A Review of Generalized Linear Models for Count Data with Emphasis on Current Geospatial Procedures

A thesis submitted in fulfillment of the requirements for the degree of

MASTER OF COMMERCE

in the

DEPARTMENT OF STATISTICS

of

RHODES UNIVERSITY

by

Justin Walter Michell

November 2014

Abstract

Analytical problems caused by over-fitting, confounding and non-independence in the data is a major challenge for variable selection. As more variables are tested against a certain data set, there is a greater risk that some will explain the data merely by chance, but will fail to explain new data. The main aim of this study is to employ a systematic and practicable variable selection process for the spatial analysis and mapping of historical malaria risk in Botswana using data collected from the MARA (Mapping Malaria Risk in Africa) project and environmental and climatic datasets from various sources. Details of how a spatial database is compiled for a statistical analysis to proceed is provided. The automation of the entire process is also explored.

The final bayesian spatial model derived from the non-spatial variable selection procedure using Markov Chain Monte Carlo simulation was fitted to the data. Winter temperature had the greatest effect of malaria prevalence in Botswana. Summer rainfall, maximum temperature of the warmest month, annual range of temperature, altitude and distance to closest water source were also significantly associated with malaria prevalence in the final spatial model after accounting for spatial correlation. Using this spatial model malaria prevalence at unobserved locations was predicted, producing a smooth risk map covering Botswana.

The automation of both compiling the spatial database and the variable selection procedure proved challenging and could only be achieved in parts of the process. The non-spatial selection procedure proved practical and was able to identify stable explanatory variables and provide an objective means for selecting one variable over another, however ultimately it was not entirely successful due to the fact that a unique set of spatial variables could not be selected.

Keywords: Spatial statistics, bayesian geostatistics, variable selection procedure

Contents

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge the support of my family and friends.

Chapter 1

Introduction

1.1 Background

Geographical variations of disease have been a subject of interest (Cibulskis *et al.*, 2007; Zayeri *et al.*, 2011) in epidemiology for a long time, as demonstrated in the monograph by Doll (1980). Doll was one of the first to study the influence of environment and lifestyle characteristics on cancer mortality. Doll stated that his hypotheses arose from studying the geographic distribution of various cancers (Richardson *et al.*, 2004). This highlights the importance as seen through history, of studying such variations. The main goal of this thesis is to present a case study of the spatial analysis of malaria count data or malaria prevalence data from survey stations across Botswana collected between 1944 and 1997. These prevalence data are point-referenced spatial data, otherwise known as geostatistical data. A historical continuous risk map of malaria in children between 1 and 15 years of age in Botswana will be the final result, including predictions of risk at unsampled sites, along with maps of significant environmental risk factors.

Malaria is a mosquito-borne infectious disease caused by the *Plasmodium falciparum* parasite (World Health Organization, 2014). Malaria is a major cause of morbidity and mortality in large areas of the developing world, especially Africa (Gosoniu, 2008). Rough calculations suggest that 250 million new cases occur globally every year (Zayeri *et al.*, 2011). National and global estimates of the burden of disease are imprecise as a result of inadequate malaria case reporting in most endemic countries and also because of the lack of nation wide malaria surveys (Cibulskis *et al.*, 2007). Accurate risk maps that describe the spatial variation and prevalence of the disease have long been recognized as instrumental for the planning of malaria prevention and control, and for estimating the disease burden (Gemperli *et al.*, 2006). In this study the terms prevalence and risk will be used interchangeably in describing

malaria risk or prevalence.

There is a growing recognition of the importance of robust handling of uncertainty. The advancement of spatial theory by authors like Cressie (1991), Diggle *et al.* (1998), and Finley and Banerjee (2013) and the increasing availability of computation facilities, for example, through open source programs like R (R Core Team, 2013) and its powerful spatial packages like sp which supports spatial data (Pebesma and Bivand, 2005), spdep which supports distance and proximity analysis (Bivand, 2013), geoRglm (Christensen and Ribeiro Jr, 2002), spBayes (Finley and Banerjee, 2013) which implement Bayesian spatial Gaussian and generalized linear regression mixed models; as well as the growing appreciation for the need of robust uncertainty handling have all contributed to a relatively recent shift in spatial thinking that utilize a special family of generalized linear models known as model-based geostatistics (MBG). MBG is generally implemented in a Bayesian framework (Diggle *et al.*, 1998; Diggle and Ribeiro, 2007, p. 15). This type of research can be helpful for the purpose of highlighting areas of elevated disease risk in the interest of prioritizing resources, especially where resources are limited (Kazembe *et al.*, 2006).

Proximity of observations, that is the closeness of survey sites, in space introduces correlations between the observations rendering the independence assumption of standard statistical methods invalid. Ignoring spatial correlation may result in underestimation of the standard error of the parameter estimates, and therefore liberal inference as the null hypothesis may be rejected too often (Mohebbi *et al.*, 2011).

This thesis will review and develop non-spatial and spatial models for the analysis of count data in space. A case study of malaria in Botswana using historical count data is presented. Initially non-spatial models are built using a staged variable selection procedure. This procedure will attempt to ensure that multicollinearity, confounding and overfitting is avoided and that the most representative variables are included in the spatial model. A spatial generalized linear mixed (SGLM) model will then be developed using the variables selected by the non-spatial analysis. This SGLM model will be built in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods available in the spBayes package (Finley and Banerjee, 2013) in R (R Core Team, 2013). The resulting spatial model will then be used to predict malaria risk everywhere in Botswana on a prediction grid. To check the accuracy of predictions cross validation between derivation and validation subsets of the data will be performed. Recommendations will also be given as to how these kinds of results can be incorporated into a geographic information system (GIS).

1.2 Research Questions

1. Is there evidence to link the incidence of malaria prevalence to environmental and climatic variables?
2. Is the non-spatial selection procedure effective? Does the procedure have an effect on selecting spatial variables?
3. Is the predictive performance in the spatial model better than the non-spatial model?
4. Are there areas of high malaria risk?
5. Are the results, particularly the predictions of risk, useful? Can they be used to develop a GIS and if so, how?
6. Are all necessary routines available in R to conduct the analyses?
7. Can the process be automated?

1.3 Research Objectives

Regression models for spatial count data, using the binomial model in a Bayesian framework are reviewed. This review will make clear why a spatial model is used so as to incorporate non-independent or correlated data. The simple linear model is extended to accommodate count data modelled using a Binomial distribution. This generalized linear model (GLM) is appropriate because of the non-normality of the errors. GLMs require independent observations. Generalized linear mixed models in a spatial context are introduced as a method to deal with correlated data. This will provide the context needed for the development of a parsimonious model that explains the spatial nature of malaria and its attributing environmental factors. Extensive non-spatial multi-stage modelling criteria are introduced to select the best set of geographic indicators. Once this is achieved, malaria risk will be predicted at sample sites in the validation subset of the data. Predictions will also be made using a suitable prediction grid all over Botswana. Details of how spatial data in R are handled are also provided. In addition an example of how a spatial database is compiled to facilitate the spatial analysis. Finally, recommendations and a discussion will be given as to how a policy maker might benefit from these resultant maps.

1.4 Structure of Thesis

This thesis is made up of four chapters and is structured as outlined below:

- Chapter 1 is the current chapter which serves as an introduction and outline of what this thesis entails.

- Chapter 2 is the literature review of previous work done on malaria, from a non-spatial and spatial point of view, and how Geographic Information Systems (GIS) with geostatistics has been used in disease mapping applications.

- Chapter 3 details the materials as well as the methods and techniques incorporated in this thesis.

- Chapter 4 discusses the results of the methods outlined in Chapter 3, as well as a discussion and conclusions, as it applies to the Botswana case study.

1.5 Definitions

- Geographic Information System (GIS): a suite of computer based tools used for the manipulation, management, analysis and capture of spatial data (Huisman and Rolf, 2009, p. 32).

- GIS software: computer software that can be used to develop tools for the spatial analysis of data (Huisman and Rolf, 2009, p. 142).

- Map projection: the mathematical transformation of the Earth's curved 3-d surface to a flat 2-d plan, that is a map (Huisman and Rolf, 2009, p. 520).

- Map coordinate system: a reference system defined on a flat, 2-d surface used to represent or locate the locations of geographic features, imagery, and observations such as GPS (Global Positioning System) locations using a particular map projection, such as azimuthal stereographic projection, as used in the Netherlands, or WGS84 (World Geodetic System 1984) which provides the current standard for locational measurement worldwide (Huisman and Rolf, 2009, p. 520).

- Geo-referenced data: refers to data defined using map coordinates in a specific map coordinate system which is referenced to a datum. A datum provides a frame of reference for measuring locations on the surface of the Earth, that is the relationship between the surface and the position of the surface relative to the center of the earth (Bernhardsen, 2002, p. 116; Lowry, 2004). Different reference surfaces are used to approximate the Earth's surface. The two main reference surfaces used are called the Geoid and the ellipsoid (Huisman and Rolf, 2009, p. 192).

- Geostatistics: a sub-branch of spatial statistics consisting of data which are a finite sample of measured values relating to an underlying spatially continuous phenomenon (Diggle and Ribeiro, 2007, p. 7). The main goal of geostatistics is to model continuous spatial variation (Ribeiro *et al.*, 2001).

- Interpolation: to estimate the value of a continuous variable given by n sampled values at some intermediate point or instant (Huisman and Rolf, 2009, p. 518).

- Euclidean distance: the standard straight line, Pythagorean distance function between locations (Huisman and Rolf, 2009, p. 515).

- Bayesian geostatistical analysis: involves the use of probability theory to find a probability distribution that quantifies knowledge about an unknown map given imperfect data, and making predictions using that probability distribution with associated precision (Patil *et al.*, 2011).

- Probability distribution: for a discrete random variable, a mathematical formula defining the probability of each value of the variable, for example a random variable following the binomial distribution. For a continuous random variable, a mathematical formula describing a curve which specifies, by means of the areas under the curve, the probability that the variable falls within a particular interval, for example a random variable following the normal distribution (Everitt, 2002, pp. 312-314).

- Likelihood: the probability of some observed outcomes given the value of some parameter or set of parameters. For example, the likelihood of a set of parameter values, $\boldsymbol{\theta}$ given a random sample of n observations, $x_1, \ldots, x_n$ with probability distribution $f(\mathbf{x}, \boldsymbol{\theta})$ is equal to the probability of those observed outcomes given those parameter values and is given by $L = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$ (Everitt, 2002, p. 232).

- Prior: the probability or uncertainty associated with an unknown variable in a model before data have been taken into account (Gelman *et al.*, 2014, p. 481).

- Posterior: the probability distribution of an unknown quantity conditional on the data. It can be derived given the prior and the likelihood using Bayes' Rule (Gelman *et al.*, 2014, p. 32).

- Posterior predictive: similar to posterior except it usually signifies that the variable considered relates to predicted data, such as in this thesis, predicting malaria risk at unsampled locations (Gelman *et al.*, 2014, p. 118).

- Markov chain Monte Carlo (MCMC): a popular and efficient algorithm for drawing samples from posterior distributions (Basáñez *et al.*, 2004). Typically MCMC methodology seeks to obtain characteristics of interest, for example the mean and variance of the marginal distribution, $f(x)$ arising from a joint distribution, $g(x, y_1, \ldots, y_q)$ as $f(x) = \int \ldots \int g(x, y_1, \ldots, y_q) \mathrm{d}y_1, \ldots, \mathrm{d}y_q$. Generally the necessary integrations to calculate $f(x)$ are extremely difficult or intractable, either analytically or numerically. MCMC methods incorporate simulation based methods in order to effectively allow for the drawing of samples from $f(x)$ without requiring $f(x)$ explicitly (Everitt, 2002, pp. 248-249).

- Bootstrap sampling: sampling with replacement to produce random samples of size n from the original data, $x_1, \ldots, x_n$. These n samples are called bootstrap samples and each sample provides an estimate of the parameter of interest (Everitt, 2002, p. 55).

Chapter 2

Review Of Previous Malaria Prevalence Studies

2.1 Applications of GIS and Mapping in Malaria Studies

GIS can be broadly described as a computer-based technology used for handling geographical data in digital form for the purpose of capturing, storing, manipulating, analyzing and displaying a wide variety of spatial or geo-referenced data (Burrough and McDonnell, 1998, p. 11).

Data such as climatic and environmental variables, distances, areas, and selections based on spatial criteria that are stored within a GIS, can provide the inputs needed for statistical modelling (Kleinschmidt, 2001). Large amounts of information are necessary for almost all aspects of malaria control programmes (Daash *et al.*, 2009). In this context GIS can be thought of as a spatial database or information management system, making large amounts of data easily accessible. Maps of interest given certain spatial criteria can be quickly retrieved and easily compiled into a document or report (Huisman and Rolf, 2009, p. 32).

Data can easily be updated and new maps can be generated to highlight hot spots of malaria prevalence in the interest of timely and focused malaria control planning. A GIS based approach in a national malaria control programme in India helped to identify hot spots of malaria prevalence and provided the inputs needed for a spatial analysis of the disease (Daash *et al.*, 2009). A GIS based approach was successfully applied to malaria research and control in South Africa (Martin *et al.*, 2002). This enabled the data to be timeously processed into usable formats. In this paper, Martin *et al.* (2002) stressed the relevance of GIS to malaria research.

Spatial statistical models yield estimated quantities of the population parameters for the purpose of quantifying the true underlying magnitudes and their associated uncertainty rather than the mere mapping of recorded data that are subject to sampling error. Spatial statistical modelling uses statistical methodologies to deal with the random nature of the processes involved. Using a purely GIS approach tends not to deal with the random nature of processes explicitly and therefore such models can produce only point estimates of processed quantities at specific locations typically located on a grid (Kleinschmidt, 2001). As a result GIS should be used in conjunction with appropriate spatial statistical methodologies.

2.2 Environmental Risk Factors

Various ecological and climatic factors affect the development and survival of the *Plasmodium falciparum* parasite and the malaria-transmitting Anopheles vector (Molineaux *et al.*, 1988). When predicting the risk of malaria infection in Africa the following environmental and climatic factors have been considered in prior studies:

- rainfall (Kleinschmidt *et al.*, 2000; Craig *et al.*, 2004; Abeku *et al.*, 2004; Kazembe *et al.*, 2006; Gemperli *et al.*, 2006);
- vegetation coverage (Hay *et al.*, 1998; Kleinschmidt *et al.*, 2000; Gemperli *et al.*, 2006; Craig *et al.*, 2007);
- distance to water bodies (Kleinschmidt *et al.*, 2000, 2001; Omumbo *et al.*, 2002; Kazembe *et al.*, 2006; Gemperli *et al.*, 2006; Craig *et al.*, 2007);
- altitude (Craig *et al.*, 1999; Omumbo *et al.*, 2002; Kazembe *et al.*, 2006);
- temperature (Craig *et al.*, 1999; Kleinschmidt *et al.*, 2000, 2001,?; Omumbo *et al.*, 2002; Craig *et al.*, 2004; Kazembe *et al.*, 2006; Gemperli *et al.*, 2006; Craig *et al.*, 2007) and
- bioclimatic variables (Kulkarni *et al.*, 2010; Chammartin *et al.*, 2013; Scholte *et al.*, 2014).

These variables are typically generated through interpolation of average monthly climate data from weather stations over a long term period, for example, a 50 years (Hijmans *et al.*, 2005). The references listed for each explanatory variable shows their wide use as predictors in malaria studies conducted in Africa. Table 2.1 below shows the source of each environmental and climatic factor used in the present analysis.

Table 2.1: Spatial databases used in this study.

Layer	Type	Resolution	Source
NDVI	raster	1 km	NASA Land Processes Distributed Active Archive Center (2001)
Temperature	raster	1 km	Hijmans *et al.* (2005)
Rainfall	raster	1 km	Hijmans *et al.* (2005)
Elevation	raster	1 km	Hijmans *et al.* (2005)
Bioclimatic Variables	raster	1 km	Hijmans *et al.* (2005)
Water bodies	raster	1 km	Gazetteer (2006)

2.3 Techniques Used for Modelling Malaria Prevalence

2.3.1 The Modelling Techniques Reviewed

Before considering the spatial aspects of the data, a non-spatial analysis is typically undertaken (Craig *et al.*, 2007; Noor *et al.*, 2009; Zacarias and Andersson, 2011). The attribute space can be explored by ignoring the coordinates and building a non-spatial Generalized Linear model (GLM). Gosoniu *et al.* (2006) first fitted a non-spatial regression model on malaria count data in Mali in order to determine which factors and possible transformations should be included in the spatial Bayesian modelling that follows.

In an influential paper by Diggle *et al.* (1998), spatial process models for non-Gaussian data within the framework of generalized linear models were discussed and implemented. Numerous studies modelling the spatial distribution of malaria and other tropical diseases in Africa and their association with environmental factors have taken this Bayesian approach, see for example Kleinschmidt *et al.* (2001), Kleinschmidt *et al.* (2002), Mabaso *et al.* (2005), Clements *et al.* (2006). These Bayesian geostatistical methods are described and implemented by Craig *et al.* (2007), Gosoniu *et al.* (2006) and others and are based on the pioneering work of Diggle *et al.* (1998).

In dealing with spatial dependence among the residuals a common solution is to add a spatially-varying model intercept that accounts for spatial association through a decreasing function of distance and perhaps direction between observed locations (Diggle *et al.*, 1998). Apart from ensuring the statistical validity of the model, adding a random spatial effect to the intercept allows conveniently for the separation of residual uncertainty into a spatial and non-spatial component. Appropriately accounting for residual uncertainty can improve inference, reveal missing explanatory variables and allow for better prediction accuracy and

precision (Diggle and Ribeiro, 2007). The properties of stationarity and isotropy in spatial data are often assumed (see Chapter 3 Section 3.6.3 on page 32 for a discussion of stationarity and isotropy) to simplify matters (Cressie, 1991, p. 57), since they cannot be subjected to formal rigorous hypothesis testing (Ver Hoef and Cressie, 2001, p. 299). These properties however can be investigated by exploratory data analysis, see Ver Hoef and Cressie (2001, p. 299): 'These assumptions are impossible to test, because it is impossible to go back in time again and again to generate the experiment each time to check whether each experimental unit has the same mean value or weather the correlation is the same for all pairs of plots that are at some fixed distance from each other.' Myers (1989, p. 348) also asserts that it is not possible to test any data set for stationarity because a data set is only one realization of the random function.

Spatial models can be fitted within a Bayesian framework using an adaptive Metropolis within Gibbs sampler (Roberts and Rosenthal, 2009). Computations can be performed in R (R Core Team, 2013) using the spGLM function in the spBayes package (Finley and Banerjee, 2013). This Bayesian routine in conjunction with a Binomial Generalized Linear Mixed Model (GLMM) on the logit scale was used in a species distribution modelling context by Swanson *et al.* (2013). Finley *et al.* (2008) implemented a Bayesian spatial logistic regression model to predict forested areas. A similar Bayesian routine in the geoRglm (Christensen and Ribeiro Jr, 2002) R package has been used by others in a spatial malaria modelling context. See for example Kazembe *et al.* (2006) and Craig *et al.* (2007).

2.3.2 A Brief Review of Prior Models Implemented

A spatial analysis in Mali was undertaken by Gosoniu *et al.* (2006). The malaria prevalence data used in this analysis are stored in the MARA database (Le Sueur *et al.*, 1997). These data were generated from surveys carried out on children between 1 and 10 years old at 89 sites between 1977 and 1995. A total of 43 492 children were surveyed. The climatic variables were aggregated into yearly averages over the months suitable for transmission following the map of Gemperli *et al.* (2006). The climate suitability criteria used in the generation of this map is an amended version of Tanser *et al.* (2003). The explanatory variables were standardized prior to model fitting. Among other considerations (such as building a non-stationarity model), a comparison of fit between the spatial and non-spatial models was performed. Surprisingly the spatial analysis yielded a positive relation between malaria risk and the distance to water. This novel result implies that malaria risk increases as the distance increases from permanent water bodies in Mali. This is surprising since Anopheles mosquitoes typically breed in water (World Health Organization, 2014).

Zacarias and Andersson (2011) implemented a hierarchical model applied to malaria count data in Maputo, Mozambique, aggregated at district level over a two years period (2001 and 2002). This period was divided into two climate conditions: rainy and dry seasons. The two years of climatic data, monthly averages, maximum temperature and rainfall were obtained from INAM (Mozambique National Meteorology Institute) and used as explanatory variables. Monthly maximum temperature and rainfall data were used in the model. This spatial analysis led to the conclusion that temperature and rainfall were significant in explaining malaria prevalence, with relative differences in importance in Winter and Summer. This study found that these explanatory variables do not explain all the variability present in the malaria data given the effect of overdispersion that is captured by regional structured and unstructured random effects. This lends support to the inclusion of a spatial random intercept to the standard GLM model in the current study in so far as it might help explain otherwise unexplained variation. Clements *et al.* (2006), who fitted Bayesian models to the parasite disease schistosomiasis, noted that adding a spatial dependence structure to the data made it evident that, notwithstanding what is known as biologically important environmental explanatory variables, the statistical relationships observed in the non-spatial models were no longer supported by the data and spurious significant relationships between the explanatory variables and malaria risk would have been accepted had spatial correlation not been considered.

Noor *et al.* (2009) built Bayesian geostatistical spatial-temporal models in their work in Kenya. *P. falciparum* parasite rate data were assembled from cross-sectional community based surveys undertaken from 1975 to 2009 and corrected to a standard age-range of 2 to less than 10 years, denoted as $PfPR_{2-10}$. After visually examining the relationships of the chosen explanatory variables in their continuous and categorical forms against $PfPR_{2-10}$ using scatter and box plots, the explanatory variables were aggregated into categories that are in line with biologically appropriate themes or categories, corresponding with the literature and expert knowledge. A non-spatial binomial logistic regression model was then fitted with the following categorical environmental factors: urbanization, minimum and maximum temperature, sets of 3 consecutive months in an average year of rainfall, enhanced vegetation index, altitude and distance to main waterbodies. Where more than one possible way of categorizing a explanatory variable presented itself, the size of the odds ratio, the Wald's p-value and the value of the Akaike's Information Criterion (AIC) score (see Chapter 3 Section 3.4.1 on page 26 for a discussion of the AIC) were used to establish the best way of categorizing the explanatory variables in order to achieve the strongest association with $PfPR_{2-10}$. No transformations on the data were considered. This non-spatial analysis showed that all the biologically selected categorized explanatory variables were statistically significant predictors of differences in $PfPR_{2-10}$. A collinearity test of all these explana-

tory variables was undertaken and if a pair had a correlation coefficient of greater than 0.9 (Clements *et al.*, 2006) the variable with the highest AIC was dropped and not used further in the analysis. This study found a reduced risk of prevalence in areas that had the following characteristics:

- urban relative to rural;
- maximum average annual temperatures of less than 25°C or greater than 30°C compared to between 25°C - 30°C;
- zero or 1-3 sets of three adjacent months of rainfall greater than 60 mm in an average year compared to corresponding rainfall patterns greater than 3 sets in an average year;
- where EVI was less than or equal to 0.3 compared to greater than 0.3;
- distance to main water bodies of greater than 12 km relative to less than or equal to 12 km.

