Modelling spatial dependence using extensions of the Poisson distribution

Charl Arthur Henry Cowley

Modelling spatial dependence using extensions of the Poisson distribution

by

Charl Arthur Henry Cowley

Submitted in partial fulfilment of the requirements for the degree

MSc (Advanced Data Analytics)

In the Department of Statistics

In the Faculty of Natural and Agricultural Science

University of Pretoria

Pretoria

Supervisor: Dr A. de Waal

2021

ABSTRACT

When modelling univariate count data, the Poisson distribution is a popular choice that is routinely studied by academics and applied by practitioners. It does not, however, allow for the modelling of dependencies found in real-world datasets. The Poisson distribution is particulary insufficient when modelling overdispersed and spatially dependent data. It is for this reason that extensions of the Poisson distribution that are known to perform well in these two areas are considered. Poisson mixture regression is effective at modelling overdispersed data and Gaussian Process/Kriging is a well-known method for capturing spatial dependence. A framework is created within which exploratory spatial metrics are categorised. Model accuracy is evaluated in terms of model fit through a residual analysis and Mean-Square Error (MSE) evaluation. The model's ability to capture spatial dependence is evaluated with a confusion matrix. This gives us a range of tools to assess in what manner an extension outperform its counterparts. We then decide which of the Poisson mixture regression and Gaussian Process/Kriging models achieve the best performance on a dataset with given spatial characteristics. Expansions to the exploratory spatial framework, modelling techniques and accuracy measures that are not considered here, are also suggested for further work.

DECLARATION

I, Charl Arthur Henry Cowley declare that the mini-dissertation, which I hereby submit for the degree MSc Advanced Data Analytics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: 

DATE: October 2021.

ACKNOWLEDGEMENTS

Firstly, I would like to dedicate this dissertation to my late grandparents, Manie and Sue du Preez. Words fail attempting to describe the magnitude of their loss.

Secondly, I would like to extend my greatest thanks to my employer, Lightstone, for the financial support to complete this degree. I am also grateful that I could use proprietary data in one of my applications. It has made so many of the exciting things I have learned over this degree and dissertation unbelievably real and tangible. I also want to thank some of my closest colleagues. Sam Viljoen - for your wisdom, calm demeanour and positive mindset. No problem is ever too big. Santelle Allgeier - I hold your attention to detail and ability to challenge even the strictest assumptions in the highest regard. Paul-Roux de Kock - for being our fearless leader and encouraging us to always pursue new ways of thinking. Finally, Lawrence Tjatjie - our newest member. You have been a timely addition to our team. Your eagerness to learn has helped me immensely over the last few months and I look forward to see what your future as Data Scientist holds.

Next, I want to acknowledge the Department of Statistics at UP for an immense course. Profs. Andriette Bekker and Inger Fabris-Rotelli have been extremely accommodating towards part-time students and their compassion throughout has been heartwarming. Drs. Frans Kanfer and Sollie Millard have always inspired me and their lectures completely rewrote the rule book of how a statistician can approach the daunting mathematics behind a new algorithm. I will always continue to look for areas where I can apply your thinking.

My biggest thanks goes out to my supervisor, Dr. Alta de Waal. I am immensely grateful to have found a supervisor who's philosophy of work aligns so well with own. Our regular check-ins have been beacons to chase in the dark and your inputs into this mini-dissertation have been invaluable. I hope there will be more to come.

Next, I want to thank my parents, Henry and Marleen Cowley, for instilling the importance of

reading and pursuing difficult goals from a very young age. It has proven to be a solid foundation for my life. My in-laws, Petrus and Tinkie van Staden, have been my biggest cheerleaders throughout the process and I am immensely grateful for your belief in me. My brothers and sisters, Gerhard and Marlise Jordaan and Theunis and Arista Botha have been great soundboards over this time and you all inspire me in different and wonderful ways.

I also want to recognise two friends who have inspired me beyond belief. Ivan Basson - my best man and the person in my life with the greatest reservoir of perseverance. Anika Wessels - my study buddy. It has been 10 years since I found a study companion in the first year Statistics class. I am always amazed at how you find ways to understand the things that simply go over my head and then explain it in a way that makes it all so obvious. It has been an amazing journey.

My biggest thanks is extended to my wife, Nadia Cowley. Your love, devotion and patience has been a rock throughout this degree. All this while being an inspirational business woman, teacher, cat mom and artist. You never cease to amaze me.

All that is left to say, is:

$\mathcal{S}.\mathcal{D}.\mathcal{G}.$

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is spatial data?

Spatial data consist of spatial and non-spatial components. The spatial component informs the size, shape and location of a spatial entity. These take on the form of latitude, longitude, elevation and polygon shape [15]. The non-spatial component provides non-spatial characterisations of the spatial entity such as the name of the spatial entity and other observed data [72].

Spatial data is therefore viewed as realisations of a stochastic process, $Z(s) : s \in D$ where s is a spatial entity and D is a random set in the spatial framework. Spatial data is often thought of as either point pattern, lattice or gesostatistical data . Point pattern data are often represented by natural phenomena, such as the position of migrating animals and are modelled by comparing the observed point patterns with random patterns generated by statistical processes. Lattice data consider the spatial entities in a spatial framework as a network of contiguous or neighbouring units [72]. Finally, geostatistical data are mostly concerned with the analysis of spatial continuity and weak stationarity [20] and provides some of the geostatistical tools that will be studied in this mini-dissertation.

1.2 Why spatial data?

Since spatial data consist of spatial and non-spatial components, spatial data tell us where data are located. Waldo Tobler stated in his First Law of Geography: "All things are related. Things that are nearby are more similar than things that are further away" [82] and thereby introduced the intuition

for spatial dependence. Practitioners and researchers studying concepts as wide and varied as the spread of diseases [52], the distribution of mining deposits [50], the movements of rhino poachers [79], the prevalence of crime [53, 69], the impact of a severe fire [76], the estimation of house prices [63] and even the success of a football team [11] rely heavily on spatial dependence to inform their decisions. At its core, the modelling of spatial dependence strives to use the spatial proximity of observations to not only tell us where things are located as accurately as possible, but also quantify the impact of this spatial "nearness".

1.3 Exploring spatial data and the Poisson distribution

In the same way that an explorer uses a map to navigate the world, a researcher and practitioner needs a guide to inform his/her modelling efforts. Exploratory Data Analysis (EDA), is an important component of the modelling process, regardless of the type of data [92]. The process that informs this in a spatial context, is called Exploratory Spatial Data Analysis (ESDA). Many textbooks have been written that discuss methods to explore spatial data [3, 9]. Another important metric - besides spatial dependence - to consider when confronted with a spatial count data modelling problem is dispersion. Dispersion measures the ratio of the variance of a variable to its mean.

Count data is often used in a spatial context. Botanists are concerned with the prevalence of certain tree species in sections of a forest and police forces measure the number of crimes in suburbs. As such, the Poisson distribution is a prime candidate to use in modelling these outcomes, since it considers the number of count outcomes in a given time period for a random variable, X.

If $X \sim Poisson(\lambda)$, its probability mass function (pmf) is given by

$$f(x) = \frac{\lambda^x \exp(-\lambda)}{x!}, x = 0, 1, 2, ..., \tag{1.1}$$

and

$$E(X) = var(X) = \lambda. \tag{1.2}$$

However, the Poisson distribution has a few notable shortcomings when modelling real-world geostatistical count data. From Eq. 1.2 it is observed that the Poisson distribution assumes an equal mean and variance and it cannot accurately fit overdispersed data. There is also nothing in its functional form that enables it to account for spatial dependence.

It is for this reason that we will study extensions of the Poisson distribution that have proven their ability to accurately model overdispersed data and effectively capture spatial dependence.

1.4 Extensions of the Poisson distribution

The Gaussian Process is a powerful, non-parametric Bayesian method which uses a covariance function as its defining component. A covariance function is related to a semivariogram which visualises the relation of variance to the distance between spatial observations. The semivariogram is in turn used as an input into the well-known Ordinary Kriging model, which produces a Best Linear Unbiased Estimator (BLUE) [19]. These models act as interpolators of data and have the added benefit of providing a variance with each estimate it produces. Since the variance is not restricted to a single value and changes in relation to the distance between spatial entities it is effective at fitting overdispersed data.

A Poisson Mixture Regression model is a supervised analogue to the Poisson Mixture Model (PMM) which has shown to be effective in modelling overdispersed data in other contexts such as text data [18, 46]. Despite not considering spatial dependence through a covariance function, it manages to capture the spatial autocorrelation structure as a result of the model-based clustering that is performed. This makes it a model worth considering for our research question.

1.5 Research question

Given a set of ESDA metrics as set out it in a framework, which extension of the Poisson distribution works best for modelling spatial dependence in a given count data set when evaluated

against a framework of accuracy metrics?

1.6 Methodology

This research question is answered by creating a framework within which exploratory spatial metrics are categorised. Categories are defined for number of observations and variables, dispersion and spatial autocorrelation. Category thresholds are defined in Chapter 5. A tabular representation of the ESDA framework is given in Table 1.1. The purpose of this framework is to give the modeller an overview of the dataset which will enable more informed modelling decisions. For example, a model with extremely low spatial autocorrelation will not require a model such as a Kriging/GP model that considers spatial autocorrelation explicitly.

ESDA Framework		**Category**				
Number of variables		Very Low	**Low**	Medium	High	Very High
Number of observations		Very Low	**Low**	Medium	High	Very High
Dispersion		Very Low	Low	**Medium**	High	Very High
Spatial autocorrelation	Global	Very Low	Low	**Medium**	High	Very High
	Local	Very Low	Low	**Medium**	High	Very High
	Bivariate	Very Low	**Low**	Medium	High	Very High

Table 1.1: ESDA framework

An example data set with categorised ESDA metrics. The data set has Low dimensionality (variables and observations), Medium dispersion, Global and Local spatial autocorrelation and Low Bivariate spatial autocorrelation.

A Poisson Generalised Linear Model (GLM) is used as a baseline model and its performance is compared to the Poisson Mixture Regression and Poisson Gaussian Process (GP)/Kriging models.

Model accuracy is evaluated in terms of model fit through a residual analysis and Mean-Square Error (MSE) evaluation. The model's ability to capture spatial dependence is evaluated with a confusion matrix. In Chapter 2 the concepts of spatial autocorrelation and a Local Indicator of Spatial Association (LISA) will be discussed. When spatially autocorrelated spatial entities group together, LISA clusters are formed. Model predictions can also be grouped into LISA clusters. The LISA clusters of the actual values are compared to the LISA clusters of the fitted values

with a confusion matrix. This gives a quick overview of where the model does not capture the spatial autocorrelation effectively by considering measures such as specificity. The purpose of the model evaluation framework is to give the modeller a range of tools to assess in what manner an extension outperforms its counterparts. We then decide which of these extensions achieves the best performance on a dataset with the given set of spatial characteristics. Further extensions to the exploratory spatial framework, modelling techniques and accuracy measures are also suggested. A tabular example of this model evaluation framework is given in Table 1.2.

Model evaluation framework		**Model**		
Model fit	Residual Analysis	GLM	**Mixture**	GP/Kriging
	Fitted values	GLM	**Mixture**	GP/Kriging
Spatial dependence structure	Global	GLM	Mixture	**GP/Kriging**
	Local	GLM	Mixture	**GP/Kriging**
	Bivariate	GLM	Mixture	**GP/Kriging**

Table 1.2: Model evaluation framework

An example of our model evaluation framework. This example shows that the Mixture and Kriging models outperformed the Poisson GLM in all categories of evaluation, with the Mixtures performing best in terms of model fit while the Kriging model captured the spatial dependence structure best.

1.7 Contribution

There are three main deliverables that are contributed to the existing body of knowledge:

1. An ESDA framework that enables practitioners/researchers to characterise spatial datasets according to ESDA metrics.

2. A modelling evaluation framework that enables practitioners/researchers to evaluate and rank the accuracy of competitor spatial dependence models.

3. A comprehensive code base from which ESDA and modelling efforts can be commenced and built upon.

4. Research article, Modelling spatial dependence using extensions of the Poisson distribution, submitted to the Southern African Conference for AI Research.

1.8 Structure

ESDA and our chosen metrics, dispersion and spatial autocorrelation will be discussed in Chapter 2. Gaussian Processes/Kriging will be explored in Chapter 3 and Poisson Mixture Regression will follow in Chapter 4. Related work will be reviewed in the individual chapters, rather than in a chapter of its own. We will apply the ESDA metrics and models to two datasets in Chapter 5 with the goal to answer the research question stated in Section 1.5. All our findings will be summarised in Chapter 6 along with suggestions for further work.

Chapter 2

Exploratory Spatial Data Analysis

2.1 Introduction

When confronted by a new dataset for a regression problem, whether it spatial or otherwise, researchers and/or practitioners first need to understand the structure of the data by conducting an Exploratory Data Analysis (EDA). This enables them to make informed decisions regarding the definition of the dependent variable - or dependent variables in a multivariate problem. They will also have a better understanding of how to address problems such as multi-collinearity between independent variables and decide whether dimensionality reduction techniques such as Principal Component Analysis and Factor Analysis are necessary for datasets with high-dimensionality. The creation of new variables - or feature engineering as it is commonly known in the Machine Learning (ML) community - is also informed by the EDA. The assumptions that need to be satisfied for a certain modelling technique to be applicable to a given problem can also be tested. Finally, an accuracy measure is also often informed by the EDA. From the EDA it will become apparent whether it is more important to optimise for accuracy by using a small error value or rather accommodate other biases that exist in the data.

The principles of the EDA extend to spatial data and is called Exploratory Spatial Data Analysis (ESDA). Many books have been written on the topic [3, 9]. This chapter will set out the elements of the ESDA that are important to our research question and lead us into the modelling techniques that will be used to model count data in a spatial context. We will consider dispersion

in Section 2.2 and spatial autocorrelation in Section 2.3 .

2.2 Dispersion

In this section, dispersion is defined in terms of absolute and relative measures. The most important relative measure of dispersion for our purposes, the variance to mean ratio, is discussed in Section 2.2.1. It is a widely used dispersion measure, but we consider other measures that have been developed to address some of the shortcomings of the variance to mean ratio in Section 2.2.2.

2.2.1 Variance to mean ratio

Dispersion measures the extent to which values in a variable are spread around a central value or over the variable's domain [28]. It is measured by absolute and/or relative measures. Absolute measures calculate dispersion in the original units of the observed variable. The most popular absolute measures are the range, quartile deviation, mean deviation and standard deviation. The range and quartile deviation are based upon the spread of values and the mean and standard deviation measures consider variation around the mean. A problem arises when using absolute measures since dispersion among variables cannot be compared when their mean values differ widely [28].

Each of the absolute measures mentioned above have relative analogues. The relative dispersion analogue of standard deviation is the variance to mean ratio, δ. Consider a random variable X with observed values $X_1, ..., X_n$ with a sample mean μ and sample standard devation σ [28]. The dispersion of X is calculated as

$$\delta = \frac{\sigma^2}{\mu}. \tag{2.1}$$

X is said to be *overdispersed* when $\delta > 1$ and *underdispersed* when $\delta < 1$. As with all relative measures of dispersion it can be used to compare variables with widely differing means.

Consider a Poisson random variable, $Y \sim Poisson(\lambda)$, with mean, $E(Y) = \lambda$, and variance, $var(Y) = \lambda$. Since the theoretical variance is equal to the theoretical mean, the theoretical variance to mean ratio of Y is equal to 1. When the variance to mean ratio is over- or underdispersed,

a Poisson distribution will not fit the data well. However, the Poisson Mixture Model (PMM) has been shown to model overdispersed data effectively [18, 46].

2.2.2 Other dispersion measures

Despite being a popular relative analogue dispersion measure of standard deviation, the variance to mean ratio does have a potential shortcoming in that non-random patterns can give rise to index values of 1 and incorrectly indicate perfect dispersion [45]. Other dispersion measures have been developed to address this problem.

Another relative dispersion measure is Green's coefficient, I_G, which is used when data is shown to be overdispersed by the variance to mean ratio. It is defined as

$$I_G = \frac{\frac{\sigma^2}{\mu} - 1}{\sum x - 1}. \tag{2.2}$$

Green's coefficient does not have a defined sampling distribution which makes it difficult to test its statistical robustness by assigning confidence limits [39].

The Morisita Index, I_d, contrastingly, does have a sampling distribution and is defined as

$$I_d = n\left(\frac{\sum x^2 - \sum x}{(\sum x)^2 - \sum x}\right) \tag{2.3}$$

where n indicates the sample size [62].

The null hypothesis that the dispersion occurred randomly is tested by

$$\chi^2 = I_d\left(\sum x - 1\right) + n - \sum x \tag{2.4}$$

with $n - 1$ degrees of freedom [62].

I_d is analagous to Lloyd's Index of Mean Crowding [56] who also found that when measures of dispersion are used in a spatial context the measures depend on the size of the spatial entities. In real-world datasets, polygon shapes and sizes are often defined by political, socio-economic and/or geographical boundaries. The work of Wieczorek et al. [93] showed that the different sizes

seen in real-world datasets can effectively capture underlying global characteristics. In other instances local nuances are lost when polygons are defined too large. A number of spatial definitions is therefore recommended when analysing dispersion and clustering across an entire spatial region.

For our purposes, the variance to mean ratio will be sufficient as a measure of dispersion. Examples of overdispersed, perfectly dispersed and overdispersed datasets are given in Figure 2.1.

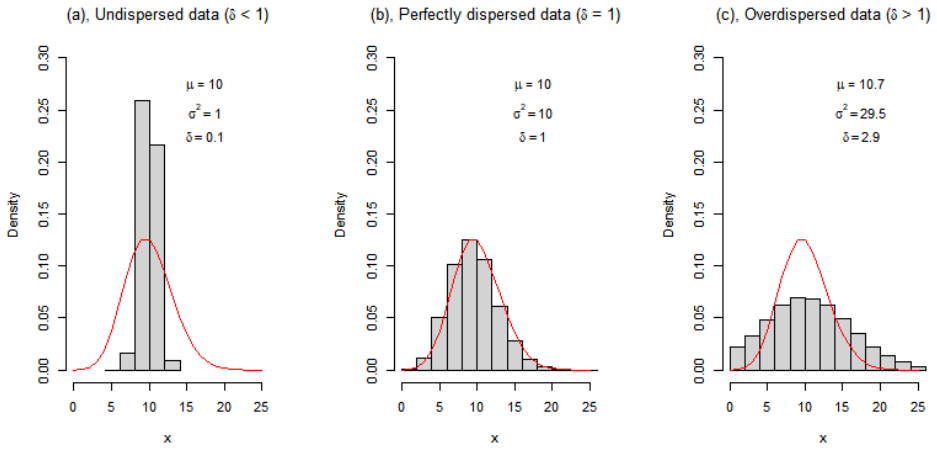

Figure 2.1: Neighbourhood definitions

All three panels show the distribution of three simulated data with a mean value, μ, close to 10. Panel (a) shows underdispersed data, Panel (b) shows a dataset with perfect dispersion and Panel (c) shows overdispersed data. The red line indicates a fitted Poisson distribution with a rate parameter, $\lambda = 10$. It is clear that the Poisson distribution fails to sufficiently fit the under-and overdispersed data.

2.3 Spatial Autocorrelation

Spatial clustering is concerned with the identification of areas where similar values are observed in close spatial proximity. When high values are observed close to other high values and low values are observed close to other low values, positive spatial autocorrelation is said to occur. When high and low values cluster together it is called negative spatial autocorrelation. Spatial autocorrelation is generally measured at global and local levels. A bivariate perspective will also be considered.

2.3.1 Univariate Global Spatial Autocorrelation

A widely accepted measure of univariate global spatial autocorrelation is the Moran's I statistic [61]. It has been discussed in various spatial statistics textbooks [70, 84, 88] and papers with

diverse applications [5, 44].

Since spatial autocorrelation is concerned with the identification of concentrations of values, the calculation of the Moran's I statistic starts by defining a set of neighbours around a spatial entity in which a variable is measured. The intuition behind this approach is that if neighbouring values are more similar than non-neighbouring values, spatial autocorrelation will be present in the data [61].

Neighbours can be assigned on a distance metric, k-nearest neighbours basis (more generally applied to point data) or by a contiguity condition (more common for spatial polygons). The *queen* spatial contiguity condition states that, if a spatial entity with a set of boundary points shares an intersection with the boundary points of another spatial entity then they are defined as neighbours. A stricter contiguity condition, called the *rook* contiguity condition states that a positive portion of their boundary be shared in order for them to be defined as neighbours [88]. These two contiguity conditions are named after the movements allowed for the queen and rook chess pieces. Examples of the different methods of creating neighbourhoods mentioned are given in Figure 2.2. The examples are based on the principles illustrated in a vignette in [88] and applied to a subset of the us_states dataset found in the R package, spData [13].

Once the neighbours of a spatial entity, i, have been defined, a weight, w, is assigned to each of the neighbours, j, according to a weighting rule. The weight for non-neighbouring entities will be set equal to zero, that is $w_{ij} = 0$. Weights can be assigned based on a simple equal weighting basis, exponential weights or more novel weighting schemes such as a variance-stabilising coding [81]. The weighted values of the measured variable, or spatially-lagged values, are then computed. Anselin [7] suggested the use of the Moran's scatterplot to estimate Moran's I by fitting an Ordinary Least Squares (OLS) model to the spatially-lagged values with the observed values being used as the explanatory variable in the model. The Moran's I statistic is obtained from the slope parameter of the fitted OLS model.

