2, ..., n$ locations, $\sum = C + \sigma^2\mathbf{I_n}$ and $\mathbf{c}_*$ is the covariance between all locations in the spatial domain D and the unobserved location $s*$; thus, $c_* = (c(s_1, s_*), \ldots, c(s_n, s_*))^T$. Under the true covariance

matrix, MSPE is given as;

$$\begin{aligned} MSPE(s*, C) &= E\Big(\hat{Y}(s*) - Y(s)\Big)^2 \\ &= E\Big((\hat{Y}(s*) - Y(s)\Big)^T \Big(E(\hat{Y}(s*) - Y(s)\Big) \\ &= E\Big(\tilde{Y}'(s_*)\tilde{Y}(s_*)\Big) - E\Big(\tilde{Y}'(s_*)Y(s_*)\Big) \\ &= C(s_*s_*) - E\Big(c_*^T \sum\nolimits^{-1} Y_n\Big) \\ &= C(s_*, s_*) - c_*^T \sum\nolimits^{-1} c_*. \end{aligned}$$

Under the reduced rank approximation or the misspecified covariance matrix, the mean square prediction error is given as;

$$\begin{aligned} MSPE(s*, \tilde{C}) &= E\Big(\hat{Y}(s*) - Y(s)\Big)^2 \\ &= E\Big((\hat{Y}(s*) - Y(s))^T (E(\hat{Y}(s*) - Y(s)\Big) \\ &= E\Big(c_*^T \widetilde{\sum}^{-1} Y_n - Y(s)\Big)\Big(c_*^T \widetilde{\sum}^{-1} Y_n - Y(s)\Big) \\ &= c_*^T \widetilde{\sum}^{-1} E\Big(Y^T(s)Y(s)\Big) \widetilde{\sum}^{-1} c_* - 2c_*^T \sum^{-1} E\Big(Y^T(s_*)Y(s)\Big) + E\Big(\tilde{Y}'(s_*)\tilde{Y}(s_*)\Big) \\ &= c_*^T \widetilde{\sum}^{-1} \sum \widetilde{\sum} c_* - 2c_*^T \sum \widetilde{\sum}^{-1} c_* + C(s_*s_*) \\ &= C(s_*, s_*) - 2c_*^T \widetilde{\sum}^{-1} c_* + c_*^T \widetilde{\sum}^{-1} \sum \widetilde{\sum}^{-1} c_*. \end{aligned}$$

2.6.5 Proof of Theorem 2.2.2

The upper bound of the difference between the mean square prediction error under the true covariance structure and the misspecified structure is given as;

$$\begin{aligned}
\left|MSPE_t(s*) - MSPE_m(s*)\right| &= \left|C(s_*, s_*) - c_*^T \sum^{-1} c_* - C(s_*, s_*) + 2c_*^T \widetilde{\sum}^{-1} c_* \right. && (2.31)\\
&\quad \left. - c_*^T \widetilde{\sum}^{-1} \sum \widetilde{\sum}^{-1} c_* \right| && (2.32)\\
&= \left|c_*^T M c_*\right| && (2.33)\\
&= \text{trace } (c_*^T M c_*) && (2.34)\\
&= \text{trace } (M c_* c_*^T) && (2.35)\\
&= < M c_* c_*^T > && (2.36)\\
&\leq \| M \|_F \| c_* c_*^T \|_F . && (2.37)
\end{aligned}$$

We know that, $\| \sum - \widetilde{\sum} \|_F \leq \Delta$, and $\widetilde{\sum}^{-1} = (C_{xx}^{lp} + \sigma^2 I)^{-1} = (R_m D_m R_m' + \sigma^2 I)^{-1} = \sigma^{-2} I - \sigma^{-4} I R_m (D_m^{-1} + \sigma^{-2} I)^{-1} R_m'$ and by Sherman–Morrison–Woodbury formula, we have the following; $\sum^{-1} = (C + \sigma^2 I)^{-1} = (R_m D_m R_m' + R_{n-m} D_{n-m} R_{n-m}' + \sigma^2 I)^{-1} = \sigma^{-2} I - \sigma^{-4} I R_m (D_m^{-1} + \sigma^{-2} I)^{-1} R_m' - \sigma^{-4} I R_{n-m} (D_{n-m}^{-1} + \sigma^{-2} I)^{-1} R_{n-m}'$. Also we note that $\|\widetilde{\sum}^{-1}\|_2$ is the largest singular value of inverse of $\widetilde{\sum}$ that is $\frac{1}{\lambda + \sigma^2}$. Now, let $M = 2\widetilde{\sum}^{-1} - \widetilde{\sum}^{-1} \sum \widetilde{\sum}^{-1} - \sum^{-1}$ and consider first bounding $\left\| \widetilde{\sum}^{-1} - \sum^{-1} \right\|_F$ and M.