Chapter 3

Methodology

3.1 Overview

Generalized linear models (GLM) are typically used to model linear relationships where it is assumed that the data in question are independent (Dobson and Barnett, 2008, p. 51). Data in spatial statistics are typically spatially correlated (Diggle and Ribeiro, 2007, p. 30). This methodology chapter will explain how a spatial database or GIS is compiled as well as explain the theory involved in studying such correlated data. Regression models for count data are introduced and explained. The simple linear model is extended to the GLM. GLMs are extended to the spatial generalized linear mixed model (SGLM). Non-spatial models are discussed. These models are extended to include a spatial component. This non-spatial model arises from a staged variable selection procedure (Craig *et al.*, 2007). How and why this procedure improves the spatial model will be discussed. The spatial models will be in a Bayesian framework using the spGLM function in the spBayes package (Finley and Banerjee, 2013) in R. These techniques and computational procedures are applied to a real data set consisting of malaria count data at different sample sites in Botswana in Chapter 4.

3.2 Geospatial Data in R

3.2.1 Overview

By the year 2000 there was a lot of activity and interest in spatial analysis. GIS software use was increasing and getting wide coverage and maps were appearing from web providers such as MultiMap (Matise *et al.*, 1994). Google Maps was still 5 years away and so as a way

to make sense and order of this, the Open Geospatial Consortium (OGC) (Open Geospatial Consortium, 1994) created a standard for spatial data and OGC protocols. The OGC Simple Features Specification defines data for points, lines, and polygons with associated attribute data. This format was implemented in R in the sp package as discussed in the book by Bivand *et al.* (2008) entitled *Applied Spatial Data Analysis in R*. The development of these spatial object classes and methods in the sp package, and its closer dependencies, was guided by the idea that users who are new to R but have GIS experience will want to see 'layers', 'coverages', 'rasters', or 'geometries'. From this point of view, sp classes should not present difficulties to GIS users. On the other hand, for statisticians using R, data are typically stored in a data.frame, a rectangular table with rows of observations on columns of variables. These classes were therefore developed to appear as GIS data models to GIS and other spatial data users and look and behave like data frames benefitting applied statisticians and other data analysts (Bivand *et al.*, 2008, p. 1).

3.2.2 Technical Details: Compiling Spatial Databases

Spatial data have coordinate values and a system of reference for these coordinates (Diggle and Ribeiro, 2007, p. 7). These data can be point locations or sites (with longitude/latitude coordinates) with attributes such as the number of people infected with malaria and the number examined at each site. These data are typically termed point-referenced data (Gemperli, 2003). Consider that if all these points of malaria risk data were to be drawn on a (flat) map, there would inevitably be a shift in the relative positions of these points. This illustrates the problem of projection, that is having to translate from the spherical longitude/latitude system to the non-spherical coordinate system (Diggle and Ribeiro, 2007, p. 7).

Gridded spatial data consisting of an array of equally sized cells arranged in rows and columns and composed of one or many attributes or bands, are known as raster images or raster layers. With these data, raster image processing and operations are required (Diggle and Ribeiro, 2007, p. 3). This process requires that georeferenced raster image layers must be acquired. Subsequently all raster layers must be in the same projection and must be precisely spatially aligned and cover exactly the same area (Hijmans, 2013). This means that all the rows and columns in all raster layers must have the same number of rows and columns and they must match pixel for pixel (Hijmans, 2013). Once aligned and in the same projection each cell in each raster layer will refer to the same position in space and the point locations can then be overlaid onto the map of the study area (Diggle and Ribeiro, 2007, p. 116). Overlay operations involve the combining of two (or more) spatial data layers comparing them position by position (Huisman and Rolf, 2009, p. 345). A spatial database

is compiled by extracting the raster values from each layer at each sample point. With the exception of the MODIS NDVI data sets, all selected data layers are unprojected in a geographic coordinate system and the datum is WGS84 at 1 km spatial resolution.

Once all layers are aligned they are combined into a spatial data structure in R, namely a spatial points dataframe (SPDF). From here a statistical analysis can follow. Below is a detailed description of the process applied in this study. References to the relevant lines of R code in the appendix are made.

- Set the current working environment or workspace. An environment is made up of a frame or a collection of named objects, and a pointer to an enclosing environment (R Core Team, 2013). For example, a frame of variables used to call a function are enclosed in the environment or workspace where the function was defined. The named objects or variables created in the current workspace can be saved and reloaded for later use (R Core Team, 2013). Refer to code lines 6 to 7. The code lines before this are comments.

- Load required packages for the spatial database compilation by running a function which checks if each package is either installed on the system or available in R's data structure or in the current working environment or workspace (R Core Team, 2013). New packages are installed. All required packages are loaded into the current working environment. Refer to code lines 9 to 23.

- Create a database connection between R and MySql (MySQL Community Server, 2011) using the RMySQL package (James and DebRoy, 2012) in R. Import the raw prevalence data including the month and year of the survey and its coordinates from a .cvs file into a newly created MySQL table (MySQL Community Server, 2011) in R. Load the relevant data obtained using a SQL query into a dataframe. Using MySQL in R allows the user to execute a SQL statement on a database connection within R (James and DebRoy, 2012). Using RMySQL is a useful way of extracting data from a large database where filters are needed in order to obtain only the data required. Refer to code lines 38 to 80.

Installing MySQL on Unix/Mac OS system from the command prompt:

1. Download appropriate file (Mac OS X 10.7 - at time of writing) from http://dev.mysql.com/downloads/mysql (MySQL Community Server, 2011).
2. Open downloaded file and double click on appropriate .pkg file to install, for example the mysql-x.x.xx-osx10.x-x86_32.pkg.
3. Go back to .dmg file and open MySQLStartupItem.pkg and install this. This second installation enables the starting of MySQL server instance when the Mac is turned on. Note that this can be done manually in the system preferences.
4. For convenience, edit the PATH variable in the terminal to ensure that the MySQL command will be recognized for future use. As a result it will then not be necessary to navigate to the full path where MySQL is installed. The PATH variable can be edited in the terminal by typing: export PATH = ${PATH} /usr/local/mysql/bin/
5. Once the above has been done it is important to save the .csv file, which contains the spatial data, in the /usr/local directory for example, so that MySQL knows where to get the data.

- Clean up the data, that is remove duplicate coordinates, remove entries where zero people were examined and across sites that have multiple counts take average count across years and months. See code lines 83 to 99.
- Read in surface water body shapefile using the readShapePoints function available in the maptools package (Bivand and Nicholas, 2014) in R. Refer to code lines 110 to 111.
- Calculate the distance to the closest surface water body at each site. In a loop, at each sample site use the spDistsN1 function available in the sp package (Bivand *et al.*, 2008) which calculates the Great Circle distance (WGS84 ellipsoid) from a single point to all surface water bodies in kilometers. The Great Circle distance takes the earth's curvature into account, for example the distance along earth's surface (Dormann *et al.*, 2007). At each site a vector of the distance to all surface water bodies is obtained. The minimum at each site is taken. Refer to code lines 104 to 122.
- Obtain a map of the boundary of Botswana for overlay purposes. Boundary maps are available in the maps package (Becker *et al.*, 2013) in R. Ensure that the spatial points dataframe (SPDF) containing the sample points is in the same unprojected geographic coordinate system as the Botswana boundary map. Keep sample points that are inside the spatial domain, that is all of Botswana. Refer to code lines 130 to 146.

- Download all available monthly NDVI images for the years 2000 - 2013 (MODIS PRODUCT MOD13A3) using functions in the "ModisDownload.R" script and "ModisLP.RData" work space (Naimi, 2014). The MODIS Terra product, MOD13A3, provides monthly data for the years 2000 to 2013 at 1 km resolution in the Sinusoidal projection with a scale factor of 0.0001. The output file is in Hierarchical Data Format (HDF) format. This format is designed to store and manage large amounts of numerical data (Qu *et al.*, 2006, p. 123). This HDF file contains 11 Scientific Data Sets (SDS) stored in array format. Extract only the relevant sub dataset, namely the mean monthly NDVI sub dataset, from all HDF files across all years. Refer to code lines 161 to 174.

- Using the raster function available in the raster package (Hijmans, 2013) read in a WorldClim raster layer for any month and crop this layer to the same extent as the Botswana shapefile. Compare this WorldClim raster layer to a MODIS mean monthly NDVI sub dataset obtained in the previous step for any month and year. In order to establish which raster layer should be the reference layer, compare the dimensions of the two raster layers. Refer to code lines 176 to 186.

- Extract NDVI sub datasets for each month across all years. Refer to code lines 188 to 227.

- Loop through each month and through all the years (2000 - 2013) of data reprojecting, merging tiles and converting to TIFF format in one step using MODIS NDVI layer as the model or reference raster. Various image blocks or tiles cover the area of Botswana and thus must be merged to span the relevant area (Naimi, 2014). This is achieved using the gdalwarp function in the gdalUtils package (Greenberg and Mattiuzzi, 2014) in R. Refer to code lines 229 to 252.

- Initialize raster stacks to be populated with climate and NDVI layers. A raster stack is a collection of raster layer objects with the same spatial extent and resolution. A raster stack can be created from raster files such as TIFF images (Hijmans, 2013). Refer to code lines 265 to 271.

- For the NDVI monthly layers apply NDVIRasterFunction to each monthly TIFF image for each year. The function involves calculating the mean NDVI value of each cell for each month across all years and multiplying each cell by the scale factor 0.0001. Refer to code lines 273 to 283.

- Read in all WorldClim climate TIFF images into a list in R. Refer to code lines 262 to 263.

- Populate each WorldClim climate raster stack in a loop for each month across all years using the raster function in the raster package (Hijmans, 2013). Multiply temperature layers by the scale factor 0.1 and crop to the extent of the NDVI reference raster. There is no scale factor used for rain layers. Crop rain layers to the extent of the NDVI layer. No loop needed for altitude since the altitude is constant across months. Refer to code lines 285 to 303.

- Resample all climate WorldClim layers to match the reference raster layers, namely the NDVI layers. Refer to code lines 306 to 312.

- Write the sample points and attributes of the Spatial Points Data Frame (SPDF) to a polygon covering Botswana using the writeOGR function available in the rgdal package (Bivand *et al.*, 2013). Read this polygon as a new SPDF object using the readOGR function available also in the rgdal package in R. Refer to code lines 328 to 337.

- Extract raster values at matching coordinates and add to the @data component of the SPDF for each climate and environmental stack then give layers column names and append them to the SPDF. Refer to code lines 339 to 370.

3.3 Regression Models for Count Data

3.3.1 Introduction

Regression models are often employed to assess the relationship between between a response variable, also called a dependent or outcome variable, and one or more explanatory variables, also called predictor variables, covariates or risk factors. Where there is only one explanatory variable the analysis is called simple linear regression, and when there are more than one the analysis is called multiple linear regression. Regression models for count data in epidemiology are often employed when a study is concerned with the count of a disease within each spatial region/unit comprising the area of interest (Lawson, 2013, pp. 6 - 13). More generally, regression analysis in a spatial context allows one to model, examine, and explore and predict spatial relationships between an outcome of interest such as malaria prevalence in Africa, and its environment. Simple linear regression is a good starting point for the spatial regression analysis that follows.

3.3.2 Simple Linear Regression

In a simple linear regression model the response variable, denoted by y_i, is modelled by a linear function of the explanatory variable, denoted by x_i, plus an error term, denoted by ϵ_i. β_0 is the intercept, that is the predicted value of y_i when x_i equals zero. β_1 is the regression coefficient. The regression coefficient represents the rate of change of the response variable as the dependent explanatory variable changes (Everitt, 2002, p. 39). In this model the subscript i denotes the observation number. This model is typically denoted as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3.1}$$

or

$$\mathrm{E}\,(y_i) = \beta_0 + \beta_1 x_i. \tag{3.2}$$

The error term, ε_i, for each observation i, is a random variable which explains the random variation or noise in the outcome y_i. The errors are assumed to be independent, normal and identically distributed random variables with an expected value of zero and a constant variance, that is $\varepsilon_i \sim N(0, \sigma^2)$. The normality in the errors implies normality in the response variable, y_i, which is continuous (Seltman, 2012, p. 215).

The distribution of the population of possible values for y_i at x_i has mean $\beta_0 + \beta_1 x_i$ and constant variance σ^2. In general for each different value of the explanatory variable a separate distribution of responses exists such that its form is the same (their distributions are identically distributed) and their variances are the same but their means differ (Christensen, 2011, p. 346).

The fundamental idea underlying a linear regression analysis is that the expected response is linear in the parameters (Seltman, 2012, p. 214).

3.3.3 The General Linear Model

The simple linear regression model represented by Equation 3.1 is extended to the multiple linear regression case to include multiple explanatory variables. This model is known as the general linear model. For responses $Y_1, \ldots, Y_n$, this can be written as

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3.3}$$

or

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \tag{3.4}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

is a vector of responses,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

is termed the design matrix and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

is a vector of parameters.

$\mathbf{X}$ is often termed the design matrix. This matrix consists of constants representing levels of categorical explanatory variables or the measured values of continuous explanatory variables (Dobson and Barnett, 2008, p. 37). For a continuous variable, for example temperature, the model has a linear component $\beta_i x_i$ for the i^{th} observation where the parameter or coefficient represents the change in the response Y_i corresponding to a one unit change in x_i when all other explanatory variables are kept constant. For categorical variables, instead of representing a measured constant value in the dependent variable, parameters are coded for different levels of the factor. These elements are chosen in $\mathbf{X}$ so as to include or exclude appropriate parameters for each observation and are known as dummy variables. If variables are only zeros or ones they are called indicator variables (Dobson and Barnett, 2008, p. 37).

Consider the function g on the vector of expected responses in Equation 3.5,

$$g[E(\mathbf{Y})] = \begin{bmatrix} g[E(Y_1)] \\ \vdots \\ g[E(Y_n)] \end{bmatrix}. \tag{3.5}$$

Note that the same function, g, applies to every element in the vector. This function, g, is needed in cases when the response data are not linear. For example when the response data are not continuous but discrete. Typically discrete data consist of counts or binary responses. Consider the problem of modelling temperature against malaria risk as a binomial proportion. In the simple linear model framework the intensity of malaria risk, which ranges between 0 and 1, is assumed to be a linear function of temperature. This assumption would be incorrect as this relationship is not linear given the binary nature of the response variable (Dobson and Barnett, 2008, p. 46). The variability of observations around the mean cannot be thought to vary about the mean according to a normal distribution (Christensen, 2011, pp. 12-14). This function g is called a link function. It preserves the linear structure of the model (Dobson and Barnett, 2008, p. 46). Its workings will be discussed in Section 3.3.7.

3.3.4 The Bernoulli Distribution

The Bernoulli distribution is defined for a variable that is binary or dichotomous in that the variable has one of two possible outcomes (Dobson and Barnett, 2008, p. 53). A binary random variable S is defined as

$$S = \begin{cases} 1 & \texttt{if the outcome is a success} \\ 0 & \texttt{if the outcome is a failure} \end{cases}$$

with probabilities

$P(S = 1) = \pi$ and $P(S = 0) = 1 - \pi$, that is $S \sim$ Bernouli(π). If there are n such independent Bernoulli random variables $S_i, \ldots, S_n$ with $Pr(S_i = 1) = \pi$, then the sum of these events,

$$Y = \sum_{i=1}^{n} S_i, \tag{3.6}$$

denotes the number of successes in n independent trials (Dobson and Barnett, 2008, p. 48).

3.3.5 The Binomial Distribution

The random variable Y denotes the sum of n independent Bernoulli trials where each probability of success, π, is the same for each trial, that is Y is composed of n Bernoulli experiments. Y has a Binomial distribution (Dobson and Barnett, 2008, p. 48). The probability of obtaining y successes and $n - y$ failures is then given by

$$Pr\{Y = y\} = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n. \tag{3.7}$$

The combinatorial coefficient is the number of ways of obtaining y successes in n independent trials. The mean of the Binomial distribution for Y is given by

$$E(Y) = \mu = \pi n$$

and the variance is

$$Var(Y) = \sigma^2 = \pi(1-\pi).$$

3.3.6 Bernoulli and Binomial Models for Count Data

Consider the context of predicting malaria risk using point-referenced malaria prevalence data. This kind of spatial data is described in Section 3.2.2. Define N random variables, that is $Y_1, \ldots, Y_N$ corresponding to the number of infections at site i. These data are count data. At each site i in the study area, as shown in Figure 4.1 in Chapter 4, there is a count of the number of people infected with malaria as well as the number who were examined. That is,

$$Y_i = \text{number infected at site } i \text{ and } n_i = \text{number examined at site } i.$$

The random variable Y_i can take the values $0, 1, ..., n_i$ associated with site i. If one person was examined at site i, he/she would be either infected or not infected with malaria. This binary response would describe one Bernoulli trial. If more than one person was examined at each site, site i consists of the sum of the independent Bernoulli trials, as shown generally in Equation 3.6.

The n_i observations at each site are assumed to be independent. Each site has the same exposure to explanatory variables as per the creation of the spatial database, that is only one

measure for each explanatory variable is associated with a site. Thus at site i all n_i have the same probability, π_i, of having the attribute of interest namely, malaria infection. It follows that the distribution of Y_i is Binomial with parameters π_i and n_i, that is $Y_i \sim Bin(n_i, \pi_i)$. See Section 3.3.7.1 for a continuation of this problem.

3.3.7 Generalized Linear Models

When the response variable $\mathbf{Y}$ is normally distributed the linear model, as described in Section 3.3.3, is appropriate. However when the response is not normal, or when the data can not be coerced by transformation to be normal, the normality assumption is not appropriate. This is typically the case for the malaria count data considered in this study (Finley *et al.*, 2007). As highlighted above, the binomial model is particularly unsuitable for a linear response model since probabilities are bounded on both ends (they must be between 0 and 1). Hence it is often more meaningful to model a function of the mean as opposed to the mean itself, in this case to ensure that the mean of $\mathbf{Y}$ (which is a probability) is between 0 and 1, as a linear combination of the unknown parameters $\boldsymbol{\beta}$ (Gotway and Stroup, 1997). A central assumption in the linear model is that the variance should be constant. In count data where the response variable is an integer and often takes the value 0, the variance will likely increase with the mean (Nelder and Wedderburn, 1972) thus violating the constant variance assumption.

Generalized Linear Models were first introduced by Nelder and Wedderburn (1972) as an extension of the general linear model for analyzing data from non-normal distributions. Consider the random variable Y that has probability density function (pdf), $f(y;\theta)$, which depends on parameter θ. If $f(y;\theta)$ takes the following form

$$f(y;\theta) = s(y)t(\theta)exp(a(y)b(\theta))$$

or equivalently after rearranging (Dobson and Barnett, 2008, p. 46)

$$f(y;\theta) = exp(a(y)b(\theta) + c(\theta) + d(y)), \tag{3.8}$$

where $c(\theta) = ln(t(\theta))$ and $d(y) = ln(s(y))$, then $f(y;\theta)$ is said to belong to the exponential family of distributions (Barndorff-Nielsen, 1978, p. 107). Here,

- $b(\theta)$, according to Dobson and Barnett (2008, p. 46), is called the natural parameter and is a function dependent only on θ;

- $c(\theta)$ is a function dependent only on θ;
- $d(y)$ is a function dependent only on y.

The expected value and variance of $a(Y)$ can be expressed as

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)} \tag{3.9}$$

and

$$Var[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^2} \tag{3.10}$$

respectively (Dobson and Barnett, 2008, p. 49). The distributional form in 3.8 is called canonical if $a(y) = y$. When the distribution is in canonical form 3.9 and 3.10 are the mean and variance for Y respectively. Working with an exponential family distribution in the canonical form is analytically convenient. Once in this form the pdf of Y can be rewritten the in terms of a single parameter, that is Y depends on a single parameter, θ (Dobson and Barnett, 2008, p. 51).

Consider a set of independent random variables, $Y_1, \ldots, Y_n$, from the exponential family of distributions (Equation 3.8) and a set of explanatory variables, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where each $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is a vector of length p. The expected value, μ_i, of Y_i is modelled as a linear function of explanatory variables, $\mathbf{x}_i$, employing the following transformation

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{3.11}$$

where $g(\centerdot)$ is called the link function and is monotonically increasing in μ_i, that is the transformation of the explanatory variables is either strictly increasing or strictly decreasing. The mean function is given by

$$\mu_i = exp(\mathbf{x}_i^T \boldsymbol{\beta}).$$

The link function relates the linear predictor to the mean. When $b(\theta)$ in Equation 3.8 is equal to the linear predictor η_i, the link function $g(\centerdot)$ in 3.11 is then known as the canonical link. A special case of the generalized linear model is the linear regression model where an identity link

$$g(\mu_i) = \mu = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{3.12}$$

is used (refer to Section 3.3.7 on the general linear model). The binomial distribution can be employed to model count data, that is when $Y_i \sim Bin(n_i, \mu_i)$. The Binomial logistic regression model is obtained via the logit link,

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{3.13}$$

that forces μ_i to be between 0 and 1. The identity link shown in 3.12, which has no transformation on the explanatory variables, is the general linear model, $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, (Equation 3.3). This model will not work because the mean response, $E(Y)$, must take values between 0 and 1. By construction the transformed mean function of Equation 3.13 forces the response to be between 0 and 1, as required.