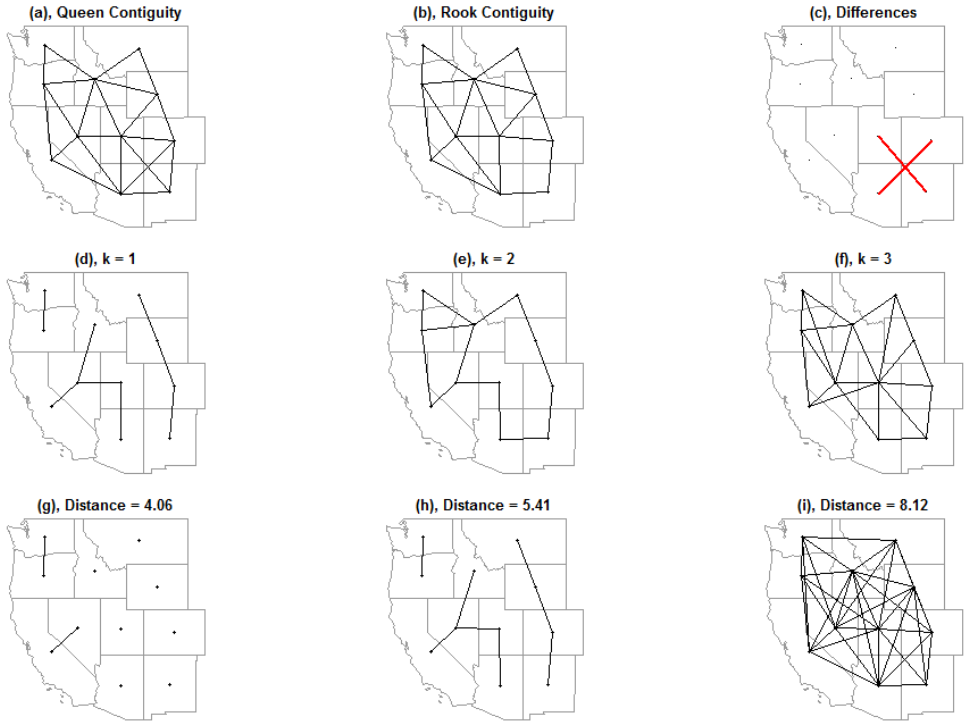

Figure 2.2: Neighbourhood definitions

The first row illustrates the difference between the Queen and Rook contiguity conditions. Panel (a) shows the neighbourhoods when the Queen contiguity condition is applied to the Western States of the us_states dataset as found in the R package, spData [13]. Panel (b) shows the neighbourhoods when defined by the Rook contiguity condition and Panel (c) shows the additional neighbourhoods defined by the more lenient Queen condition. The second row shows the neighbourhoods when they are defined by a k-nearest neighbours definition. Panel (d) gives the neighbourhoods where k = 1, (e) k = 2 and (f) k = 4. The third row shows the neighbourhoods when they are defined by a distance measure. The k = 1 definition in (d) gives the minimum distance for which all areas will have one distance-based neighbour. For this dataset, this distance is 5.41 units, as shown in Panel (h). A distance less than this will lead to some areas not having neighbours, as can be seen in Panel (g). Finally, Panel (i) shows a dense neighbourhood which is created by a large distance measure.

More formally, the global Moran's I is defined as:

$$I = n \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.5}$$

where w_{ij} represents the weight assigned to each neighbour.

The Moran's I is often a value between -1 and 1, and is interpreted as any other correlation measure. Perfect negative spatial autocorrelation is observed when $I = -1$, perfect positive spatial autocorrelation is observed when $I = 1$. No spatial autocorrelation is said to occur when

$I = 0$. Despite the similarities to other correlation measures, it is not generally true that the range of Moran's I is between -1 and 1 [58, 88]. It has been shown that the extreme values of Moran's I are not generally bounded between -1 and 1 when the weights matrix is neither symmetrical nor row-normalised. This occurs when spatial entities have different numbers of neighbours. This is sometimes negated by assigning a weighting scheme between contiguous (or bordering) entities that ensures that the length of the spatial boundary and the consequent impact of the number of neighbours a spatial entity can have, is negated [21].

The statistical significance of the Moran's I is tested against an alternative hypothesis that the spatial clustering did not occur by chance and depend on the observed values of the neighbours. This is done by an analytical method where the expected value and variance of the Moran's I are calculated by the following expressions

$$E(I) = \frac{-1}{n-1} \tag{2.6}$$

and

$$Var(I) = \frac{nS_4 - S_3S_5}{(n-1)(n-2)(n-3)W^2} - (E(I))^2, \tag{2.7}$$

where

$$\begin{aligned} S_1 &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2 \\ S_2 &= \textstyle\sum_{i=1}^{n}(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji})^2 \\ S_3 &= \frac{n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\ S_4 &= (n^2 - 3n + 3)S_1 - nS_2 + 3W^2 \\ S_5 &= (n^2 - n)S_1 - 2nS_2 + 6W^2, \end{aligned}$$

with W as the matrix of individual weights, w_{ij} [61].

A Z-statistic is then calculated as

$$Z = \frac{I - E(I)}{\sqrt{var(I)}} \sim \mathcal{N}(0,1)), \tag{2.8}$$

from which confidence limits are calculated and a decision is made whether the null hypothesis ought to be rejected. This is a fast method to determine significance of Moran's I, but it can be

influenced by polygons that feature irregular values [37].

A better, more robust, albeit slower approach to testing the null hypothesis is by using a permutation bootstrap or Monte Carlo approach. In this approach, variable values are randomly assigned to spatial entities. Using the previously-defined neighbourhood structure, a new Moran's I is calculated. When repeated multiple times, this process produces a simulated sampling distribution of Moran's I values under the null hypothesis that there is no spatial autocorrelation. The observed Moran's I is compared to the simulated sampling distribution to calculate a pseudo p-value, p^*, which is calculated as

$$p^* = \frac{N_{ext} + 1}{N + 1}, \tag{2.9}$$

where N_{ext} is the number of samples that are more extreme than the observed Moran's I and N is the number of simulated samples [37]. An illustration of the Monte Carlo approach applied to a dataset from the spData package in R [13] is shown in Figure 2.3.

The Moran's I does have some notable limitations. Due to the distance matrix that is created, large datasets can be computationally intensive and lead to slow performance. Alternative, faster methods have been developed that do not rely on the creation of a large distance matrix to detect spatial autocorrelation [2]. While Moran's I statistic is effective at identifying spatial autocorrelation in a dataset, it is not able to identify where spatial clusters (hotspots) occur in the dataset.

2.3.2 Univariate Local Spatial Autocorrelation

The concept of a Local Indicator of Spatial Association (LISA) was defined to assist with the identification of local clusters [4]. A LISA has two defining characteristics. Firstly, a LISA provides a statistic of spatial autocorrelation at each spatial location and secondly, it creates a proportional relationship between the sum of local statistics and their global counterpart. A LISA is essentially a decomposition of the global statistic into its individual components, thereby giving the contribution to the global statistic of each individual location.

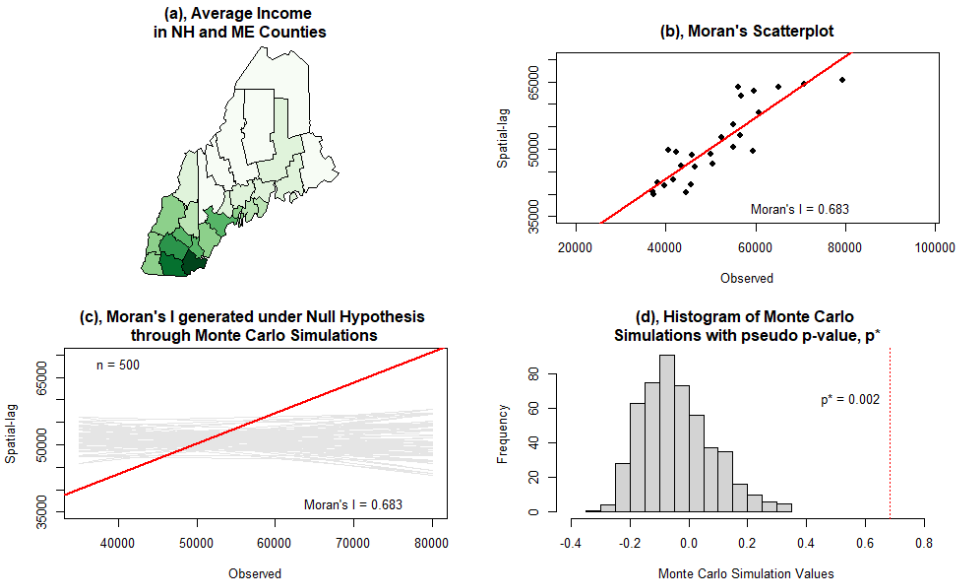

Figure 2.3: Global Moran's I

Panel (a), shows the Average Income of counties in New Hampshire and Maine. Counties with higher Average Income are shown with darker green, while counties with lower Average Income are shown with lighter green. From this, we can deduce that the counties to the South of the region have higher Average Income. This is confirmed by Panel (b) which shows the Moran's scatterplot with a Moran's I value of 0.683. This suggests that there is strong positive spatial autocorrelation present in the dataset. The statistical significance is tested by a Monte Carlo simulation or permutation test by randomising the observed values over the counties. This is to simulate values under the null hypothesis that there is no spatial autocorrelation present in the data. Panel (c) shows the Moran's I values generated by 500 Monte Carlo simulations (in grey) against the observed Moran's I. Panel (d), shows a Histogram of the simulated Moran's I values, which shows that no values greater than the observed Moran's I was generated by the Monte Carlo simulations. This generates a very small pseudo p-value, which leads us to reject the null hypothesis and conclude that there is positive spatial autocorrelation present in the data.

The local Moran's I is accordingly defined as

$$I_i = (x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}) \tag{2.10}$$

This is defined for each location, i, thereby satisying the first characteristic of a LISA [4]. The sum of all local Moran's I statistics is given by

$$\sum_{i=1}^{n} I_i = \sum_{i=1}^{n}(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}) \tag{2.11}$$

which is proportional to the Moran's I defined in Equation 2.5, thereby satisfying the second

characteristic of a LISA. It can also be shown that the average of all local Moran's I's is equivalent to the global Moran's I [4].

Assessing significance of local Moran's I's by using a permutation bootstrap approach is similar to the global case. The only difference is that significance is assessed for each entity in turn. This requires the value of one of the n entities to be fixed and the other $n - 1$ values to be randomised. Once significance has been determined, local clusters, or LISA clusters, can be determined. This is done by standardising the variable and its spatially lagged neighbouring values around their respective means. A scatterplot is then generated which is divided into four quadrants. Values in the first quadrant are classified into *high-high* clusters which indicate hotspots where high values cluster together. Values in the fourth quadrant are classified into *low-low* clusters which indicate coldspots where low values cluster together. Values in the first and third quadrants are classified into *low-high* and *high-low* clusters. These are generally seen as outlier values or clusters of negative spatial autocorrelation. An important factor to consider is that values aren't assigned to a cluster if they are not significant [4]. An illustration of this process is given in Figure 2.4

2.3.3 Bivariate Spatial Autocorrelation

The Moran's I statistic as defined in Section 2.3.2 is insufficient when trying to detect spatial autocorrelation in a multivariate context, since it is formulated in a similar way to a bivariate correlation statistic [58]. This makes it difficult to separate the correlation between variables from the spatial autocorrelation that arises due to them being in close geographical proximity.

Efforts to extend the global Moran's I to a multivariate setting originally centred around principal component analysis (PCA) [25, 91]. Attempts focusing on the local Moran's I are predominantly concerned with developing a unique case of geographically weighted regression [16, 54]. Another method considered the distance in geographical and variable spaces and derived a correlation measure that is high for values that are in close proximity in both geographical and multivariate spaces [6].

When one makes a few simplifying assumptions, a bivariate spatial correlation coefficient

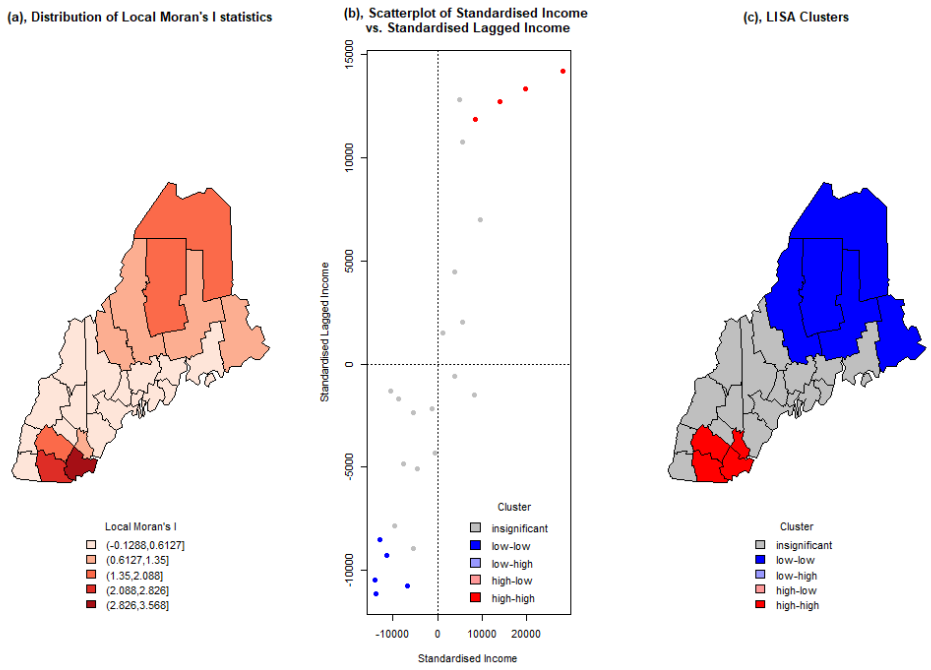

Figure 2.4: Local Moran's I

Panel (a), shows the local Moran's I values for Income in New Hampshire and Maine counties. From this, we see that values range from -0.1288 to 3.568. A clear pattern of two clusters (one in the south and one in the north) is observed. In Panel (b) standardised income values are plotted against standardised lagged income values. The cluster of an observation is determined by i.) the quadrant in which it falls and ii.) whether its local Moran's I is statistically significant. We notice many observations in the *high-high* quadrant, that are not significant and are consequently not assigned to the *high-high* cluster. Panel (c) shows the LISA clusters generated by the significant values observed in Panel (b). This confirms the observations from Panel (a) of two clusters in the north and south. In addition, we note that the cluster in the south is a hotspot (high values clustered together) and the cluster in the north is a coldspot (low values clustered together).

that resembles the Moran's I can be used to split out the spatial and Pearson correlation components. However, these assumptions are simply too strong for practical application [51]. A simpler bivariate extension of the local Moran's I describes the relationship between the value for one standardised variable at a location, i, x_i, and the spatially-lagged values for another standardised variable,y_j, $\sum_j w_{ij}y_j$, at the same location. Both x and y have zero mean and unit variance. The statistic is defined as the product of x and the spatially-lagged value of y.

$$I_i^B = x_i \sum_j w_{ij} y_j \tag{2.12}$$

where w_{ij} are the spatial weights as defined before. The global bivariate Moran's I is a summation

over the individual local bivariate Moran's I's [51].

An illustration of the bivariate Moran's I is given in Figure 2.5.

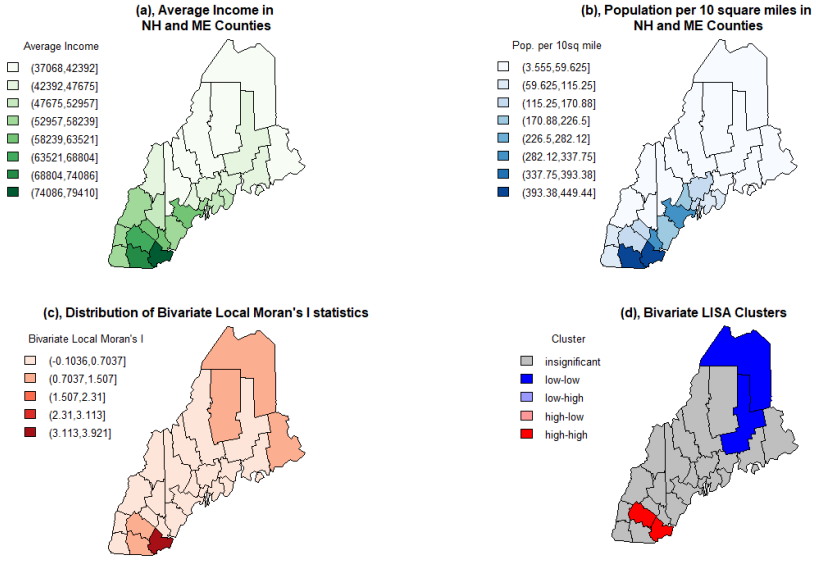

Figure 2.5: Bivariate local Moran's I

Panels (a) and b, show the Average Income and Population per 10 square miles in New Hampshire and Maine counties, respectively. This shows an indication of higher values for both variables clustering in the south. The global bivariate Moran's I is calculated as 0.54, indicating fairly strong positive spatial autocorrelation. The bivariate local Moran's I values are shown in Panel (c), ranging between -0.1036 (indicating slight negative spatial autocorrelation) and 3.921 (indicating very strong positive spatial autocorrelation). Panel (d) shows the ensuing bivariate LISA clusters that are formed by the counties that have statistically significant bivariate local Moran's I values. The procedure to calculate the LISA clusters is identical to the process illustrated in Figure 2.4.

2.4 Summary

In this chapter, two important topics regarding ESDA were considered. The first is dispersion which measures the extent to which values in a variable are spread around a central value or over the variable's domain. We explored popular measures of dispersion such as the variance to mean ratio, Green's coefficient and Morisita's Index.

Next, we explored spatial autocorrelation from global, local and bivariate perspectives. The global perspective is concerned with the identification of spatial dependence in a dataset. Local spatial autocorrelation attempts to pinpoint where the spatial dependence can be found in the

dataset and the bivariate case considers whether two or more variables cluster in similar regions. The calculation of Moran's I for each of these cases was considered. It was also illustrated how to test the statistical significance of an observed Moran's I statistic by means of a fast analytical and/or slower, but more robust Monte Carlo approach.

We have therefore defined the metrics that will be used in our ESDA framework.

Chapter 3

Gaussian Processes

This chapter is an introduction to Gaussian Processes (GPs). We start by describing the rationale for GPs in Section 3.1 and see how GPs extend linear regression and Bayesian linear regression from the realm of vectors into that of functions in Section 3.2. A simple example that illustrates the estimation procedure is created. We then delve deeper into the defining components of the GP in Section 3.3 and explore covariance functions in Section 3.4 by describing how they are chosen and optimised for a given problem. We then turn our attention to the famous geostatistical method known as Kriging [50] and understand how they are related to GPs in Section 3.5.2. The chapter concludes with Poisson GP/Kriging extensions and applications in Section 3.6.

3.1 Why GPs?

The GP model was originally proposed as a method to assist mining professionals in determining the available tonnage of a mine from extracted ore samples [50]. It assumes a joint probability distribution of known (the sampled data) and unknown data (the prospective mining locations) and conditions the unknown data on the known data to derive a posterior distribution of ore tonnage at the unknown locations. The methodology has since become a ubiquitous method in mining and other geostatistical applications such as meteorology [24] and epidemiology [85]. Alternative formulations have even extended as far as time series modelling [33, 55].

The reason for the wide application of GPs can be broadly explained by two important characteristics. A GP assumes the known and unknown data to be Gaussian distributed, which leads to

convenient conditioning properties and tractable likelihood expressions [94]. While this appeals to the academic desirability of the model, GPs have recently become more desirable to Machine Learning (ML) practitioners since they provide predictive error-bars. They serve the function of being a natural measure of confidence that differs depending on the amount of available known data. Further advances in computing ability, the development of open-source software [67] and flexible estimation methods that relax some Gaussian assumptions [75] have made GPs a popular method for non-linear Bayesian regression applications.

3.2 From linear regression to GPs

In a simple linear regression example, we have that a dependent variable y can be modelled as a linear function of an independent variable x and an identical, independent error term, ϵ, by stating that

$$y = f(x) + \epsilon \tag{3.1}$$

where

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2) \tag{3.2}$$

and

$$f(x) = \beta_0 + \beta_1 x \tag{3.3}$$

with β_0 and β_1 defining intercept and slope parameters, respectively. For this linear regression model, we attempt to find point estimates for the parameters using Ordinary Least Squares (OLS). The Bayesian perspective on this problem takes on a probabilistic approach which assumes a prior distribution on the parameters that are updated when new data are observed. This gives rise to a distribution of point estimates of the parameters [47].

GPs are an extension from the Bayesian approach to linear regression in that it assumes a distribution of functions - all of which are consistent with the observed data - and gives the posterior distribution over these functions.

GPs are extremely flexible since the number of estimated functions, can be infinite. This means that a GP is a non-parametric approach to Bayesian regression. However, not all possible functions

will be consistent with observed data and one needs to define a set of constraints to ensure that a sensible model is fitted [94]. These constraints can include limiting the eligible functions to the domain of a given regression problem as well as defining the strength or type of relationship between observed data points. A GP is described by its mean and covariance functions. The mean function is often assumed to be the zero function. The covariance function is where the constraints are defined in terms of the relationship between observed data points [94]. A popular covariance function is called the squared-exponential function which defines the covariance between two data points x_1 and x_2 as

$$k(x_1, x_2) = \exp(-\frac{(x_1 - x_2)^2}{l^2}) \tag{3.4}$$

where l is a hyperparameter known as the characteristic length scale. The length scale can be tuned to the length which suits the observed data best. For example, if observed data are close together, a shorter length scale can be used (given rise to a rougher function) while a longer length scale is often used where data are further apart (leading to a smoother function). Examples of three covariance functions with different length scales can be seen in Figure 3.1. Notice also that the squared-exponential covariance function defined here is also known as the radial-basis kernel function (RBF) [94].

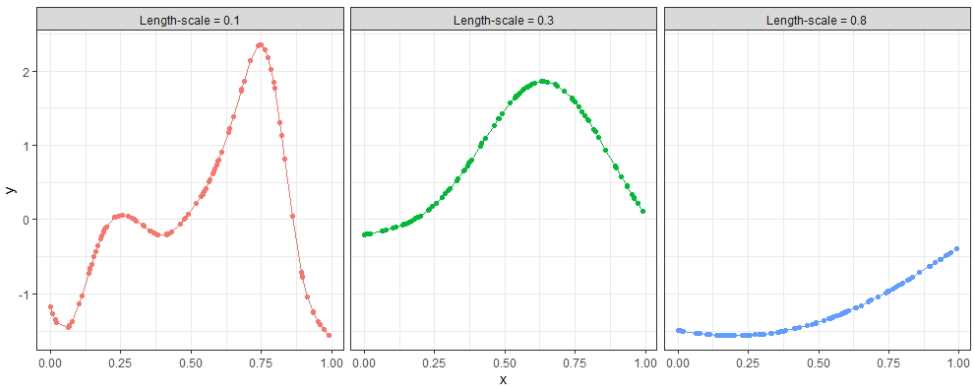

Figure 3.1: Different length scales for the squared-exponential covariance function

The covariance functions generated here are based on simulated data. The x-axis shows the distance between two When data that are close together are highly related, a shorter length-scale (0.1) will be more suitable.