Therefore

$$\begin{aligned}
\left\|\widetilde{\sum}^{-1} - \sum^{-1}\right\|_F &= \Big\|\sigma^{-2}I - \sigma^{-4}IR_m(D_m^{-1} + \sigma^{-2}I)^{-1}R_m'\sigma^{-2}I && (2.38)\\
&\quad -\sigma^{-4}IR_m(D_m^{-1} + \sigma^{-2}I)^{-1}R_m' && (2.39)\\
&\quad -\sigma^{-4}IR_{n-m}(D_{n-m}^{-1} + \sigma^{-2}I)^{-1}R_{n-m}'\Big\|_F && (2.40)\\
&= \left\|\sigma^{-4}IR_{n-m}(D_{n-m}^{-1} + \sigma^{-2}I)^{-1}R_{n-m}'\right\|_F && (2.41)\\
&= \sigma^{-4}\left\|R_{n-m}(D_{n-m}^{-1} + \sigma^{-2}I)^{-1}R_{n-m}'\right\|_F && (2.42)\\
&= \sigma^{-4}\sqrt{\text{trace}\left(R_{n-m}(D_{n-m}^{-1} + \sigma^{-2}I)^{-1}R_{n-m}'\right)^T} && (2.43)\\
&\quad \times\sqrt{\left(R_{n-m}(D_{n-m}^{-1} + \sigma^{-2}I)^{-1}R_{n-m}'\right)} && (2.44)\\
&= \sigma^{-4}\sqrt{\text{trace}\left(D_{n-m}^{-1} + \sigma^{-2}I\right)^{-2}} && (2.45)\\
&= \sigma^{-4}\sqrt{\sum_{i=m+1}^{n}\left(D_{n-m}^{-1} + \sigma^{-2}I\right)^{-2}} && (2.46)\\
&= \sigma^{-4}\sqrt{\sum_{i=m+1}^{n}\left(\frac{1}{\lambda_i} + \frac{1}{\sigma^2}\right)^{-2}} && (2.47)\\
&= \sigma^{-2}\sqrt{\sum_{i=m+1}^{n}\frac{\lambda_i^2}{(\sigma^2+\lambda_i)^2}} && (2.48)\\
&\leq \sigma^{-2}\sqrt{\sum_{i=m+1}^{n}\frac{\lambda_i^2}{\lambda_i^2}} && (2.49)\\
&= \frac{\sqrt{n}}{\sigma^2}. && (2.50)
\end{aligned}$$

Also,

$$
\begin{aligned}
\|M\|_F &= \left\| 2\widetilde{\sum}^{-1} - \widetilde{\sum}^{-1}\sum\widetilde{\sum}^{-1} - \sum^{-1} \right\|_F && (2.51)\\
&= \left\| \widetilde{\sum}^{-1} - \sum^{-1} + \widetilde{\sum}^{-1} - \widetilde{\sum}^{-1}\sum\widetilde{\sum}^{-1} \right\|_F && (2.52)\\
&= \left\| \widetilde{\sum}^{-1} - \sum^{-1} + \widetilde{\sum}^{-1}\sum\sum^{-1} - \widetilde{\sum}^{-1}\sum\widetilde{\sum}^{-1} \right\|_F && (2.53)\\
&= \left\| \widetilde{\sum}^{-1} - \sum^{-1} + \widetilde{\sum}^{-1}\sum\left(\widetilde{\sum}^{-1} - \sum^{-1}\right) \right\|_F && (2.54)\\
&= \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right)\left(I - \widetilde{\sum}^{-1}\sum\right) \right\|_F && (2.55)\\
&= \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right)\left(\widetilde{\sum}^{-1}\widetilde{\sum} - \widetilde{\sum}^{-1}\sum\right) \right\|_F && (2.56)\\
&= \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right)\widetilde{\sum}^{-1}\left(\widetilde{\sum} - \sum\right) \right\|_F && (2.57)\\
&\leq \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right) \right\|_F \left\| \widetilde{\sum}^{-1}\left(\widetilde{\sum} - \sum\right) \right\|_F && (2.58)\\
&\leq \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right) \right\|_F \left\| \widetilde{\sum}^{-1} \right\|_F \left\| \left(\widetilde{\sum} - \sum\right) \right\|_F && (2.59)\\
&\leq \left\| \left(\widetilde{\sum}^{-1} - \sum^{-1}\right) \right\|_F \left\| \widetilde{\sum}^{-1} \right\|_2 \left\| \left(\widetilde{\sum} - \sum\right) \right\|_F && (2.60)\\
&= \frac{\sqrt{n}}{\sigma^2} \times \frac{1}{\sqrt{\lambda + \sigma^2}} \times \Delta && (2.61)\\
&\leq \frac{\sqrt{n}}{\sigma^2} \times \frac{1}{\sigma} \times \Delta && (2.62)\\
&= \frac{\sqrt{n\Delta^2}}{\sigma^3}. && (2.63)
\end{aligned}
$$

Equation 2.61 is achieved by considering equation 2.50 and the upper bound for $\| \sum - \widetilde{\sum} \|_F$ and $\|\widetilde{\sum}^{-1}\|_2$. Now from equation 2.37, the upper bound of it is given as

$$
\begin{aligned}
\left| MSPE_t(s*) - MSPE_m(s*) \right| &\leq \| M \|_F \| c_* c_*^T \|_F && (2.64)\\
&= \frac{\sqrt{n\epsilon^2}}{\sigma^3} \| c_* c_*^T \|_F . && (2.65)
\end{aligned}
$$