The binomial distribution belongs to the exponential family of distributions in the canonical form since the pdf given by Equation 3.7 can be expressed as

$$\begin{aligned} Pr\{Y_i &= y_i\} = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \\ &= exp\left(y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i log(1-\pi_i) + log\binom{n_i}{y_i}\right) \end{aligned}$$

where $a(y_i) = y_i$, $b(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$, $c(\pi_i) = n_i log(1-\pi_i)$ and $d(y_i) = log\binom{n_i}{y_i}$. From Equation 3.9, the expected value of Y_i is given by

$$\begin{aligned} \mu_i &= E[Y_i] = -\frac{c'(\pi_i)}{b'(\pi_i)} \\ &= -\frac{-n_i(1-\pi_i)^{-1}}{(1-\pi_i)^{-1}(\pi)^{-1}} \\ &= n_i\pi_i \end{aligned}$$

The canonical link typically used is the logit link:

$$g(\pi_i) = logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}_i^T\boldsymbol{\beta} \tag{3.14}$$

where the logit link maps probabilities from the open interval (0,1) to the whole real line.

The likelihood function for independent responses $Y_1, \ldots, Y_n$ in the canonical form of the exponential family of distributions can be written as

$$L(\boldsymbol{\theta};\mathbf{y}) = \prod \exp(y_i b(\theta_i) + c(\theta_i) + d(y_i))$$

and the log-likelihood function, log(L($\boldsymbol{\theta}$;$\mathbf{y}$)), is given by

$$l(\boldsymbol{\theta};\mathbf{y}) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i).$$

The global maximum of the log-likelihood function $l(\boldsymbol{\theta};\mathbf{y})$ is given uniquely by solving $\frac{\partial l}{\partial \theta} = 0$ or $\frac{\partial l}{\partial \beta} = 0$ under certain regularity conditions (Cox and Hinkley (1979) as cited in Dobson and Barnett (2008, p. 74)).

3.3.7.1 Binomial Model for Count Data in a Spatial Context

GLMs focus on analyzing the linear relationship between the transformation of the expectation of the response variable, via a link function, and the explanatory variables. It is assumed that the observations are independent (Dobson and Barnett, 2008, p. 51). Spatial data are typically spatially correlated (Diggle and Ribeiro, 2007, p. 30). This means that the GLM should be modified to incorporate the spatial dependence that is often inherent in spatial data. The spatial section of this thesis, in particular Section 3.6.4, will show how this can be done.

3.4 Goodness of Fit Statistics

3.4.1 Akaike's Information Criterion (AIC)

The likelihood function, L, can be defined as the probability or likelihood of the data given a model. Define p as the number of free parameters in the model. The Akaike Information Criterion (AIC) (Akaike, 1973) is defined as

$$AIC = -2\ln(L) + 2p \tag{3.15}$$

The AIC model selection method is used in this study. The AIC is a criterion that looks for a model that has a good fit but with few parameters (Dobson and Barnett, 2008, p. 137). The goodness of fit of a statistical model is determined by how well it fits a set of observations (Jha *et al.*, 2011). Under the AIC criterion, the model with the best fit is the one with the smallest AIC. The AIC penalizes the fitted value of $-2ln(L)$ (a positive value), and adds a penalty that depends on the number of fitted parameters, as shown in Equation 3.15.

3.4.2 Cross-Validation

Cross-validation involves splitting the data at random into two sets, namely a modelling or derivation set and a validation subset (Hyndman and Koehler, 2006). Model building proceeds on the derivation set. The goodness of fit of a model can be assessed by summarizing the discrepancy between observed values and the values expected given the model (Craig *et al.*, 2007). Two such measures of goodness of fit are the mean prediction error (MPE) and the absolute mean prediction error (AMPE) (Hyndman and Koehler, 2006). Following the convention of Zeng *et al.* (2013); Noor *et al.* (2014), the MPE, given by

$$MPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

and the AMPE, given by

$$AMPE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|$$

is used in this study to assess the accuracy of predictions between the non-spatial and spatial model at validation sites.

Another common cross-validation measure (Valle, 2011) is the mean squared prediction error (MSPE) given by

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

3.5 Non-spatial Model Selection Procedure

A staged variable selection procedure employed by Craig *et al.* (2007) will be implemented in this thesis. Craig *et al.* (2007) note that variable selection is a major obstacle in spatial modelling due to analytical problems caused by over-fitting, confounding and non-independence in the data. These authors argue that although spatial dependence in the response data has been modelled successfully using Bayesian spatial modelling, variable selection remains an issue of concern. Variable selection can affect the predictions greatly especially when faced with a large number of potential risk factors (Craig *et al.*, 2007). In order to establish which variables should be included in the spatial analysis, a systematic, practical and repeatable

staged process of variable elimination is adopted. Following the study by Craig *et al.* (2007) all the available explanatory variables are split into climatic and environmental themes in which collinearity among variables is tested per theme. In the current study, these themes include a rain, a temperature and a NDVI theme. Unthemed variables include variables which do not fit into rain, temperature or NDVI themes. In this study, the unthemed variables are not tested for collinearity because they are deemed unrelated. The reference study had more themes and variables and were able to allocate a theme to all variables. The process includes a stepwise bootstrap method described by Austin and Tu (2004). Following this variable selection process, the resulting smaller subset of variables are fitted in a Bayesian geostatistical model so as to achieve the primary aim, which in the context of this study is mapping historical malaria prevalence data and making predictions at unsampled sites across Botswana. This variable selection process involves 6 stages. Each stage will here be described in detail.

3.5.1 Stage 1

In Stage 1 the dataset is split randomly into derivation and validation subsets. Univariate logistic regression is used to identify the best univariate predictors using the derivation data and all the potential explanatory variables. If an explanatory variable is insignificant at the 5% level of significance, it is excluded.

3.5.2 Stage 2

In Stage 2 the variables that were significant in Stage 1, are ranked based on each model's AIC score. Those variables that are strongly correlated with each other, Spearman's $r > 0.85$, with higher-ranking variable/s belonging to the same theme are excluded. Individual scatter plots of the remaining variables against logit(p), the logistic response, are then prepared.

3.5.3 Stage 3

Stage 3 involves running 1 000 bootstrap samples from the derivation data and running an automated backward exclusion procedure on each sample, that is automated backwards stepwise elimination in conjunction with bootstrap resampling (Austin and Tu, 2004). The automatic backward exclusion procedure involves starting with all candidate variables, testing whether each variable should be deleted using the AIC criterion, deleting the variable, if any, that improves the model the most by being deleted and continuing this process until there is no further improvement possible. In each bootstrap iteration, the coefficients and the frequency with which each candidate variable is selected are recorded.

3.5.4 Stage 4

In Stage 4 a non-spatial multiple variable model is derived in a manual forward stepwise fashion. This process starts by including the most frequently selected variable in Stage 3 and adding further variables in order of selection frequency. The manual forward stepwise regression continues as long as all entered variables remain significant at the 5% probability level. If a previously entered variable becomes non-significant with the inclusion of another variable, this process keeps the variable that was more frequently selected in Stage 3 above the other.

3.5.5 Stage 5

Stage 5 involves revisiting those variables that were excluded in Stage 2 because of high correlation. Excluded variables from Stage 2 are allowed to re-enter the Stage 4 candidate list of variables in a stepwise-bootstrap sample per theme, recording selection frequency as above. This stepwise-bootstrap procedure is the same procedure used in Stage 3. If the added variable is significant in the bootstrap sample and selected more frequently than the original variable in the Stage 4 candidate list, then the added variable replaces the original variable. Otherwise the original variable in the Stage 4 candidate list remains. This modified model is the non-spatial model.

3.5.6 Stage 6

In the final stage, Stage 6, the explanatory variables from Stage 5 are incorporated into a generalized geostatistical spatial model (or a SGLM model) using MCMC simulation methods. The details of Stage 6 will be discussed in the spatial modelling section (see Section 3.6 on the following page).

3.5.7 Implementation of Stage 2 of the Non-Spatial Variable Selection Procedure in R

Initially all of the above stages were performed manually. This took a lot of manual work, particularly Stage 2. Therefore, the Stage 2 process was automated in a loop as shown below. The terms used in the psuedocode and the psuedocode are given below (refer to code lines 511 to 646 in the appendix of Chapter 5):

- DF: Themed dataframe in which the order of univariate AIC rankings is preserved from lowest to highest);

- N: Number of variables in DF;
- X: Proposed variable (first variable in DF);
- Y: Variable/s correlated with X ;
- Condition 1: X not correlated with remaining variables in theme;
- List1: List of variables kept. The variable tested is always kept because by design it has the lowest univariate AIC ranking;
- List2: List of variables removed.

```
while N > 1 do
    invoke correlation test function;
    add X to List1;
    if Condition 1 then
        add NA to List2;
        remove X, Y from DF;
    else
        add Y to List2;
        remove X, Y from DF;
    end
end
```

Algorithm 1: Iterative algorithm to keep track of which variables are correlated with each other and which are kept and removed based on AIC rankings

3.6 Spatial Modelling

3.6.1 Overview

Spatial data contain information about both the attribute of interest as well as its location. Examples are found in ecology, geology, epidemiology, geography, image analysis, meteorology, forestry, and geosciences (Haran, 2011). Refer to Section 3.2.2 on page 14 for a description of the type of spatial data used.

The spatial component of the Botswana case study that will be presented in this thesis will draw on Bayesian geostatistical methods described and implemented by Craig *et al.* (2007), Gosoniu *et al.* (2006), and others; and pioneered by Diggle *et al.* (1998). In particular Stage

6 of the process will be discussed. This section will start by giving some background on spatial data, in particular the kind of spatial data we are dealing with, namely geostatistical data. Guided by Diggle *et al.* (1998), common objectives in spatial geostatistical modelling will also be discussed, including an example of what typical geostatistical data looks like. The theory behind linear gaussian random field models for geostatistical data is discussed, followed by details of the Bayesian framework used for estimation and prediction. Finally details on the Bayesian implementation of the spatial model in R will be given.

3.6.2 Spatial Statistics

A key property of spatial statistical analysis is that it is assumed that the data are auto-correlated in that observations in close proximity tend to be related or more similar than observations that are far apart. This assumption is known as Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970).

Spatial statistical methods incorporate spatial correlation according to the manner in which geographical proximity is defined (Gemperli, 2003). Proximity also depends on the geographical information, which is either at an aggregate area level (areal) or at a point-location level. Areal unit data are aggregated over contiguous units partitioning the whole study region. Their neighbouring structure defines proximity in space. Point-referenced or geostatistical data are collected at fixed locations, for example households or villages, over a continuous study region (Gemperli, 2003). In geostatistical data the distance between sample locations determines proximity (Gemperli, 2003). Work done by Krige (1951) and Matheron (1963) laid the foundation for this field of study. Geostatistics can be viewed as a hybrid of statistics, mathematics, mining engineering and geology (Bolin, 2009). It has become a branch of statistics that specializes in the analysis and interpretation of geographically referenced data (Goovaerts, 1997, p. 3). Many geostatistical methods are fundamental in spatial data analysis (Bolin, 2009). Cressie (1991, p. 8) views geostatistics as one of the three scientific fields specialized in the analysis of spatial data. The other two are point pattern analysis, which deals with point objects, and lattice statistics or areal analysis, which deals with pixel data. Point pattern analysis is concerned with where events of interest occur. A fundamental question in this type of spatial analysis is whether or not the points of interest occur at random, or whether or not there is evidence of clustering, or patterns of regularity. Lattice statistics typically requires data at a regularly spaced set of points. Irregular lattice type data is also possible. Lattice data are typically in the form of pixels. Pixels are small rectangularly shaped regions, which are often the result of remote sensing from satellites or aircrafts. The observed data in lattice statistics are typically aggregations within boundaries

of interest, such as population counts. The three scientific objectives of geostatistics (Diggle and Ribeiro, 2007, p. 12) are:

1. Model estimation, that is estimation of the the model parameters;
2. Prediction, that is prediction of the unobserved values of the target variable;
3. Hypothesis testing.

Most applications are concerned with the first two objectives (Diggle and Ribeiro, 2007, p .12). Estimation deals with inference about the parameters of a stochastic model for the data. Generally, a stochastic model is comprised of a family of random variables running over a suitable index (Pinsky and Karlin, 2010, p. 4). In a geostatistics context the random variable is the spatial process at work at each sample site. The sample sites denote the index. These concepts will be discussed further in the subsequent section. Following estimation, one can focus on prediction and/or hypothesis testing. Parameters of direct scientific interest such as those defining a regression relationship between a response and an explanatory variable may be explored, or parameters of indirect interest, such as those defining the covariance structure of a model may also be explored.

Spatial prediction refers to the prediction of unknown quantities, $Z(s_0)$, based on sample data, $Z(s_i)$, and assumptions regarding the form of the trend of $\mathbf{Z}$ and its variance and spatial correlation. Hypothesis testing can also appear in geostatistical problems, although it is typically not a primary concern (Diggle and Ribeiro, 2007, p. 13).

3.6.3 Geostatistics

Geostatistics is concerned with the analysis of random fields, $\mathbf{Z}(\mathbf{s})$, with $\mathbf{Z}$ random and $\mathbf{s}$ the non-random spatial index. A random variable measured at a set of locations is called a random field (Cressie, 1991, p. 8). A random field is a generalization of a stochastic process in that the underlying parameter takes values that are multidimensional vectors, or points on some manifold or two-dimensional surface in three-dimensional space (Adler, 2004). Typically analysis occurs at a limited number of sometimes arbitrarily chosen locations (Diggle and Ribeiro, 2007, p. 10). Variability in the measured response is a result of the random realization of the spatial field, not the randomness of the sampling locations. Thus two sources of variation should be distinguished, namely the spatial variation underlying the target surface, that is the random realization of the spatial field and the statistical variation given that surface (Diggle and Ribeiro, 2007, p. 3). Measurements on $\mathbf{Z}$ at these sample locations are available, and prediction and interpolation of $\mathbf{Z}$ is required at non-observed

locations S_0, or the mean of $\mathbf{Z}$ is required over a specific region, B_0. Geostatistical analysis deals with the estimation and modelling of spatial correlation or covariance or semivariance and evaluating whether simplifying assumptions such as stationarity can be justified or need refinement (Bivand *et al.*, 2008, p. 192). The problem can be defined more explicitly as follows: let a set of observations of a target variable $\mathbf{Z}$ be denoted as $z(s_1)$, $z(s_2)$, ... ,$z(s_n)$, where $s_i = (x_i, y_i)$ is a location and x_i and y_i are the measured coordinates in geographical space and n is the number of observations. The geographical domain of interest, for example area or land surface or object, can be denoted as $\mathbf{D}$. Only one reality or realization of a process ($\mathbf{Z} = \mathbf{Z}(\mathbf{s}), \forall \mathbf{s} \in \mathbf{D}$) is assumed. The domain is in continuous space so this process could have created many realities, that is the number of locations at which observations can be made is not countable. In most applications the random field is assumed to be Gaussian or normal and hence statistical properties are completely determined by the mean value function, $\mu(\mathbf{s})$ and the covariance function, $C(s_i, s_j) = C(||s_i - s_j||)$ (Diggle and Ribeiro, 2007, p. 47). A Gaussian random field is a Markov random field of continuous states and with a joint Gaussian distribution over those states (Riedl *et al.*, 2010). Markov processes are discussed in Section 3.6.5 on page 38. The form of the covariance function should be chosen so as to fit the particular dataset. Stationarity of the covariance function is often assumed in order to simplify calculations, such that it is a function of distance between points only. Further simplifying assumptions are also sometimes made where one assumes that the covariance function only depends on distance and not direction. The covariance function and random field are then called isotropic (Bolin, 2009).

3.6.3.1 An example of Geostatistical data- Rongelap data

These data was first analyzed by Diggle *et al.* (1998). It was collected from Rongelap Island in the South Pacific, which forms part of the Marshall Islands in America. Nuclear weapon testing generated heavy fallout over the island in the 1950's and since 1985 it has been uninhabited. The Rongelap Island data consists of 157 sampling locations. It is based on a sampling design which consists of a primary grid covering the island at a spacing of 200 meters and four secondary 5 by 5 sub-grids at a spacing of 50 meters. At each location, photon emission counts that are as a result of radioactive caesium were measured. The data have the form $(\mathbf{x}_i, m_i, t_i) : i = 1, \ldots, 157$, where $\mathbf{x}_i$ denotes spatial location, m_i denotes the photo emission count at that location, and t_i is the time (in seconds) over which m_i was accumulated (Diggle *et al.*, 1998).

Using the observed emission counts per unit time $\frac{m_i}{t_i}$ as a response variable z_i, the Rongelap data can be transformed into the basic format of geostatistical data,

$$(\mathbf{x}_i, z_i) : i = 1, \ldots, n$$

where each $z_i = \frac{m_i}{t_i}$ is a realization of a random variable Z_i whose distribution depends on an underlying unobservable spatially continuous stochastic process $\mathbf{Z}(\mathbf{x})$. The set of values $\mathbf{Z}(\mathbf{x}), \mathbf{x} \in \mathbf{D}$, where $\mathbf{D}$ is the domain in continuous space, can be understood as one draw in an infinite set of random variables (Diggle and Ribeiro, 2007, p. 9).

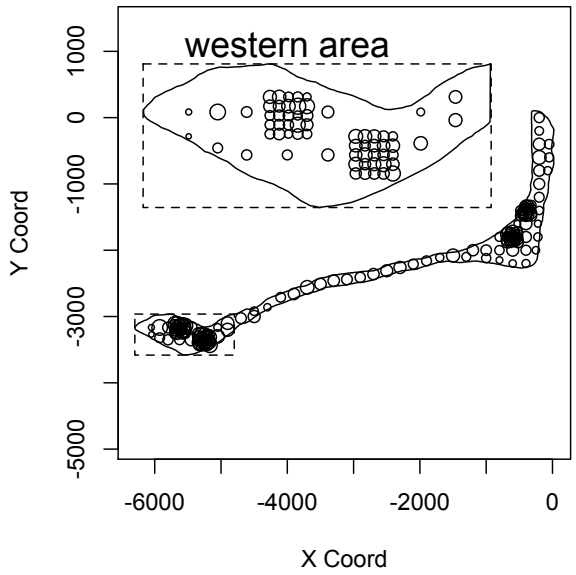

Figure 3.1: Circle plot for Rongelap island data. Circles represent sampling locations and radii are proportional to observed emission counts per unit time. The unit of distance is 1 metre. The broken lined box represents an enlargement of the western extremity of the island.

3.6.4 Linear Spatial Models

GLMs, as discussed in Section 3.3.7 on page 23, focus on analyzing data under the assumption that the observations are independent. As discussed above spatial data typically violate this assumption. This means that the dependence structure underlying the spatial data is some function of location information and must be accounted for. It is well known that ignoring spatial dependence in the data when employing regression models will result in biased estimates of variation and inefficient statistical inference (Cressie (1991) as cited in

Li (2008)). The GLM can be extended to accommodate dependent responses by introducing unobservable random effects into the linear predictor. As a result the model specification of the logit function of a GLM in Equation 3.11 on page 24 is modified to

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + w_i,$$

where $\mathbf{w} = (w_1, \ldots, w_n)$ follows a zero-mean multivariate distribution. The w_i are called random effects (Diggle and Ribeiro, 2007, p. 80). The random effects relate to the variance component of the model, that is, the random effects explicitly models the between-subject variation in the data (Dobson and Barnett, 2008, p. 221). This kind of model is typically called a spatial generalized linear model (SGLM) or a spatial generalized linear mixed model since the specification of spatial dependence via a generalized linear model framework always involves random effects (Breslow and Clayton, 1993; Lee and Nelder, 1996; Haran, 2011). Typically, $\mathbf{w}$ is specified as a multivariate Gaussian random variable with a particular covariance structure imposed in order to describe the spatial dependence or error structure in the data (Diggle and Ribeiro, 2007, p. 80).

Considering sample sites $\mathbf{s} = s_1, \ldots s_n$, in this class of models $\mathbf{w}(\mathbf{s}) = (w(s_1), \ldots, w(s_n))$ is a stationary Gaussian process. This process is stochastic and is a Gaussian model if the joint distribution of $w(s_1), \ldots, w(s_n)$ is multivariate Gaussian for any integer n and set of locations s_i. The process is stationary if the expectation of $\mathrm{S}(\mathbf{x})$ is the same for all $\mathbf{x}$, the variance of $\mathrm{S}(\mathbf{x})$ is the same for all $\mathbf{x}$ and the correlation between $\mathrm{S}(\mathbf{x}_i)$ and $\mathrm{S}(\mathbf{x}_j)$ depends only on $u = ||\mathbf{x}_i - \mathbf{x}_j||$, the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Linear Gaussian random field models for geostatistical data will be discussed, both for normal data and count data. Diggle and Ribeiro (2007, p. 80) proposed and described how the spatial dependence or error structure for SGLMs can be modeled via Gaussian processes for point-level, geostatistical data.

3.6.4.1 Linear Gaussian Process Models - Normal Case

As discussed the domain is a continuous space, that is these two spatially continuous stochastic processes could have created many realities. Let the spatial process at locations $\mathbf{s} \in \mathbf{D}$, where $\mathbf{D}$ is the domain of interest, be defined as

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \mathbf{w}(\mathbf{s}), \text{for } \mathbf{s} \in \mathbf{D}, \tag{3.16}$$

where $\mathbf{Z}(\mathbf{s})$ is the response vector as a function of sites, $\mathbf{s}$, such that $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))$, $\mathbf{X(s)}$ is the set of explanatory variables associated with each site s_i, and $\boldsymbol{\beta}$ is a p-dimensional

vector of coefficients. Spatial dependence can be imposed by modeling $\{\mathbf{w(s)} : \mathbf{s} \in \mathbf{D}\}$ in Equation 3.16 as a zero-mean multivariate stationary Gaussian process specified by

$$\mathbf{w}(\mathbf{s}) = \text{MVN}(\mathbf{0}, \mathbf{\Sigma}(\mathbf{\Theta})), \tag{3.17}$$

where $\mathbf{\Sigma}(\mathbf{\Theta})$ is the variance-covariance matrix of the n-dimensional normal density with unknown parameters, namely the spatial decay, ϕ and variance, σ^2. In order for the distribution given by Equation 3.17 to be proper $\mathbf{\Sigma}(\mathbf{\Theta})$ must be symmetric and positive definite. If $\mathbf{\Sigma}(\mathbf{\Theta})$ is specified by a positive definite parametric covariance function, these conditions are satisfied (Haran, 2011). The covariance function for a pair of locations, s_i and s_j, separated by the Euclidean distance, h, can be written as a product of the variance parameter σ^2 and a positive definite correlation function

$$\rho(h) : C(h) = \sigma^2 \rho(h).$$

The exponential correlation function is a positive definite correlation function and takes the following form

$$\rho(h) = \exp(-\phi h). \tag{3.18}$$

The exponential correlation function is a special case of the more flexible Maťern family (Handcock and Stein, 1993). This covariance structure assumes that the covariance and hence dependence between two locations decreases as the distance between the locations increases, that is for small distances the correlation between sites is large and decreases as distance increases. The Maťern correlation function is specified as follows

$$\rho(h) = \frac{1}{2^{v-1}\Gamma(v)}(\phi h)^v K_v(\phi h),$$

where v is known as the smoothness parameter and $K_v(x)$ is a modified Bessel function of order v (Abramowitz and Stegun, 1964, p. 358). $K_v(x)$ controls the smoothness of the function. As v increases, the process becomes increasingly smooth (Haran, 2011). The Maťern correlation function reduces to the Exponential correlation function when v is an integer plus $\frac{1}{2}$ (Genton, 2002).