A classic defintion of a GP states that it is a collection of random variables (RVs), any finite

number of which have a joint Gaussian distribution [94]. This implies that we can get the conditional probability of any one of the variables given the others. This is known as the marginalisation or consistency condition of Gaussian distributions. Since we can obtain the conditional probability of a single RV, we can derive the posterior functions from our set of prior functions and observed data on a given domain by using Bayes' Rule [94].

3.3 Theoretical Overview

GPs are often applied in spatial applications where data ,f , observed in certain locations, x, as $(f(\boldsymbol{x}))$ are used to estimate a function value ($f(\boldsymbol{x}_*)$) in areas where no measurements were observed $(\boldsymbol{x}_*)$ [85]. The posterior of a GP $p(\boldsymbol{f}_* \mid \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f})$ is therefore defined as the joint probability of these observed and unobserved data (often referred to as *test* data).

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*' & \boldsymbol{K}_{**} \end{pmatrix} \right)$$

where K defines the variance function of the observed data, K_* defines the covariance function between the observed and test data and K_{**} defines the variance function of the test data. The test data are often sampled from the domain of the observed data. Test data can also be sampled from outside the domain of the observed data, but this can lead to high variance of the estimates. This is to be discussed in Section 3.4. In certain instances, the GP can interpolate the training data and is then said to be fitted to *noise-free* observations and that by standard condition rules for Gaussian distributions, the posterior has the following form [66]

$$p(\boldsymbol{f}_* \mid \boldsymbol{x}_*, \boldsymbol{x}, \boldsymbol{f}) \sim \mathcal{N}(\boldsymbol{f}_* \mid \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

where

$$\boldsymbol{\mu}_* = \boldsymbol{\mu}(\boldsymbol{X}_*) + \boldsymbol{K}_*' \boldsymbol{K}_*^{-1} (\boldsymbol{f} - \boldsymbol{\mu}(\boldsymbol{X})) \tag{3.5}$$

and

$$\boldsymbol{\Sigma}_* = \boldsymbol{K}_{**} - \boldsymbol{K}_*' \boldsymbol{K}_*^{-1} \boldsymbol{K}_* \tag{3.6}$$

where $\boldsymbol{\mu}(\boldsymbol{X})$ is the mean function of the observed data and $\boldsymbol{\mu}(\boldsymbol{X}_*)$ is the mean function of the test data.

The process of fitting a GP to noise-free data is illustrated in Figure 3.2. The domain is defined after which the number of functions to sample from the prior distribution of functions is specified. The covariance function is then specified with a suitable length scale. In machine learning, this is often *tuned* using a hyperparameter search. Panel (a) shows five samples drawn from a GP prior with a squared-exponential covariance function and Panel (b) shows five data points that the GP will be fitted on. At this point, the mean function is regularly defined as the zero function. The posterior predictive is then derived using Equations 3.5 and 3.6 as shown in Panel (c). A prediction and confidence intervals based on the mean of the samples are given in Panel (d) [94].

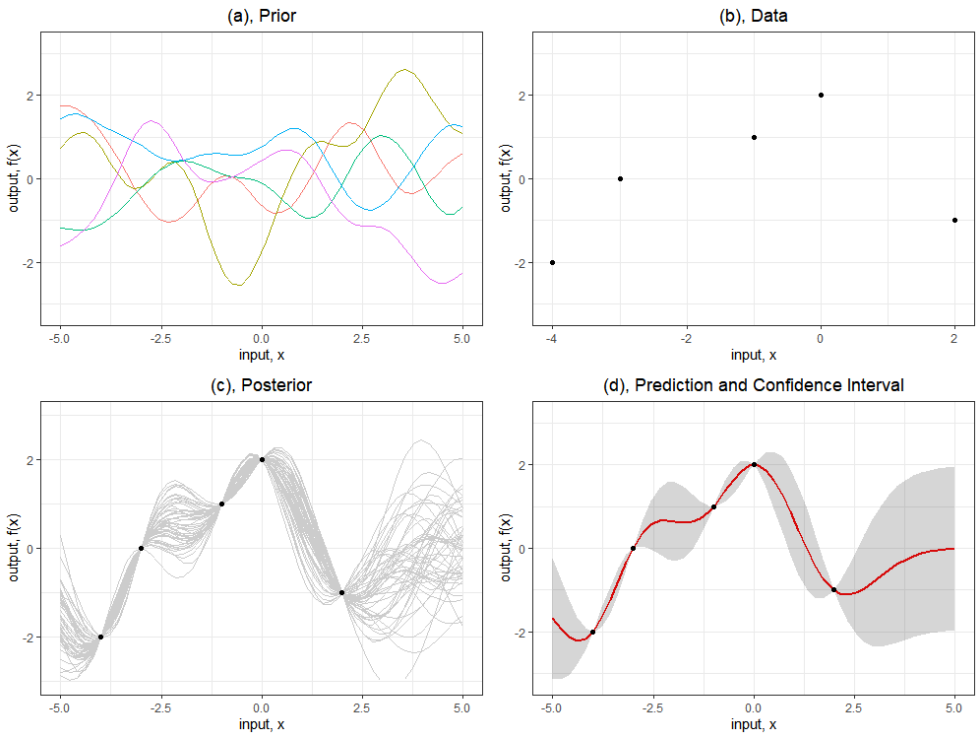

Figure 3.2: Gaussian Process noise-free data example

Panel (a) shows five samples drawn from a GP prior with a squared-exponential covariance function. Panel (b) shows five data points that the GP will be fitted on. Panel (c) shows fifty samples fitted to the data. Panel (d) shows the predictions from the fifty samples with 95% confidence intervals. Note how the predicitions *interpolate* the data points and how there are higher uncertainty between data points. The example is based on graphs in Chapter 2 of [94].

Unfortunately, this noiseless assumption is not often a true reflection of reality. A more prac-

tical application of GPs is when we assume some measure of uncertainty around our training data and fit a model to the data accordingly. In this instance, a GP is said to be fitted to *noisy* observations and we assume that the fitted model will not interpolate the training data. The joint density of the observed data and is then instead given by

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{y}} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{'} & \boldsymbol{K}_{**} \end{pmatrix} \right)$$

where

$$\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_y^2)) \tag{3.7}$$

$$\boldsymbol{K}_{\boldsymbol{y}} = \boldsymbol{K} + \sigma_y^2 \boldsymbol{I} \tag{3.8}$$

This means that the noise terms are added to each training observation, independently, and that the posterior predictive is then of the form

$$p(\boldsymbol{f}_* \mid \boldsymbol{x}_*, \boldsymbol{x}, \boldsymbol{f}) \sim \mathcal{N}\left(\boldsymbol{f}_* \mid \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*\right)$$

where

$$\boldsymbol{\mu}_* = \boldsymbol{\mu}(\boldsymbol{X}_*) + \boldsymbol{K}_*^{'} \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y} \tag{3.9}$$

and

$$\boldsymbol{\Sigma}_* = \boldsymbol{K}_{**} - \boldsymbol{K}_*^{'} \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{K}_* \tag{3.10}$$

An illustration of a GP fit to noisy training data is shown in Figure 3.3.

Applying GPs to real regression problems, however, causes some practical problems. Inverting the covariance matrices can be slow and unstable, since a covariance matrix that has small eigenvalues, cannot be inverted. This can occur in both the noise-free and noisy cases. This causes the need to apply a Cholesky decomposition, which ensures that the covariance matrix is positive semi-definite (PSD) and invertible. We also need to make assumptions around whether the model needs to accommodate any noise in the data, as well as on the length scale used in the covariance

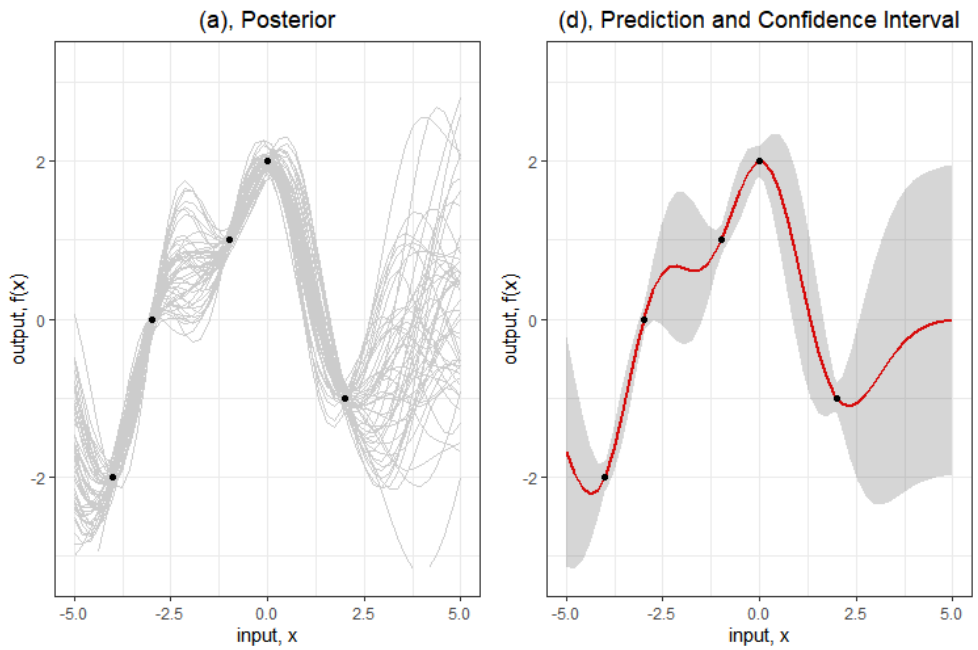

Figure 3.3: Gaussian Process noisy data example

Panel (a) shows 50 GPs fitted to the data with the assumption of noisy observations. Notice how the fitted models do not interpolate the training points. Panel (b) shows the shows the predictions from the fifty samples. Even though the prediction is similar to that obtained in the *noise-free* case, notice the 95% confidence intervals around the training data which were absent in the *noise-free* example.

functions. Having made these assumptions, the mean and variance can be calculated as in Equations 3.5 and 3.6 (for the noise-free case) and 3.9 and 3.10 (for the noisy case) and a log-likelihood can then be calculated.

The log-likelihood function in the noiseless case takes on the form

$$\log p(\boldsymbol{y} \mid X) = -\frac{1}{2}\boldsymbol{y}' K^{-1}\boldsymbol{y} - \frac{1}{2}\log \mid K \mid -\frac{n}{2}\log(2\pi) \tag{3.11}$$

where the first term indicates the data fit and the second term indicates a penalty given to the model's complexity, which is a function of the covariance function. To obtain the marginal log-likelihood in the noisy case, simply add the error variance σ_y^2 to the covariance matrix, K. This process is repeated until the marginal log-likelihood function is maximised. The marginal log-likelihood is rather maximised since it reduces the chances of overfitting, because the marginalisation only considers the known data [66].

We will now consider the covariance function in depth. We will focus specific attention on the requirements of a valid covariance function, how to choose the right covariance function for a regression problem and explore the Matérn covariance function, since it is a popular covariance function in the spatial context.

3.4 Covariance Functions

A covariance function is a mapping of similarity between two data. A general name for such a mapping, $k(\boldsymbol{x}, \boldsymbol{x}^{'})$, is called a *kernel*. A kernel is said to be symmetric when

$$k(\boldsymbol{x}, \boldsymbol{x}^{'}) = k(\boldsymbol{x}^{'}, \boldsymbol{x}) \tag{3.12}$$

A covariance function is by definition symmetric [94].

In addition, we can compute a Gram matrix K where the entries K_{ij} are specified by $k(\boldsymbol{x}_i^{'}, \boldsymbol{x}_j)$. If k is a covariance function, K is a covariance matrix. K is defined to be positive semi-definite (PSD) if and only if it does not have any negative eigenvalues. The Gram matrix that corresponds to a legitimate covariance function is positive semi-definite, even though it need not be the case for a general kernel function [94]. A covariance function can therefore be deemed legitimate when it gives rise to a positive semi-definite covariance matrix.

When the covariance function was introduced in Section 3.3, we noticed that the squared-exponential covariance function is equivalent to the RBF kernel. In general, the covariance function of a GP is specified by a positive-definite kernel function [66].

Five other popular covariance functions will now be discussed. The Periodic kernel function [57] allows for the modelling of exactly repeating functions and has a hyperparameter that specifies the distance between exact repetitions in addition to the characteristic length scale. The functional forms of all covariance functions discussed are given in Table 3.1. Examples of two samples drawn from a GP prior with the different covariance functions is shown in Figure 3.4.

Covariance Function	Functional Form
RBF	$k_{RBF}(x, x') = \sigma^2 \exp(-\frac{(x-x')^2}{2\ell^2})$
Rational Quadratic	$k_{RQ}(x, x') = \sigma^2(1 + \frac{(x-x')^2}{2\alpha\ell^2})^{-\alpha}$
Periodic	$k_{PER}(x, x') = \sigma^2 \exp(-\frac{2\sin^2(\pi(x-x')/p)}{\ell^2})$
Locally Periodic	$k_{LP}(x, x') = \frac{k_{RBF}(x,x')k_{PER}(x,x')}{\sigma^2}$
Linear	$k_{LIN}(x, x') = \sigma_b^2 + \sigma_\nu^2(x - \ell)(x' - \ell)$
Matérn	$k_{MAT}(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\frac{(x-x')^2}{\ell})^\nu K_\nu(\sqrt{2\nu}\frac{(x-x')^2}{\ell})$

Table 3.1: Functional forms of Covariance Functions

For all covariance functions, ℓ signifies the length scale. The relative weighting of large small-scale variations for the Rational Quadratic is determined by α. The distance between periods in the Periodic and Locally Periodic covariance functions is denoted by p. In the Linear covariance, c denotes the x-coordinate through which all lines in the posterior pass through (in the noiseless case) and σ_b^2 is the y value of the function where $x = 0$. The Matérn is formulated in more detail in Equation 3.13.

A kernel function has additive and multiplicative properties which means that new kernel functions can be created by adding and multiplying kernel functions. By adding kernels, the resulting kernel will have a high (or low) value when one of the kernels has a high (or low) value. By multiplying them, the ensuing kernel will have a high (or low) value when both kernels have high values. They can also be added across dimensions. The posterior over functions that used an additive kernel can also be decomposed back into its constituent parts [27].

The Radial Quadratic is an addition of RBFs with differing characteristic length scales. The Locally Periodic kernel is a more flexible version of the Periodic kernel that adds or multiplies a RBF to the Periodic kernel and allows for the modelling of functions that have varying periodic components over time. The linear kernel reduces a GP to Bayesian linear regression, but has the ability to create useful covariance functions when you have different types of variables and by adding or multiplying them to other kernel functions [27].

Another covariance function is the Matérn class of covariance functions. It was named after the Swedish forestry statistician Bertil Matérn by Michael L. Stein.

The Matérn covariance function is defined as

$$k_{MAT}(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\frac{(x-x')^2}{\ell})^\nu K_\nu(\sqrt{2\nu}\frac{(x-x')^2}{\ell}) \tag{3.13}$$

where Γ is the gamma function, $\nu > 0$, is a parameter of the covariance, $\ell > 0$ is the length scale

and K_ν is the Bessel function of the second kind [94].

In his book [77], Stein suggests the Matérn covariance function as a more realistic alternative to the RBF for physical processes since it is not infinitely differentiable. This means that it can accommodate local discrepancies in the covariance between two points which are not captured in infinitely differentiable (or smooth) covariance functions such as the RBF. This can be seen in its rougher form when compared to the RBF in Figure 3.4. It has consequently become a popular covariance function in spatial applications. The Matérn has the added benefit that when ν, tends to infinity it converges to the RBF. It is therefore seen as a more realistic generalisation of the RBF.

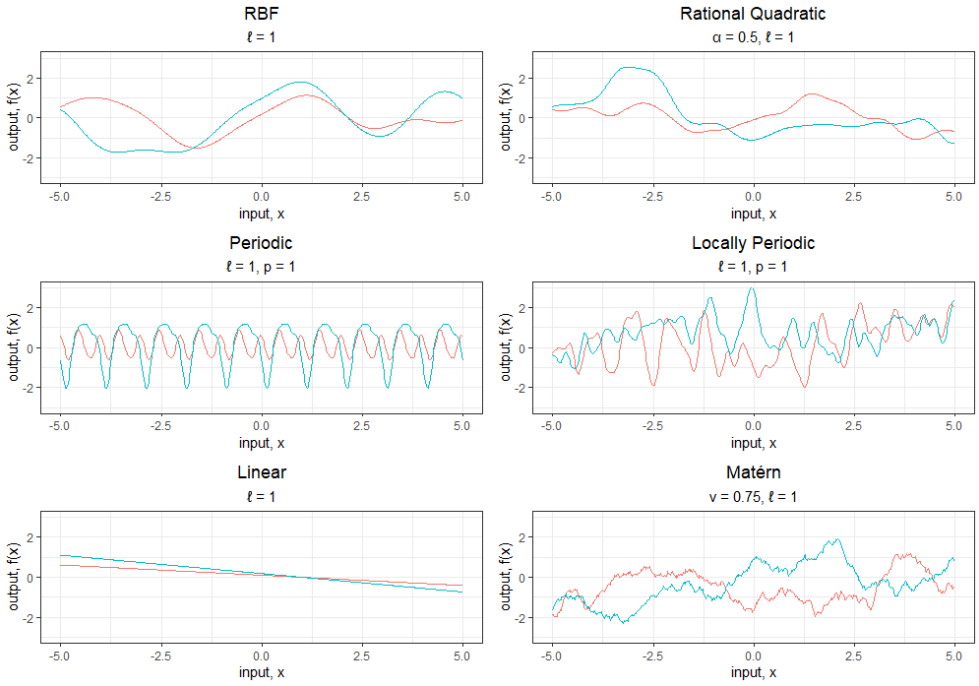

Figure 3.4: Examples of different covariance functions

Note the different levels of smoothness over the domain of the Rational Quadratic (determined by the additional hyperparameter, α) where the RBF has a constant smoothness. Note the increasing trend in the Locally Periodic which accommodates local differences, whereas the Periodic expects perfect repetition. The Matérn function is a rough curve for the given parameters and is popular in spatial applications.

Given the available set of covariance functions, one needs to choose the correct covariance function for a given regression problem. When confronted by the choice of the right covariance function, one can easily be overwhelmed by choice, especially in modern machine learning applications that have high dimensionality. A standard process has been to develop multiple models

with different covariance functions. These models each need to have optimal hyperparamerers. These can be tuned through various techniques. An exhaustive grid search (which can be slow), employing a Bayesian Model Selection criteria and limiting the error on unseen data by means of a probably approximately correct (PAC) framework [94] are three possible methods.

From the tuned models, it is then standard practice to choose the model that maximises the marginal log-likelihood. Testing all possible covariance functions can become an endless task. Especially once all combinations of additive and multiplicative kernels are considered. An intelligent search methodology has been proposed in [26] where the researchers built a set of kernels by addition and multiplication and formulated a greedy search methodology. The method has the added benefits that it identifies data structure and also imitates the scientific discovery process.

3.5 GPs in a spatial context

The term Kriging was first coined by Mathéron [59] and is named after the South African mining engineer, Danie Krige, who developed the methodology in his Master's dissertation to determine the tonnage of mining deposits at unobserved locations [50]. Kriging is also the term used for GPs in a spatial context. In this section, the two steps of Kriging estimation are discussed.

Kriging estimation requires that the estimation variable be spatially autocorrelated. Metrics that calculate spatial autocorrelation are discussed in Chapter 2. The relationship between sample information and the distance between spatial entities is therefore a critical component in Kriging estimation. This is represented by the empirical semivariogram. A semivariogram model is fitted to the empirical semivariogram which describes the distance at which spatial autocorrelation is no longer observed. In Section 3.5.1 the semivariogram model is defined. Its relationship to a covariance function is also described which shows how Kriging and GPs are related. In Section 3.5.2 the process of fitting a Kriging model is summarised.

3.5.1 Semivariogram

Consider a random variable, Z, observed at n spatial locations, u.

$$\{Z(\boldsymbol{u}_\alpha) = z_\alpha, \alpha = 1, ..., n\} \tag{3.14}$$

A formulation of semivariance exists which measures spatial dependence between all observations of Z as a function of the magnitude of the distance, h, between them.

$$\gamma(h) = \frac{1}{2}E\{Z(\boldsymbol{u}) - Z(\boldsymbol{u} + h)\} \tag{3.15}$$

Calculating the semivariance for all observations of a variable can be computationally expensive since each combination of two observations has a unique distance. It is also not a practical consideration, since sample information is not necessarily available at all possible locations. It is therefore standard practice to group distances between observations into lag bins and to calculate the average semivariance over the lag bins instead. This produces the empirical semivariogram, $\hat{\gamma}(h)$. The empirical semivariogram can then be plotted which shows how semivariance changes as the distance between lag bins increases [20], [23]. Formally, the empirical semivariogram is defined as

$$\hat{\gamma}(h) = \frac{1}{2 \mid N(h) \mid} \sum_{i=1}^{N(h)} [Z(\boldsymbol{u}_i) - Z(\boldsymbol{u}_i + h)]^2 \tag{3.16}$$

where $N(h)$ represents all the combinations of pairwise Euclidean distances between lag bins and $\mid N(h) \mid$ is the number of distinct pairs in $N(h)$. If the distance, h, were to consider both magnitude and direction, it can be represented by a vector, $\mathbf{h}$. When the semivariogram is considered to be a function of only the distance, h, between two spatial entities, it is said to be isotropic.

There are three components that define a semivariogram model, namely the Range, Sill and Nugget. Theoretical semivariance is equal to zero when there is no distance between two spatial entities. The Nugget effect,$\gamma_0(h)$, occurs when semivariance greater than zero is observed at an infinitesimally small distance between two spatial entities. It is observed when spatial sources of variation occur at distances smaller than what the lag bins are grouped by and/or when measurement errors occur.

$$\gamma_0(h) = \begin{cases} 0, h = 0 \\ \\ \alpha, otherwise \end{cases} \qquad \alpha > 0, \tag{3.17}$$

When spatial autocorrelation is present, semivariance increases up to a certain distance. Beyond this distance the spatial autocorrelation becomes negligible and the semivariance remains constant. This distance is called the Range, a_0. The semivariance at the Range,$\gamma(a_0)$ is called the Sill and is equal to variance , $C(0) = \sigma^2$, where C represents a covariance function. That is,

$$\gamma(a_0) = C(0) = \sigma^2 \tag{3.18}$$

An isotropic semivariogram is related to the covariance function, discussed in Section 3.4, by the following equality.

$$\gamma(h) = C(0) - C(h) = \gamma(a_0) - C(h) = \sigma^2 - C(h) \tag{3.19}$$

The proof of this equality is beyond the scope of this dissertation, but has been explored under various conditions in [35].