To find the frobenius of $c_* c_*^T$, consider

$$\begin{aligned}
\|c_* c_*^T\|_F &= \sqrt{tr(c_* c_*^T)^T (c_* c_*^T)} && (2.66)\\
&= \sqrt{tr(c_* c_*^T c_* c_*^T)} && (2.67)\\
&= \sqrt{tr(c_*^T c_* c_*^T c_*)} && (2.68)\\
&= \sqrt{(\sum_{i=1}^{n} c_*^T c_*} && (2.69)\\
&= \sqrt{\left(\sum_{i=1}^{n} c_*^T c_*^2\right)} && (2.70)\\
&= c_1 + c_2 + ... + c_n && (2.71)\\
&= nc_1 && (2.72)\\
&\leq n\sigma^2. && (2.73)
\end{aligned}$$

Equation 2.73 is achieved by using the fact that $c_1 = cov(x, y)$ and that each c_i $i = 1, ..., n$ are identical thus;

$$\begin{aligned}
c_1 = cov(x, y) &= \sqrt{var(x) var(y)} && (2.74)\\
&= \sqrt{\sigma^4} && (2.75)\\
&= \sigma^2. && (2.76)
\end{aligned}$$

Plugging Equation 2.73 into Equation 2.65 we obtain
$\left|MSPE_t(s*) - MSPE_m(s*)\right| = \frac{\Delta\sqrt{n^3}}{\cdot}\sigma$

2.6.6 Proof of Proposition 2.2.2

For fixed sample size n and varying rank m,

$$\begin{aligned}
\lim_{m\to n}\frac{MSPE_m(s*)}{MSPE_t(s*)} &= \frac{C(s_*,s_*)-2c_*^T\widetilde{\sum}^{-1}c_*+c_*^T\widetilde{\sum}^{-1}\sum\widetilde{\sum}^{-1}c_*}{C(s_*,s_*)-c_*^T\sum^{-1}c_*}\\
&= \frac{C(s_*,s_*)-2c_*^T\sum^{-1}c_*+c_*^T\widetilde{\sum}^{-1}\sum\widetilde{\sum}-1c_*}{C(s_*,s_*)-c_*^T\sum^{-1}c_*}\\
&= \frac{C(s_*,s_*)-2c_*^T\sum^{-1}c_*+c_*^T\sum^{-1}c_*}{C(s_*,s_*)-c_*^T\sum^{-1}c_*}\\
&= \frac{C(s_*,s_*)-c_*^T\sum^{-1}c_*}{C(s_*,s_*)-c_*^T\sum^{-1}c_*}\\
&= 1.
\end{aligned}$$

This was achieved by using the idea of continuous mapping, that is; for $m\to n$, $\quad\widetilde{\sum}^{-1}\to\sum^{-1}$.

2.6.7 Additional Results on Simulation 1

Table 2.11: RMSE in eigenvalues and eigenvectors for Matérn with $\nu = 0.5$ and RSV of 90%

Sample Size	Target rank	RMSE.Ev[1]	RMSE.Evc [2]	Time Partial (Original) [3]
500	10	1.95E-0.5	3.23E-05	0.04 (0.23)
	20	0.005522	0.001011	0.06 (0.23)
	30	0.021483	0.00320	0.09 (0.23)
1000	20	0.002203	0.0000350	0.22 (1.75)
	40	0. 052937	0.007674	0.39 (1.75)
	60	0.047023	0.002964	0.55 (1.75)
2000	40	0.040736	0.003741	1.48 (13.99)
	80	0.055339	0.00530	3.14 (13.99)
	120	0.052318	0.007593	4.33 (13.99)
3000	60	0.064240	0.003469	4.71 (46.23)
	120	0.064130	0.005763	9.11 (46.23)
	180	0.068105	0.007297	13.81 (46.23)
4000	80	0.075530	0.002938	9.86 (109.39)
	160	0.069846	0.005801	21.30 (109.39)
	240	0.075465	0.006923	32.67 (109.39)
5000	100	0.071001	0.003144	20.40 (208.27)
	200	0.073155	0.005161	40.61 (208.27)
	300	0.0813788	0.006292	63.06 (208.27)

[1] RMSE.Ev is the RMSE for the difference in Eigenvalues.
[2] RMSE.Evc is the RMSE for the difference in Eigenvectors.
[3] Time Partial (Original) is time spent on the calculation when considering partial and full eigenvalues respectively.