Stein (1999) (as cited in Haran (2011)) recommends the Maťern structure because it is flexible enough to allow the smoothness of the process to also be estimated. This author cautions against Gaussian process models with gaussian correlations due to the fact that

they are overly smooth, that is they are infinitely differentiable. Generally the smoothness, v, may be hard to estimate from data (Haran, 2011). A popular default is to use the exponential covariance structure for spatial data where the physical process yielding the realizations is not likely to be smooth and a gaussian covariance for modeling output from computer simulations or other data where the associated smoothness assumption may be reasonable (Haran, 2011).

3.6.4.2 Linear Gaussian Process Models - Binomial Case

In cases where the linear Gaussian assumption provides a poor fit to the data and transforming the data in an attempt to make it normal via, say the Box-Cox family of transformations, is unsatisfactory, SGLMs can be employed (Haran, 2011).

Let $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ and $\{\mathbf{w}(\mathbf{s})\colon \mathbf{s} \in \mathbf{D}\}$, be two spatial processes on $\mathbf{D} \subset \mathbb{R}^d (d \in \mathbb{Z}^+)$. Here it is assumed that $Z(s_i)$ conditionally follow a common distributional form, for example the binomial in this case for count data, and $Z(s_i)$ are conditionally independent given $w(s_1), \ldots, w(s_n)$ where $s_1, \ldots, s_n \in \mathbf{D}$, and

$$E(Z(s_i)|w_i) = g\{\mu_i(s_i)\}, \text{for } i = 1, \ldots, n. \tag{3.19}$$

A known link function, g, is chosen as described in Section 3.3.7, so that $\eta(\mathbf{s}) = g\{\mu(\mathbf{s})\}$, for example where $g(\mathbf{.})$ is the logit link (see Equation 3.14). Further assume that

$$\eta(\mathbf{s}) = \frac{s_i}{1 - s_i} = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \mathbf{w}(\mathbf{s}), \tag{3.20}$$

where $\mathbf{X}(\mathbf{s})$ is a set of p explanatory variables corresponding with each site $\mathbf{s}$, and $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients. Spatial dependence is handled by modelling $\{\mathbf{w}(\mathbf{s})\colon \mathbf{s} \in \mathbf{D}\}$ as a stationary Gaussian process, that is $\mathbf{w} = (w(s_1), \ldots, w(s_n))^T$ is a multivariate normal distribution defined as per Equation 3.17.

Notice the identity link function is used in Equation 3.16 for the normal conditional distribution of $\mathbf{Z}(\mathbf{s})$. This result in the normal case, described in Section 3.6.4.1 is obtained as a special case (Haran, 2011).

3.6.5 Hierarchical Bayesian Inference and Estimation

3.6.5.1 Overview

Statistical inference is the process of making decisions about some unknown aspect of the population from which the data were drawn (Christensen, 2011, p. 131). With respect to the current study, interest lies in making inference on unobserved sample points, that is, sample sites in the derivation set of the data. In a frequentist paradigm the estimation of $\boldsymbol{\theta}$, which is treated as a fixed unknown quantity of interest, typically proceeds via the maximum likelihood approach which is based on a model for data $f(\mathbf{y}|\boldsymbol{\theta})$. Inference is then based on the notion of repeated sampling. The distribution of the MLE $\hat{\boldsymbol{\theta}}$ induced by repeatedly sampling of the data is considered under identical conditions as n approaches ∞. The concept underlying Bayesian spatial modeling is Bayes' theorem (Gelman *et al.*, 2014, p. 8). In this theorem both the distributions of the data and the unknown coefficient estimates are considered.

Bayesian inference works by assigning or fitting a probability model to the observed data. The results are summarized by a probability distribution on the unknown parameters $\boldsymbol{\theta}$ or unobserved data $\widetilde{\mathbf{y}}$. In the Bayesian paradigm inferences are made in terms of probability statements conditional on the observed data $\mathbf{y}$ (Gelman *et al.*, 2014, p. 1). The Bayesian paradigm offers attractive advantages over the frequentist approach for modeling spatial data (Banerjee *et al.*, 2004, p. 97). This point can be made by noting four distinct advantages as highlighted by Banerjee *et al.* (2004, p. 97).

1. The Bayesian method allows the modeller to model spatial correlation explicitly among random effects through prior distributions;

2. The marginal likelihood function can be complex and multidimensional and is generally not tractable in closed form and must be approximated numerically, which can be computationally difficult. MCMC simulation methods in a Bayesian setting can be used to overcome difficulties associated with computing posterior distributions as discussed in Section 3.6.5.3 on page 42;

3. It is possible to specify a complicated model for non-Gaussian data, as in the malaria count data presented in this thesis, in a hierarchical Bayesian fashion. In this way the data and parameters of interest can be specified through different layers which can be easily understood and computed;

4. In a Bayesian setting the uncertainty of the model and parameters is explicitly taken into account.

In addition, in conventional frequentist geostatistical interpolation when the response data is Gaussian the covariance structure is estimated first, and then the estimated covariance is used for interpolation and, unlike in the Bayesian approach, the effect of the uncertainty in the covariance structure on subsequent predictions is often ignored (Stein (1999) as cited in Li (2008)).

3.6.5.2 Bayesian Inference Framework

A Bayesian statistical model is composed of a sampling distribution, namely the likelihood function denoted by $p(\mathbf{y}|\boldsymbol{\theta})$, for the observed data conditional on the unknown parameters $\boldsymbol{\theta}$, and a prior distribution denoted by $p(\boldsymbol{\theta})$ (Everitt, 2002, p. 36; Gelman *et al.*, 2014, p. 1). The prior distribution is a reflection of the degrees of belief on the likely values of the unknown parameters (Everitt, 2002, p. 313). With these two distributions, the joint distribution, also known as a full probability model, can be written as

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

and, via Bayes' rule (Everitt, 2002, p. 36; Gelman *et al.*, 2014, p. 1), the posterior is obtained as follows

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{3.21}$$

where $p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ for discrete $\boldsymbol{\theta}$ and for the continuous case $p(\mathbf{y}) = \int L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Because $p(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$ it can be considered a constant with fixed $\mathbf{y}$ and can thus be factored out and Equation 3.21 can be obtained up to a normalizing constant that is proportional to the likelihood function times the prior written as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.22}$$

Equation 3.22 ensures that model estimation using numerical methods (see Section 3.6.5.3 on page 42) are easier since computing the normalizing constant, which is not easy to obtain, is avoided.

Hierarchical modeling results from a simple fact from probability, namely that the joint distribution of a collection of random variables can be decomposed into a series of conditional models (Arab *et al.*, 2008). For example, consider random variables a, b and c. Basic probability allows the factorization:

$$[a, b, c] = [a|b, c][b|c][c],$$

where the notation [.] is used to denote a probability distribution (Arab *et al.*, 2008). In this way complex models can be built through the specification of several simple stages. Bayesian hierarchical modelling involves setting up a multi-level stochastic model. Such structuring of the model is well-suited for incorporating a priori knowledge, allowing prior knowledge to be inserted at various levels of the modeling, where appropriate.

More specifically, a hierarchical Bayesian model (Gelman *et al.*, 2014, p. 101) involves decomposing the prior probability distribution, $p(\boldsymbol{\theta})$, into several conditional levels of distributions:

$$p_1(\boldsymbol{\theta}|\boldsymbol{\theta_1}), p_2(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}), \ldots, p_n(\boldsymbol{\theta_n}|\boldsymbol{\theta_{n-1}})$$

and a marginal distribution

$$p_{n+1}(\boldsymbol{\theta}_n)$$

such that

$$p(\boldsymbol{\theta}) = \int_{\Theta_1 \times \cdots \times \Theta_n} p_1(\boldsymbol{\theta}|\boldsymbol{\theta}_1) p_2(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \ldots p_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1}) p_{n+1}(\boldsymbol{\theta}_n) d\boldsymbol{\theta_1} \ldots \boldsymbol{\theta_n}.$$

The parameters, $\boldsymbol{\theta_i}$, are called hyperparameters of level i, for $1 \leq i \leq n$. Hyperparameters are the parameters of the prior distributions to distinguish them from parameters of the model of the underlying data. In most hierarchical Bayesian problems the number of levels, n, is equal to 2 (Gelman *et al.*, 2014, p. 101). At the first stage, a likelihood function for the data given the parameters is specified. At the second stage the prior distributions for the parameters given the hyperparameters are specified and distributions for the hyperparameters are specified at the third stage. The first stage can be defined as the data-level uncertainty as it is made up of the study-specific likelihood that may, for example, incorporate uncertain linear restrictions on the parameters of a regression model, whereas the prior distributions at the second level correspond to the more subjective information that accounts for the imprecision or uncertainty at the first stage (Raudenbush, 2002, p. 415). Hierarchical modeling improves the robustness of the resulting Bayes estimators, since uncertainty regarding the model structure can be incorporated into additional prior distributions (Raudenbush, 2002, p. 415). The decomposition of the prior distribution into its components

simplifies Bayesian computation and facilitates a simpler and more intuitive approximation of posterior quantities by simulation (Raudenbush, 2002, p. 415).

The choice of prior distribution is critical for Bayesian inference, especially when the sample size is small or when the sample is sparse (Gelman *et al.*, 2014, p. 167). Prior information from external experts in a field can be incorporated in the construction of a prior distribution for the unknown parameters, although the process of converting prior information to prior probability distributions is often not clear (Winkler, 1967). Prior information will also typically not yield a unique prior distribution. When there is little prior information regarding model unknowns, as is often the case, a noninformative or vague prior distribution can be employed. These priors typically are from a parametric distribution with large or infinite variance, thus expressing the associated uncertainly or lack of knowledge (Winkler, 1967). For large data sets the likelihood will dominate the prior, and inference will be primarily data-driven and so such an approach is reasonable. For small data sets however, inference will be far more sensitive to prior choice and more caution is needed in specifying the priors (Winkler, 1967; Li, 2008). An important aspect of Bayesian modelling is the notion of a conjugate prior (Gelman *et al.*, 2014, p. 36). A prior is called a conjugate prior when the posterior distribution follows the same parametric form as the prior distribution. Probability distributions belonging to the exponential family of distributions always have conjugate prior distributions (Gelman *et al.*, 2014, p. 36).

Suppose $f(\mathbf{y}|\boldsymbol{\theta})$ are from the exponential family of distributions with the form as in Equation 3.8 on page 23 for $i = 1, \ldots, n$. The likelihood function for a random sample is given by

$$L(\mathbf{y}, \boldsymbol{\theta}) = \phi(y)t(\boldsymbol{\theta})^n exp(w(\mathbf{y})b(\boldsymbol{\theta})),$$

where

$$\phi(\mathbf{y}) = \prod_{i=1}^{n} s(y_i) \qquad \text{and} \qquad w(\mathbf{y}) = \sum_{i=1}^{n} (y_i).$$

If the prior distribution is specified as

$$p(\boldsymbol{\theta}) \propto t(\boldsymbol{\theta})^{\eta} exp(b(\boldsymbol{\theta})v),$$

then the posterior distribution is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto t(\boldsymbol{\theta})^{n+\eta} exp(b(\boldsymbol{\theta})(w(\mathbf{y}) + v))$$

which has the same density form as the prior distribution. This choice of prior density is conjugate and is often called the natural conjugate prior (Gelman *et al.*, 2014, p. 44).

In the current context we can envisage a three stage hierarchical specification (Bernardo, 1996):

At the first stage the likelihood is conditional on the spatial random effects

$$\text{Level 1}: \ \mathbf{Z}(\mathbf{s})|\mu(\mathbf{s}) \sim Bin(\mu(\mathbf{s}))$$

with $g(\mu(\mathbf{s})) = X(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s})$ where g is the logit link.

At the second stage $w(\mathbf{s})$ provides the process model

$$\text{Level 2}: \ \mathbf{w}(\mathbf{s}) \mid \Theta \sim N(\mathbf{0}, \Sigma(\boldsymbol{\Theta})) \tag{3.23}$$

At the third stage priors are placed on the parameters

$$\text{Level 3}: \ \text{priors on } (\boldsymbol{\beta}, \boldsymbol{\Theta}),$$

where Θ denotes the unknown spatial decay, ϕ and variance, σ^2 parameters.

3.6.5.3 MCMC Methodology

In general a stochastic process (Gamerman and Lopes (2006) as cited in Li (2008)) can be defined as a collection of random variables, denoted by $\theta^{(n)} \in S$ where $n \in T$. T takes nonnegative integers and S is called the state space. A discrete-time stochastic process denoted by $\{\theta^{(n)} : n \geq 0\}$ on a countable set S is a set of random variables defined on a probability space denoted by (Ω, $\mathcal{F}$, P). The probability space (Loève, 1955, p. 149) is made up of:

1. a sample space, Ω, which defines all possible outcomes of a random trial;
2. σ-algebra $\mathcal{F}$ of measurable subsets of Ω, where σ-algebra $\mathcal{F}$ is the collection of events, $\mathcal{F}$, where each event is a set containing zero or more outcomes and
3. a probability measure, $P : \mathcal{F} \to [0, 1]$, where $0 \leq P(A) \leq 1$ is the probability that the event $A \in \mathcal{F}$ occurs.

The convention of not displaying the probability space (Ω, $\mathcal{F}$, P) when random variables or processes are introduced is taken in this study.

Assume that the state space, S, is discrete for the following discussion. A Markov chain is a special type of stochastic process in which the past and future states are conditionally independent given the current state. A stochastic process $\boldsymbol{\theta} = \{\theta^{(n)} : n \geq 0\}$ on a countable state space S is a Markov Chain (Serfozo (2009) as cited in Li (2008)) if, for any $x, y \in S$ and $n \geq 0$,

$$P\{\theta^{(n+1)} = y|\theta^{(0)}, \ldots, \theta^{(n)}\} = P\{\theta^{(n+1)} = y|\theta_n\}, \tag{3.24}$$

$$P\{\theta_{n+1} = y|\theta_n = x\} = P(x, y). \tag{3.25}$$

$P(x, y)$ denotes the probability that the Markov chain jumps from state x to state y. $P(x, y)$ is known as a transition probability (Serfozo (2009) as cited in Li (2008)). These transition probabilities satisfy

1. $P(x, y) > 0 \quad \forall x, y \in S$;
2. $\sum_{y \in S} P(x, y) = 1 \quad \forall x \in S$.

The condition denoted by Equation 3.24 is called the Markov property. This property states that at any time n, the next state X_{n+1} is conditionally independent of the past $X_0, \ldots, X_{n-1}$ states given the present state X_n (Serfozo (2009) as cited in Li (2008)). Alternatively put, the next state is dependent on the past and present only via the present state, that is, the chain jumps around the parameter space remembering only where it has been in the last period or iteration. The condition denoted by Equation 3.25 states that the transition probabilities do not depend on the time parameter n.

The matrix, $\mathbf{P}$, for discrete state spaces, $S = \{x_1, x_2 \ldots\}$, with the $(i, j)^{th}$ element given by $P(x_i, x_j)$ is called the transition matrix of the chain (Gamerman and Lopes (2006) as cited in Li (2008)). If S is finite with r elements the transition matrix, $\mathbf{P}$, is given by

$$\mathbf{P} = \begin{bmatrix} P(x_1, x_1) & \ldots & P(x_1, x_r) \\ \vdots & \ddots & \vdots \\ P(x_r, x_1) & \ldots & P(x_r, x_r) \end{bmatrix}.$$

To arrive at the transition probability from state i to state j over exactly m steps the matrix product of $\mathbf{P}$ m-times is taken and written as $P^m(x, y)$ (Gamerman and Lopes (2006) as cited in Li (2008)). A row vector containing marginal probabilities associated with realization $\theta^{(n)}$ is denoted by $\pi^{(n)}$ with components $\pi^{(n)}(x_i) = P(\theta^{(n)} = x_i)$. The

recursive relationship between successive marginal distributions of the chain can be written as $\pi^{(n)} = \pi^{(0)}P^{n-1}P = \pi^{(n-1)}P$. When $n = 0$ this is the initial distribution of the chain. In matrix notation the initial distribution is given by $\pi^{(n)} = \pi^{(0)}P^n$. The probability of an event $A \subset S$ for a Markov chain starting at x is denoted by $P_x(A)$. The hitting time of event A is defined as $T_A = \min\{n \geq 1 : \theta^{(n)} \in A\}$ if $\theta^{(n)} \in A$ for $n > 0$, otherwise $T_A = \infty$. Concerning the state space, S, and the transition matrix, P, two important quantities needed in the discussion that follows is defined below (Gamerman and Lopes (2006) as cited in Li (2008)):

1. The probability of the chain, starting from state x and in subsequent steps reaching state y, is denoted by $\rho_{xy} = P_x(T_y < \infty)$;

2. The number of visits of a chain to state y is denoted by $N(y) = \#\{n > 0 : \theta^{(n)} = y\} = \sum_{n=1}^{\infty} I(\theta^{(n)} = y)$, where I is the identity matrix.

A state $y \in S$ is called recurrent if $\rho_{yy} = 1$, that is, the chain returns to y with probability one. A state is called transient if $\rho_{yy} < 1$. For a recurrent state y, if $\mathrm{E}\left[T_y|\theta^{(0)}\right] < \infty$ where $T_y = min\left\{n \geq 1 : \theta^{(n)} = y\right\}$ is the hitting time of y, the state is then positive recurrent. This is an important property for obtaining limiting results. For iterative simulation algorithms, the asymptotic behavior of the chain as the number of iterations $n \to \infty$ can be considered the most important area of the Markov chain theory (Gamerman and Lopes (2006) as cited in Li (2008)).

A distribution, $\boldsymbol{\pi}$, is called a stationary distribution of a chain with transition probabilities $P(x, y)$ if

$$\sum_{x \in S} \boldsymbol{\pi}(x)P(x, y) = \boldsymbol{\pi}(y) \quad \forall y \in S.$$

In matrix form this can be stated as $\boldsymbol{\pi} = \mathbf{P}\boldsymbol{\pi}$. If the stationary distribution $\boldsymbol{\pi}$ exists and

$$\lim_{n\to\infty} P^{(n)}(x, y) = \boldsymbol{\pi}(y),$$

then the sequence of marginal distributions $\pi^{(n)}$ will approach $\boldsymbol{\pi}$ as $n \to \infty$, whatever the initial distribution of the chain may be. As such $\boldsymbol{\pi}$ may be referred to as the limiting distribution. There are cases where stationarity holds but the limiting distributions does not exist (Gamerman and Lopes (2006) as cited in Li (2008)). In order to establish limiting results the concept of periodicity needs to be introduced. The period of a state x is the largest common divisor of the set $\left\{n \geq 1 : P^{(n)}(x, x) > 0\right\}$ denoted by d_x. A state is called ergodic if the state is positive recurrent and aperiodic if $d_x = 1$. Also, a chain is ergodic if

all its states are ergodic. Suppose that $\theta^{(n)}$ is ergodic with stationary distribution $\boldsymbol{\pi}$ and $t(\theta)$ a real valued function $\mathrm{E}[t(\boldsymbol{\theta})] < \infty$, then the ergodic average is

$$\bar{t_n} = \left(\frac{1}{n} \sum_{i=1}^{n} t(\theta^{(i)}) \right) \overset{a.s}{\to} \mathrm{E}_{\pi}[t(\theta)] \quad \text{as} \quad n \to \infty.$$

In this case the Markov chain follows the strong law of large numbers (Feller (1950) as cited in Li (2008)), that is ergodic averages satisfy central limit theorems are needed to estimate posterior quantities. As a result the use of MCMC for estimating expectations taken with respect to the posterior distribution for Bayesian inference is justified (Li, 2008; Doss and Hobert, 2010).

In most real world applications, when using Markov Chain simulation to fit statistical models in a Bayesian framework, the state space S will not be discrete. Recall that in the present Botswana case study only one of many possible realizations of a geostatistical process in continuous space is obtainable. However the ergodic theorem described above can be extended and applied more generally, namely in continuous space. When S is a continuous state space the transition kernel is defined through a conditional probability density function

$$p(x, y) = \frac{\partial P(x, y)}{\partial y}$$

where

$$P(x, y) = Pr(\theta^{(n+1)} \leq y | \theta^{(n)} = x) = Pr(\theta^{(1)} \leq y | \theta^{(0)} = x), \quad \text{for} \quad x, y \in S.$$

Then the continuous version can be written as

$$\boldsymbol{\pi}(y) = \int_{-\infty}^{\infty} \boldsymbol{\pi}(x) p(x, y) dx$$

where $\boldsymbol{\pi}$ is the stationary distribution of the chain. Following these definitions, the limiting results considered in the discrete case can be applied to the continuous case (Gamerman and Lopes (2006) as cited in Li (2008)).