Examples of valid isotropic semivariograms are given in Table 3.2. Note that c_0 represents the Sill and a_0 represents the Range. Validity conditions that are required for a semivariogram to be used in a Kriging model, such as conditional-negative definiteness have been studied at length in [20].

Semivariogram	Functional Form
Spherical	$\gamma(h) = c_0[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}(\frac{h}{a_0})^3] if h \leq a_0$
Exponential	$\gamma(h) = c_0[1 - \exp(-\frac{h}{a_0})]$
Power	$\gamma(h) = c_0 h^{a_0}$
Gaussian	$\gamma(h) = c_0[1 - \exp(-\frac{h^2}{a_0^2})]$

Table 3.2: Functional Forms of Isotropic Semivariograms

A semivariogram model is fitted to the empirical semivariogram using a valid semivariogram function and estimates the Range, Sill and Nugget by Least Squares and Maximum Likelihood methods. Goodness of fit can be assessed by Sum of Squared Errors (SSE). More refined methods

such as jackknifing cross validation are also used and are often performed to decrease estimate bias [20], [59]. An example that shows the process of fitting a semivariogram to an empirical semivariogram of Zinc observations from the meuse dataset from the sp package in R [14] is shown in Figure 3.5.

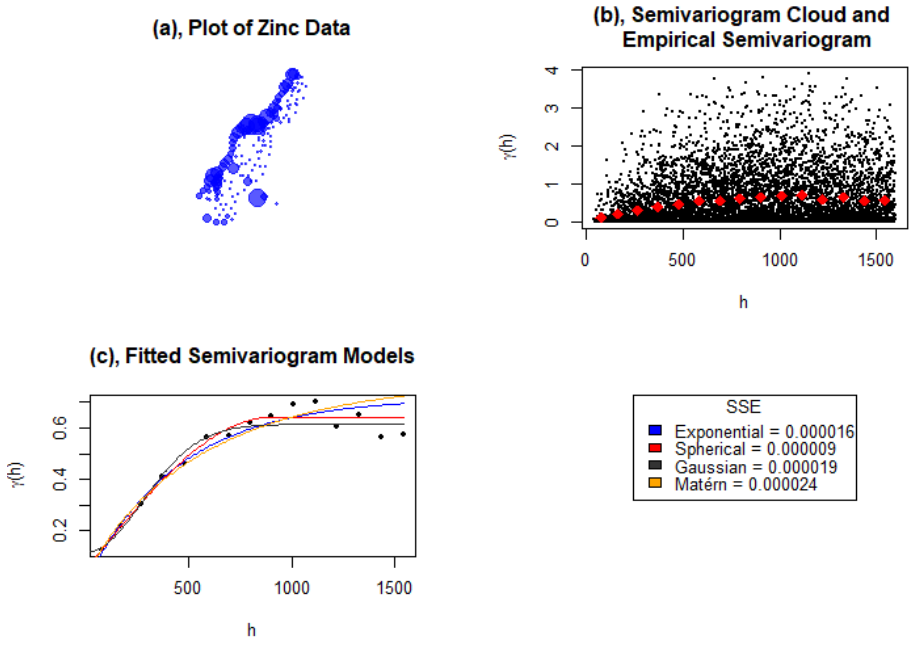

Figure 3.5: Semivariogram example

The meuse dataset from the sp package in R is used for this example. Zinc concentrations across the Netherlands were measured and are represented spatially in Panel (a) . A high concentration of Zinc is seen close to the coastline with a few outlying areas also having high values. Panel (b) shows the semivariance, $\gamma(h)$ at all the possible combinations of distances in the Semivariogram Cloud (black dots). No clear pattern of spatial autocorrelation can be deduced from this plot. Panel (b) also shows the empirical semivariogram calculated at a number of lag bins. From this we see a clear increase in semivariance that plateaus just below 1000 metres. Panel (c) shows the empirical semivariogram with four fitted semivariogram models, Exponential, Spherical, Gaussian and Matérn. SSEs of the four models are shown in the legend. The model that fits the empirical semivariogram best, is the Spherical, with a SSE of 0.000009. The fitted model parameters indicate a Sill of 0.59, obtained at a Range of 897 metres and a Nugget of 0.05.

3.5.2 Kriging estimation

One of the goals of a geostatistical model is to estimate samples of a random variable, Z at unsampled locations, Z^*. A Kriging model does this by computing a linear combination of n nearby observed data.

$$Z^*(\boldsymbol{u}) - \mu = \sum_{i=1}^{n} \lambda_i (z_i - \mu) \tag{3.20}$$

where μ represents a mean function which is constant over all spatial entities and λ_i represent individual weights assigned to neighbours. Owing to the relationship between the semivariogram as shown in Equation 3.19, a covariance function can be determined from a valid semivariogram where the off-diagonal elements of the covariance matrix indicate the weights assigned to each of the i components of the linear combination.

By assuming an unknown mean function and semivariogram, an *Ordinary Kriging* predictor of $Z^*(\boldsymbol{u})$ can be derived.

The weights are optimised to ensure minimum estimation bias and variance. The Kriging estimate is therefore a Best Linear Unbiased Estimate (BLUE) [20, 23]. An example of Ordinary Kriging is shown in Figure 3.6.

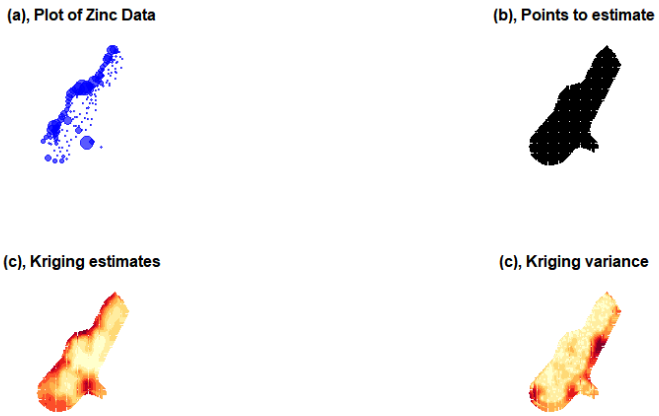

Figure 3.6: Kriging estimation example

Panel (a) shows Zinc observations in the Netherlands from the meuse dataset from the sp package in R. A high concentration of Zinc is seen close to the coastline with a few outlying areas also having high values. Panel (b) shows a grid of points at which Zinc estimates need to be estimated. Panel (c) shows the Kriging estimates with high values concentrated near the coastline. Panel (d) shows the associated variance with larger variance observed in areas with lower Zinc concentrations.

3.6 Kriging for count data

A simple method of modelling a count variable with a GP/Kriging model, X, is to apply a log transformation and add a small constant, $c > 0$ to ensure that the estimated values are positive and

$$Y = \log(X + c), c > 0 \tag{3.21}$$

The transformation can then be reversed to get the prediction in the original units [85].

More complex applications and variations on Poisson Kriging than the simple log transformation proposed in Equation 3.21 can be found in the work of Goovaerts [1, 38] that consider the Poisson distribution as a critical component of the model formulation.

Another method for modelling count data with a GP/Kriging approach was suggested by [85] for modelling sparse data, as is often seen in disease mapping applications. The data is separated into n polygon areas, with centroid co-ordinates, $\mathbf{X}$, number of fatalities (the modelling variable) $\mathbf{y}$ and populations, $\mathbf{p}$. An expected number of fatalities is then calculated from explanatory variables of the population and fatalities in each area. The number of fatalities and the background population are separated into R groups according to the explanatory variables. This leads to a standardised expected number of fatalities

$$e_i = \sum_{r=1}^{R} \left(\frac{\sum_{i=1}^{n} y_{i,r}}{\sum_{i=1}^{n} p_{i,r}}\right) p_{i,r} \tag{3.22}$$

$\mathbf{y}$ is modelled as a Poisson process

$$y_i \sim Poisson(e_i \mu_i) \tag{3.23}$$

where μ_i represents the relative risk in the ith of the n polygon areas. The authors then proceed to create a GP prior as well as a hyperprior which searches for optimal hyperparameters instead of a simpler grid search methodology. Due to the Poisson assumption on the outcome variable, the posterior inference becomes much harder, since tractable solutions cannot be found. It is for this reason that approximation such as the Laplace method are proposed [94] in tandem with a Monte Carlo Markov Chain (MCMC) approach. While these approximations increase

speed by decreasing the number of estimations, the methods scale $O(n^3)$ in the number of data, n and conditional methods are proposed that reduce dependency of computational time to $O(nm^2)$ where $m < n$ [85].

Thus far, the impact of auxiliary variables on Kriging model performance has not been considered. When an auxiliary variable is used as part of the Kriging model, it is known as Co-Kriging. This allows the auxiliary variable to be modelled along with the explanatory variable as a multivariate Kriging model. This is often done where the auxiliary variable is more densely sampled than the explanatory variable [40]. An interesting extension of the model defined the Co-Kriging model as a Generalised Linear Mixed Model (GLMM) which yielded promising results as an alternative to the traditional formulation [74].

3.7 Summary

In this chapter GPs and Kriging models were described. Particular care was taken in the formulation of GPs as an extension of linear regression from the realm of vectors into that of functions. The theory of both models were reviewed and the connection between the models through the covariance function and semivariogram was shown in Equation 3.19. A detailed discussion of covariance functions was given along with many illustrative examples. The chapter concluded with a short review of some of the literature on Poisson Kriging.

Chapter 4

Poisson Mixture Regession

4.1 Why Mixtures?

A mixture model is a model-based clustering technique which ensures that the groups that are formed manage to capture unobserved heterogeneity in different mixing proportions and is appropriate when the data is originated from different groups. They have been shown to perform well on overdispersed data [18, 46]. Mixture models are called mixture regressions in a supervised context.

In a spatial context, a GP/Kriging model takes spatial dependence into account by quantifying the impact of the distance between two spatial entities by means of a covariance function or semivariogram. Contrastingly, a mixture regression model is not *aware* of the spatial dependence in the same way. It rather clusters similar values into the same mixtures. The spatial dependence is captured as a consequence of the clustering. The different approach to capturing spatial dependence is the reason why we want to explore the use of mixture regressions alongside a GP/Kriging model.

The general definition and estimation of a finite mixture model is reviewed in Section 4.2. In Section 4.3 we see how the general definition applies to Poisson mixtures and then discuss how these concepts translate to a Poisson mixture regression model in Section 4.4. The chapter concludes with a review of previous research of mixture models on spatial data in Section 4.5.

4.2 Mixture Definition and Estimation

A finite mixture model with K mixtures/components is defined as

$$h(y \mid x, \phi) = \sum_{k=1}^{K} \pi_k f(y \mid x, \theta_k) \tag{4.1}$$

where y is a univariate dependent variable with h as its conditional density function, x is a vector of explanatory variables, π_k represents the mixing/prior probability of the kth mixture and θ_k is the set of parameters for a density function, f. The complete vector of all parameters is given by $\phi = (\pi_1, ..., \pi_K, \theta_1^{'}, ..., \theta_K^{'})$. Finally, $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

The mixture model can be used to assign a data observation (x, y) to the mixture component, j, that obtains a maximum posterior probability [31]. The posterior probability that the observation belongs to the jth component is:

$$p(j \mid x, y, \phi) = \frac{\pi_j f(y \mid x, \theta_j)}{\sum_K \pi_k f(y \mid x, \theta_k)} \tag{4.2}$$

While a Bayesian approach through Monte Carlo Markov Chain (MCMC) methods is a possible estimation method [32], the maximum posterior probability is more often obtained with a frequentist Maximum Likelihood Estimation (MLE) of ϕ via the EM algorithm [22] by first using Equation 4.2 to derive prior mixing/component probabilities for N sample observations $(x_1, y_1), ..., (x_N, y_N)$

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_{nk} \tag{4.3}$$

and then estimating the posterior mixing/component probabilities for each observation

$$\hat{p}_{nk} = p(k \mid x_n, y_n, \hat{\phi}) \tag{4.4}$$

in the Estimation (E) step.

The log-likelihood of these observations is given by

$$\log(L) = \sum_{n=1}^{N} \log h(y_n \mid x_n, \phi) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k f(y_n \mid x_n, \theta_k)). \tag{4.5}$$

In the Maximisation (M) step the log-likelihood for each mixture/component is maximised individually with respect to θ by using the posterior probabilities from Equation 4.4 as weights

$$max(\sum_{n=1}^{N} \hat{p}_{nk} f(y_n \mid x_n, \theta_k)). \tag{4.6}$$

The E and M steps are repeated until either a maximum number of iterations is reached or the log-likelihood is maximised by reaching an improvement threshold [42].

4.3 Poisson Mixtures

The mixture model defined in Equation 4.1 does not assume a specific density function. By substituting the density function with the Poisson pmf as defined in Equation 1.1 the EM algorithm can be used to estimate the mixing/component probability, π_k and rate parameter, λ_k, for each of the $k = 1, ..., K$ components of the following mixture model

$$h(y \mid x, \phi) = \sum_{k=1}^{K} \pi_k f(y \mid x, \lambda) \tag{4.7}$$

where $\phi = (\pi_1, ..., \pi_K, \lambda_1^{'}, ..., \lambda_K^{'})$. By differentiating the log-likelihood function with regards to each parameter, separately, the parameter estimates are

$$\hat{\lambda}_k = \frac{\sum_{n=1}^{N} p(k \mid x_n, y_n, \hat{\phi}) x_n}{Z(k)} \tag{4.8}$$

and

$$\hat{\pi}_k = \frac{Z(k)}{N} \tag{4.9}$$

where $Z(k) = \sum_{n=1}^{N} p(k \mid x_n, y_n, \hat{\phi})$ [12, 83].

The EM algorithm is a good estimation method and is useful not only for mixture models, but for any model with missing data [42]. It is, however, sensitive to the correct choice of initial values

as convergence can be influenced by local optima and if the likelihood function is unbounded [41].

4.4 Poisson Mixture Regression

Mixture models are used in a supervised context by estimating a Generalised Linear Model (GLM) for each mixture component.

A standard finite mixture regression model assumes identical error and link functions, but different linear operators for each mixture/component and is given by

$$f(y \mid x, \Theta) = \sum_{k=1}^{K} \pi_k f_k(y \mid x, \beta_{0k}, \beta_k). \tag{4.10}$$

where $\Theta = (\beta_{0k}, \beta_k)$ and y follows a distribution from the exponential family, conditional upon mixture/component, k with an expected value

$$E(y \mid x) = g^{-1}(\beta_{0k} + x^{'}\beta_k) \tag{4.11}$$

where $g(.)$ is a link function.

A Poisson distribution-based version of this model assumes that g is a log link function and that

$$f_j(y_i, \theta_{ij}) = \frac{\exp(-\lambda_{ij})(\lambda_{ij})^{y_i}}{y_i!} I_A(y_i) \tag{4.12}$$

with $log(\lambda_{ij}) = \beta_j^T x_i$, $i = 1, ..., n$ and $j = 1, ..., k$. This model is maximised by the EM algorithm and has been studied in [60].

The choice of K needs to be determined before the model is fit. A general method is to fit a range of values for K and choosing the corresponding model with the lowest Akaike's Information Criterion (AIC) and/or Bayesian Information Criterion (BIC) value [42]. An example of a fitted Poisson mixture regression Model is shown in Figure 4.1.

The goodness of fit of a Poisson Mixture Regression Model is often assessed by using a rootogram [42]. A rootogram shows the square root of counts, similar to a histogram, of the

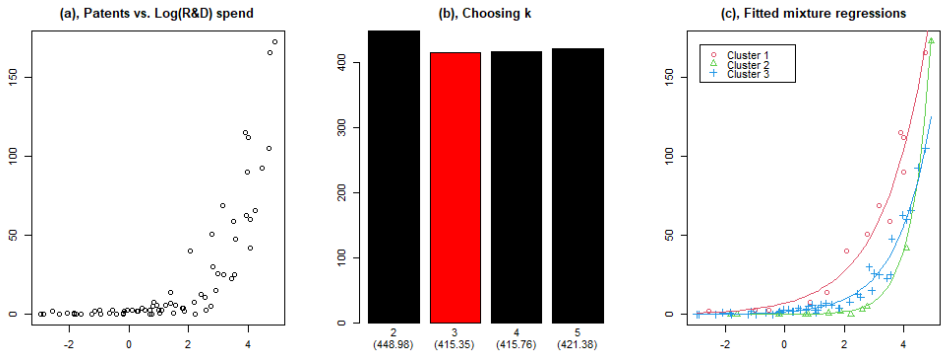

Figure 4.1: Example of a fitted Poisson Mixture Regression Model

Panel (a) shows the Number of Patents created against the log of Research and Development spending. The data is obtained from the flexmix package in R [42]. Panel (b) shows the optimal number of clusters, k, to be 3, based on a minimum AIC value of 415.35. The fitted mixture regression models along with the associated clusters are shown in Panel (c).

posterior probabilities of the observations that belong to each component. To prevent the bar at 0 probability from dominating the plot a threshold is used to exclude very small probabilities. A peak close to 1 indicates that a mixture is well-separated from other mixtures, while no peak and a cluster of values in the middle of a distibution indicates that the mixtures do not separate the data well [42]. An example of a rootogram is shown in Figure 4.2 which is based on the model fitted in Figure 4.1.

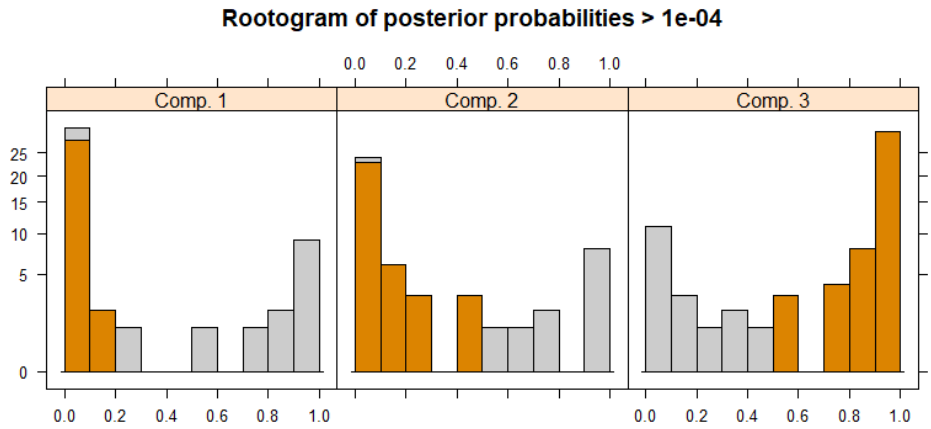

Figure 4.2: Example of a Rootogram

The third mixture regression fitted in Figure 4.1 is well separated since there is a high peak close to 1. For the other two mixtures, there is a slight overlap, since the mass of the rootogram is closer to 0.

4.5 Mixtures and Spatial Data

It has been found that a data-driven approach such as a mixture model cannot always account for all the spatial autocorrelation that might be present in a dataset [87]. Despite mixture models not considering spatial dependence explicitly, they have been used to model spatial problems. Frequentist approaches have considered spatio-temporal movements [80] and have estimated general mixture models by developing particle learning methods to create more efficient cluster allocations [17]. Extreme values are commonly seen in overdispersed data and Bayesian approaches to mixture models have been focused on extreme value analyses [49]. Other Bayesian explorations have compared the performance of different distributions to obtain the best-performing weakly informative prior [86]. Mixtures and GPs have also been combined [29].

A common problem with spatial data is the prevalence of excess zeros relative to standard distributions. This is known as sparsity. Zero-inflated mixture models are a popular class of models for dealing with such data in other contexts such as heart disease prediction [64], dental caries [34] and microbial data [36] and one might expect the findings to translate to a spatial context.

4.6 Summary

Poisson mixtures were explored in this chapter. A general definition of mixture models was given and then applied to the Poisson distribution, whereupon the Poisson-specific parameters obtained through the EM algorithm were given. Mixture regressions, mixture models in a supervised context, were again generally defined with a Poisson-based formulation then given. A method of choosing the correct number of mixtures (AIC) was given along with a popular goodness-of-fit tool (rootogram). The chapter concluded with a short review of mixture research conducted in a spatial context. Frequentist and Bayesian approaches were described along with research that focused on two prevalent problems of spatial data - overdispersion and sparsity.

Chapter 5

Application

In Section 5.1 an ESDA framework is formulated. The framework is based on the exploratory metrics discussed in Chapter 2, namely dispersion and spatial autocorrelation (global, local and bivariate). Data dimensionality in terms of the number of observations and the number of modelling variables, is also incorporated in the framework as these can impact the estimation methods to be used by the models. For each metric, categories are defined based on thresholds for VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH values. The measures are then calculated for two datasets that are described in Section 5.2. Each dataset is then classified in terms of the categorised ESDA measures. Kriging and Poisson Mixture Regression models are then trained on the data - the parameters of which are set up in Section 5.3. They are compared to a baseline Poisson Generalised Linear Model (GLM). Model accuracy answers 1.) how well does the model fit the dataset and 2.) how well does it capture the spatial dependence structure? These questions are expanded in Section 5.4 by building these questions into an accuracy metric framework.

Finally, the suitability of each of the models for a dataset with a given set of ESDA metrics is discussed. This enables academics to combine our framework with a study that covers a list of spatial models for count datasets [89] and develop a new research framework. These results, along with the code provided in the author's personal github repository (link to repo here) will enable practitioners to begin spatial count modelling applications with a concise set of tools. From it they will be able to a.) define ESDA metrics, b.) classify them according to known standards, c.) build models that are known to work well for a given dataset and d.) measure its suitability in terms of model fit and structural dependence.

5.1 ESDA Framework

5.1.1 Dimensionality

5.1.1.1 Number of variables

The number of variables, p, are classified accordingly and are motivated by classifications used in [46].

Variables	Classification
$p < 2$	VERY LOW
$3 < p \leq 4$	LOW
$5 < p \leq 10$	MEDIUM
$10 < p \leq 25$	HIGH
$p > 25$	VERY HIGH

Table 5.1: ESDA Classification of Number of Variables

5.1.1.2 Number of observations

The number of spatial entities or observations, n, in a dataset can have significant impact on the method in which a model is trained. These entities are often signified by spatial polygons. GPs/Kriging models in particular are sensitive to large data and need to be be adapted where datasets become very large [94].