Table 2.12: RMSE in eigenvalues and eigenvectors for Matérn with $\nu = 1.5$ and RSV of 90%

Sample Size	Target rank	RMSE.Ev[1]	RMSE.Evc [2]	Time Partial (Original) [3]
100	2	1.93E-12	2.16E-09	0.02 (0.10)
	4	6.11E-06	0.000115	0.04 (0.10)
	6	0.000143	0.000679	0.07 (0.10)
500	10	0.00133	0.000936	0.05 (0.25)
	20	0.112587	0.020857	0.06 (0.25)
	30	0.090984	0.026822	0.08 (0.25)
1000	20	0.124195	0.011982	0.20 (1.49)
	40	0.100422	0.019979	0.34 (1.49)
	60	0.0721447	0.022255	0.50 (1.49)
2000	40	0.148164	0.013136	1.25 (11.82)
	80	0.083292	0.015948	2.43 (11.82)
	120	0.057460	0.016918	3.52 (11.82)
3000	60	0.0125821	0.012082	3.97 (37.78)
	120	0.071521	0.013692	7.77 (37.78)
	180	0.050524	0.014303	11.70 (37.78)
4000	80	0.116656	0.010840	8.93 (89.18)
	160	0.064742	0.0121230	17.62 (89.18)
	240	0.045666	0.012550	27.28 (89.18)
5000	100	0.107766	0.010023	17.17 (172.02)
	200	0.060068	0.010984	34.30 (172.02)
	300	0.042321	0.011307	52.73 (172.02)

[1] RMSE.Ev is the RMSE for the difference in Eigenvalues.

[2] RMSE.Evc is the RMSE for the difference in Eigenvectors.

[3] Time Partial (Original) is time spent on the calculation when considering partial and full eigenvalues respectively.

2.6.8 Additional Results on Simulation 2

Table 2.13: Results of ratio of MSPEs at fixed N and varying M with an RSV of 70%

Sample Size	Rank ($\nu = 1.5$)	M.Ratio[1]	Rank ($\nu = 0.5$)	E.Ratio[2]
1000	13	1.007832	50	1.014111
	15	1.002448	60	1.008473
	17	1.001369	70	1.006558
	20	1.001200	80	1.005847
	25	1.00104	90	1.005205
2000	13	1.036543	50	1.037145
	15	1.004556	60	1.018247
	17	1.004461	70	1.010042
	20	1.002923	80	1.007384
	25	1.000912	90	1.005971
3000	13	1.105287	50	1.08447
	15	1.014915	60	1.02708
	17	1.012993	70	1.017968
	20	1.012744	80	1.014557
	25	1.001153	90	1.008477
4000	13	1.186012	50	1.25936
	15	1.025066	60	1.074737
	17	1.023595	70	1.047202
	20	1.023321	80	1.027155
	25	1.001945	90	1.014855
5000	13	1.454199	50	1.416339
	15	1.039332	60	1.143963
	17	1.038489	70	1.096908
	20	1.031912	80	1.04883
	25	1.003115	90	1.028716

[1] M.Ratio is ratio between the MPSEs for Matérn with $\nu = 1.5$.

[2] E.Ratio is ratio between the MPSEs for Matérn with $\nu = 0.5$.

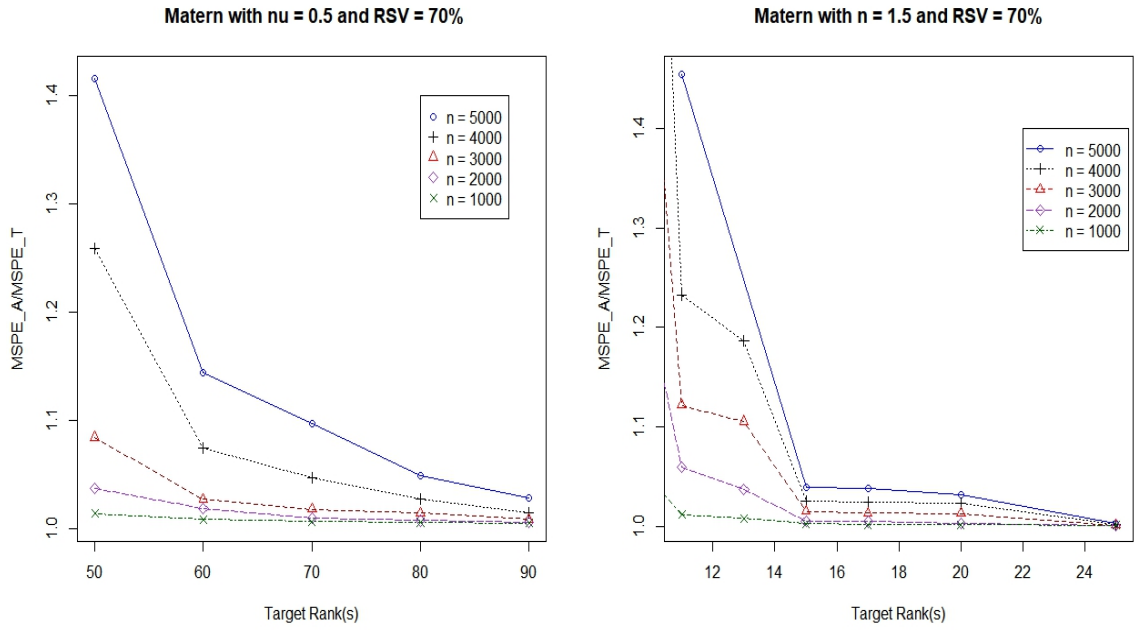

Figure 2.7: Results Graph Left: Matérn with $\nu = 0.5$ and RSV of 70% and Right: Matérn with $\nu = 1.5$ and RSV of 70%