The goal of MCMC simulation for Bayesian inference is to simulate realizations $\theta^{(0)}, \theta^{(1)}, \ldots$ from an ergodic Markov chain whose stationary distribution is the posterior distribution of interest. From an initial state $\theta^{(0)}$, realizations of the chain are generated successively until the chain 'forgets' this initial state and exhibits steady state behavior. At this point, call it,

T, the set of sampled values, $\theta^{(0)}, \ldots, \theta^{(T)}$, is discarded as a 'burn-in' period and realizations after this point, $\theta^{(T+1)}, \theta^{(T+2)}, \theta^{(T+3)}, \ldots$, are approximate draws from the posterior distribution. These realizations, namely $\theta^{(T+1)}, \theta^{(T+2)}, \theta^{(T+3)}, \ldots$, are the stationary distribution of the Markov chain. Bayesian inference can proceed by summarizing the posterior distribution which is made up of the J draws after the burn-in period. There are various ways to construct the required Markov chain needed for a given Bayesian inference problem, for example the two most widely used methods, the Gibbs sampler and the Metropolis-Hastings algorithm (Li, 2008). The method used in this study is an adaptive Metropolis-Hastings algorithm. This algorithm will be discussed in Section 3.6.5.4.

3.6.5.4 An Adaptive Metropolis-Hastings Method

Generally the transition probability matrix, $\mathbf{P}$, of the Markov chain depends on the tuning of associated parameters such as the proposal variances or the parameters estimating spatial decay and smoothness. The choice of tuning parameters is crucial to the success of the MCMC procedure (Roberts and Rosenthal, 2009).

For high-dimensional problems, that is problems with many fitted parameters, the favoured choice is MCMC (Li, 2008). However, these methods can be slow to converge, making practical implementation difficult (Brooks and Gelman, 1998; Gelman *et al.*, 2014, p. 294, p. 393). Good performance can be obtained from an adaptive MCMC algorithm, see Roberts and Rosenthal (2009) implemented in the spBayes package (Finley and Banerjee, 2013) in R. The adaptive method adjusts the tuning parameters for the jumping function based on the local acceptance/rejection of the new parameters which speeds up convergence. It should be noted that the method has no effect on the model or final result, but does improve the speed and accuracy of the fitted values. Various diagnostics are used to test for convergence (see Section 3.6.5.6).

3.6.5.5 Bayesian Prediction

Bayesian prediction entails sampling from the posterior predictive distribution. The posterior predictive distribution is the distribution of a new data point, say, y_0, marginalized over the posterior:

$$P(y_0|\mathbf{y}) = \int P(y_0|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})d\beta d\theta \quad \text{for} \quad y_0, \mathbf{y} \in S.$$

3.6.5.6 Bayesian Implementation of SGLM in R

In the spBayes package (Finley and Banerjee, 2013) in R, specifically using the spGLM function, 3 MCMC chains are fitted and various tuning and prior settings and batch sizes and lengths are considered with the aim of getting the 3 chains to converge. As per the theory of Markov chains, a chain is expected to eventually converge to the stationary distribution provided that the length of the chain is long enough (Larget and Simon, 1999). However convergence is not guaranteed after a given number of draws. Convergence is therefore assessed visually to assess the mixing of each chain using trace plots as well using Gelman and Rubin's convergence diagnostic (Brooks and Gelman, 1998). Using the Gelman and Rubin's convergence diagnostic, approximate convergence is diagnosed when the upper confidence limit is close to 1. Gelman and Rubin's convergence diagnostic is obtained using the gelman.diag function in the Coda package (Plummer *et al.*, 2006). A trace plot plots the iteration number against the value of the draw of the parameter at each iteration (Plummer *et al.*, 2006). Trace plots combining all 3 chains are inspected for each parameter to see how well each chain is mixing in the parameter space as well as to see if multiple chains are converging. Once the model has converged the spPredict function in spBayes package is used for the prediction of malaria prevalence at unsampled sites across Botswana, that is, sites in the validation subset of the data. A prediction grid will also be created whereby each grid cell will have an associated value for each explanatory variable so that a prediction at each grid cell or pixel can be made, also using the spPredict function in R. Various grid resolutions will be tested so that a balance between computational efficiency and sufficient accuracy can be achieved.

Chapter 4

Modelling Malaria Prevalence in Botswana

4.1 Study Area

Figure 4.1 shows the 122 survey sites used in this study. It can be seen that the distribution of survey sites is sparse in south western Botswana. Typically sparse data are characterized by large variability (Gosoniu *et al.*, 2010). As a result it is difficult to detect the underlying spatial correlation. Therefore it was anticipated that prediction might prove more difficult in this area (Howes *et al.*, 2012).

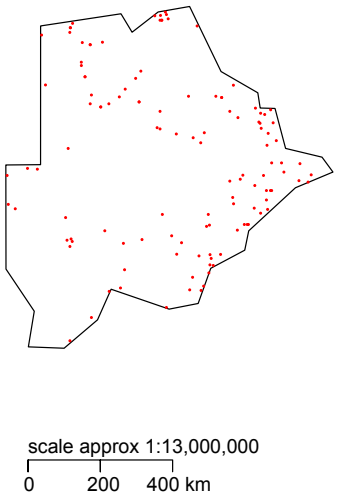

Figure 4.1: Distribution of sample sites in Botswana conatined in the MARA database.

4.2 Malaria Data

The malaria count data were extracted from the MARA/ARMA (Mapping Malaria Risk in Africa) database (Le Sueur *et al.*, 1997). MARA is the most comprehensive database on malaria compiling data from 1900 to present. It was initiated to provide a malaria risk atlas by collecting published and unpublished data from over 10 000 surveys across Africa (Gosoniu *et al.*, 2006). Malaria count data from surveys carried out on 47 171 children between 1 and 15 years old at 129 unique sites in Botswana are used in the present research (see Figure 4.1). Surveys which erroneously reported no sample size were excluded, leaving 122 prevalence surveys available for modelling. Historical data was used and hence might be outdated due to intervention programmes that have may have been implemented, although using such data has the advantage of including all the data in the model. Further, although generating risk maps using historical data must be interpreted with caution, Bayesian geostatistical risk mapping provides information relating to the uncertainty of the model-based estimates (Raso *et al.*, 2012). Such a historical approach has been undertaken by Gosoniu *et al.* (2006); Craig *et al.* (2007) and Raso *et al.* (2012).

4.3 Climate and Environmental Data

See Chapter 2, Table 2.1 on page 9 for the source of each climatic and environmental variable. NDVI data were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS). The TERRA MODIS satellite collects data about the earth's changing climate. In particular, MODIS vegetation indices product MOD13A3 was downloaded (NASA Land Processes Distributed Active Archive Center, 2001). MOD13A3 has a temporal resolution of one month and a spatial resolution of 1 km. These NDVI data are monthly data for the years 2000 to 2013. Long term temperature, rainfall and elevation are grid data were extracted from the WorldClim - Global Climate Data website (Hijmans *et al.*, 2005) based on data extracted between 1950 and 2000 at 1 km resolution. All raster layers must be in the same projection and must be precisely spatially aligned and cover exactly the same area in order for a statistical analysis to take place (Hijmans, 2013). Although the stated resolution of each of these layers is 1 km the layers come from different sources and could as a result differ for example, in accuracy (Huisman and Rolf, 2009, p. 312). Thus, after ensuring that these layers are in the same projection or have the same coordinate reference system, these layers could possibly still be out of sync or not aligned. It was observed that each NDVI layer had fewer cells than the WorldClim layers. It is always better to decrease the resolution through resampling methods rather than to increase the resolution (Gotway and Young, 2002). The NDVI layer was used as a reference layer and all WorldClim layers were resampled using the bilinear method so as to achieve a matching extent and resolution, that is to ensure that all raster layers are aligned. Surface Water Body Features were extracted from GEOnet Gazetteer (Gazetteer, 2006) as a shapefile data layer comprised of 46 591 derivative point gazetteer features based on 1:250 000 data. In the final composition of the spatial database only those data points at the locations of sample sites were considered, that is for each sample site an attribute for each predictor was known. Subsequently these data were imported into R and manipulated for analysis (see Section 3.2.2 on page 14 in Chapter 3 for a description of this process).

A survey of the research in similar malaria studies across Africa served as a guide as to which variables should be considered. A list of all the variables used at the start of the model building process with their full name, is provided in Table 4.1 and calculations used to obtain some of the variables not fully specified are provided in the Appendix.

4.4 Basic Exploratory Data Analysis

Figure 4.2 shows a plot of the observed malaria prevalence at each sample site in the study. This plot shows that the observed malaria prevalence in Botswana is relatively low and

uniform for most parts of the country. In the north higher prevalence and more variation can be seen.

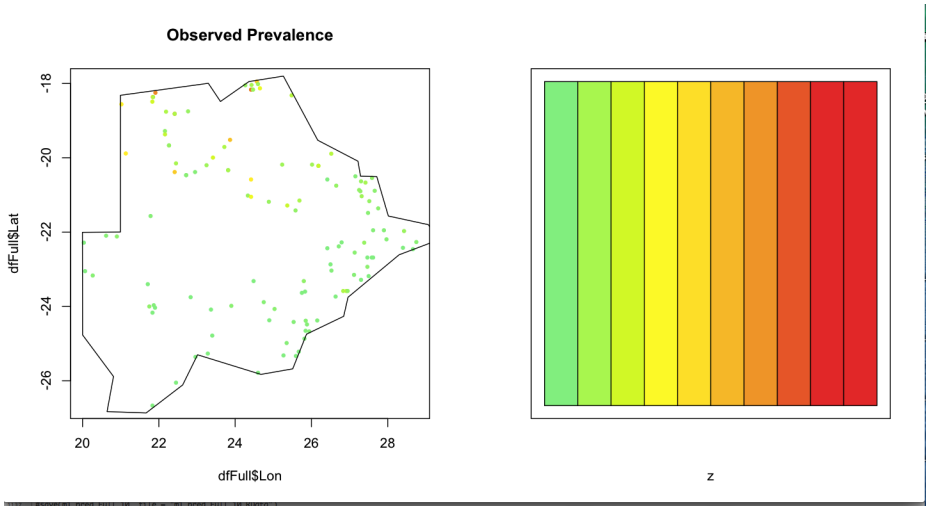

Figure 4.2: A plot of observed malaria prevalance at sample sites in Botswana from data conatined in the MARA database.

An informal test of spatial dependency and association between observed prevalence or risk of malaria was performed by plotting a bubble plot of sample sites in Botswana. Figure 4.3 shows the proportion of malaria cases out of the number examined multiplied by 4 over the maximum of this ratio.

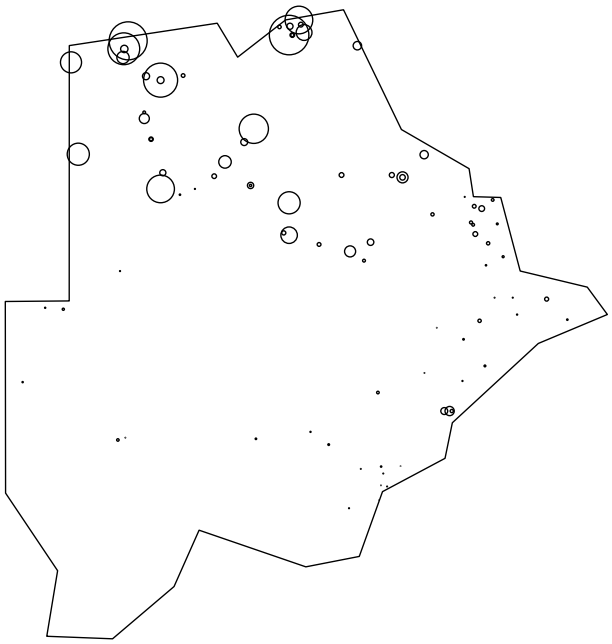

Figure 4.3: A bubble plot of sample sites in Botswana representing the proportion of malaria cases out of the number examined multiplied by 4 over the maximum of this ratio. This ratio is represented by the size of the circle.

These ratios, represented by circles, suggest that there is some association between big circles and other big ones that are close together. This is especially apparent in northern Botswana. Generally circles close together exhibit similar malaria risk intensities. Patterns of the attribute of interest, namely observed malaria risk, appear not to be random. Although, it must be noted that this is an informal test used to obtain a general sense of the observed spatial association between sample sites (Hengl, 2009).

4.5 Non-Spatial Model

A spatial database was compiled so that at each sample site a value for each climatic and environmental explanatory variable could be obtained (see Section 3.2.2 on page 14 in Chapter 3). Table 4.1 shows the explanatory variables used in the non-spatial modelling with a

description of each variable.

Table 4.1: Variables by theme used in non-spatial model building.

Varaible	Description	Theme
bio1	Annual Mean Temperature	Temperature
bio2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	
bio3	Isothermality (bio2/bio7) (* 100)	
bio4	Temperature Seasonality (standard deviation*100)	
bio5	Max Temperature of Warmest Month	
bio6	Min Temperature of Coldest Month	
bio7	Temperature Annual Range (bio5-bio6)	
bio8	Mean Temperature of Wettest Quarter	
bio9	Mean Temperature of Driest Quarter	
bio10	Mean Temperature of Warmest Quarter	
bio11	Mean Temperature of Coldest Quarter	
summerTemp	Summer Temperature (months 12, 1, 2, 3)	
winterTemp	Winter Temperature (months 4-10)	
SDTemp	Standard Deviation of Annual Temperature	
bio12	Annual Rainfall	Rain
bio13	Rainfall of Wettest Month	
bio14	Rainfall of Driest Month	
bio15	Rainfall Seasonality (Coefficient of Variation)	
bio16	Rainfall of Wettest Quarter	
bio17	Rainfall of Driest Quarter	
bio18	Rainfall of Warmest Quarter	
bio19	Rainfall of Coldest Quarter	
q	Mean Peak Month where Rainfall is Concentrated	
rCIndex	Rainfall Concentration Index	
totRain	Annual Total Rainfall	
summerRain	Summer Rainfall (months 12, 1, 2, 3)	
winterRain	Winter Rainfall (months 4-10)	
SDRain	Standard Deviation of Annual Rainfall	
summerNDVI	Summer NDVI (months 12, 1, 2, 3)	NDVI
winterNDVI	Winter NDVI (months 4-10)	
SDNDVI	Standard Deviation of Annual NDVI	
NDVI	Annual NDVI	
DstTClW	Distance to Closest Water Surface	Unthemed
altitude	Altitude	

In Stage 1, for cross-validation purposes, as discussed in Section 3.4.2 on page 27 in Chapter 3, the malaria prevalence dataset were randomly split into derivation (n = 104) and validation (n = 18) subsets. All model building proceeded on the derivation dataset. Stage 1 also involved selecting variables that were good predictors of malaria prevalence. Each variable was tested in a univariate logistic regression model. Of the 34 potential explanatory variables, all were significantly associated with malaria prevalence in univariate logistic regression (see Table 4.2).

Table 4.2: Significant variables associated with malaria prevalence in univariate logistic regression in Stage 1 ranked from lowest AIC to highest - Stage 2. The P(z) column represents the P value, or significance level. The smaller the P value, and if it is less than a threshold probability, the stronger the evidence, in this case, against the exclusion of a variable in the univariate logistic regression.

Independent Variable	AIC	P(z)
bio9	676.60	0.00
bio11	677.06	0.00
winterTemp	680.21	0.00
rCIndex	714.96	0.00
bio15	726.85	0.00
bio6	727.45	0.00
bio1	736.56	0.00
bio17	753.62	0.00
winterRain	813.10	0.00
SDTemp	820.10	0.00
bio4	820.16	0.00
bio13	843.60	0.00
bio19	851.35	0.00
bio5	857.91	0.00
SDRain	861.68	0.00
bio14	887.53	0.00
bio16	888.25	0.00
summerRain	900.93	0.00
bio10	960.73	0.00
summerTemp	1008.64	0.00
totRain	1020.07	0.00
bio12	1020.07	0.00
bio8	1023.37	0.00
altitude	1027.90	0.00
bio18	1059.81	0.00
summerNDVI	1066.56	0.00
NDVI	1075.83	0.00
SDNDVI	1076.26	0.00
winterNDVI	1079.44	0.00
DstTClW	1084.40	0.00
bio3	1088.01	0.01
q	1088.39	0.01
bio7	1089.27	0.02
bio2	1092.31	0.01

In Stage 2 variables that were significant in Stage 1 were ranked based on each model's AIC score and then tested within each theme for correlation among the variables. Variables were tested within three themes, namely temperature, rain and NDVI. Variables that were strongly correlated, Spearman's $r > 0.85$, with a higher-ranking (AIC) variable belonging to the same theme were excluded. In the temperature theme mean temperature of driest quarter (bio9), standard deviation of annual temperature (SDTemp), maximum temperature of warmest month (bio5), summer temperature , isothermality (bio3) and annual temperature range (bio7) were selected. In the rain theme rainfall concentration index (rCIndex), precipitation of wettest month (bio13), precipitation of coldest quarter (bio19), precipitation of driest month (bio14), total rain (totRain), precipitation of warmest quarter (bio18) were selected. In the NDVI theme summer NDVI, winter NDVI, annual NDVI, standard deviation of annual NDVI and NDVI were all correlated with summer NDVI having the lowest AIC. The remaining unthemed variables representing unrelated explanatory variables, namely distance to closest water source (DstTClW) and altitude, were added to the selected variables from each theme in Stage 2. Individual scatter plots of logit(p) against these 15 variables selected at Stage 2 for further analysis, are shown in Figure 4.4.

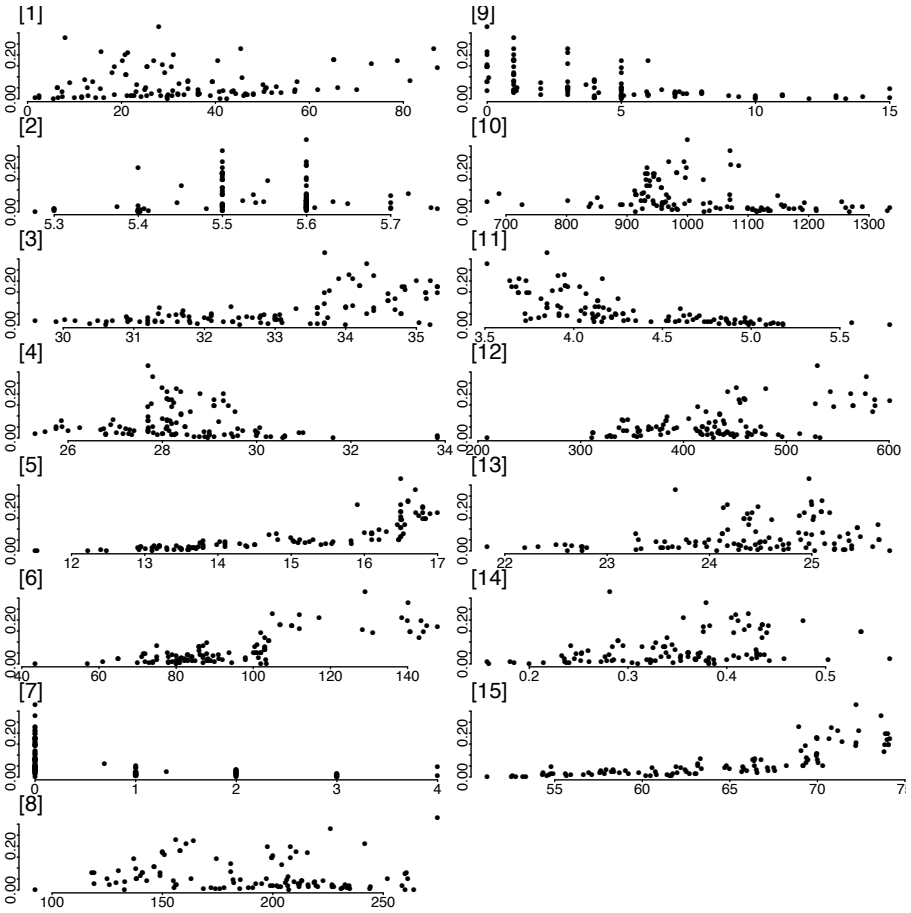

Figure 4.4: Scatter plots of candidate explanatory variables selected in Stage 2 to be used in step-wise procedures. Malaria prevalence in 1 to 14 year old children in Tanzania based on historical MARA data is reprsented by the Y axis on a logit scale. On the X axis are the following variables (see table for variable description): [1] DstTClW in km, [2] bio3 in °C, [3] bio5 in °C, [4] bio7 in °C, [5] bio9 in °C, [6] bio13 in mm, [7] bio14 in mm, [8] bio18 in mm, [9] bio19 in mm[10], altitude in m above sea level, [11] SDTemp in °C, [12] totRain in mm, [13] summerTemp in °C [14] summerNDVI ratio [0,1], [15] rCIndex percentage between 0 and 100.

In Stage 3, 1 000 bootstrap samples from the derivation data were run on the list of variables that survived Stage 2 and an automated backward exclusion procedure on each sample was performed, that is automated backwards stepwise elimination in conjunction with bootstrap resampling (Austin and Tu, 2004). The automatic backward exclusion procedure involved starting with all candidate variables, testing whether each variable should be deleted using the AIC criterion, deleting the variable, if any, that improved the model the most by being deleted. This process continued on each bootstrap sample until no further improvement was possible. The frequency, a number out of a thousand, with which a candidate variable was selected as an independent predictor in each bootstrap sample was recorded as well as the frequency of the sign of the coefficients, as shown in Table 4.3.

Table 4.3: Results of bootstrap backward step-wise procedure models in Stage 3 and Stage 5 against 1000 bootstrap samples of the malaria prevalence data, yielding a candidate list of variables to be analysed in remaing stages. The selection frequency is preseneted as well as the rate of change of the sign of the coefficient for each variable as a percentage. Coef+ and Coef− represent the frequency with which a coefficient is positive and negative respectively.

Stage 3				Stage 5			
Candidate List	Freq	Coef+	Coef−	Final List	Freq	Coef+	Coef−
bio9[1]	997.00	100.00	0.00	winterTemp[2]	993.00	100.00	0.00
altitude[1]	987.00	100.00	0.00	altitude	900.00	99.00	1.00
bio5[1]	941.00	0.00	100.00	bio5[3]			
bio7 [1]	914.00	99.00	1.00	bio7[4]	796.00	93.00	7.00
summerTemp	875.00	3.00	97.00				
summerNDVI	873.00	5.00	95.00				
SDTemp	869.00	98.00	2.00				
DstTClW[1]	865.00	99.00	1.00	DstTClW	843.00	100.00	0.00
bio18[1]	723.00	86.00	14.00	bio18	804.00	88.00	12.00
rCIndex	663.00	5.00	95.00				
bio3	657.00	11.00	89.00				
totRain[1]	653.00	23.00	77.00	totRain[5]	745.00	21.00	79.00
bio13	617.00	75.00	25.00				
bio19	563.00	32.00	68.00				
bio14	476.00	17.00	83.00				

[1] Variables selected into Stage 4 model.