Observations	Classification
$n \leq 100$	VERY LOW
$100 < n \leq 1,000$	LOW
$1,000 < n \leq 10,000$	MEDIUM
$10,000 < n \leq 100,000$	HIGH
$n > 100,000$	VERY HIGH

Table 5.2: ESDA Classification of Number of Observations

5.1.2 Dispersion

In Chapter 2, dispersion measures were discussed. The variance to mean ratio, δ, is used as the dispersion index. It is know that overdispersion is observed when $\delta > 1$. From [46] the following thresholds for defining categories of dispersion values are inferred. The categorisations aren't based on thresholds calculated in literature, but provide the practitioner with a rough guide to interpret dispersion values and classify them accordingly. It is also important to remember that dispersion values are always greater than zero.

Dispersion, δ	**Classification**
$0 \leq \delta < 1$	VERY LOW (underdispersion)
$\delta = 1$	LOW (variance is equal to mean or no dispersion)
$1 < \delta \leq 5$	MEDIUM
$5 < \delta \leq 10$	HIGH
$\delta > 10$	VERY HIGH

Table 5.3: ESDA Classification of Dispersion

5.1.3 Global spatial autocorrelation

The global Moran's I, I, is used as the measure of global spatial autocorrelation. The following thresholds for defining categories of spatial autocorrelation values are proposed and are similar to general interpretations of correlation measures [61]. The same thresholds apply for both positive and negative spatial autocorrelation, hence the absolute value of I being used.

Global Moran's I	**Classification**
$0 <\mid I \mid \leq 0.10$	VERY LOW
$0.10 <\mid I \mid \leq 0.30$	LOW
$0.30 <\mid I \mid \leq 0.50$	MEDIUM
$0.50 <\mid I \mid \leq 0.75$	HIGH
$0.75 <\mid I \mid \leq 1$	VERY HIGH

Table 5.4: ESDA Classification of Global Spatial Autocorrelation

5.1.4 Local and bivariate spatial autocorrelation

For local and bivariate spatial autocorrelation, LISA clusters are calculated. The proportion of spatial entities that belong to LISA clusters is used to categorise these spatial autocorrelation

measures.

The following thresholds are proposed:

Proportion of spatial entities that belong to LISA clusters	Classification
0% to 10%	VERY LOW
10% to 20%	LOW
20% to 30%	MEDIUM
30% to 40%	HIGH
More than 40%	VERY HIGH

Table 5.5: ESDA Classification of Local and Bivariate Spatial Autocorrelation

5.2 The datasets

The first dataset describes an aggregated view of the *lansing* dataset in the spatstat package [8] in R, which considered the locations of 2,251 trees split into five different tree species (Maple, Hickory, Red Oak, White Oak and Black Oak). The dataset that is used can be found as the *Lansing_Trees* dataset in the gcKrig package [43]. The original plots have been rescaled to a unit square and the number of different types of trees within squares of length 1/16 has been counted and can be seen in Figure 5.1. Due to the rescaling, the dataset consists of 256 spatial entities.

The second dataset considers suburb-level Vehicle and Residential crimes observed in 2020 in 2,309 suburbs in Gauteng, South Africa. The data was originally captured by the South African Police Service and was subsequently processed and provided by Lightstone to reflect the distribution of Police District crime data (a large spatial definition) into a suburb (a smaller spatial definition). The distribution can be seen in Figure 5.2. This dataset was ethically cleared for use (ethics number: NAS240/2021).

5.3 Models

A baseline Poisson Regression Generalised Linear Model (GLM) is fit to compare the performance of the Kriging and Mixture Regressions by using the *glm* function from the stats package in R. This package is a loaded onto the default R installation. 10-fold cross-validation is performed

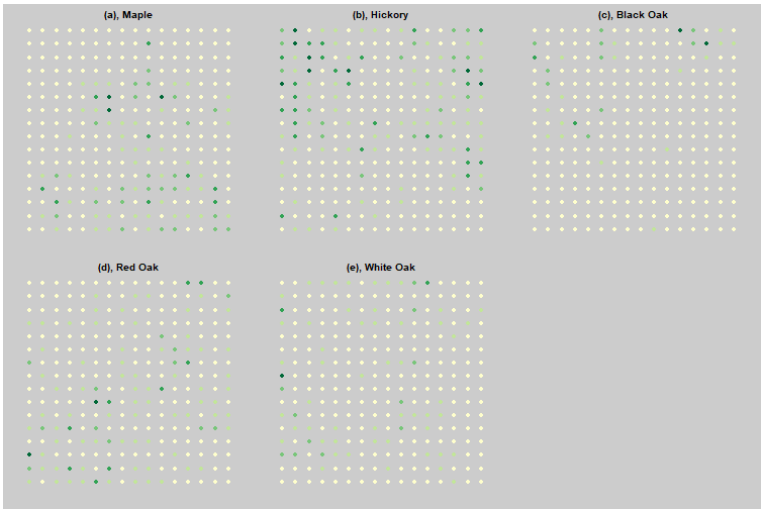

Figure 5.1: Plot of Lansing Trees dataset

Yellow indicates low concentration of trees in a cell, while Green indicates higher values. The colour gradient of each plot has been graded according to the values of each individual species. This is done to highlight where high values occur for a specific species.

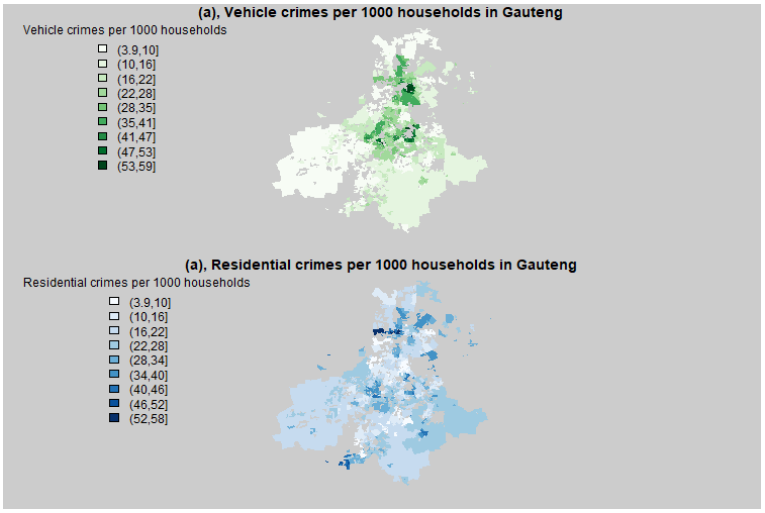

Figure 5.2: Plot for Gauteng Crime dataset

to prevent overfitting.

The semivariogram to be used by the Kriging model, will be determined by fitting Exponential, Spherical, Gaussian and/or Matérn semivariogram models. The best fitting model will then be chosen based on MSE. The Kriging model is then fit using the *krige.cv* function from the gstat package in R [40].

As with all mixture problems, the optimal number of mixtures, K, first needs to be determined. This is done by fitting the Poisson Mixture Regression model with 2 to 10 mixtures and choosing the K which yields the lowest value for the Akaike Information Criterion (AIC). The model is fit using the *stepFlexmix* and *FLXMRglm* functions from the flexmix package [42].

5.4 Accuracy measures

The model accuracy metrics have been chosen to answer two specific questions:

1. How well does the model fit the dataset?
2. How well does it capture the spatial dependence structure?

The first question is evaluated by performing a residual analysis. Four graphs are considered:

- Histogram of residuals with Mean-Square Error (MSE), $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$.

  A low MSE indicates a good model fit.

- Histogram of normalised residuals with Mean-Square Error (MSE).

  For normalised residuals, a MSE close to 1 indicates a good model fit.

- Scatter plot of actual values vs. fitted values with Pearson's correlation coefficient.

  A correlation coefficient close to 1 indicates a good model fit.

- Scatter plot residuals vs. fitted values with Pearson's correlation coefficient.

  A correlation coefficient close to 0 indicates a good model fit.

More sophisticated alternatives have been proposed [90] while Maximum Mean Discrepancy (MMD) has been used in other Poisson studies [46]. Spatial adaptations to MSE have also been developed to cater for small-area estimation problems [68]. We have chosen to use the standard formulation of MSE due to its widely accepted use and ease of computation and interpretability.

The second question is answered from a global and local perspective. The global perspective is done by calculating the absolute difference between the observed spatial autocorrelation and the estimated spatial autocorrelation obtained from the estimated model samples for the global, I Moran's I values, $D^I = \mid I - \hat{I} \mid$. This is a spatial analogue to the Spearman's correlation coefficient-based metric used in [46].

For the local, I_i and bivariate I_i^B Moran's I, LISA clusters are calculated from the fitted values. The bivariate case will only be considered for the Gauteng Crime dataset. The LISA clusters of the fitted values are then compared to the actual LISA clusters by means of a confusion matrix. This shows how many of the spatial entities are classified into the correct LISA cluster. While confusion matrices provide many accuracy metrics, we will consider overall accuracy as the main metric of interest.

5.5 Lansing Trees

5.5.1 ESDA

A table summarising the ESDA metrics of the Lansing Trees data is given in 5.6. The dataset consists of 256 cells and 5 different tree species. They are Maple, Hickory, Black Oak, White Oak and Red Oak. This suggests LOW observations and MEDIUM number of variables.

Data: Lansing Trees					*n = 256*	*p = 5*
Variable	**Dispersion**	**global Moran's I**	**Local Moran's I**			
			L-L	**L-H/H-L**	**H-H**	**I**
Maple	2.8203	0.8809	15.70%	0.00%	12.28%	72.02%
Hickory	2.3770	0.9550	18.52%	0.00%	17.18%	64.30%
Black Oak	1.7833	0.6078	0.00%	0.00%	12.11%	87.89%
Red Oak	1.6317	0.2474	5.47%	0.00%	6.64%	87.89%
White Oak	1.4521	0.2027	5.47%	0.00%	7.03%	87.50%

Table 5.6: Summary of ESDA metrics for Lansing Trees dataset

5.5.1.1 Dispersion

The dispersion indices for the five tree species range between 1.4521 for White Oak and 2.8203 for Maple. This translates to a classification of MEDIUM overdispersion.

5.5.1.2 Spatial Autocorrelation

The global Moran's I values are LOW for Red Oak and White Oak, HIGH for Black Oak and VERY HIGH for Maple and Hickory. All of the values are statistically significant. The values are reflected in the Moran's scatterplot in Figure 5.3.

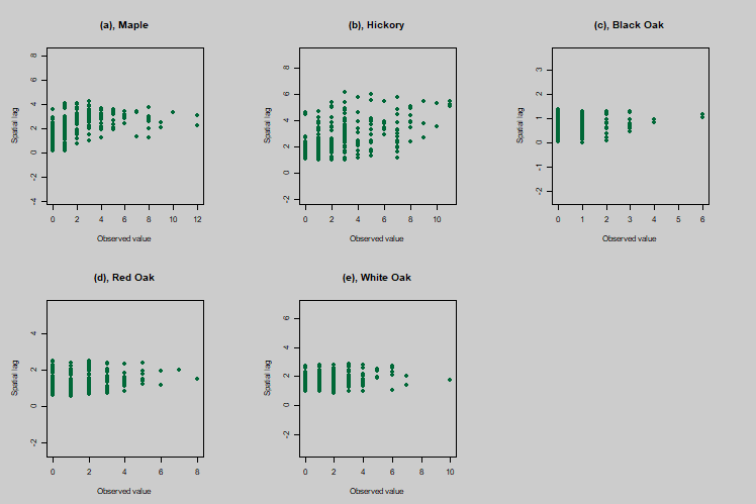

Figure 5.3: global Moran's Scatter Plot for Lansing Trees dataset

A stronger relationship between the spatially-lagged values and the observed values is seen for Maple and Hickory compared to Red Oak and White Oak.

The local Moran's I values are LOW for Black Oak, Red Oak and White Oak and HIGH for Maple and Hickory. A plot showing the LISA clusters is shown in Figure 5.4.

Due to the higher dimensionality (p = 5), a bivariate Moran's I classification is not being considered here.

5.5.1.3 Conclusion

A summary of the ESDA metrics for the Lansing Trees dataset is given in Table 5.7.

ESDA Framework		**Category**				
Number of Variables		Very Low	Low	**Medium**	High	Very High
Number of Observations		Very Low	**Low**	Medium	High	Very High
Dispersion		Very Low	Low	**Medium**	High	Very High
Spatial autocorrelation	global	Very Low	Low	Medium	**High**	Very High
	Local	Very Low	Low	Medium	**High**	Very High

Table 5.7: Summary of ESDA framework for Lansing Trees dataset

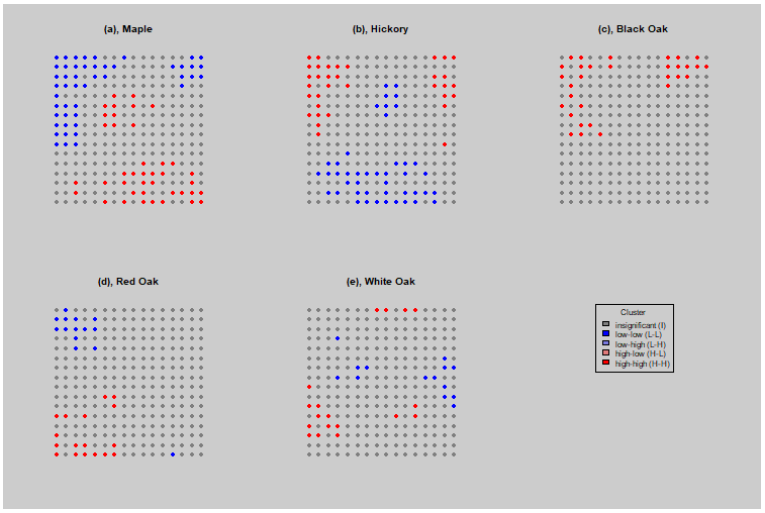

Figure 5.4: Local LISA clusters Lansing Trees dataset

Note the higher proportion of cells belonging to LISA clusters for Maple (34.38%) and Hickory (30.47%).compared to Red Oak, Black Oak (both 12.11%) and White Oak (12.5%). Also note how the L-L clusters of Maple and H-H clusters of Hickory are grouped together. It is also worth noting that Black Oak has no L-L clusters.

5.5.2 Modelling

In all of the models, we are modelling the number of Maple trees per cell. For the Baseline and Poisson Mixture Regression, we are using the number of Hickory trees as an explanatory variable, while the Kriging model does not use any explanatory variables. We are modelling in this manner, since a Kriging model with a covariate will be a Co-Kriging model, which is a multivariate technique. We can adapt the methodology to model any other of the five tree species in the dataset. Multivariate techniques can also be used, but is beyond the scope of this experiment. Maple and Hickory are being considered since they have the highest number of counts of the available tree species. A summary of the number of trees for each species is given in Table 5.8.

Maple	Hickory	Black Oak	Red Oak	White Oak	Other
514	703	135	346	448	105

Table 5.8: Number of trees per species

5.5.2.1 Baseline

5.5.2.1.1 Residual Analysis In Figure 5.5 we see the residual analysis of the baseline model for the Lansing Trees dataset and learn that:

- The model has a very low MSE, (Panel (a)).
- The MSE of the normalised residuals is close to 1.
- The actual and fitted values have a fairly strong correlation of 0.3821.
- The residuals and fitted values are not strongly correlated.

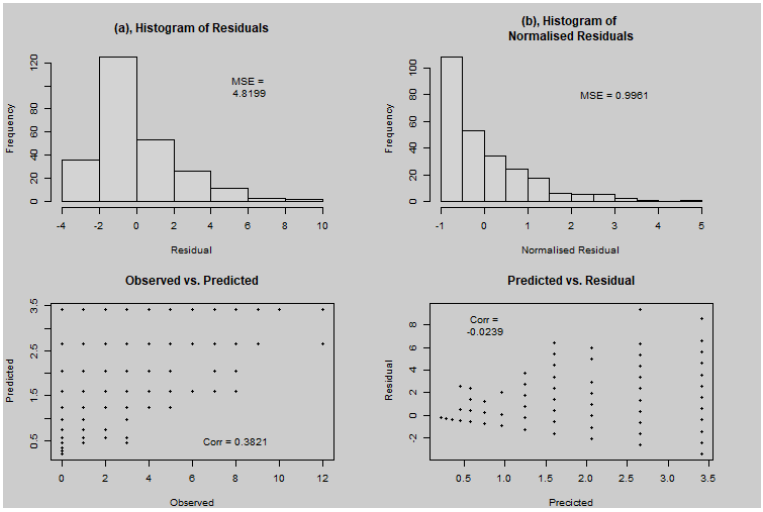

Figure 5.5: Residual analysis of the Lansing Trees dataset - baseline Poisson Regression model

5.5.2.1.2 Model Fit From the Model Fit plots in Figure 5.6 we see that the model appears to be slightly too simple to accurately fit the data over the entire domain, with no values greater than 4 being estimated. The spatial graph also suggests that the model doesn't accurately capture the correlation structure, since the graph does not reflect a similar pattern to the one observed in Panel (a) of Figure 5.1.

We can therefore conclude that the baseline model achieved a fairly fit, with a few notable shortcomings.

5.5.2.1.3 Global Correlation Structure In Figure 5.7 we see that the global Moran's I of the fitted values differs by a mere 0.0097 from the global Moran's I of the actual values.

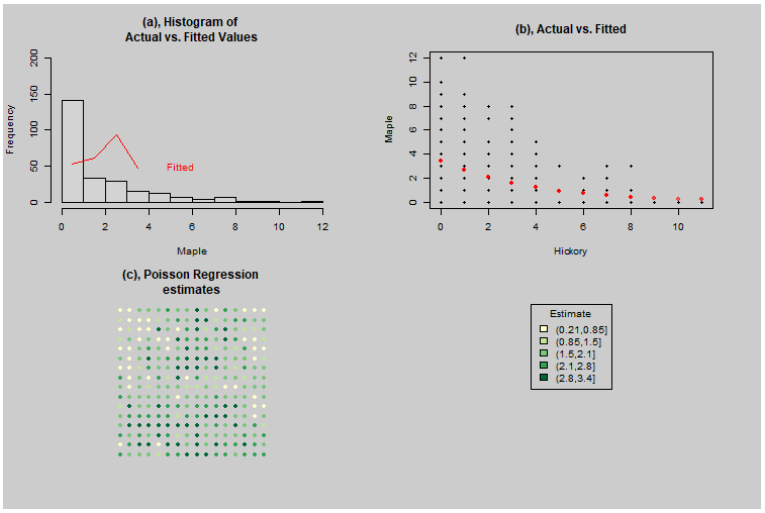

Figure 5.6: Model fit plots of the Lansing Trees dataset - baseline Poisson Regression model

In Panel (b), the red dots indicate fitted values. In Panel (c), the fitted values are visualised in the spatial format.

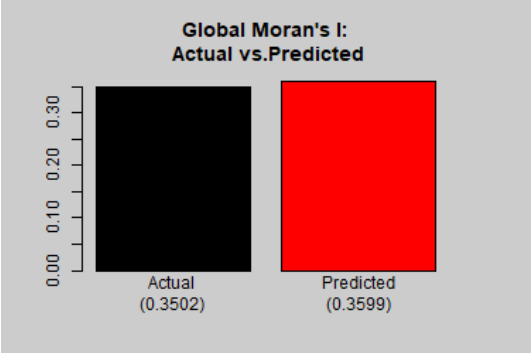

Figure 5.7: global Moran's I difference of the Lansing Trees dataset - baseline Poisson Regression model

5.5.2.1.4 Local Correlation Structure From Figure 5.8 we see that the local Moran's I values based on the fitted values of the baseline model are not correlated with the local Moran's I of the actual values. While it seems to accurately capture the presence of spatial autocorrelation with the global Moran's I, it is not very effective at identifying where the local spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows that Accuracy of 0.6680 is achieved. This means that only two thirds of the fitted values were classified into their actual LISA cluster by the model.

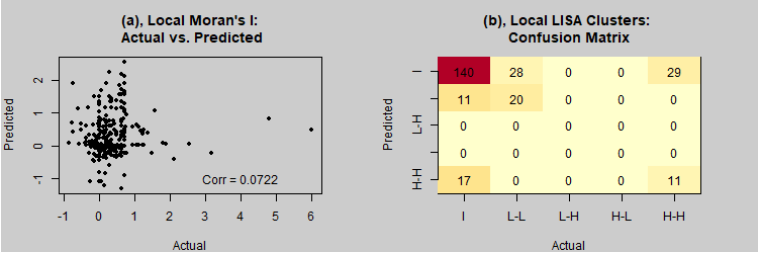

Figure 5.8: Local Moran's I difference of the Lansing Trees dataset - baseline Poisson Regression model

5.5.2.2 Kriging

5.5.2.2.1 Semivariogram In Figure 5.9 we see that the Exponential semivariogram fits the empirical semivariogram best. Note, however that the empirical semivariogram exhibits strange behaviour for cells that are further than 0.4 units apart. It appears that the semivariogram models do not accommodate this value very well.

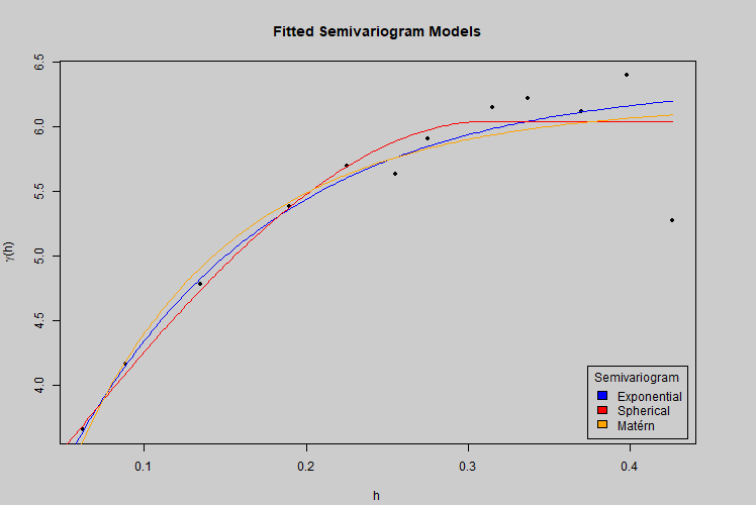

Figure 5.9: Semivariograms of the Lansing Trees dataset - Kriging model

5.5.2.2.2 Residual Analysis In Figure 5.10 we see the residual analysis of the Kriging model for the Lansing Trees dataset and learn that:

- The model has a higher MSE (7.3535) than the baseline model, with many values being overpredicted. This is reflected in the high number of estimates with a residual less than 0.