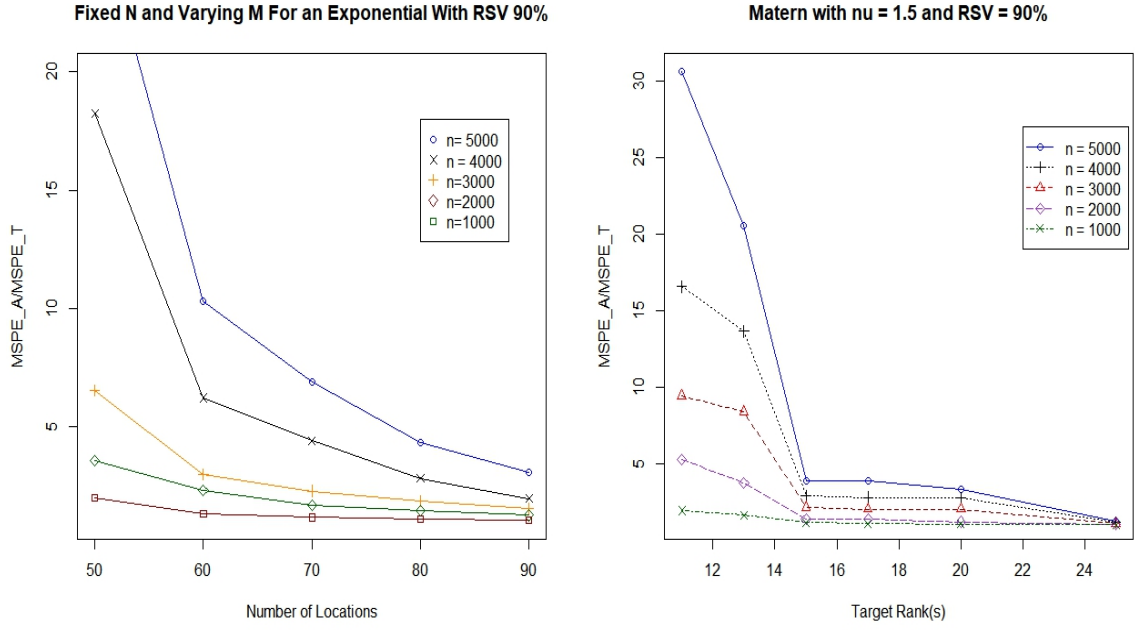

Figure 2.8: Results Graph Left: Matérn with $\nu = 0.5$ and an RSV of 90% and Right: Matérn with $\nu = 1.5$ and an RSV of 90%