[2] Previously excluded variable selected more frequently than bio9 in bootstrap procedure.

[3] Selected more frequently than previously excluded bio10 in bootstrap procedure.

[4] Selected more frequently than previously excluded bio2 in bootstrap procedure.

[5] Selected more frequently than previously excluded bio12 in bootstrap procedure.

Given this candidate list of variables from Stage 3, Stage 4 involved performing manual stepwise tests for inclusion starting with the most frequently selected variable from Stage 3. The manual forward stepwise regression continued as long as all entered variables remained significant at the 5% probability level. If a previously entered variable became non-significant with the addition of another, the one more frequently selected was retained. The marked variables, denoted by the superscript equal to 1 in Table 4.3 were selected into the Stage 4 model.

Stage 5 involved adjusting the Stage 4 model by re-assessing variables that were previously excluded at Stage 2 using further bootstrapping procedures. The excluded correlated variables in each theme corresponding to the favoured variable chosen at Stage 2 is allowed to re-enter the model in order to compete for selection in the Stage 5 model. Selection is based on frequency of selection in bootstrapped samples and if a variable was not excluded by the stepwise algorithm. Except for winter temperature which re-entered the model replacing bio9, none of the previously excluded variables that re-entered improved the model based on these criteria. The variables denoted by superscripts 2 to 5 in Table 4.3 are the variables that survived after re-assessing the bootstrap model with previously excluded variables.

The variables that survived in Stage 5 were used to fit a logistic regression model. As per the theory of generalized linear models (GLM) as described in Section 3.3.7 on page 23 in Chapter 3, since the response is in the form of count data a link function was required to ensure that the expectation of the response was a linear function of the Stage 5 explanatory variables. The logit link was used for this purpose. In R, the glm function in the stats package, which is a standard package supplied with R (R Core Team, 2013), was used to perform the logistic regression. The results of the Stage 5 non-spatial logistic model are presented in Table 4.4.

Table 4.4: Stage 5 non-spatial model results. Odds ratios, and corresponding confidence interval estimated from non-spatial regression against seven variables, fitted on derivation data only (n = 106, AIC = 613.8).

	Variable	Odds Ratio	P(z)	95% CI
1	winterTemp	6.01	0.001	(3.657,9.931)
2	altitude	1.00	0.001	(1.002,1.005)
3	bio5	0.48	0.01	(0.297,0.768)
4	bio7	1.53	0.01	(1.154,2.044)
5	DstClW	1.01	0.01	(1.003,1.014)
6	bio18	1.01	0.01	(1.005,1.013)
7	totRain	1.00	0.001	(0.994,0.999)

4.6 Spatial Model

In Stage 6 the variables that survived in Stage 5 were used to fit a spatial generalized linear model (SGLM), also known as a generalized linear mixed model. As discussed in Section 3.6.4 on page 34 in Chapter 3, the SGLM extends the GLM by allowing additional sources of variability that occur due to unobservable random effects (Christensen *et al.*, 2000). Diggle *et al.* (1998) proposed and described how the spatial dependence or error structure for SGLMs can be modeled via Gaussian processes for point-level, geostatistical data. Following the primary reference paper (Craig *et al.*, 2007) of this study, stationarity is assumed and the exponential covariance function is also the assumed covariance structure. This covariance structure is imposed in order to describe the spatial dependence or error structure among the observations (Diggle and Ribeiro, 2007). The spGLM function in the spBayes package (Finley and Banerjee, 2013) was used to fit this SGLM. The probability that $y(s_i) = 1$ or 0, that is the probability of an individual between 1 and 15 years of age being infected with malaria or not, is given by

$$p(s_i) = \frac{\exp(\mathbf{x}(s_\mathbf{i})^T\boldsymbol{\beta} + \mathbf{w}(s_i))}{1 + \exp(\mathbf{x}(s_\mathbf{i})^T\boldsymbol{\beta} + \mathbf{w}(s_i))}$$

where it is assumed the sample sites $s_1 \dots s_n \in \mathbf{D}$ where $\mathbf{D}$ a fixed subset of $\mathbb{R}^2$. The explanatory variables from Stage 5 are included in the transposed vector $\mathbf{x}(s_\mathbf{i})^T$ associated with each site s_i, and $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients. These coefficients from the non-spatial model served as starting points and the Cholesky square root of the regression parameters estimated covariance were used as Metropolis tuning values in the spGLM function (Finley and Banerjee, 2013). As per the details of the hierarchical Bayesian model setup specified detailed in Chapter 3 in Equation 3.23 on page 42, the second stage specifies the association in the random effects. A Gaussian process specifies the random effect, denoted by $\mathbf{w}(\mathbf{s}) = \text{MVN}(0, \boldsymbol{\Sigma}(\boldsymbol{\Theta}))$, with Θ denoting the variance, σ^2, and the decay, ϕ, parameters as defined in Chapter 3 in Section 3.6.4.1 on page 35. Staring values for these parameters in the spGLM function were specified as follows, $\phi = \frac{3}{0.5(\text{d})}$, $\sigma^2 = 1$, $w = 0$. A non-informative flat prior for the regression effects $\boldsymbol{\beta}$ that is $p(\boldsymbol{\beta}) \propto 1$ was assigned. For the variance parameter, σ^2, an inverse Gamma prior was assigned with shape and scale parameters 2 and 1 respectively. Prior distributions assigned to the decay parameters are typically set relative to the size of their domains (Finley *et al.*, 2007). The approximate effective range, r, which is the range at which the magnitude of correlation decays to 5% of its maximum value, is given by solving the equation for the exponential correlation function given in Chapter 3, 3.18 on page 36, that is solving, $\exp(-\phi r) = 0.05$, to give $r \approx \frac{3}{\phi}$ (Finley *et al.*, 2007). Hence the effective range is represented by the denominator of the fraction

with the numerator being 3. As a result for the decay parameter, ϕ uniform priors were assigned with an upper and lower range of $\frac{3}{\mathrm{d}}$ and $\frac{3}{0.1(\mathrm{d})}$ respectively where d represents the maximum intersite Euclidean distance and equates to approximately 1010 km. Scanning the literature where the spGLM function has been used various variants of the starting values and ranges of priors were attempted. For example the range of the spatial decay parameter, ϕ, was tested using a larger range than currently used, namely $\frac{3}{\mathrm{d}}$ and $\frac{3}{0.01(\mathrm{d})}$. This resulted in a lack of convergence and a much longer running time. The results of the Stage 6 spatial model are presented in Table 4.5 and the results of the mean error and mean absolute error of the spatial and non-spatial prediction at validation sites are presented in Table 4.6.

Table 4.5: Stage 6 spatial model results. Odds ratios, and corresponding credibility interval derived from 70000 bayesian simulations, fitted on all data (n = 122).

	Variable	Odds Ratio	95% Credibility Interval
1	winterTemp	30.54	(1.550, 1244.975)
2	altitude	1.84	(0.947, 4.057)
3	bio5	0.31	(0.009, 5.451)
4	bio7	1.91	(0.293, 19.954)
5	DstTClW	1.09	(0.794, 1.453)
6	bio18	1.95	(1.428, 2.716)
7	totRain	0.68	(0.378, 1.158)

Table 4.6: Mean error and mean absolute error of spatial and non-spatial prediction at validation sites.

	Measure	Spatial	NonSpatial
1	Mean Error	-0.15	-0.46
2	Mean Absolute Error	0.20	0.46

The spatial maps of mean predicted malaria prevalence as well as the associated standard deviation of predicted malaria prevalence in Botswana resulting from the Stage 6 spatial model at a 20 km resolution are presented in Figures 4.5 and 4.6.

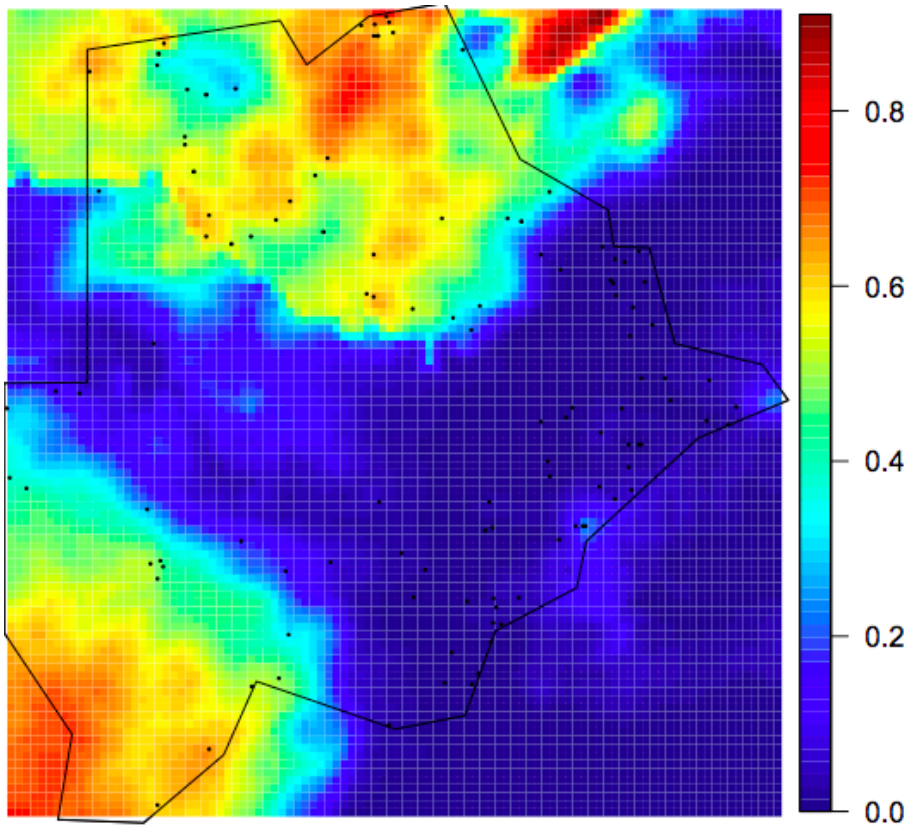

Figure 4.5: Map of mean predicted malaria prevalence in Botswana resulting from the Stage 6 spatial model at a 20 km resolution.

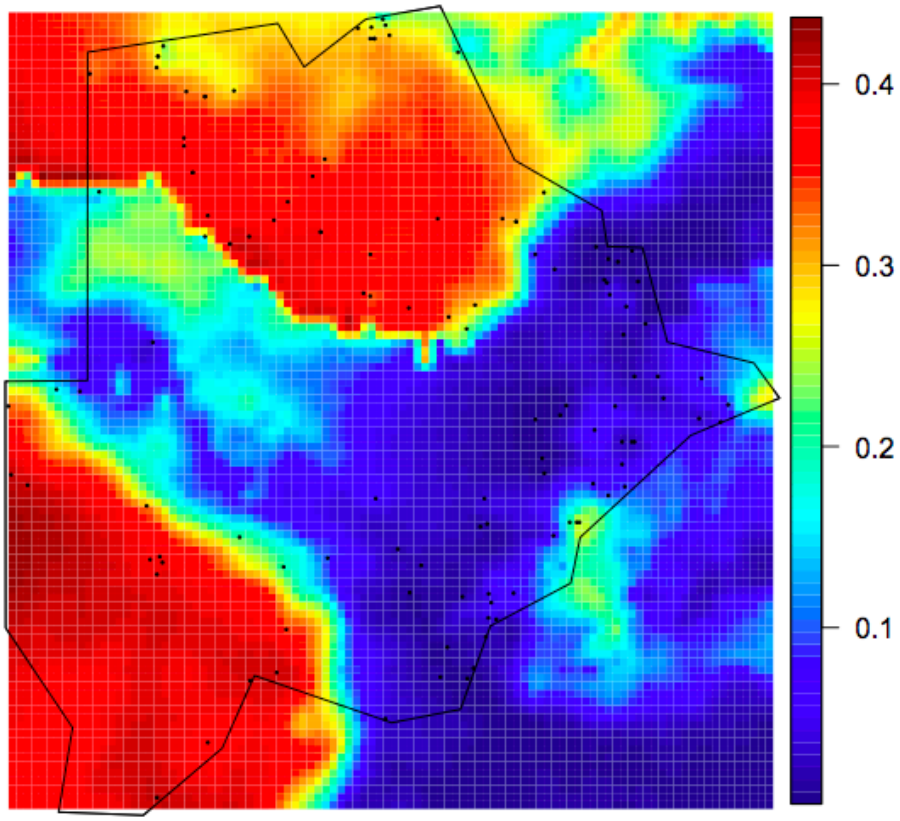

Figure 4.6: Map of associated standard deviation of predicted malaria prevalence in Botswana resulting from the Stage 6 spatial model at a 20 km resolution.

4.7 Discussion

As more variables are tested against a certain data set, there is a greater risk that some will explain the data merely by chance, but will fail to explain new data (Craig *et al.*, 2007). Selecting a small subset of variables for spatial modelling from a large number of potential candidates is a major challenge and can easily become arbitrary (Craig *et al.*, 2007). The ideal solution would be to test every possible combination of variables in a Bayesian spatial framework. However, from a computing point of view this is unfeasible, if not impossible (Craig *et al.*, 2007). In the interest of finding the most practical and parsimonious solution the list of candidate variables was reduced using non-spatial selection methods before moving to the spatial context. The small subset of variables derived in this manner, although each independently associated with the response, may possibly have been spurious because the spatial correlation was not yet acknowledged. For this reason in Stage 6 this subset of variables was fitted in a Bayesian geostatistical model. The spatial model derived from the observed locations was used to predict prevalence of malaria infection in children 1–14 years old at unobserved map locations across the whole of Botswana.

Correlation among predictors compromises the identification of consistent predictors (Craig *et al.*, 2007). As a result if more than one correlated variables compete for entry into a model, a strong, reliable predictor may ultimately be selected less frequently than a weaker predictor (Austin and Tu, 2004). Given this it was crucial that the candidate list contained only variables that are slightly correlated. This was achieved in Stage 2 where the candidate list was reduced from 34 to 15 variables.

A set of predictors are reliable if they not only explain a particular data set, but are associated consistently with the response (Craig *et al.*, 2007). The bootstrapping of Stage 3 aimed to identify such predictors because those that consistently explain different subsets of the data will more likely do a better job at explaining new data (Austin and Tu, 2004). The step-wise bootstrap procedures ensure that variables which explained the most observations would be selected most frequently while those that only explained few observation would be selected only when these observations appeared in the bootstrap sample. The effect of individual observations, in particular outlying observations, on variable selection is thus minimized.

Univariate ranking (Stage 1 and 2) can lead to a problem known as "data peeking" (Babyak, 2004). The phrase "data peeking" refers to the process of examining the relation between an explanatory variable and the response variable, in isolation, in order to select which variables to include or exclude from a regression model (Babyak, 2004). As a result the data is artificially set up for success in that undeclared testing and discarding of variables, as was

done in these early stages, may lead to illegitimately high model fit. Furthermore at Stage 2 variables were excluded based on low univariate correlation with the response variable. This says nothing of their predictive power which may be different when other variables are accounted for (Craig *et al.*, 2007). For example, variables tested on their own in a univariate setting may behave differently with respect to the response variable when they are considered simultaneously with one or multiple other variables. If there is a suppressor variable present, for example, the relation between a variable and a response variable may not appear to be important when tested in isolation, but may become important after including other explanatory variables (Babyak, 2004). Conger (1974) describes a suppressor variable as a variable which when included in a regression model increases the predictive validity of another variable or multiple variables. Stage 5 sought to address these issues, by giving each variable excluded in Stage 2, in favour of its surviving counterpart in Stage 4, a chance to re-enter and possibly outperform its competitors in a multiple-variable context. At the same time the bootstrap sub-sampling reduces the influence of the data set on this process (Craig *et al.*, 2007). Winter temperature was such a variable that re-entered when allowed to re-compete in a multiple variable context.

The Stage 3 bootstrap-stepwise procedures also provided useful information regarding the frequency distributions of coefficients in the 1 000 stepwise models. An insight into the reliability of a predictor can be seen in this way. A variable whose coefficient varies widely, or one that is sometimes positive and sometimes negative, is not reliable and should be considered cautiously (Concato *et al.*, 1993). Austin and Tu (2004) found that 60% was an optimal cut-off level for including the predictors in the final equation. Austin and Tu (2004) also note that if a coefficient is positive half the time and negative the other then that is an indication of instability in the model.

Consider the results of Stage 3 to Stage 5 presented in Table 4.3. It can be seen that some variables were unstable, having positive coefficients in some models and negative coefficients in others. The variable depicting precipitation of coldest quarter (bio19) was the most unstable. It was also selected second to last frequently in the bootstrap samples. The benefits of Stage 3 can be seen with the variable altitude. In Stage 2 it performed only reasonably well- it's univariate ranking positioned it somewhere in the bottom half of Table 4.2. However in Stage 3 in the bootstrapped multiple variable context it proved to be the second most frequently selected variable and it progressed to Stage 4 and 5. Altitude, bio5 and bio7 were selected most frequently, apart from bio9 which was replaced by winterTemp in Stage 5, and were all selected into Stage 4 and Stage 5 and all variables which progressed had stable coefficients. These results confirm the usefulness of Stage 3 as a way of selecting the most important predictors.

Exponentiation of the model parameters, in the non-spatial and spatial models, gives the odds ratio for each explanatory variable. The odds ratio indicates whether there are negative or positive relationships and the strength of relationships between the explanatory and outcome variables (Dobson and Barnett, 2008, p. 152). Testing model parameter significance in the spatial model was based on 95% credibility intervals (C_rI). If the value zero is not in 95% of the C_rI then the estimated parameter of the model is significant. Consider the credibility interval column in Table 4.5: it can be seen that all parameters are significant. All of these explanatory variables, in the final spatial model, or transformations of these, have been successfully implemented in previous geostatistical modelling approaches employed in other African countries (see Section 2.2 on page 8 in Chapter 2).

Consider the trace plot of winter temperature (winterTemp) as shown in Figure 4.7. As described in Section 3.6.5.6 on page 47 in Chapter 3, this is a trace plot of all 3 MCMC chains combined for the parameter winter temperature. This plot shows erratic movement in the parameter space. The trace plots for all parameters are shown in Figures 4.7, 4.8, 4.9, 4.10 on pages 72 to 74. Inspecting these plots, as was seen with the trace plot of winter temperature, it is not clear to see that the 3 chains have converged. The inspection of trace plots of each parameter are as a result coupled with the assessment of the Gelman and Rubin's convergence diagnostic score (Brooks and Gelman, 1998). Approximate convergence is diagnosed when the upper confidence limit is close to 1. An upper confidence limit of 1.07 was obtained which suggests that the spatial model in Stage 6 has approximately converged adequately.

It should be noted the differences between the studies can be attributed to the fact that the reference study had more themes and variables available and the prevalence data used does not refer to the same time period. Craig *et al.* (2007) implemented the same type of non-spatial and spatial models using MARA data over survey years 1961 to 1962. The present study used data also from MARA but spanning from 1944 to 1997. With respect to their temperature theme, they found annual mean temperature to be significant in their final spatial model. This research found maximum temperature of warmest month (bio5), annual range of temperature (bio7), mean winter temperature to be significant. In terms of their rain theme, they found total summer rainfall to be significant. This research found total annual precipitation and precipitation of warmest quarter (bio18) to be significant in final spatial model. They found altitude to be significant in their final spatial model. This research also found altitude and distance to closest water surface (DstTClW) to be significant in the final spatial model (DstTClW variable was not part of their initial list of variables tested). Gosoniu *et al.* (2010) found temperature, altitude, distance to the nearest water surface to be significantly associated with malaria prevalence in Angola. This model included socio-economic index and indoor residual spraying variables which were not tested

in the current study because they could not easily be obtained for Botswana. At the time of writing, it was found in this study that some GIS data was not as easily available or obtainable as other GIS data.

High rainfall during the hot summer months, as reflected by bio18 - precipitation of warmest quarter, allows rapid breeding and population expansion of the mosquito vectors (Craig *et al.*, 2007). Zacarias and Andersson (2010) found that in Mozambique, malaria transmission is higher in the wet season with both temperature and rainfall positively related to malaria. Annual total rainfall is positively associated with malaria risk and it is conceivable for it to also influence malaria breeding and risk (Zacarias and Andersson, 2010). High temperatures, reflected by bio5 - maximum temperature of warmest month and influenced by bio7 - annual range of temperature, maximizes the maturation rate of the parasite in its exothermic arthropod host (Molineaux *et al.*, 1988). Warmer winters, as reflected by a positive association between winter temperature and malaria risk, reduces the die-back of mosquitoes and parasites, in this way increasing the reservoir for the following season (Molineaux *et al.*, 1988). In general, the warmer the climate, the better chance the mosquito has for survival (World Health Organization, 2014). A major finding in this study is that winter temperature has by far the greatest effect on malaria risk. Referring to Table 4.5, it can be see that the odds ratio for winter temperature is about 15 times greater than any other predictor. To see what effect would be had on malaria risk without winter temperature the spatial model was run without this variable. Malaria risk was virtually zero all over the country without accounting for winter temperature.

Scarcity of data or sparse data in certain areas can introduce large prediction errors (Gosoniu *et al.*, 2010). The spatial maps resulting from the Stage 6 spatial model at a 20 km resolution include a map of mean predicted malaria prevalence as well as the associated standard deviation of predicted malaria prevalence in Botswana (see Figures 4.5 and 4.6). Considering these spatial maps generated in this study large errors can be seen where there are few data points. There is evidence of this in the south western region of Botswana where the model in this study over predicts malaria risk greatly and presents a picture of high malaria risk where the reference paper (Craig *et al.*, 2007) followed predicts no risk in that region. This region happens to have the fewest observations and also accordingly has high uncertainty as seen in corresponding map of standard deviation (see Figure 4.6). Two different resolutions were attempted in the generation of these maps, 10 km and 20 km, respectively. The memory resources on the computer used for computations, namely a 1.8 GHz Intel Core i5 with 4 GB RAM, were exceeded when predicting risk on the 10 km resolution grid. No problems were experienced using a 20 km resolution grid. Determining a suitable balance between computer capabilities and map precision, by experimenting with varying grid sizes, is a common goal in geostatistics (Swanson *et al.*, 2013).