- The MSE of the normalised residuals is also not as close to 1 as that of the baseline model.
- The actual and fitted values have a fairly strong correlation of 0.4283, which is sightly better than the baseline model.
- The residuals and fitted values are not strongly correlated.

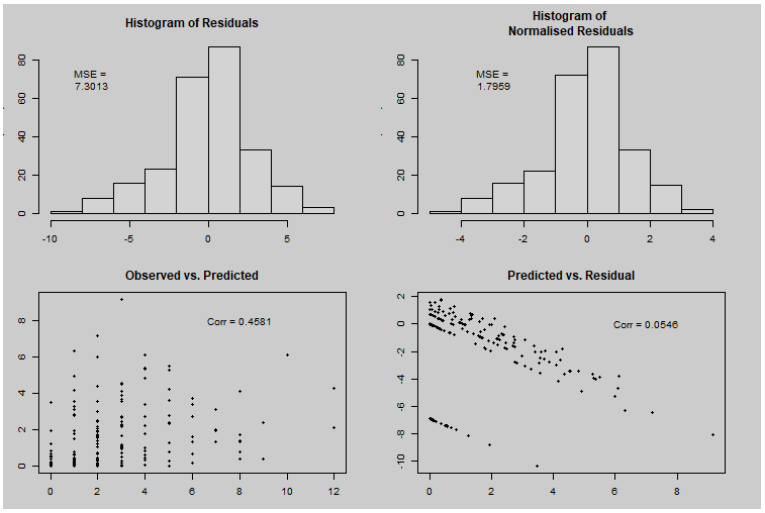

Figure 5.10: Residual analysis of the Lansing Trees dataset - Kriging model

5.5.2.2.3 Model Fit From the Model Fit plots in Figure 5.11 we see that the model appears to the data well when considering the Histogram with fitted values in Panel (a). The spatial graph also suggests that the model captured the correlation structure well, since the graph shows a similar pattern to the one observed in Panel (a) of Figure 5.1.

5.5.2.2.4 Global Correlation Structure In Figure 5.12 we see that the global Moran's I of the fitted values differs by 0.2013 from the global Moran's I of the actual values. It did not capture the global Spatial Autocorrelation effectively.

5.5.2.2.5 Local Correlation Structure From Figure 5.13 we see that the local Moran's I values based on the fitted values of the Kriging model are fairly strongly correlated with the local Moran's I of the actual values. It did not accurately capture the presence of spatial autocorrelation

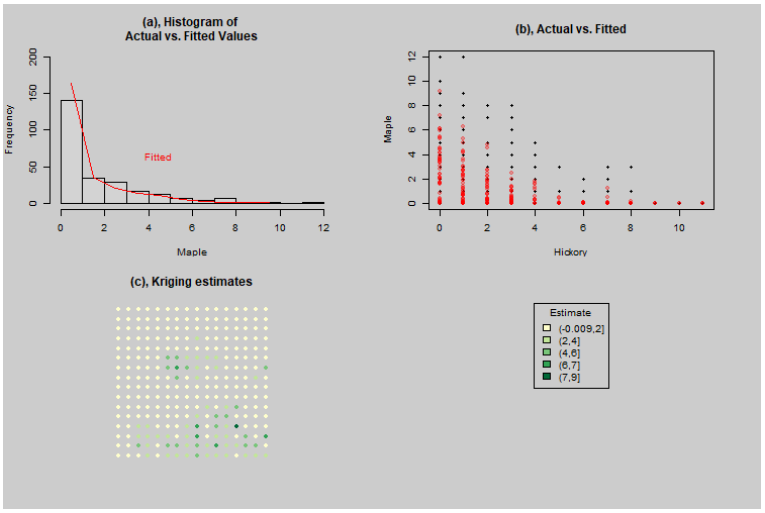

Figure 5.11: Model fit plots of the Lansing Trees dataset - Kriging model

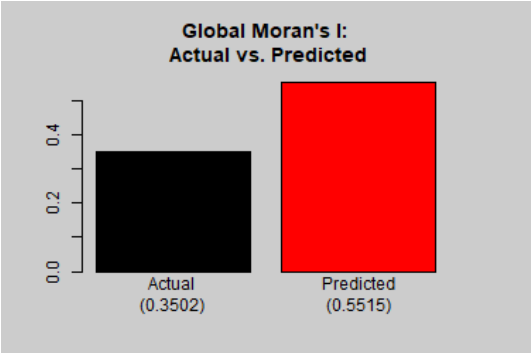

Figure 5.12: Global Moran's I difference of the Lansing Trees dataset - Kriging model

with the global Moran's I and was only partially successful in identifying where the local spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows that Accuracy of 0.6791 is achieved, which is slightly better than the baseline model. However, the Kriging model did not succeed in correctly classifying any of the low-low clusters and estimated them all to be Insignificant. This can be attributed to the tendency of the model to overpredict for low values.

5.5.2.3 Poisson Mixture Regression

5.5.2.3.1 Number of Mixtures In Figure 5.14 we see that two mixtures were chosen since they achieved the lowest AIC value.

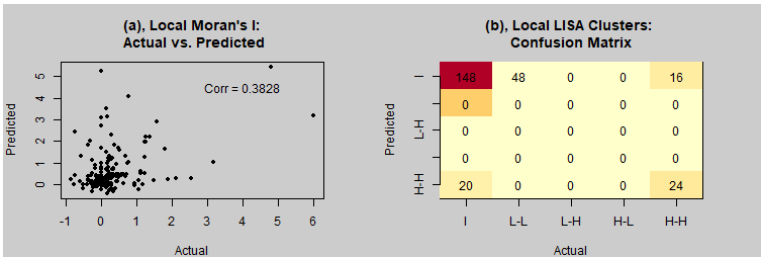

Figure 5.13: Local Moran's I difference of the Lansing Trees dataset - Kriging model

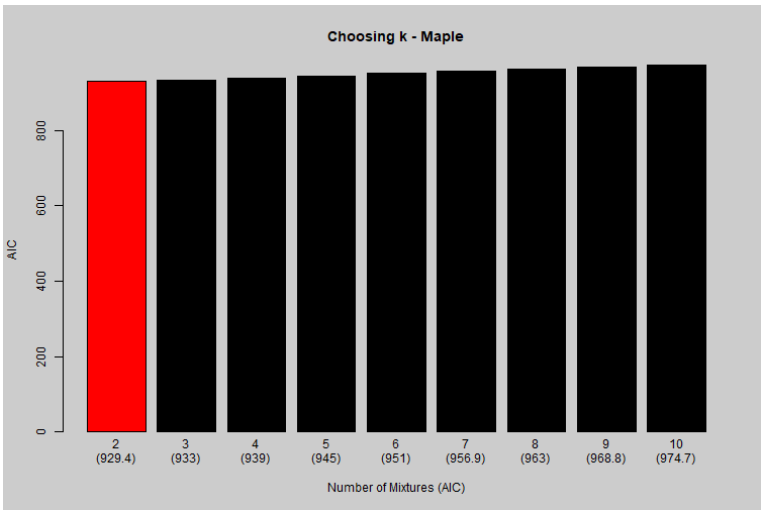

Figure 5.14: Choosing the number of mixtures of the Lansing Trees dataset - Poisson Mixture Regression model

5.5.2.3.2 Residual Analysis In Figure 5.15 we see the residual analysis of the Kriging model for the Lansing Trees dataset and learn that:

- The model has the lowest MSE (2.7729) of all three models, with many values again being overpredicted. This is reflected in the high number of estimates with a residual less than 0.
- The MSE of the normalised residuals is very close to 1, thereby outperforming the Kriging model.
- The actual and fitted values have a strong correlation of 0.8017, which outperforms the other models.
- The residuals and fitted values are, however, fairly strongly correlated.

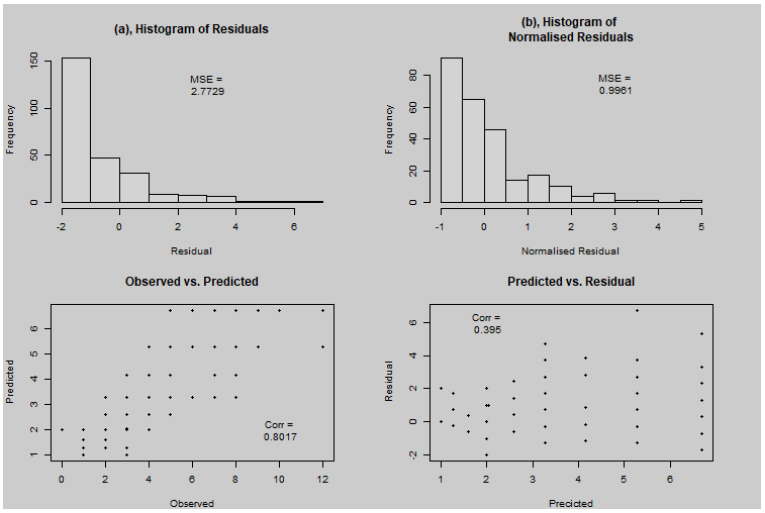

Figure 5.15: Residual analysis of the Lansing Trees dataset - Poisson Mixture Regression model

5.5.2.3.3 Model Fit From the Model Fit plots in Figure 5.16 we see that the model appears to not fit the data well for actual values between 1 and 2 and does not fit any values greater than 8. It does, however, achieve a good fit in the centre of the distribution. The spatial graph also suggests that the model captured the correlation structure well, since the graph shows a similar pattern to the one observed in Panel (a) of Figure 5.1.

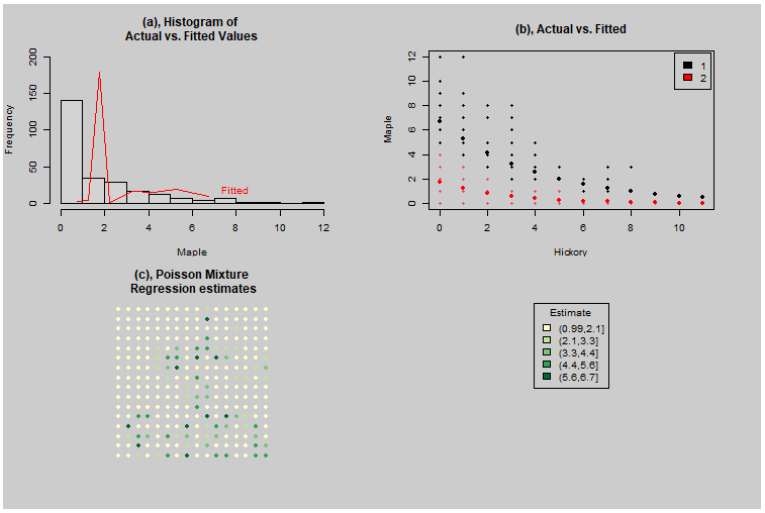

Figure 5.16: Model fit plots of the Lansing Trees dataset - Poisson Mixture Regression model

In Panel (b), the bold black and red values indicate the fitted values for the two mixture regression lines.

5.5.2.3.4 Global Correlation Structure In Figure 5.17 we see that the global Moran's I of the fitted values differs by 0.2037 from the global Moran's I of the actual values. It did not capture the global Spatial Autocorrelation effectively and underestimated the amount of global spatial autocorrelation.

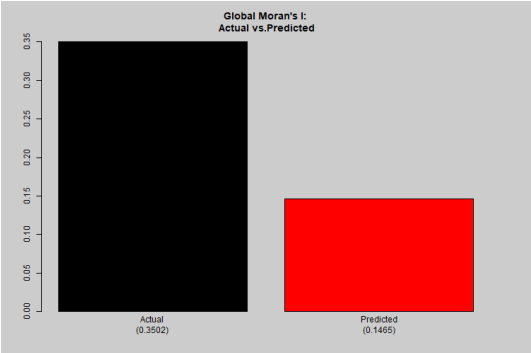

Figure 5.17: Global Moran's I difference of the Lansing Trees dataset - Poisson Mixture Regression model

5.5.2.3.5 Local Correlation Structure From Figure 5.18 we see that the local Moran's I values based on the fitted values of the Poisson Mixture Regression model are correlated with the local Moran's I of the actual values. Despite not capturing the extent of the spatial autocorrelation with the global Moran's I, it is more effective at identifying where the local spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows that Accuracy of 0.6953 is achieved, which is better than the comparative models. However, as with the Kriging, the Poisson Mixture Regression model did not succeed in correctly classifying any of the low-low clusters and estimated them all to be Insignificant. This can again be attributed to the tendency of the model to overpredict for low values. It is also not successful at classifying high-high clusters since it predicted no high values

5.5.3 Conclusion

The Lansing Trees dataset has the following ESDA metrics:

- LOW number of observations
- MEDIUM number of variables

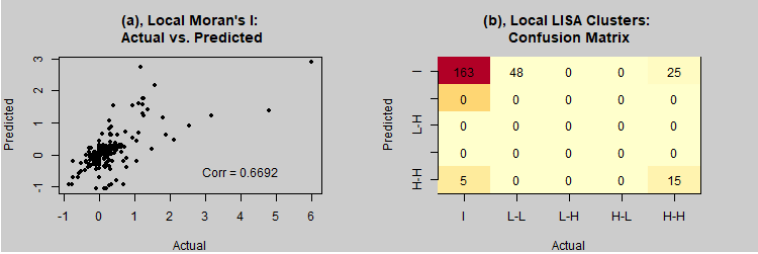

Figure 5.18: Local Moran's I difference of the Lansing Trees dataset - Poisson Mixture Regression model

- MEDIUM overdispersion
- HIGH global spatial autocorrelation
- HIGH local spatial autocorrelation

Given these metrics, we want to understand which model fits this dataset best. A summary of the models that performed best for each accuracy measure is given in Table 5.9.

Model evaluation framework		**Model**		
Model fit	Residual Analysis	GLM	**Mixture**	GP/Kriging
	Fitted values	GLM	**Mixture**	GP/Kriging
Spatial dependence structure	Global	**GLM**	Mixture	GP/Kriging
	Local	GLM	**Mixture**	GP/Kriging

Table 5.9: Summary of accuracy measures for Lansing Trees dataset

A different model can be chosen depending on the outcome that needs to be optimised. If MSE, correlation between actual and fitted values and overall ability to capture local Spatial Autocorrelation needs to be captured, a Poisson Mixture Regression Model performs best. A baseline Poisson Regression model captures global Spatial Autocorrelation best, while a Kriging model is a good model when an overall perspective of the accuracy metrics is considered.

Despite their good performance, the Kriging and Poisson Mixture Regression models did not correctly classify local Moran's I values into LISA clusters for low values. This was attributed to the tendency of both models to overpredict for low values.

5.6 Gauteng Crime

5.6.1 ESDA

A table summarising the ESDA metrics of the Gauteng Crime dataset is given in Table 5.10. The data consists of 2,309 cells/suburb polygons and 2 different types of crime, Vehicle and Residential. This suggests MEDIUM count and LOW dimensionality.

Data: Gauteng Crime		*n = 2,309*	*p = 2*
Variable		**Residential crime**	**Vehicle crime**
Dispersion		3.124	7.081
Global Moran's I		0.3319	0.3680
Local Moran's I	L-L	15.76%	18.45%
	L-H/H-L	0.00%	0.00%
	H-H	12.21%	17.15%
	I	72.02%	64.40%
Bivariate Local Moran's I	L	12.43%	
	L-H/H-L	3.17%	
	H-H	8.58%	
	I	75.83%	

Table 5.10: Summary of ESDA metrics for Gauteng Crime data

5.6.1.1 Dispersion

The Dispersion indices for the Gauteng Crime dataset is 3.124 and 7.081 for Residential and Vehicle crime, respectively. This suggests MEDIUM to HIGH overdispersion.

5.6.1.2 Spatial Autocorrelation

The Moran's scatterplot in Figure 5.19 shows the strong spatial autocorrelation for both Vehicle and Residential crime. The global Moran's I values for both variables is around 0.35, suggesting MEDIUM positive global spatial autocorrelation. Both values are statistically significant. The strong relationships are reflected in the scatterplot.

Considering the local Moran's I values for Residential crime, we see that 15.70% of suburbs belong to low-low LISA clusters. This means that these suburbs belong to a LISA cluster that has very low Residential crime. The suburbs that form part of the largest LISA cluster are in the Northern suburbs of Johannesburg region that is a high income area.

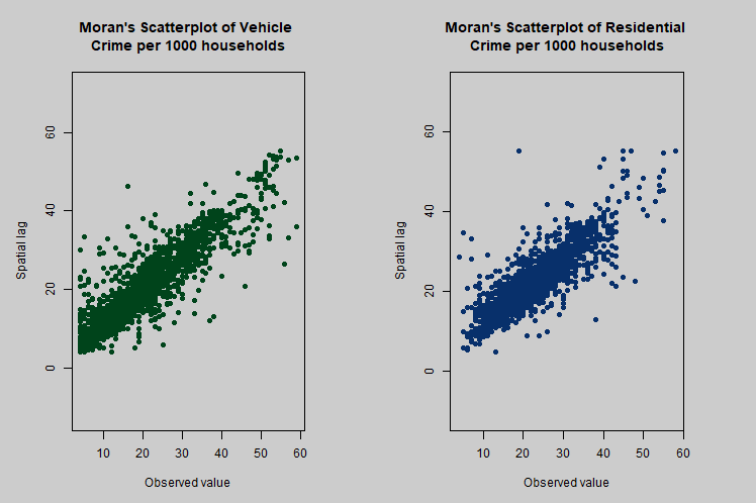

Figure 5.19: Global Moran's Scatter Plot for Gauteng Crime dataset

A strong relationship between the spatially-lagged values and the observed values is seen for both Vehicle and Residential crime.

Contrastingly, 12.28% (or 283) of the suburbs belong to high-high LISA clusters, or clusters with very high Residential crime. Some of the most prominent suburbs to feature in these clusters are Kameeldrift and outer areas of Pretoria North that are notorious crime hotspots and are lower income areas. We conclude that almost a third of the data for Residential crime consists of LISA clusters.

Considering the local Moran's I values for Vehicle crime, 18.52% of the suburbs belong to low-low clusters. A visual representation of the LISA clusters for both variables is shown in Figure 5.20. The suburbs that form part of the larger clusters here are centered around the poorer Non-Urban areas of Johannesburg and the East Rand. 17.18% of the suburbs belong to high-high clusters. The suburbs that feature prominently here are suburbs in the wealthier areas of Pretoria East. This suggests that more than a third of the data consists of either cold-(low-low) or hotspots (high-high) and that Vehicle crime is more prevalent - per 1000 households - in wealthier suburbs than in poorer suburbs. This is reasonable, since there are also more vehicles located in wealthier suburbs. We conclude that the Gauteng Crime dataset has HIGH local spatial autocorrelation.

From the bivariate Moran's I values, we learn that 11.93% of the suburbs belong to low-low bi-

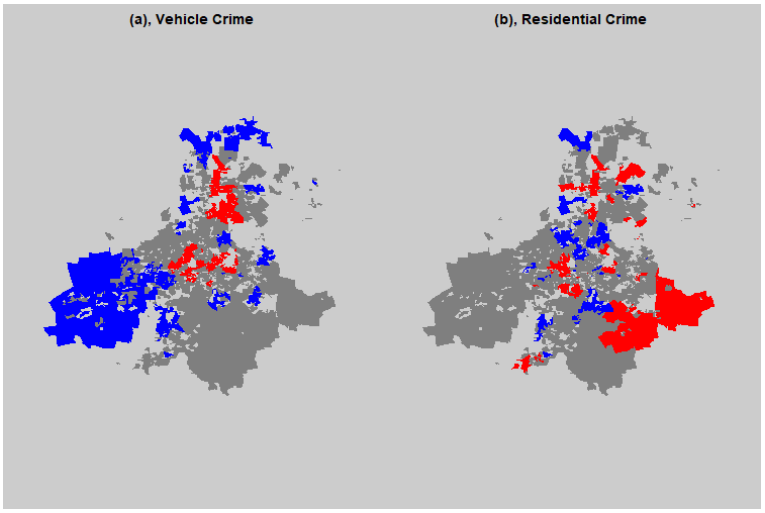

Figure 5.20: Local LISA clusters Gauteng Crime dataset

A high proportion of suburbs belong to LISA clusters for both Vehicle and Residential crime.

variate LISA clusters. Prominent suburbs in this cluster are Waterfall City and Steyn City. These suburbs are prominent due to their high income demographic. This means that low Residential crimes are grouped close to low Vehicle crimes in these suburbs. Only 8.89% of the suburbs belong to high-high bivariate LISA clusters. The suburbs that make up this unfortunate group are a combination of wealthy suburbs such as Brooklyn and Menlo Park in Pretoria and poorer suburbs such as Kempton Park in the Ekurhuleni metropolitan area. This suggests that high Residential crime values do not often group close to high Vehicle crime values and do not form purely among income groups. The clusters can be seen in Figure 5.21. We also note some outlier values (3.42%) which can be interpreted as high Residential crime grouping close to low Vehicle crime and vice versa. Outlying areas such as the aformentioned Kameeldrift and Boschkop tend to have high Residential crime and low Vehicle crime cluster together, while wealthier areas such as Morningside and River Club in Johannesburg have higher Vehicle crime and lower Residential crime cluster together. We conclude from these figures that the data has MEDIUM bivariate spatial autocorrelation.

5.6.1.3 Conclusion

A summary of the ESDA metrics for the Gauteng Crime dataset is given in Table 5.11.