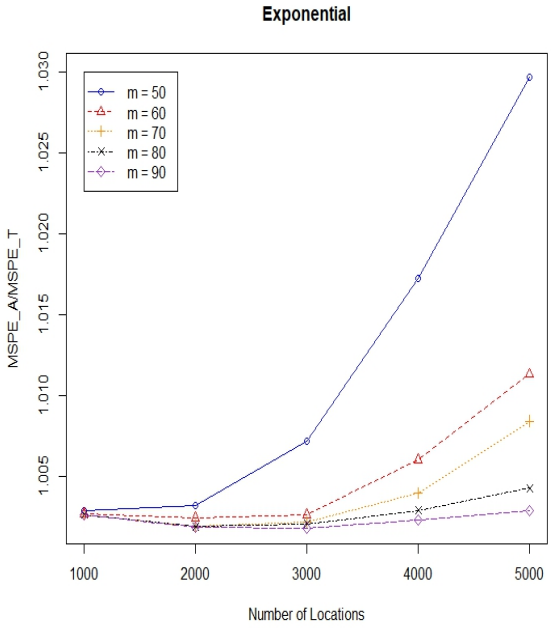

Figure 2.9: Results Graph: Matérn with $\nu = 0.5$ and an RSV of 50%

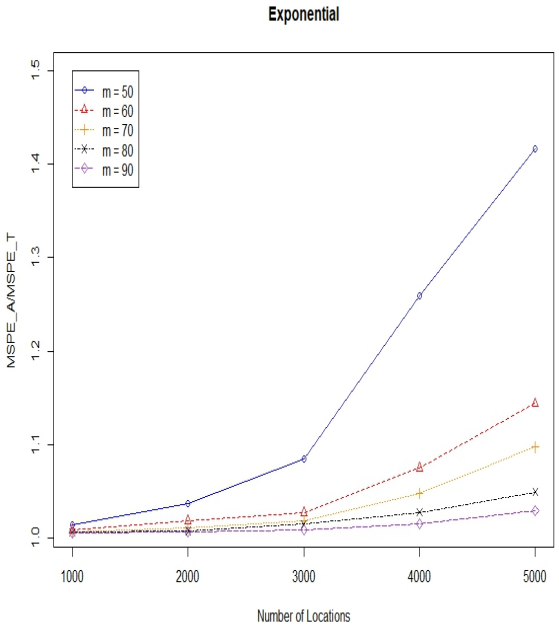

Figure 2.10: Results Graph: Matérn with $\nu = 0.5$ and an RSV of 70%

2.6.9 Additional Results on Simulation 3

Table 2.14: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 0.5$ and an RSV: 90%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000005	1.000222	1.033227	1.063252
	Rank	82	50	24	17
	Time.True[1]	0.02	0.02	0.02	0.02
	Time.Miss[2]	0.12	0.02	0.03	0.01
500	Ratio	1.000007	1.002097	1.116252	1.500523
	Rank	315	119	49	33
	Time.True[1]	0.21	0.21	0.21	0.21
	Time.Miss [2]	2.2	0.62	0.22	0.16
1000	Ratio	1.00001	1.003476	1.081364	1.44597
	Rank	519	161	66	46
	Time.True[1]	1.38	1.38	1.38	1.38
	Time.Miss [2]	12.15	2.75	1.01	0.78
2000	Ratio	1.000024	1.005936	1.392816	2.009402
	Rank	772	216	89	63
	Time.True[1]	12.92	12.92	12.92	12.92
	Time.Miss[2]	67.25	13.39	5.23	3.65
3000	Ratio	1.000005	1.003679	1.450703	2.448229
	Rank	870	270	112	78
	Time.True[1]	40.85	40.85	40.85	40.85
	Time.Miss [2]	153.11	35.91	14.31	10.11
4000	Ratio	1.000003	1.006319	1.353902	2.710149
	Rank	889	320	132	91
	Time.True[1]	94.43	94.43	94.43	94.43
	Time.Miss[2]	249.97	74.47	27.41	19.15
5000	Ratio	1.000004	1.010529	1.252496	2.655185
	Rank	888	365	152	104
	Time.True [1]	174.70	174.70	174.70	174.70
	Time.Miss [2]	382.31	122.79	48.38	33.18

Table 2.15: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1$ and : RSV: 90%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000312	1.024681	1.351251	14.08195
	Rank	16	8	5	4
	Time.True[1]	0.00	0.00	0.00	0.00
	Time.Miss [2]	0.02	0.03	0.02	0.05
500	Ratio	1.000767	1.013961	2.131363	2.161326
	Rank	27	14	8	7
	Time.True [1]	0.13	0.13	0.13	0.13
	Time.Miss[2]	0.12	0.14	0.09	0.08
1000	Ratio	1.001281	1.065786	6.286708	6.291397
	Rank	34	17	10	9
	Time.True[1]	0.99	0.99	0.99	0.99
	Time.Miss[2]	0.70	0.61	0.56	0.61
2000	Ratio	1.00132	1.116296	3.858155	25.29093
	Rank	42	21	13	10
	Time.True[1]	9.25	9.25	9.25	9.25
	Time.Miss[2]	3.45	2.90	2.77	2.64
3000	Ratio	1.003384	1.074725	2.090425	8.228058
	Rank	47	23	15	12
	Time.True[1]	31.92	31.92	31.92	31.92
	Time.Miss[2]	10.25	9.06	8.77	8.57
4000	Ratio	1.003649	1.136181	2.803713	13.18204
	Rank	52	25	16	13
	Time.True[1]	78.86	78.86	78.86	78.86
	Time.Miss[2]	16.95	14.35	13.31	12.97
5000	Ratio	1.002465	1.222514	3.51821	4.630569
	Rank	56	27	17	14
	Time.True [1]	153.8	153.8	153.8	153.8
	Time.Miss[2]	31.44	27.14	24.78	25.00

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.16: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1.5$ and : RSV: 50%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000009	1.009884	1.019341	1.019341
	Rank	7	4	2	2
	Time.True[1]	0.00	0.00	0.00	0.00
	Time.Miss[2]	0.06	0.03	0.03	0.03
500	Ratio	1.000593	1.000737	1.001442	1.634426
	Rank	10	7	5	4
	Time.True[1]	0.23	0.23	0.23	0.23
	Time.Miss[2]	0.03	0.05	0.12	0.08
1000	Ratio	1.000039	1.004345	1.038748	1.038748
	Rank	15	9	5	5
	Time.True[1]	1.15	1.15	1.15	1.15
	Time.Miss [2]	0.31	0.30	0.26	0.24
2000	Ratio	1.000222	1.002477	1.025798	1.060442
	Rank	16	11	7	5
	Time.True[1]	10.26	10.26	10.26	10.26
	Time.Miss[2]	1.27	1.11	1.03	1.14
3000	Ratio	1.000807	1.006814	1.065251	1.070792
	Rank	15	12	7	6
	Time.True [1]	34.39	34.39	34.39	34.39
	Time.Miss [2]	2.58	2.50	2.14	2.05
4000	Ratio	1.001132	1.012287	1.121105	1.123732
	Rank	20	13	8	7
	Time.True[1]	80.11	80.11	80.11	80.11
	Time.Miss [1]	5.50	4.94	4.48	4.31
5000	Ratio	1.000126	1.003285	1.196391	1.201402
	Rank	25	14	9	7
	Time.True [1]	149.74	149.74	149.74	149.74
	Time.Miss [2]	9.23	7.80	7.09	6.77