4.8 Conclusions

The objectives in this thesis were primarily to: 1) assess whether there is evidence to link the incidence of malaria prevalence to environmental and climatic variables operating in the area, 2) assess whether the non-spatial selection procedure is effective and whether it has had an effect on selecting spatial variables 3) assess the predictive performance of the non-spatial versus the spatial model, 4) ascertain if there are any areas of high malaria risk, 5) assess whether the predictions of prevalence are useful and whether they can they be used to develop a GIS, 6) determine if all the necessary routines are available in R to conduct the analyses and 7) assess whether the process can be automated.

All of the explanatory variables in the final spatial model were positively associated with malaria prevalence and all that survived to Stage 6 were significant in the final model. There was evidence to suggest that winter temperature had the greatest effect on malaria prevalence in Botswana given the data in this study. Evidence for this can be seen in Table 4.5, where the odds ratio for winter temperature is about 5 times greater than any other predictor. Leaving winter temperature out of the spatial model malaria risk was virtually zero all over the country.

The non-spatial staged variable selection process proved to be practical, although not necessarily optimal. On repeating the procedure a second time some new variables were added and some variables in the current model were excluded. This suggests that some variables explain the data merely by chance, but will fail to explain new data. Multiple bootstrap samples drawn from the data allowed for the identification of consistent and stable explanatory variables. The selection frequency criterion provided an objective means for choosing between two variables, and to choose between variables that were strongly correlated. Although this non-spatial selection procedure proved practical and able to identify stable explanatory variables and also able to provide an objective means for selecting one variable over another, ultimately its efficacy is questionable due to the fact that a unique set of spatial variables could not be selected.

The mean prediction error measure suggests that the non-spatial model overestimated malaria prevalence at sample sites 3 times more so than it did in the spatial model. The mean prediction absolute error measure suggests that the average magnitude of prediction errors is also less in the spatial than in the non-spatial model (see Table 4.6). As a result, there is evidence in the current study to suggest that the spatial model's predictive performance was better than the non-spatial model.

The smoothed spatial map presented in this study, namely Figure 4.5, is similar over large portions than that of the reference paper followed in this study (Craig *et al.*, 2007). In the

east of Botswana malaria prevalence is fairly similar. Both maps present low prevalence in the east which gradually increases northwards. High malaria prevalence can be see in parts of the north of Botswana and also in the south west. Where the maps differ most is in the south western region of Botswana. The model in this research over predicts greatly in this region and presents a high picture of malaria risk where the reference paper predicts no risk in that region. This region happens to have the fewest observations and also accordingly has high uncertainty as seen in corresponding map of standard deviation (see Figure 4.6). Sparse data is difficult to predict and the reliability of the map depends on the data available to derive the model (Gosoniu *et al.*, 2010). Therefore it is conceivable for such weaknesses to exist in the map across data-sparse regions (Howes *et al.*, 2012). A more similar picture of malaria risk, and a more accurate one, might have been achieved had a more comprehensive list of predictors been used. Craig *et al.* (2007) had at their disposal 50 variables to begin with, in this research only 34 variables could be obtained, and the prevalence data was also not over the same period.

R as an open source program, with its wide array of geospatial packages, proved to provide all the necessary routines needed to conduct the analyses. Automating the analyses proved more difficult than expected. For example, working with MODIS NDVI data required a technical understanding of the data which are Hierarchical Data Format (HDF) files. Manipulating, reprojecting, merging and aligning raster data proved challenging and required a fair amount of programming to accomplish. Compiling the spatial database can be automated but not without careful thinking and a good understanding of all the checks involved. The variable selection procedure involves many steps and comparisons such as the AIC criterion, the frequency selection criterion and the correlation criterion. Determining which variables in each theme, one variable at a time, were correlated and which of these had the lowest AIC, was automated. The rest of the procedure had to be completed each staged at a time. The automation of all these steps in both compiling the spatial database and the variable selection procedure seemed unnecessarily difficult given the current scope. The length of time for the running of each MCMC chain is also a drawback. At least 24 hours was needed to run 3 chains of 350 000 simulations each. If the chains were not mixing well 3 chains of 350 000 simulations each would take even longer than 24 hours to run. Experimenting with different parameters and settings in order to achieve convergence in a reasonable time period is thus a lengthy process making the spatial analysis more difficult to automate.

Revisiting and extending this study in the future may reveal that ignoring spatial correlations during the non-spatial variable selection procedure could prove to be a major weakness, leading to sub-optimal variable selection results. As computers get more powerful and as statistical software packages are further developed, a variable selection procedure within a spatial framework may be viable for the non-expert researcher.

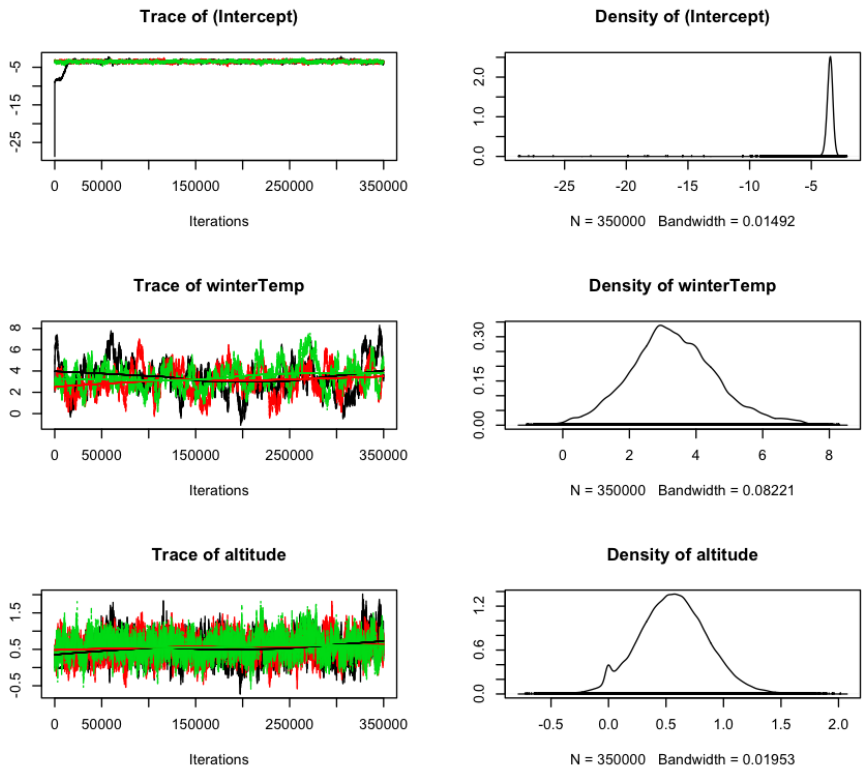

Figure 4.7: MCMC chain trace plots.

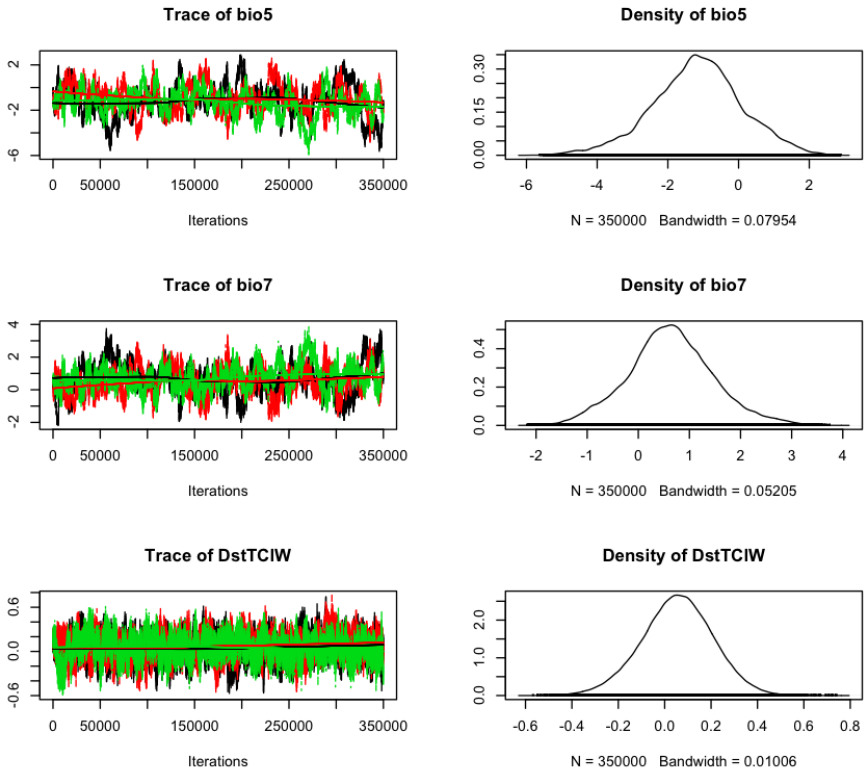

Figure 4.8: MCMC chain trace plots.

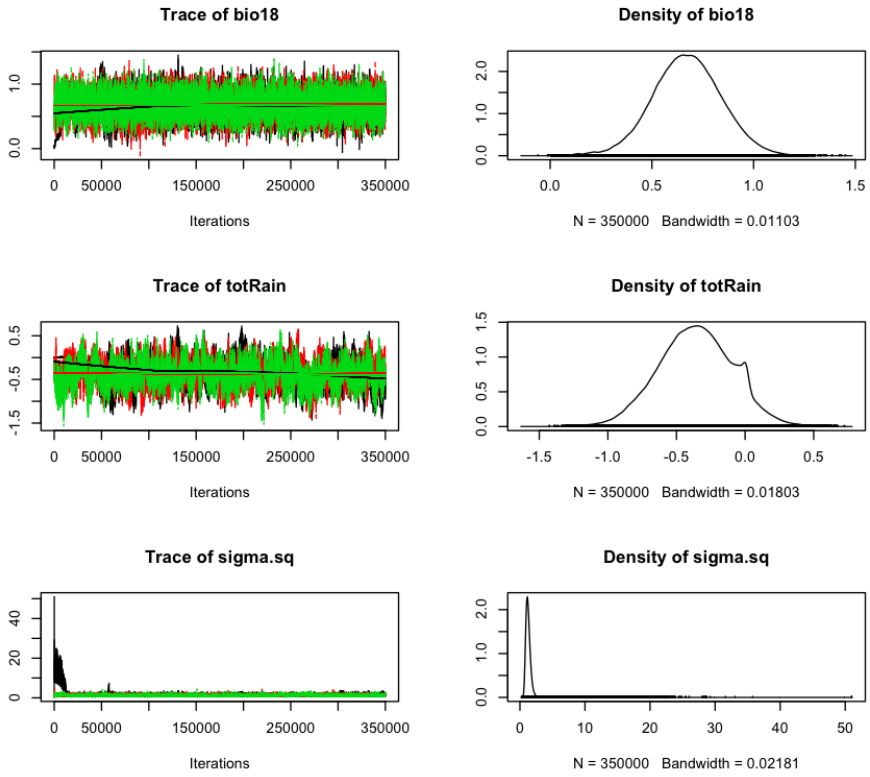

Figure 4.9: MCMC chain trace plots.

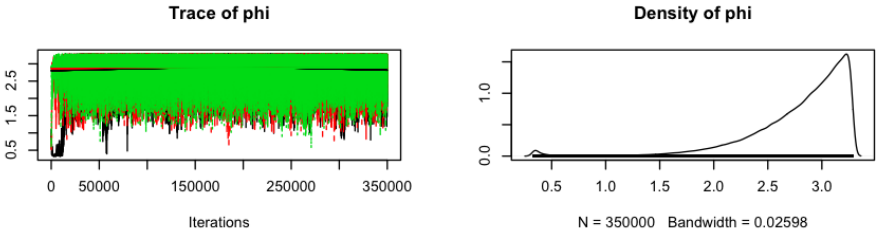

Figure 4.10: MCMC chain trace plots.

Bibliography

Abeku, T., Hay, S., Ochola, S., Langi, P., Beard, B., de Vlas, S., and Cox, J. (2004). Malaria Epidemic Early Warning and Detection in African Highlands. *Trends in Parasitology*, *20*(9), 400–405.

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, Volume 55. Dover Publications.

Adler, R. (2004). Gaussian Random Fields on Manifolds. In *Seminar on Stochastic Analysis, Random Fields and Applications IV*, 3–19. Springer.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*, 267–281. Akademinai Kiado.

Arab, A., Hooten, M., and Wikle, C. (2008). Hierarchical Spatial Models. In S. Shekhar and H. Xiong (Eds.), *Encyclopedia of GIS*, 425–431. New York: Springer.

Austin, P. and Tu, J. (2004). Bootstrap Methods for Developing Predictive Models. *The American Statistician*, *58*(2), 131–137.

Babyak, M. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-type Models. *Psychosomatic Medicine*, *66*(3), 411–421.

Banerjee, S., Gelfand, A., and Carlin, B. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families: In Statistical Theory*. Wiley New York.

Basáñez, M., Marshall, C., Carabin, H., Gyorkos, T., and Joseph, L. (2004). Bayesian Statistics for Parasitologists. *Trends in Parasitology*, *20*(2), 85–91.

Becker, A., Wilks, A., Brownrigg, R., and Minka, T. (2013). *maps: Draw Geographical Maps*. R package version 2.3-6.

Bernardo, J. (1996). *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994*, Volume 5. Oxford University Press, USA.

Bernhardsen, T. (2002). *Geographic Information Systems: An Introduction*. John Wiley and Sons.

Bivand, R. (2013). *spdep: Spatial Dependence: Weighting Schemes, Statistics and Models*. R package version 0.5-68.

Bivand, R., Keitt, T., and Rowlingson, B. (2013). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.8-10.

Bivand, R. and Nicholas, L. (2014). *maptools: Tools for Reading and Handling Spatial Objects*. R Package Version 0.8-29.

Bivand, R., Pebesma, E., and Virgilio, G. (2008). *Applied Spatial Data Analysis with R* (First ed.). Springer.

Bolin, D. (2009). *Computationally Efficient Methods in Spatial Statistics*. Ph. D. thesis, Lund University.

Breslow, N. and Clayton, D. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Brooks, S. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Burrough, P. and McDonnell, R. (1998). *Principles of Geographical Information Systems*, Volume 333. Oxford University Press Oxford.

Chammartin, F., Scholte, R., Guimarães, L., Tanner, M., Utzinger, J., and Vounatsou, P. (2013). Soil-transmitted Helminth Infection in South America: A Systematic Review and Geostatistical Meta-analysis. *The Lancet Infectious Diseases*, *13*(6), 507–518.

Christensen, O., Møller, J., and Waagepetersen, R. (2000). Analysis of Spatial Data Using Generalized Linear Mixed models and Langevin-type Markov Chain Monte Carlo. Technical Report R-00-2009, Department of Mathematical Sciences, Aalborg University, Aalborg.

Christensen, O. and Ribeiro Jr, P. (2002). geoRglm - A Package for Generalised Linear Spatial Models. *R News*, *2*(2), 26–28. ISSN 1609-3631.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer.

Cibulskis, R., Bell, D., Christophel, E., Hii, J., Delacollette, C., Bakyaita, N., and Aregawi, M. (2007). Estimating Trends in the Burden of Malaria at Country Level. *The American Journal of Tropical Medicine and Hygiene*, *77*(6 Suppl), 133–137.

Clements, A., Lwambo, N., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi, S., Webster, J., Fenwick, A., and Brooker, S. (2006). Bayesian Spatial Analysis and Disease Mapping: Tools to Enhance Planning and Implementation of a Schistosomiasis Control Programme in Tanzania. *Tropical Medicine & International Health*, *11*(4), 490–503.

Concato, J., Feinstein, A., and Holford, T. (1993). The Risk of Determining Risk with Multivariable Models. *Annals of Internal Medicine*, *118*(3), 201–210.

Conger, A. (1974). A Revised Definition for Suppressor Variables: A Guide to their Identification and Interpretation. *Educational and Psychological Measurement*, *34*(1), 35–46.

Cox, D. and Hinkley, D. (1979). *Theoretical Statistics*. CRC Press.

Craig, M., Kleinschmidt, I., Nawn, J., Le Sueur, D., and Sharp, B. (2004). Exploring 30 Years of Malaria Case Data in KwaZulu-Natal, South Africa: Part I. The Impact of Climatic Factors. *Tropical Medicine & International Health*, *9*(12), 1247–1257.

Craig, M., Sharp, B., Mabaso, M., and Kleinschmidt, I. (2007). Developing a Spatial-statistical Model and Map of Historical Malaria Prevalence in Botswana using a Staged Variable Selection Procedure. *International Journal of Health Geographics*, *6*(1), 44–59.

Craig, M., Snow, R., and le Sueur, D. (1999). A Climate-based Distribution Model of Malaria Transmission in Sub-Saharan Africa. *Parasitology Today*, *15*(3), 105–111.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley and Sons.

Daash, A., Srivastava, A., Nagpal, B., Saxena, R., Gupta, S., *et al.* (2009). Geographical Information System (GIS) in Decision Support to Control Malaria: A Case Study of Koraput District in Orissa, India. *Journal of Vector Borne Diseases*, *46*, 72–74.

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics*. Springer.

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based Geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 299–350.

Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models* (Third ed.). Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC.

Doll, R. (1980). The Epidemiology of Cancer. *Cancer*, *45*(10), 2475–2485.

Dormann, F., Carsten, McPherson, J., Araújo, M., Bivand, R., Bolliger, J., Carl, G., Davies, R., Hirzel, A., Jetz, W., Kissling, D., *et al.* (2007). Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review. *Ecography*, *30*(5), 609–628.

Doss, H. and Hobert, J. (2010). Estimation of Bayes Factors in a Class of Hierarchical Random Effects Models Using a Geometrically Ergodic MCMC Algorithm. *Journal of Computational and Graphical Statistics*, *19*(2), 295–312.

Everitt, B. (2002). *The Cambridge Dictionary of Statistics* (Third ed.). Cambridge University Press, Cambridge, UK.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications: Volume One*. John Wiley and Sons.

Finley, A. and Banerjee, S. (2013). *spBayes: Univariate and Multivariate Spatial-temporal Modeling*. R package version 0.3-7.

Finley, A., Banerjee, S., and Carlin, B. (2007). spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models. *Journal of Statistical Software*, *19*(4), 1–24.

Finley, A., Banerjee, S., and McRoberts, R. (2008). A Bayesian Approach to Multi-source Forest Area Estimation. *Environmental and Ecological Statistics*, *15*(2), 241–258.

Gamerman, D. and Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press.

Gazetteer, G. (2006). Surface Water Body Features. http://www.fao.org/geonetwork/srv/en/main.home. Accessed: 2013-08-13.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2014). *Bayesian Data Analysis* (Third ed.), Volume 2. Taylor & Francis.

Gemperli, A. (2003). *Development of Spatial Statistical Methods for Modelling Point-referenced Spatial Data in Malaria Epidemiology*. Ph. D. thesis, University of Basel.

Gemperli, A., Vounatsou, P., Sogoba, N., and Smith, T. (2006). Malaria Mapping Using Transmission Models: Application to Survey Data From Mali. *American Journal of Epidemiology*, *163*(3), 289–297.

Genton, M. (2002). Classes of Kernels for Machine Learning: A Statistics Perspective. *The Journal of Machine Learning Research*, *2*, 299–312.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

Gosoniu, L. (2008). *Development of Bayesian Geostatistical Models with Applications in Malaria Epidemiology*. Ph. D. thesis, University of Basel.

Gosoniu, L., Veta, A., and Vounatsou, P. (2010). Bayesian Geostatistical Modeling of Malaria Indicator Survey data in Angola. *PloS one*, *5*(3), e9322.

Gosoniu, L., Vounatsou, P., Sogoba, N., and Smith, T. (2006). Bayesian Modelling of Geostatistical Malaria Risk Data. *Geospatial Health*, *1*(1), 127–139.

Gotway, C. and Stroup, W. (1997). A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(2), 157–178.

Gotway, C. and Young, L. (2002). Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, *97*(458), 632–648.

Greenberg, J. and Mattiuzzi, M. (2014). *gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities*. R package version 0.3.1.

Handcock, M. and Stein, M. (1993). A Bayesian Analysis of Kriging. *Technometrics*, *35*(4), 403–410.

Haran, M. (2011). Gaussian Random Field Models for Spatial Data. In S. Brooks, A. Gelman, G. Jones, and X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Chapter 18, 449–478. Chapman & Hall/CRC.

Hay, S., Snow, R., and Rogers, D. (1998). From Predicting Mosquito Habitat to Malaria Seasons Using Remotely Sensed Data: Practice, Problems and Perspectives. *Parasitology Today*, *14*(8), 306–313.

Hengl, T. (2009). A Practical Guide to Geostatistical Mapping. Technical report, University of Amsterdam, Amsterdam.

Hijmans, R. (2013). *raster: Geographic Data Analysis and Modeling*. R package version 2.1-49.

Hijmans, R., Cemeron, S., Parra, J., Jones, P., and Jarvis, A. (2005). Very High Resolution Interpolated Climate Surfaces for Global Land Areas. http://www.worldclim.org/tiles.php. Accessed: 2014-07-15.

Howes, R., Piel, F., Patil, A., Nyangiri, O., Gething, P., Dewi, M., Hogg, M., Battle, K., Padilla, C., Baird, J., *et al.* (2012). G6PD Deficiency Prevalence and Estimates of Affected Populations in Malaria Endemic Countries: A Geostatistical Model-based Map. *PLoS Medicine*, *9*(11), e1001339.

Huisman, O. and Rolf, A. (2009). *Principles of Geographic Information Systems*. ITC Press, Enschede, The Netherlands.

Hyndman, R. and Koehler, A. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

James, D. and DebRoy, S. (2012). *RMySQL: R Interface to the MySQL Database*. R package version 0.9-3.

Jha, S., Dwivedi, A., and Tiwari, A. (2011). Necessity of Goodness of Fit Tests in Research and Development. *International Journal of Computer Science and Technology*, *2*, 135–141.

Kazembe, L., Kleinschmidt, I., Holtz, T., and Sharp, B. (2006). Spatial Analysis and Mapping of Malaria Risk in Malawi Using Point-Referenced Prevalence of Infection Data. *International Journal of Health Geographics*, *5*(1), 41–49.

Kleinschmidt, I. (2001). *Spatial Statistical Analysis, Modelling and Mapping of Malaria in Africa*. Ph. D. thesis, University of Basel.