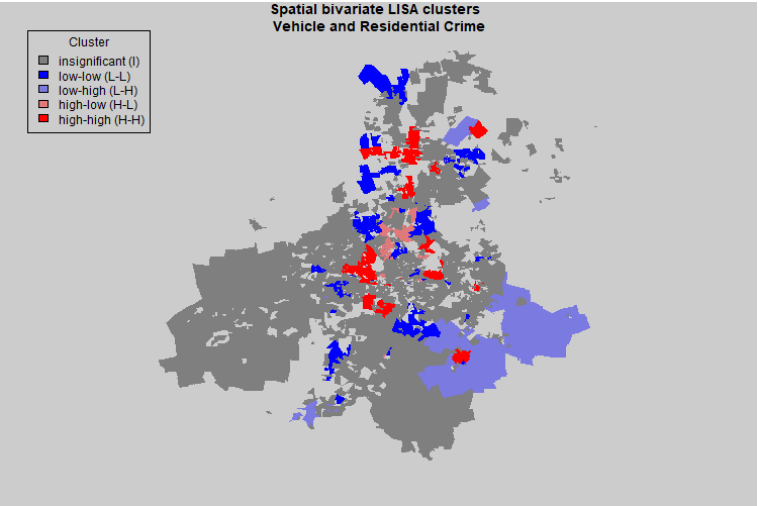

Figure 5.21: Bivariate LISA clusters for Gauteng Crime dataset

Note the higher prevalence of low-low and high-high clusters in comparison to low-high and high-low clusters, which means that high Residential crime groups closer to high Vehicle crime more frequently than it does to low Vehicle crime.

ESDA Framework		Category				
Number of variables		**Very Low**	Low	Medium	High	Very High
Number of Observations		Very Low	**Low**	Medium	High	Very High
Dispersion		Very Low	Low	Medium	**High**	Very High
Spatial Autocorrelation	Global	Very Low	Low	**Medium**	High	Very High
	Local	Very Low	Low	Medium	**High**	Very High
	Bivariate	Very Low	Low	**Medium**	High	Very High

Table 5.11: Summary of ESDA framework for Gauteng Crime dataset

5.6.2 Modelling

5.6.2.1 Baseline

5.6.2.1.1 Residual Analysis In Figure 5.22 we see the residual analysis of the baseline model for the Gauteng Crime dataset and learn that:

- The model has a very high MSE (101.5707). There is also a clear pattern of underprediction, with many residuals greater than zero.
- The MSE of the normalised residuals is 0.9996. Which indicates a good model fit.
- The actual and fitted values are correlated fairly strongly with a correlation of 0.5025.
- The residuals and fitted values are not strongly correlated with a correlation of -0.0824.

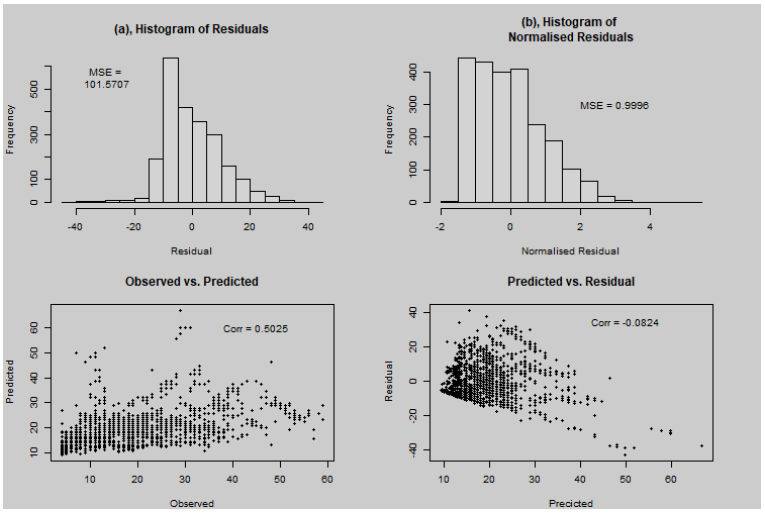

Figure 5.22: Residual analysis of the Gauteng Crime dataset - baseline Poisson Regression model

5.6.2.1.2 Model Fit From the model fit plots in Figure 5.23 we see that the model appears to not fit the data very well when considering the histogram with fitted values in Panel (a). From Panel (b) it is clear that the model is too simple to accurately fit the data. The spatial graph also suggests that the model did not capture the correlation structure well, since the graph shows a very different pattern to the one observed in Panel (a) of Figure 5.2.

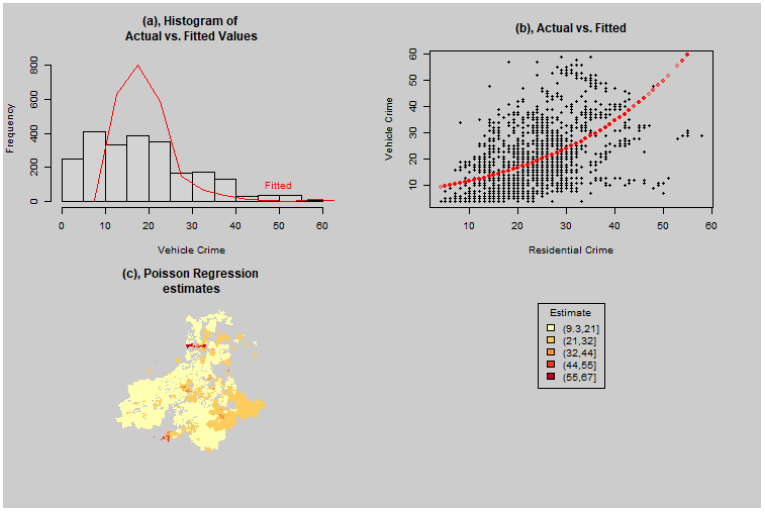

Figure 5.23: Model fit plots of the Gauteng Crime dataset - baseline Poisson Regression model

5.6.2.1.3 Global Correlation Structure In Figure 5.24 we see that the global Moran's I of the fitted values differs by 0.0359 from the global Moran's I of the actual values. It captured the global Spatial Autocorrelation effectively.

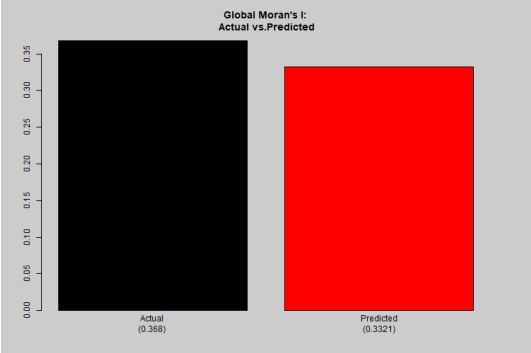

Figure 5.24: Global Moran's I difference of the Gauteng Crime dataset - baseline Poisson Regression model

5.6.2.1.4 Local Correlation Structure From Figure 5.25 we see that the local Moran's I values based on the fitted values of the baseline model are not correlated with the local Moran's I of the actual values. It is not effective at identifying where the local spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows Accuracy of 0.6955. It is effective at identifying where the bivariate spatial autocorrelation occurs in the dataset which is confirmed by the confusion matrix, with an Accuracy of 0.9493.

5.6.2.2 Kriging

5.6.2.2.1 Semivariogram In Figure 5.26 we see that the Spherical semivariogram fits the empirical semivariogram best. The fitted semivariogram has a Nugget of 0.0445, a Sill of 0.4492 and a Range of 0.2414. Note that the empirical semivariogram exhibits a periodical behaviour for cells that are further than 0.35 units apart. None of our possible semivariograms capture this effectively. A more complex combination of a Spherical and Periodic/Locally periodic kernel [27] might be more suited to this dataset, but is beyond the scope of this dissertation.

5.6.2.2.2 Residual Analysis In Figure 5.27 we see the residual analysis of the Kriging model for the Gauteng Crime dataset and learn that:

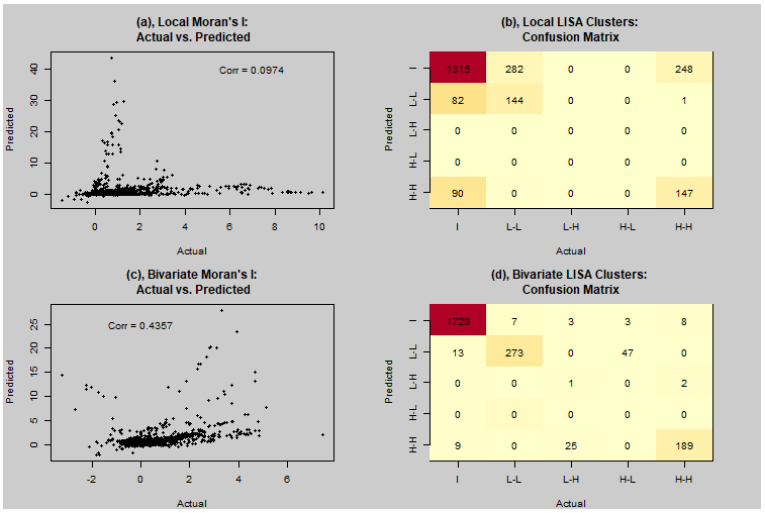

Figure 5.25: Local and Bivariate Moran's I difference of the Gauteng Crime dataset - baseline Poisson Regression model

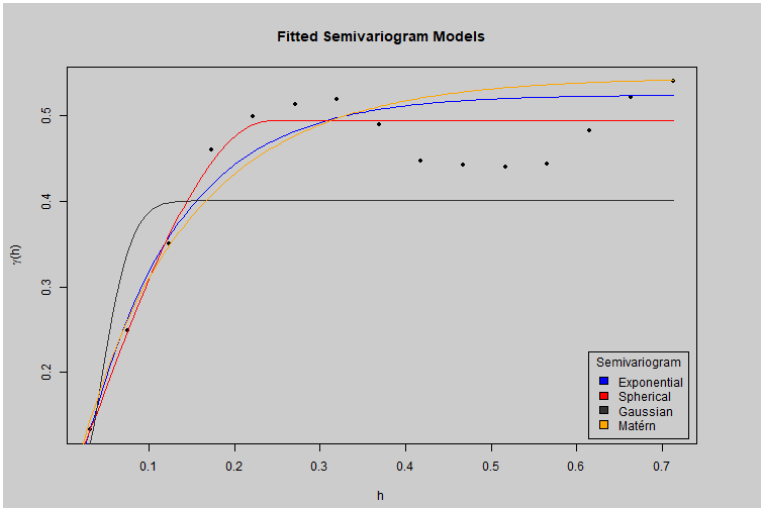

Figure 5.26: Semivariograms of the Gauteng Crime dataset - Kriging model

- The model has a MSE (20.8546) which is a fifth of the baseline model's MSE. There is also no clear pattern of either over or underprediction, with the residual distribution being fairly symmetric.
- The MSE of the normalised residuals is 0.7606. Which indicates a good model fit.
- The actual and fitted values are strongly correlated with a correlation of 0.9222.
- The residuals and fitted values are not strongly correlated with a correlation of -0.005.

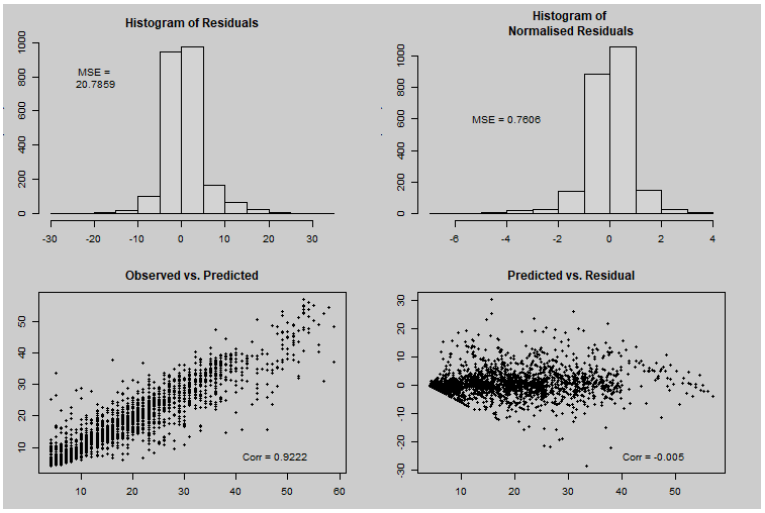

Figure 5.27: Residual analysis of the Gauteng Crime dataset - Kriging model

5.6.2.2.3 Model fit From the Model Fit plots in Figure 5.28 we see that the model appears to fit the data very well when considering the Histogram with fitted values in Panel (a). The model's ability to interpolate is clearly shown in the way that the fitted values are plotted. The spatial graph also suggests that the model captured the correlation structure well, since the graph shows a similar pattern to the one observed in Panel (a) of Figure 5.2.

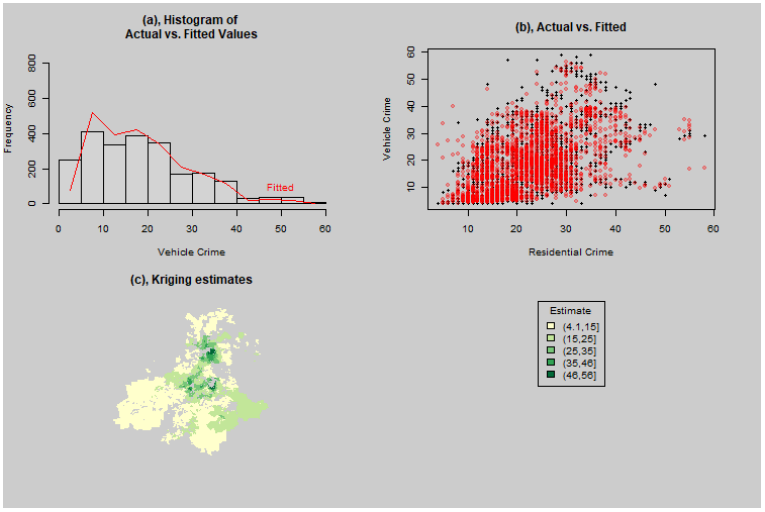

Figure 5.28: Model fit plots of the Gauteng Crime dataset - Kriging model

5.6.2.2.4 Global Correlation Structure In Figure 5.29 we see that the global Moran's I of the fitted values differs by 0.0399 from the global Moran's I of the actual values. It captured the global Spatial Autocorrelation effectively.

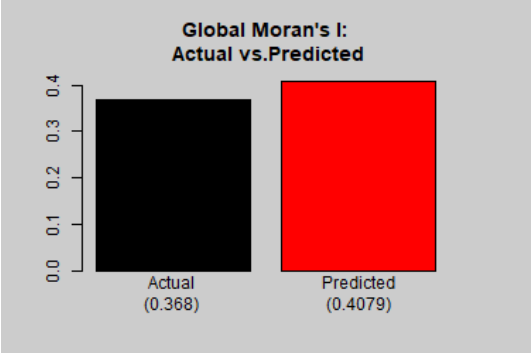

Figure 5.29: Global Moran's I difference of the Gauteng Crime dataset - Kriging model

5.6.2.2.5 Local and Bivariate Correlation Structure From Figure 5.30 we see that the local and bivariate Moran's I values based on the fitted values of the Kriging model are strongly correlated with the local Moran's I of the actual values. It is very effective at identifying where the local and bivariate spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows that Accuracy of 0.9281 and 0.9675 is achieved for the local and bivariate Moran's I, respectively. This shows a clear improvement over the baseline model. Due to the model not over or underpredicting, it could also correctly classify low-low clusters, unlike the Kriging model in the Lansing Trees dataset.

5.6.2.3 Poisson Mixture

5.6.2.3.1 Mixtures In Figure 5.31 we see that eight mixtures achieve the best AIC value.

5.6.2.3.2 Residual Analysis In Figure 5.32 we see the residual analysis of the Kriging model for the Gauteng Crime dataset and learn that:

- The model has a MSE (7.5204) which is a third of the Kriging model's MSE and outperforms the baseline model's MSE by an order of magnitude. There is also no clear pattern of either over or underprediction, with the residual distribution being fairly symmetric.

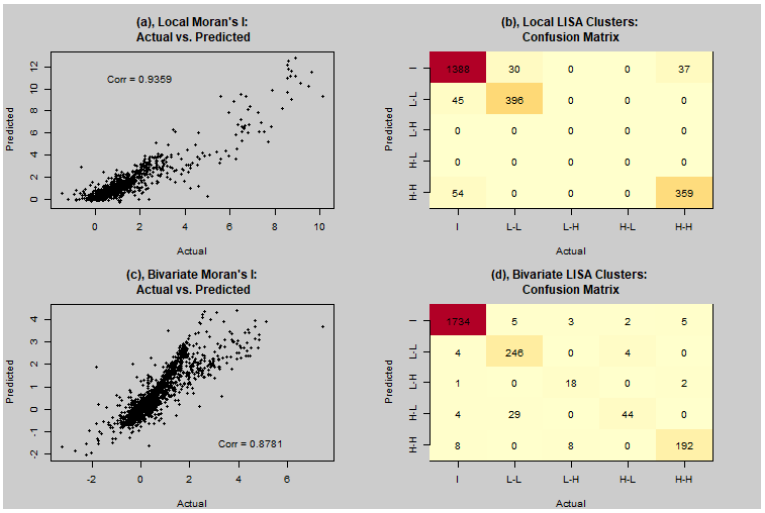

Figure 5.30: Local and bivariate Moran's I difference of the Gauteng Crime dataset - Kriging model

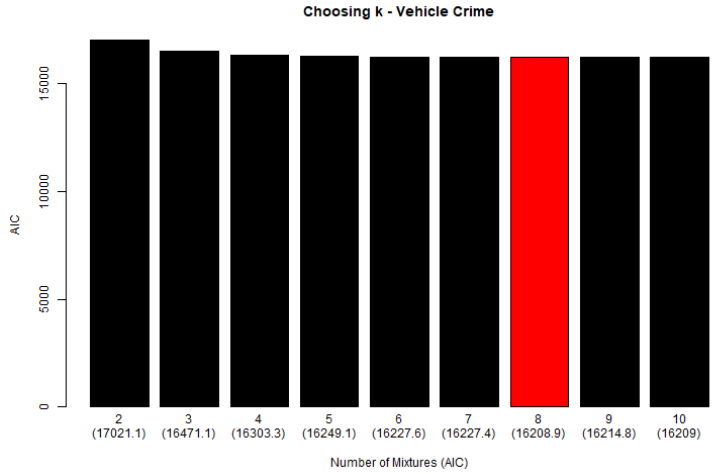

Figure 5.31: Choosing the number of mixtures of the Gauteng Crime dataset - Poisson Mixture Regression model

- The MSE of the normalised residuals is 0.9996. Which indicates a very good model fit.
- The actual and fitted values are strongly correlated (0.9722).
- The residuals and fitted values are not strongly correlated (0.0972).

5.6.2.3.3 Model fit From the Model Fit plots in Figure 5.33 we see that the model appears to the data very well when considering the Histogram with fitted values in Panel (a). All the fitted

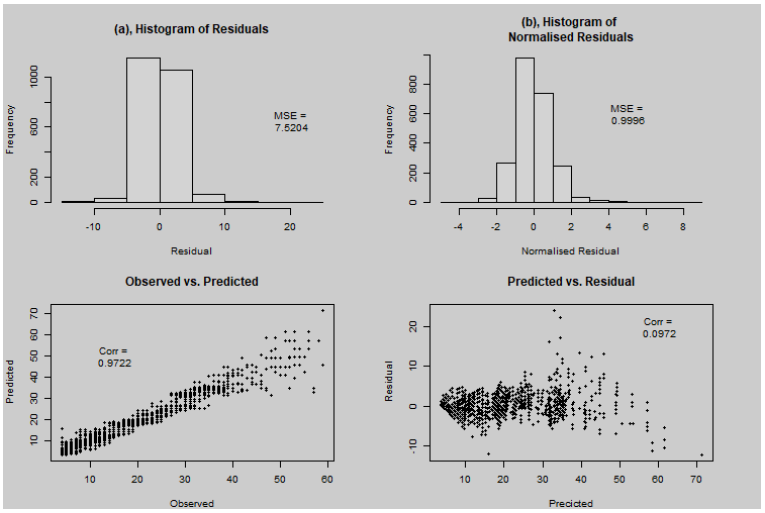

Figure 5.32: Residual analysis of the Gauteng Crime dataset - Poisson Mixture Regression model

mixture regression lines are shown in Panel (b). The spatial graph also suggests that the model captured the correlation structure well, since the graph shows a similar pattern to the one observed in Panel (a) of Figure 5.2.

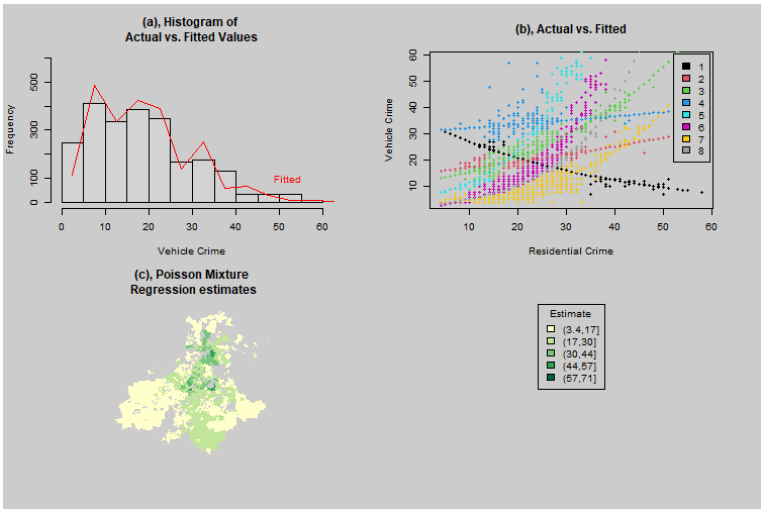

Figure 5.33: Model fit plots of the Gauteng Crime dataset - Poisson Mixture Regression model

5.6.2.3.4 Global Correlation Structure In Figure 5.34 we see that the global Moran's I of the fitted values differs by 0.0101 from the global Moran's I of the actual values. It captured the global

Spatial Autocorrelation very effectively.

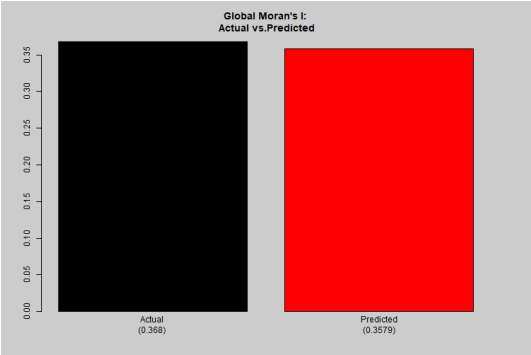

Figure 5.34: Global Moran's I difference of the Gauteng Crime dataset - Poisson Mixture Regression model

5.6.2.3.5 Local and Bivariate Correlation Structure From Figure 5.35 we see that the local and bivariate Moran's I values based on the fitted values of the Poisson Mixture Regression model are strongly correlated with the local Moran's I of the actual values. It is very effective at identifying where the local and bivariate spatial autocorrelation occurs in the dataset. This is confirmed by the confusion matrix in Panel (b), which shows that Accuracy of 0.9324 and 0.971 is achieved for the local and bivariate Moran's I, respectively. This shows a clear improvement over the baseline model and some improvement over the Kriging model. Due to the model not over or underpredicting, it could also correctly classify low-low clusters, unlike the Poisson Mixture Regression model in the Lansing Trees dataset.