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.17: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1.5$ and : RSV: 70%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000041	1.02695	1.02695	1.02695
	Rank	5	2	2	2
	Time.True[1]	0.00	0.00	0.00	0.00
	Time.Miss [2]	0.03	0.01	0.03	0.02
500	Ratio	1.000041	1.000282	1.368246	1.373425
	Rank	7	5	3	2
	Time.True[1]	0.13	0.13	0.13	0.13
	Time.Miss[2]	0.08	0.08	0.08	0.08
1000	Ratio	1.000246	1.01041	2.621826	2.622457
	Rank	8	5	4	3
	Time.True[1]	0.95	0.95	0.95	0.95
	Time.Miss[2]	0.40	0.41	0.39	0.34
2000	Ratio	1.001479	1.013525	8.146825	8.146825
	Rank	9	5	4	4
	Time.True[1]	9.13	9.13	9.13	9.13
	Time.Miss [2]	1.48	1.39	1.37	1.35
3000	Ratio	1.00403	1.005187	1.018616	17.50018
	Rank	10	6	5	4
	Time.True[1]	31.42	31.42	31.42	31.42
	Time.Miss [2]	2.91	2.65	2.61	2.60
4000	Ratio	1.000399	1.0083	1.025799	30.68005
	Rank	11	7	5	4
	Time.True [1]	73.31	73.31	73.31	73.31
	Time.Miss[2]	5.91	5.48	5.31	5.36
5000	Ratio	1.000762	1.013901	1.035083	1.035083
	Rank	11	7	5	5
	Time.True [1]	143.02	143.02	143.02	143.02
	Time.Miss[2]	10.36	9.84	9.48	9.50

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Table 2.18: Ratio of the MSPE under the True and Misspecified covariance matrix for Matérn with $\nu = 1.5$ and : RSV: 90%

		Tolerance Level = Δ			
Sample		$\Delta_1 = 0.001F^3$	$\Delta_2 = 0.01F^3$	$\Delta_3 = 0.05F^3$	$\Delta_4 = 0.1F^3$
100	Ratio	1.000157	1.448038	2.92835	2.92835
	Rank	6	4	2	2
	Time.True[1]	0.00	0.00	0.00	0.00
	Time.Miss [2]	0.00	0.02	0.01	0.00
500	Ratio	1.004585	1.023504	25.44646	25.44646
	Rank	9	5	4	4
	Time.True[1]	0.18	0.18	0.18	0.18
	Time.Miss[2]	0.07	0.07	0.08	0.06
1000	Ratio	1.005641	1.029703	1.704783	106.4379
	Rank	11	7	5	4
	Time.True [1]	1.31	1.31	1.31	1.31
	Time.Miss[2]	0.50	0.45	0.42	0.44
2000	Ratio	1.006857	1.139147	1.902936	1.902936
	Rank	12	8	5	5
	Time.True[1]	11.66	11.66	11.66	11.66
	Time.Miss[2]	2.94	2.86	2.79	2.78
3000	Ratio	1.001665	1.338112	1.415787	2.2502
	Rank	14	9	6	5
	Time.True[1]	38.75	38.75	38.75	38.75
	Time.Miss[2]	9.11	8.87	8.76	8.62
4000	Ratio	1.002853	1.628332	1.784121	2.733497
	Rank	15	9	6	5
	Time.True[1]	89.78	89.78	89.78	89.78
	Time.Miss[2]	9.86	9.20	8.90	8.79
5000	Ratio	1.005162	2.009171	2.040137	3.347959
	Rank	15	9	7	5
	Time.True[1]	173.05	173.05	173.05	173.05
	Time.Miss[2]	17.89	16.84	16.43	16.25

[1] Time.True[1] is the time to calculate the MPSE under the true covariance matrix.

[2] Time.Miss[2] is the time to calculate the MPSE under the misspecified covariance matrix.

[3] F is the first eigenvalue of the true covariance matrix.

Bibliography

[1] Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Institute of Mathematical Statistics.

[2] Arango Argoti, M. A. (2013). *Nitrous oxide emissions: measurements in corn and simulations at field and regional scale*. PhD thesis, Kansas State University.

[3] Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex.

[4] Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2012). Efficient Gaussian process regression for large datasets. *Biometrika*, 100(1):75–89.

[5] Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

[6] Becker, R. (2018). *The new S language*. CRC Press.

[7] Bloemerts, M. and De Vries, W. (2009). *Relationships between nitrous oxide emissions from natural ecosystems and environmental factors*, volume 1853. Alterra, Wageningen.

[8] Brown, B., Chui, M., and Manyika, J. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly*, 4(1):24–35.

[9] Capotondi, A., Wittenberg, A., and Masina, S. (2006). Spatial and temporal structure of tropical pacific interannual variability in 20th century coupled simulations. *Ocean Modelling*, 15(3-4):274–298.

[10] Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons.

[11] Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

[12] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

[13] De Vries, W., Butterbach-Bahl, K., Denier van der Gon, H., and Oenema, O. (2007). The impact of atmospheric nitrogen deposition on the exchange of carbon dioxide, nitrous oxide and methane from european forests. *Greenhouse Gas Sinks*.

[14] Du, J., Zhang, H., Mandrekar, V., et al. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Annals of Statistics*, 37(6A):3330–3361.

[15] Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, 53(8):2873–2884.

[16] Franses, P. H. (1996). Periodicity and stochastic trends in economic time series. *OUP Catalogue*.