5.6.3 Conclusion

The Gauteng Crime dataset has the following ESDA metrics:

- MEDIUM counts
- MEDIUM to HIGH overdispersion
- MEDIUM global spatial autocorrelation
- HIGH local spatial autocorrelation
- MEDIUM bivariate spatial autocorrelation

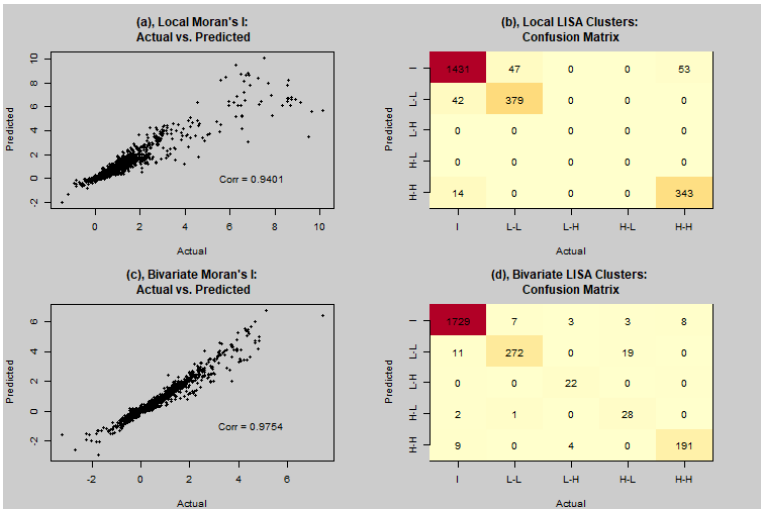

Figure 5.35: Local and bivariate Moran's I difference of the Gauteng Crime dataset - Poisson Mixture Regression model

Given these metrics, we want to understand which model fits this dataset best. A summary of the models that performed best for each accuracy measure is given in Table 5.12.

Model evaluation framework		**Model**		
Model fit	Residual Analysis	GLM	**Mixture**	GP/Kriging
	Fitted values	GLM	Mixture	GP/Kriging
Spatial dependence structure	Global	GLM	**Mixture**	GP/Kriging
	Local	GLM	**Mixture**	GP/Kriging
	Bivariate	GLM	**Mixture**	GP/Kriging

Table 5.12: Summary of accuracy measures for Gauteng Crime dataset

The Poisson Mixture Regression model and Kriging outperformed the baseline model. The Poisson Mixture Regression Model performed slightly better than the Kriging model when considering most accuracy metrics. However, the Kriging model does provide an estimate of prediction variance (as can be seen in Figure 5.36) which the Poisson Mixture Regression model does not provide. The final model decision can thus be made between whether the goal is to achieve a higher model accuracy and ability to capture the global, local and bivariate spatial autocorrelation strucure, or to provide some uncertainty measure around a given estimate.

Consdering both datasets, we see that the results for both datasets confirm that Poisson Mixture Regressions are excellent models for overdispersed data, based on most accuracy measures.

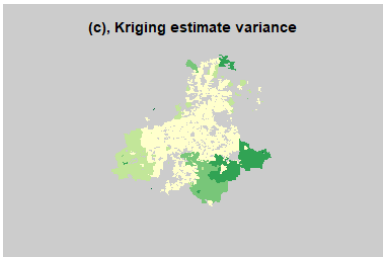

Figure 5.36: Variance of the Kriging model estimates of the Gauteng Crime dataset

This confirms the results of [18] and [46] in a spatial context. The GP/Kriging models also outperformed the baseline models, which suggests that consideration of spatial autocorrelation leads to better results. Finally, the ESDA framework proposed allowed us to learn valuable insights into the dataset structure and can lead to better informed model selections in future work.

Chapter 6

Conclusion

In this final chapter, all previous chapters are summarised. The application results are reviewed and the dissertation is concluded with suggestions for future work.

6.1 Exploratory Spatial Data Analysis

In this chapter, two important topics regarding ESDA were considered. The first is dispersion which measures the extent to which values in a variable are spread around a central value or over the variable's domain. We explored popular measures of dispersion such as the variance to mean ratio, Green's coefficient and Morisita's Index.

Next, we explored spatial autocorrelation from Global, Local and bivariate perspectives. The Global perspective is concerned with the identification of spatial dependence in a dataset. Local spatial autocorrelation attempts to pinpoint where the spatial dependence can be found in the dataset and the bivariate case considers whether two or more variables cluster in similar regions. The calculation of Moran's I for each of these cases was considered. It was also illustrated how to test the statistical significance of an observed Moran's I statistic by means of a fast analytical and/or slower, but more robust Monte Carlo approach.

6.2 Gaussian Processes

In this chapter GPs and Kriging models were described. Particular care was taken in the formulation of GPs as an extension of linear regression from the realm of vectors into that of functions. The theory of both models were reviewed and the connection between the models through the covariance function and semivariogram was shown in Equation 3.19. A detailed discussion of covariance functions was given along with many illustrative examples. The chapter concluded with a short review of some of the literature on Poisson Kriging.

6.3 Poisson Mixture Models

Poisson mixtures were explored in this chapter. A general definition of mixture models was given and then applied to the Poisson distribution, whereupon the Poisson-specific parameters obtained through the EM algorithm were given. Mixture regressions, mixture models in a supervised context, were again generally defined with a Poisson-based formulation then given. A method of choosing the correct number of mixtures (AIC) was given along with a popular goodness of fit tool (rootogram). The chapter concluded with a short review of mixture research conducted in a spatial context. Frequentist and Bayesian approaches were described along with research that focused on two prevalent problems of spatial data - overdispersion and sparsity.

6.4 Application

In Section 5.1 an ESDA framework was formulated. The framework is based on the exploratory metrics discussed in Chapter 2, namely dispersion and spatial autocorrelation (Global, Local and bivariate). For each metric, categories are defined based on thresholds for VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH values. The measures were then calculated for two datasets that are described in Section 5.2. The first dataset looks at the distribution of trees in Lansing Forest, while the second considers Crime in Gauteng. Each dataset is then classified in terms of the categorised ESDA measures. Summaries of the ESDA metrics for each dataset can be found in Table 5.7 and Table 5.11, respectively.

Kriging and Poisson Mixture Regression models are then trained on the data and are compared

to a baseline Poisson Generalised Linear Model (GLM) in terms of model fit as well as the model's ability to capture Global, Local and bivariate spatial autocorrelation. Model fit is evaluated by means of a residual analysis and MSE. The model's ability to capture Global spatial autocorrelation is measured by observing the absolute difference between the Actual Global Moran's I and the Global Moran's I obtained from the model estimates. Local and bivariate spatial autocorrelation is evaluated by a confusion matrix and observing how many spatial entities were classified into the same LISA clusters as the Actual Local and bivariate Moran's I values.

For both datasets, Poisson Mixture Regression performed best which confirmed the results of [18] and [46] that Mixture Regression performs well on overdispersed data in a spatial context.

In terms of the ability of the different models to capture spatial autocorrelation, the Poisson Mixture Regression outperformed the Kriging model by a slight margin in both datasets. This is an important learning, since the models approach the problem of spatial autocorrelation from different perspectives. Mixture regression captures spatial autocorrelation as a result of the clustering that is performed while a Kriging/GP model explicitly considers the distance between spatial entities. The slightly poorer performance of the Kriging/GP model in this regard can be offset by the model's ability to provide a confidence interval that is based on actual sample data around the given prediction. This is something which the Poisson Mixture Regression model does not provide as standard.

The results do not provide a definitive decision on the best model to use on a dataset with the given set of ESDA metrics. We do not suggest that each of the different combinations of ESDA metrics will necessarily lead to a different model being chosen as the best. Instead, it provides a systematic process of analysing ESDA metrics and assessing model accuracy for competitor models.

6.5 Shortcomings and further work

This paper is concluded with a short summary of different shortcomings of our methodology as well as remediations that can be made to address these shortcomings

6.5.1 ESDA

Further research opportunities and practical application are plentiful in the expansion of the ESDA framework. Firstly, the category thresholds of the chosen ESDA metrics are based on arbitrary boundaries that make intuitive sense. Formal boundaries for the category thresholds of the chosen ESDA metrics can be formulated. A potential approach will be to estimate the metrics on a comprehensive sample and make a decision on boundaries based on the distribution of the ESDA metrics that were calculated.

Other ESDA metrics can be added. Sparsity is a common problem in spatial problems where rare events need to be estimated and can have a severe impact on model estimation techniques [85]. Stationarity is an important condition for Ordinary Kriging models and has not been considered here, but can also be included. An ensemble approach can be used to first find stationary clusters with a method such as a Mixture Model and then apply Kriging models on these stationary clusters. The evaluation of semivariograms can also be explored in more rigorous detail than has been done here. Multivariate dispersion indices [48] can also be considered.

A detailed study of ESDA [3] can be consulted for additional metrics.

6.5.2 Modelling

Our experiments only considered two extensions of the Poisson distribution. They were chosen for their different approaches to addressing dispersion and spatial autocorrelation. Other models can therefore be included. Classic spatial models such as spatial autoregressive (SAR) models and a Poisson log-normal random effects model where the random effects are modelled by a conditional autoregressive (CAR) model [78] are viable options. Other novel approaches such as hierarchical models [10] and Bayesian spatial Poisson-lognormal models [65] can also be compared. The Conway-Maxwell Poisson (COM-Poisson) model that caters for both underdispersion as well as overdispersion is another possible alternative [73]. Finally, the problem of modelling spatial count data can also be considered from the Negative Binomial perspective as an alternative to the Poisson GLM used here [71].

6.5.3 Model evaluation

While the model evaluation framework presented addresses the ability of competitor models to capture ESDA metrics, alternative accuracy measures can be studied depending on the ESDA metrics that are considered [90]. Newer goodness-of-fit metrics applicable to count regressions, such as rootograms [42] can be included in the model assessments. The confusion matrix approached proposed in the evaluation of estimated LISA clusters can be adapted to optimise for different confusion matrix metrics such as Specificity - instead of pure Accuracy - where the cost of misclassification can have a high impact.

6.5.4 Multivariate models and covariates

We did not consider multivariate models and the covariates that were used for the baseline and Poisson Mixture Regression models were not rigorously examined for their predictive ability. The modelling framework can therefore also be expanded to include multivariate models. In particular, Co-Kriging and extensions thereof offer many research opportunities [74]. Finally, Multivariate Analysis of Variance (MANOVA) tests can be implemented in R [30] to determine the impact of covariates and whether variables need to be included in the model and provide a potentially rich addition to the proposed framework.

Bibliography

[1] Mohammad Ali, Pierre Goovaerts, Nushrat Nazia, M Zahirul Haq, Mohammad Yunus, and Michael Emch. Application of Poisson Kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh. *International journal of health geographics*, 5(1):1–11, 2006.

[2] Anar Amgalan, Lilianne R Mujica-Parodi, and Steven S Skiena. Fast Spatial Autocorrelation. *arXiv preprint arXiv:2010.08676*, 2020.

[3] Natalia Andrienko and Gennady Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006.

[4] Luc Anselin. Local indicators of spatial association—LISA. *Geographical analysis*, 27(2):93–115, 1995.

[5] Luc Anselin. Spatial Econometrics. *A companion to theoretical econometrics*, 310330, 2001.

[6] Luc Anselin. A local indicator of multivariate spatial association: extending Geary's C. *Geographical Analysis*, 51(2):133–150, 2019.

[7] Luc Anselin. *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. Routledge, 2019.

[8] Adrian Baddeley and Rolf Turner. spatstat: An R package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[9] Simon Baechler and Stefano Caneppele. Exploratory Spatial Data Analysis Methodologies (ESDA). *The Routledge International Handbook of Forensic Intelligence and Criminology*, page 19, 2017.

[10] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2003.

[11] Fabian Barthel and Christian-Mathias Wellbrock. Regional competition and knowledge spillovers-spatial dependence in international football success. *Available at SSRN 1635008*, 2010.

[12] Jeff A Bilmes et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.

[13] Roger Bivand, Jakub Nowosad, and Robin Lovelace. *spData: Datasets for Spatial Analysis*, 2020. R package version 0.3.8.

[14] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied Spatial Data Analysis with R, Second edition*. Springer, NY, 2013.

[15] Paul Bolstad. *GIS fundamentals: A first text on geographic information systems*. Eider (PressMinnesota), 2016.

[16] Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443, 1998.

[17] Carlos M Carvalho, Hedibert F Lopes, Nicholas G Polson, and Matt A Taddy. Particle learning for general mixtures. *Bayesian Analysis*, 5(4):709–740, 2010.

[18] Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[19] Noel Cressie. The origins of Kriging. *Mathematical geology*, 22(3):239–252, 1990.

[20] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 2015.

[21] P De Jong, C Sprenger, and F Van Veen. On extreme values of moran's i and geary's c. *Geographical Analysis*, 16(1):17–24, 1984.

[22] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[23] JL Deutsch, J Szymanski, and CV Deutsch. Checks and measures of performance for Kriging estimates. *Journal of the Southern African Institute of Mining and Metallurgy*, 114(3):223–223, 2014.

[24] Hartwig Dobesch, Pierre Dumolard, and Izabela Dyras. *Spatial interpolation for climate data: the use of GIS in climatology and meteorology*. John Wiley & Sons, 2013.

[25] Stéphane Dray, Sonia Saïd, and Françis Débias. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of vegetation science*, 19(1):45–56, 2008.

[26] David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[27] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013.

[28] Brian Everitt and Anders Skrondal. *The Cambridge Dictionary of Statistics*, volume 106. Cambridge University Press Cambridge, 2002.

[29] Carmen Fernández and Peter J Green. Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the royal statistical society: series B (Statistical methodology)*, 64(4):805–826, 2002.

[30] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019.

[31] Chris Fraley and Adrian E Raftery. Model-based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[32] Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, 2006.

[33] Andrej Gajdos, Martina Hancova, and Jozef Hanc. Kriging methodology and its development in forecasting econometric time series. *Statistika-Statistics and Economy Journal*, 97(1):59–73, 2017.

[34] Mark S Gilthorpe, Morten Frydenberg, Yaping Cheng, and Vibeke Baelum. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine*, 28(28):3539–3553, 2009.

[35] Tilmann Gneiting, Zoltán Sasvári, and Martin Schlather. Analogies and correspondences between variograms and covariance functions. *Advances in Applied Probability*, 33(3):617–630, 2001.

[36] Ursula Gonzales-Barron, Marie Kerr, James J Sheridan, and Francis Butler. Count data distributions and their zero-modified equivalents as a framework for modelling microbial data with a relatively high occurrence of zero counts. *International Journal of Food Microbiology*, 136(3):268–277, 2010.

[37] Phillip I Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Science & Business Media, 2006.

[38] Pierre Goovaerts. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson Kriging. *International Journal of Health Geographics*, 5(1):1–31, 2006.

[39] Roger H Green. Measurement of non-randomness in spatial distributions. *Researches on Population Ecology*, 8(1):1–7, 1966.

[40] Benedikt Gräler, Edzer Pebesma, and Gerard Heuvelink. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016.

[41] Bettina Grün and Friedrich Leisch. Finite Mixtures of Generalized Linear Regression Models. In *Recent advances in linear models and related areas*, pages 205–230. Springer, 2008.

[42] Bettina Grün and Friedrich Leisch. Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, 2008.

[43] Zifei Han and Victor De Oliveira. gcKrig: An R package for the Analysis of Geostatistical Count Data Using Gaussian copulas. *Journal of Statistical Software*, 87(13):1–32, 2018.

[44] John S Heywood. Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics*, 22(1):335–355, 1991.

[45] Stuart H Hurlbert. Spatial distribution of the Montane unicorn. *Oikos*, pages 257–271, 1990.

[46] David I. Inouye, Eunho Yang, Genevera I. Allen, and Pradeep Ravikumar. A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398, 2017.

[47] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[48] Célestin C Kokonendji and Pedro Puig. Fisher dispersion index for multivariate count distributions: A review and a new proposal. *Journal of Multivariate Analysis*, 165:180–193, 2018.

[49] Athanasios Kottas and Bruno Sansó. Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137(10):3151–3163, 2007.

[50] Daniel G Krige. *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. PhD thesis, University of the Witwatersrand, 1951.

[51] Sang-Il Lee. Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. *Journal of geographical systems*, 3(4):369–385, 2001.

[52] Justin Lessler, Henrik Salje, M Kate Grabowski, and Derek AT Cummings. Measuring spatial dependence for infectious disease epidemiology. *PLOS One*, 11(5):e0155249, 2016.

[53] Michael T Light and Casey T Harris. Race, space, and violence: Exploring spatial dependence in structural covariates of white and black violent crime in us counties. *Journal of Quantitative Criminology*, 28(4):559–586, 2012.

[54] Jie Lin. A Local Model for Multivariate Analysis: Extending Wartenberg's Multivariate Spatial Correlation. *Geographical Analysis*, 52(2):190–210, 2020.

[55] Heping Liu, Jing Shi, and Ergin Erdem. Prediction of wind speed time series using modified Taylor Kriging method. *Energy*, 35(12):4870–4879, 2010.

[56] Monte Lloyd. Mean Crowding. *The Journal of Animal Ecology*, 36:1–30, 1967.

[57] David JC MacKay. Introduction to Gaussian processes. *NATO ASI series. Series F: Computer and Systems Sciences*, 168:133–166, 1998.

[58] Yuzo Maruyama. An alternative to Moran's I for spatial autocorrelation. *arXiv e-prints*, pages arXiv–1501, 2015.

[59] Georges Matheron. Principles of Geostatistics. *Economic Geology*, 58(8):1246–1266, 1963.

[60] GJ McLachlan. On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research*, 6(1):76–98, 1997.

[61] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[62] Masaaki Morisita. I σ-Index, a measure of dispersion of individuals. *Researches on Population Ecology*, 4(1):1–7, 1962.

[63] Julie M Mueller and John B Loomis. Spatial dependence in hedonic property models: Do different corrections for spatial dependence result in economically significant differences in estimated implicit prices? *Journal of Agricultural and Resource Economics*, 33:212–231, 2008.

[64] Chipo Mufudza and Hamza Erol. Poisson Mixture Regression Models for Heart Disease Prediction. *Computational and Mathematical Methods in Medicine*, 2016, 2016.

[65] Sirajum Munira, Ipek N Sener, and Boya Dai. A Bayesian spatial Poisson-lognormal model to examine pedestrian crash severity at signalized intersections. *Accident Analysis & Prevention*, 144:105679, 2020.

[66] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[67] Anand Patil, David Huard, and Christopher J Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4):1, 2010.

[68] N. G. N. Prasad and J. N. K. Rao. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

[69] Frédéric Puech. How do criminals locate? Crime and spatial dependence in Minas Gerais. In *Documento presentado en la conferencia International Economic Policies in the New Millenium, Coimbra, Portugal. Disponible en: http://www4. fe. uc. pt/30years/papers/68. pdf*, 2004.

[70] Brian D Ripley. *Spatial Statistics*, volume 575. John Wiley & Sons, 2005.

[71] Kimberly F Sellers and Galit Shmueli. A flexible regression model for count data. *The Annals of Applied Statistics*, pages 943–961, 2010.

[72] Shashi Shekhar, Pusheng Zhang, Yan Huang, and Ranga Raju Vatsavai. Trends in spatial data mining. *Data mining: Next generation challenges and future directions*, pages 357–380, 2003.

[73] Galit Shmueli, Thomas P Minka, Joseph B Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

[74] Lynette M Smith, Walter W Stroup, and David B Marx. Poisson Co-Kriging as a Generalized Linear Mixed Model. *Spatial Statistics*, 35:100399, 2020.

[75] Edward Lloyd Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, UCL (University College London), 2007.

[76] Michael C Stambaugh, Richard P Guyette, Joseph M Marschall, and Daniel C Dey. Scale dependence of oak woodland historical fire intervals: contrasting The Barrens of Tennessee and Cross Timbers of Oklahoma, USA. *Fire Ecology*, 12(2):65–84, 2016.

[77] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 2012.

[78] Hal Stern and Noel A Cressie. Inference for extremes in disease mapping. *Disease Mapping and Risk Assessment for Public Health*, 1:63–84.

[79] M Subedi and R Subedi. Identification and mapping of risk areas of rhino poaching; a geospatial approach: a case study from eastern sector of Chitwan National Park, Nepal. *Banko Janakari*, 27(2):12–20, 2017.

[80] Matthew A Taddy. Autoregressive mixture models for dynamic spatial poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105(492):1403–1417, 2010.

[81] Michael Tiefelsdorf, Daniel A Griffith, and Barry Boots. A Variance-Stabilizing Coding Scheme for Spatial Link Matrices. *Environment and Planning A*, 31(1):165–180, 1999.

[82] Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(1):234–240, 1970.

[83] Carlo Tomasi. Estimating Gaussian mixture densities with EM – A Tutorial. *Duke University*, pages 1–8, 2004.

[84] Graham Upton, Bernard Fingleton, et al. *Spatial Data Analysis by example. Volume 1: Point Pattern and Quantitative Data*. John Wiley & Sons Ltd., 1985.

[85] Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607, 2010.

[86] Valérie Viallefont, Sylvia Richardson, and Peter J Green. Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics*, 14(1-2):181–202, 2002.

[87] Jon Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

[88] Lance A Waller and Carol A Gotway. *Applied Spatial Statistics for Public Health Data*, volume 368. John Wiley & Sons, 2004.

[89] Yiyi Wang, Kara Kockelman, and Amir Jamali. A synthesis of spatial models for multivariate count responses. In *Regional Research Frontiers-Vol. 2*, pages 221–237. Springer, 2017.

[90] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

[91] Daniel Wartenberg. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17(4):263–283, 1985.

[92] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.", 2016.

[93] William F Wieczorek, Alan M Delmerico, Peter A Rogerson, and David WS Wong. Clusters in irregular areas and lattices. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):67–74, 2012.

[94] Christopher K I Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT Press, Cambridge, Massachusetts, 2006.