[17] Franses, P. H. (1998). *Time series models for business and economic forecasting*. Cambridge university press.

[18] Franses, P. H. and Paap, R. (1994). Model selection in periodic autoregressions. *Oxford Bulletin of Economics and Statistics*, 56(4):421–439.

[19] Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

[20] Getis, A. and Ord, J. K. (1996). Local spatial statistics: an overview. *Spatial Analysis: Modelling in a GIS Environment*, 374:261–277.

[21] Giltrap, D. L., Li, C., and Saggar, S. (2010). Dndc: a process-based model of greenhouse gas fluxes from agricultural soils. *Agriculture, Ecosystems and Environment*, 136(3-4):292–300.

[22] Gladyshev, E. (1963). Periodically and almost-periodically correlated random processes with a continuous time parameter. *Theory of Probability and its Applications*, 8(2):173–177.

[23] Griffis, T. J., Chen, Z., Baker, J. M., Wood, J. D., Millet, D. B., Lee, X., Venterea, R. T., and Turner, P. A. (2017). Nitrous oxide emissions are enhanced in a warmer and wetter world. *Proceedings of the National Academy of Sciences*, 114(45):12081–12085.

[24] Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

[25] Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*. Springer.

[26] Jones, R. H. and Brelsford, W. M. (1967). Time series with periodic structure. *Biometrika*, 54(3-4):403–408.

[27] Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15(23):2539–2560.

[28] Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18.

[29] Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

[30] Laird, N. M., Ware, J. H., et al. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

[31] Lee, J., De Gryze, S., and Six, J. (2011). Effect of climate change on field crop production in california's central valley. *Climatic Change*, 109(1):335–353.

[32] Li, H., Qiu, J., Wang, L., and Yang, L. (2011). Advance in a terrestrial biogeochemical model—dndc model. *Acta Ecologica Sinica*, 31(2):91–96.

[33] Lund, R. and Li, B. (2009). Revisiting climate region definitions via clustering. *Journal of Climate*, 22(7):1787–1800.

[34] Lund, R., Poplin, C., and McCarthy, K. (1995). Preliminary analysis of the interrelationships of some paleozoic actinopterygii. *Geobios*, 28:215–220.

[35] Luo, G., Kiese, R., Wolf, B., and Butterbach-Bahl, K. (2013). Effects of soil temperature and moisture on methane uptake and nitrous oxide emissions across three different ecosystem types. *Biogeosciences*, 10(5):3205–3219.

[36] Martin, N. and Maes, H. (1979). *Multivariate analysis*. Academic press London.

[37] McLeod, A. I. and Li, W. K. (1983). Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, 4(4):269–273.

[38] Mearns, L. O., Arritt, R., Biner, S., Bukovsky, M. S., McGinnis, S., Sain, S., Caya, D., Correia Jr, J., Flory, D., Gutowski, W., et al. (2012). The north american regional climate change assessment program: overview of phase i results. *Bulletin of the American Meteorological Society*, 93(9):1337–1362.

[39] Merriam, D., L. Brady, L., and David Newell, K. (2011). Kansas energy sources: A geological review. *Natural Resources Research*, 21.

[40] Merriam, D. F. (2009). Kansas energy, environment, and conservation: a geological overview. *Environmental Geology*, 56(8):1697–1706.

[41] Noakes, D. J., McLeod, A. I., and Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, 1(2):179–190.

[42] Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System*, 1(4):335–358.

[43] Openshaw, S. and Perree, T. (1996). User centred intelligent spatial analysis of point data. *Innovations in GIS*, 3:119–134.

[44] Pagano, M. et al. (1978). On periodic and multiple autoregressions. *The Annals of Statistics*, 6(6):1310–1317.

[45] Pandya, A. S. and Macy, R. B. (1995). *Pattern recognition with neural networks in C++*. CRC press.

[46] Pathak, H., Li, C., and Wassmann, R. (2005). Greenhouse gas emissions from indian rice fields: calibration and upscaling using the dndc model. *Biogeosciences*, 2(2):113–123.

[47] Penalba, O. C. and Robledo, F. A. (2010). Spatial and temporal variability of the frequency of extreme daily rainfall regime in the la plata basin during the 20th century. *Climatic Change*, 98(3-4):531–550.

[48] Prather, M. (1995). Other trace gases and atmospheric chemistry. *Climate Change*.

[49] Raziei, T., Daryabari, J., Bordi, I., and Pereira, L. S. (2014). Spatial patterns and temporal trends of precipitation in iran. *Theoretical and Applied Climatology*, 115(3-4):531–540.

[50] Reay, D. S., Davidson, E. A., Smith, K. A., Smith, P., Melillo, J. M., Dentener, F., and Crutzen, P. J. (2012). Global agriculture and nitrous oxide emissions. *Nature Climate Change*, 2(6):410.

[51] Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

[52] Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science*. IEEE.

[53] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

[54] Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

[55] Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19.

[56] Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566.

[57] Walthall, C. L., Anderson, C. J., Baumgard, L. H., Takle, E., and Wright-Morton (2013). *Climate Change and Agriculture in the United States: Effects and adaptation*.

[58] Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86(4):815–829.

[59] Williams, C. K. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*.