Kleinschmidt, I., Bagayoko, M., Clarke, G., Craig, M., and Le Sueur, D. (2000). A Spatial Statistical Approach to Malaria Mapping. *International Journal of Epidemiology*, *29*(2), 355–361.

Kleinschmidt, I., Omumbo, J., Briet, O., Van De Giesen, N., Sogoba, N., Mensah, N., Windmeijer, P., Moussa, M., and Teuscher, T. (2001). An Empirical Malaria Distribution Map for West Africa. *Tropical Medicine & International Health*, *6*(10), 779–786.

Kleinschmidt, I., Sharp, B., Clarke, G., Curtis, B., and Fraser, C. (2001). Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-area Malaria Incidence Rates in KwaZulu Natal, South Africa. *American Journal of Epidemiology*, *153*(12), 1213–1221.

Kleinschmidt, I., Sharp, B., Mueller, I., and Vounatsou, P. (2002). Rise in Malaria Incidence Rates in South Africa: A Small-Area Spatial Analysis of Variation in Time Trends. *American Journal of Epidemiology*, *155*(3), 257–264.

Krige, D. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, *52*, 119–139.

Kulkarni, M., Desrochers, R., and Kerr, J. (2010). High Resolution Niche Models of Malaria Vectors in Northern Tanzania: A New Capacity to Predict Malaria Risk? *PLoS One*, *5*(2), e9396.

Larget, B. and Simon, D. (1999). Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Molecular Biology and Evolution*, *16*, 750–759.

Lawson, A. (2013). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, Volume 32. CRC Press.

Le Sueur, D., Binka, F., Lengeler, C., De Savigny, D., Snow, B., Teuscher, T., and Toure, Y. (1997). An Atlas of Malaria in Africa. *Africa Health*, *19*(2), 23–24.

Lee, Y. and Nelder, J. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 619–678.

Li, H. (2008). *Bayesian Hierarchical Models for Spatial Count Data with Application to Fire Frequency in British Columbia*. Ph. D. thesis, University of Victoria.

Loève, M. (1955). *Probability Theory. Foundations. Random Sequences*. New York: D. Van Nostrand Company.

Lowry, J. (2004). WGS - AGD - GDA: Selecting the Correct Datum, Coordinate System and Projection for North Australian Applications. Technical Report 473, Department of the Environment and Heritage, Australian Government.

Mabaso, M., Craig, M., Vounatsou, P., and Smith, T. (2005). Towards Empirical Description of Malaria Seasonality in Southern Africa: The Example of Zimbabwe. *Tropical Medicine & International Health*, *10*(9), 909–918.

Martin, C., Curtis, B., Fraser, C., and Sharp, B. (2002). The Use of a GIS-based Malaria Information System for Malaria Research and Control in South Africa. *Health & Place*, *8*(4), 227–236.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology*, *58*(8), 1246–1266.

Matise, T., Perlin, M., and Chakravarti, A. (1994). Automated Construction of Genetic Linkage Maps Using an Expert System (MultiMap): A Human Genome Linkage Map. *Nature Genetics*, *6*(4), 384–390.

Mohebbi, M., Wolfe, R., and Jolley, D. (2011). A Poisson Regression Approach for Modelling Spatial Autocorrelation between Geographically Referenced Observations. *BMC Medical Research Methodology*, *11*(1), 133–143.

Molineaux, L., Wernsdorfer, W., McGregor, I., *et al.* (1988). The Epidemiology of Human Malaria as an Explanation of its Distribution, Including Some Implications for its Control. *Malaria: Principles and Practice of Malariology*, *2*, 913–998.

Myers, D. (1989). To Be or Not to Be... Stationary? That Is the Question. *Mathematical Geology*, *21*(3), 347–362.

MySQL Community Server (2011). http://dev.mysql.com/downloads/mysql. Accessed: 2014-06-20.

Naimi, B. (2014). ModisDownload: An R Function to Download, Mosaic, and Reproject the MODIS Images. http://r-gis.net/?q=ModisDownload. Accessed: 2014-07-15.

NASA Land Processes Distributed Active Archive Center (2001). MOD13A3. https://lpdaac.usgs.gov/products/modis Accessed: 2014-07-15.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*, 370–384.

Noor, A., Gething, P., Alegana, V., Patil, A., Hay, S., Muchiri, E., Juma, E., and Snow, R. (2009). The Risks of Malaria Infection in Kenya in 2009. *BMC Infectious Diseases*, *9*, 180–193.

Noor, A., Kinyoki, D., Mundia, C., Kabaria, C., Mutua, J., Alegana, V., Fall, I., and Snow, R. (2014). The Changing Risk of Plasmodium falciparum Malaria Infection in Africa: 2000–10: A Spatial and Temporal Analysis of Transmission Intensity. *The Lancet*, *383*(9930), 1739–1747.

Omumbo, J., Hay, S., Goetz, S., Snow, R., and Rogers, D. (2002). Updating Historical Maps of Malaria Transmission Intensity in East Africa Using Remote Sensing. *PE & RS-Photogrammetric Engineering & Remote Sensing*, *68*(2), 161–166.

Open Geospatial Consortium (1994). Making Location Count. http://www.opengeospatial.org. Accessed: 2014-03-17.

Patil, A., Gething, P., Piel, F., and Hay, S. (2011). Bayesian Geostatistics in Health Cartography: The Perspective of Malaria. *Trends in Parasitology*, *27*(6), 246–253.

Pebesma, E. and Bivand, R. (2005). Classes and Methods for Spatial Data in R. *R News*, *5*(2), 9–13.

Pinsky, M. and Karlin, S. (2010). *An Introduction to Stochastic Modeling*. Academic Press.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA Convergence Diagnosis and Output Analysis for MCMC. *R news*, *6*(1), 7–11.

Qu, J., Gao, W., Kafatos, M., Murphy, R., and Salomonson, V. (2006). *Earth Science Satellite Remote Sensing*, Volume 2. Springer.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raso, G., Schur, N., Utzinger, J., Koudou, B., Tchicaya, E., Rohner, F., N'Goran, E., Silue, K., Matthys, B., Assi, S., Tanner, M., and Vounatsou, P. (2012). Mapping Malaria Risk Among Children in Cote d'Ivoire Using Bayesian Geostatistical Models. *Malaria Journal*, *11*(1), 160–170.

Raudenbush, S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Volume 1. Sage.

Ribeiro, J., Paulo, J., and Diggle, P. (2001). geoR: A Package for Geostatistical Analysis. *R News*, *1*(2), 14–18.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004). Interpreting Posterior Relative Risk Estimates in Disease-mapping Studies. *Environmental Health Perspectives*, *112*, 1016–1025.

Riedl, T., Singer, A., and Choi, J. (2010). Learning in Gaussian Markov Random Fields. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference*, 3070–3073. IEEE.

Roberts, G. and Rosenthal, J. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367.

Scholte, R., Gosoniu, L., Malone, J., Chammartin, F., Utzinger, J., and Vounatsou, P. (2014). Predictive Risk Mapping of Schistosomiasis in Brazil Using Bayesian Geostatistical Models. *Acta Tropica*, *132*, 57–63.

Seltman, H. (2012). Experimental Design and Analysis. Technical report, Carnegie Mellon University, Pittsburgh.

Serfozo, R. (2009). *Basics of Applied Stochastic Processes*. Springer.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Swanson, A., Dobrowski, S., Finley, A., Thorne, J., and Schwartz, M. (2013). Spatial Regression Methods Capture Prediction Uncertainty in Species Distribution Model Projections Through Time. *Global Ecology and Biogeography*, *22*, 242–251.

Tanser, F., Sharp, B., and Le Sueur, D. (2003). Potential Effect of Climate Change on Malaria Transmission in Africa. *The Lancet*, *362*(9398), 1792–1798.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240.

Valle, D. Clark, J. Z. K. (2011). Enhanced Understanding of Infectious Diseases by Fusing Multiple Datasets: A Case Study on Malaria in the Western Brazilian Amazon Region. *PloS one*, *6*(11), e27462.

Ver Hoef, J. and Cressie, N. (2001). Spatial Statistics: Analysis of Field Experiments. In S. M. Scheiner and J. Gurevitch (Eds.), *Design and Analysis of Ecological Experiments*, Chapter 15, 289–307. New York: Oxford University Press.

Winkler, R. (1967). The Assessment of Prior Distributions in Bayesian Analysis. *Journal of the American Statistical Association*, *62*(319), 776–800.

World Health Organization (2014). Malaria Fact Sheet Number 94. http://www.who.int/mediacentre/factsheets/fs094/en/. Accessed: 2014-08-28.

Zacarias, O. and Andersson, M. (2010). Mapping Malaria Incidence Distribution that Accounts for Environmental Factors in Maputo Province-Mozambique. *Malaria Journal*, *9*(1), 79–88.

Zacarias, O. and Andersson, M. (2011). Spatial and Temporal Patterns of Malaria Incidence in Mozambique. *Malaria Journal*, *10*(189), 23–32.

Zayeri, F., Salehi, M., and Pirhosseini, H. (2011). Geographical Mapping and Bayesian Spatial Modeling of Malaria Incidence in Sistan and Baluchistan Province, Iran. *Asian Pacific Journal of Tropical Medicine*, *4*(12), 985–992.

Zeng, Z., Lei, L., Guo, L., Zhang, L., and Zhang, B. (2013). Incorporating Temporal Variability to Improve Geostatistical Analysis of Satellite-observed CO2 in China. *Chinese Science Bulletin*, *58*(16), 1948–1954.

Appendix A: R Code

```r

#
# Compliling of Spatial Database #
#

# set workspace
setwd("/Volumes/JUSTJUBBA/Spatial")

# check if package is installed, install if not load otherwise
packageInstallLoad <- function(x){
  for( i in x ){
    #  require returns TRUE invisibly if it was able to load package
    if( ! require( i , character.only = TRUE ) ){
      #  If package was not able to be loaded then re-install
      install.packages( i , dependencies = TRUE )
      #  Load package after installing
      require( i , character.only = TRUE )
    }
  }
}

# packages needed for spatial db comilation and non-spatial model building
packageInstallLoad(c("raster", "RMySQL", "sp", "rgdal", "gdalUtils", "maptools
    ", "maps", "plyr", "stats", "glm2", "bootStepAIC", "texreg", "xtable", "
    tables", "data.table", "diagram", "caret", "bootStepAIC"))

# packages needed for spatial analysis
packageInstallLoad(c("spBayes", "MBA", "fields", "raster", "coda", "fields"))

# connect to MySQL
con <- dbConnect(MySQL(),
        user     = "root",
        password = "hons123",
        dbname   = "MySql",
        host     = "localhost")
```

```
34
35 # delete table if exists
36 dbSendQuery(con, 'DROP TABLE IF EXISTS botsTable;' )  #decimal(9,6)
37
38 # create table
39 # MARA Botswana dataset
40 dbSendQuery(con, 'CREATE TABLE botsTable (Lat             float(10,8),
41                                           Lon              float(10,8),
42                                           Start_Mnth       int,
43                                           Start_Yr         int,
44                                           End_Mnth         int,
45                                           End_Yr           int,
46                                           AgeGroup_Lower   int,
47                                           AgeGroup_Upper   int,
48                                           Numb_Positive    int,
49                                           Numb_Examined    int);')
50
51 # import data from csv file into newly created table - save csv where root
       user has default access
52 dbSendQuery(con, 'LOAD DATA LOCAL INFILE "/usr/local/BotsMARA.csv"
53                                 INTO TABLE botsTable
54                       FIELDS TERMINATED BY ","
55                        LINES TERMINATED BY "\n"
56                            IGNORE 1 LINES (Lat,
57                                            Lon,
58                                            Start_Mnth,
59                                            Start_Yr,
60                                            End_Mnth,
61                                            End_Yr,
62                                            AgeGroup_Lower,
63                                            AgeGroup_Upper,
64                                            Numb_Positive,
65                                            Numb_Examined );')
66
67 # select all childern between 1 and 15
68 rs <-  dbSendQuery(con, 'SELECT *
69                            FROM botsTable
70                           WHERE AgeGroup_Upper <= 15')
71
72
73 # retrieve data from MySQL
74 sitesMultiple <- fetch(rs, n = 6000)
75
76 # clear previous MySQL transaction from memory
77 dbClearResult(rs)
78
```

```
79 # close connection to MySQL
80 dbDisconnect(con)
81
82
83 # clean up data #
84
85 # average over multiple sites to get unique sets
86 sitesDf <- ddply(sitesMultiple, .(Lon, Lat), summarise, Month = round(mean(
       Start_Mnth)), Pos = round(mean(Numb_Positive)), Examined = round(mean(Numb
       _Examined)))
87
88
89 sitesDf <- ddply(sitesMultiple, .(Lon, Lat), summarise, Month = round(mean(
       Start_Mnth)), Pos = round(mean(Numb_Positive)), Examined = round(mean(Numb
       _Examined)))
90
91 # extract coords
92 coordsSitesDf <- sitesDf[,c("Lon","Lat")]
93
94 # remove any duplicates
95 if (any(duplicated(coordsSitesDf)))
96   sitesDf <- sitesDf[!duplicated(coordsSitesDf),]
97
98 # remove obs where 0 ppl examined
99 sitesDf <- subset(sitesDf, sitesDf$Examined > 0)
100
101 # extract coords from clean data
102 coordsSitesDf <- as.matrix(sitesDf[,c("Lon","Lat")])
103
104 # Perform calculations to find distance to closest surface water body and find
       where these are #
105
106 # read shapefile into SpatialPointsDataFrame
107 surfWaterSpdf <- readShapePoints("Surface_water/gns_swb/gns_swb")
108
109 # convert SPDF to DF
110 surfWaterDf <- as.data.frame(surfWaterSpdf)
111
112 # 129 sites
113 locs <- SpatialPoints(sitesDf[,1:2],  proj4string=CRS("+proj=longlat +datum=
       WGS84"))
114
115 # 46591 water sources
116 src  <- SpatialPoints(surfWaterDf[,1:2],  proj4string=CRS("+proj=longlat +
       datum=WGS84"))
117
```

```
#Â distances of 46576 water sources per site
distances <- lapply(1:length(locs), function(i) spDistsN1(src, locs[i],
    longlat=TRUE))

# get min distance per site in km
sitesDf$DstTClW <- sapply(distances, min)

# get index of min distance sites to find their coords
minPos <- sapply(distances, which.min)

sitesDf$LonWater <- surfWaterDf[minPos,1]
sitesDf$LatWater <- surfWaterDf[minPos,2]

# get map boundary of Botswana explore the sample site data #

# get boundary shape for Botswana
data(wrld_simpl)
botswana <- wrld_simpl[wrld_simpl$NAME == "Botswana",]

# extract coords
coordsSitesDf <- sitesDf[,c("Lon","Lat")]

# convert DF to SPDF
sitesSpdf <- SpatialPointsDataFrame(coordsSitesDf, sitesDf)

# assign projection to SPDF
projection(sitesSpdf)   <- projection(botswana)

# only keep sample sites that are in Botswana
sitesSpdf <- sitesSpdf[botswana,]  # 122 obs

# plot sample sites representing proportion of malaria cases out of no.
    examined
# over the maximum of this ratio by the size of the circle
plot(botswana)
plot(sitesSpdf, add = T, asp = 1, cex = 4 * sitesSpdf$Pos/sitesSpdf$Examined/
    max(sitesSpdf$Pos/sitesSpdf$Examined), pch = 1)

# Notes: There seems to be some association btw big circles and other big ones
#        that are close together. Informal test showing that circles close
    together exhibit similiar malaria intensitie although.
#        Patterns of attribute seem not to be random.


# WorldClim climate layers & MODIS NDVI raster processing #

```

```
# get and process modis data .hdf files
#Â download all available monthly images for the years 2000 - 2013. MODIS PRODUCT: MOD13A3, Terra, Vegetation Indices, Tile, 1000m, Monthly
ModisDownload(x="MOD13A3",h=c(21,22),v=c(8,9),dates=c('2000.01.01','2013.12.31'),mosaic=F,proj=F)

load("ModisLP.RData")
source("ModisDownload.R")

# set wd to where .hdf files live
setwd("/Volumes/JUSTJUBBA/Spatial/NDVI_MODIS_Botswana")

# put each .hdf file in list
out.files <- list.files(getwd(), pattern="hdf$", full.names = F)

# get list of subdatasets from .hdf files, choose subdataset 1 - mean monthly NDVI
sdsList <- sapply(X= out.files, FUN = function(out.files){get_subdatasets(out.files)[1]})

# see which raster data has more rows/columns between MODIS & WorldClim- must resample to grid with smallest no. rows/cols
 gdalwarp(srcfile= sdsList[1], t_srs="+proj=longlat +datum=WGS84 +no_defs", dstfile = "/Volumes/JUSTJUBBA/Spatial/NDVIBotsTest.tiff", te = c(bbox(botswana)[1],bbox(botswana)[2],bbox(botswana)[3],bbox(botswana)[4]))

NDVIRast <- raster("NDVIBotsTest.tiff")
worldClimRast <- crop(raster("WorldClim_botswana/tmean1_36.tif"), extent(botswana))

# compare resolution
ncell(NDVIRast); ncell(worldClimRast)

# Notes: NDVI has less cells so re-align WorldClim rasters to that of NDVI rasters
#       Use NDVI as the model raster to get correct dimensions for WorldClim and NDVI processing

# get sub dataset for each month
# then remove the null entries to get length of list

janList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("001", "002")){return(x)}})
janList<-janList[!sapply(janList, is.null)]

```

```
194 febList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("032", "033
    "))){return(x)}})
195 febList<-febList[!sapply(febList, is.null)]
196
197 marList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("061", "062
    "))){return(x)}})
198 marList<-marList[!sapply(marList, is.null)]
199
200 aprList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("091", "092
    "))){return(x)}})
201 aprList<-aprList[!sapply(aprList, is.null)]
202
203 mayList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("122", "123
    "))){return(x)}})
204 mayList<-mayList[!sapply(mayList, is.null)]
205
206 junList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("152", "153
    "))){return(x)}})
207 junList<-junList[!sapply(junList, is.null)]
208
209 julList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("182", "183
    "))){return(x)}})
210 julList <-julList[!sapply(julList, is.null)]
211
212 augList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("213", "214
    "))){return(x)}})
213 augList <-augList[!sapply(augList, is.null)]
214
215 sepList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("244", "245
    "))){return(x)}})
216 sepList <-sepList[!sapply(sepList, is.null)]
217
218 octList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("274", "275
    "))){return(x)}})
219 octList <-octList[!sapply(octList, is.null)]
220
221 novList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("305", "306
    "))){return(x)}})
222 novList <-novList[!sapply(novList, is.null)]
223
224 decList <- llply(sdsList, function(x){if (substr(x, 32, 34) %in% c("335", "336
    "))){return(x)}})
225 decList <-decList[!sapply(decList, is.null)]
226 # put above lists into another list for processing
227 NDVIMonths <- list(janList, febList, marList, aprList, mayList, junList,
    julList, augList, sepList, octList, novList, decList)
```

```
228
229 # loop through each month, looping again through all the years (2000 - 2013)
      of data reprojecting, merging tiles and converting to tiff
230 # format for each available month and year in one step
231
232 for (j in 1 : 12){
233   i <- 0
234   while (i < 14){
235    if (i < 10 & length(NDVIMonths[[j]][grep(paste0("A200", "", i), NDVIMonths
         [[j]])]) > 0){
236     gdalwarp(srcfile= NDVIMonths[[j]][grep(paste0("A200", "", i), NDVIMonths[[
          j]])], t_srs="+proj=longlat +datum=WGS84 +no_defs", dstfile = paste0("
          NDVI", j, "200", i, ".tiff"), te = c(bbox(NDVIRast)[1],bbox(NDVIRast)
          [2],bbox(NDVIRast)[3],bbox(NDVIRast)[4]), tr = c(xres(NDVIRast), yres(
          NDVIRast)))
237     i <- i + 1
238    }
239    else if (i < 10){
240     print(paste0("no data available for month ", j,  " and year 200", i))
241     i <- i + 1
242    }
243    else if (i > 9 & length(NDVIMonths[[j]][grep(paste0("A20", "", i),
         NDVIMonths[[j]])]) > 0){
244     gdalwarp(srcfile= NDVIMonths[[j]][grep(paste0("A20", "", i), NDVIMonths[[j
          ]])], t_srs="+proj=longlat +datum=WGS84 +no_defs", dstfile = paste0("
          NDVI", j, "20", i, ".tiff"),te = c(bbox(NDVIRast)[1],bbox(NDVIRast)
          [2],bbox(NDVIRast)[3],bbox(NDVIRast)[4]), tr = c(xres(NDVIRast), yres(
          NDVIRast)))
245     i <- i + 1
246    }
247    else if (i > 9 & i < 14){
248     print(paste0("no data available for month ", j,  " and year 20", i))
249     i <- i + 1
250    }
251  }
252 }
253
254 # (1) intialise empty stack for each climate and environmental factor
255 # (2) for each climate stack except NDVI merge monthly tiles, do appropriate
      raster calculations and crop to extent of model NDVI stack then add each
      monthly raster layer to the stack in a loop (no
256 # loop needed for altitude)
257 # (3) for the NDVI montly layers apply NDVIRasterFunction to each monthly tiff
       image for each year, function involves calculating the mean NDVI value
      for each month across all years and cropping to
258 # same extent as botswana using NDVI as model raster
```

```
259
260 # Notes: temperature layers must be divided by 10 and NDVI layers must be
        divided by 10000
261
262 # put all WorldClim images in a list for processing
263 climFilelist <- list.files("/Volumes/JUSTJUBBA/Spatial/WorldClim_Botswana",
        pattern="tif$", full.names=T)
264
265 # initialize all climatic and environmental stacks
266 meanTempStack <- stack()
267 maxTempStack  <- stack()
268 minTempStack  <- stack()
269 rainStack     <- stack()
270 bioStack      <- stack()
271 NDVIStack     <- stack()
272
273 NDVIRasterFunction<- function(y){
274   x <- stack(list.files("/Volumes/JUSTJUBBA/Spatial/NDVI_MODIS_Botswana/",
          pattern = y, full.names=T))
275   x <- calc(x, function(z) z*0.0001)
276   x <- mean(x, na.rm = T)
277   x <- crop(x, extent(NDVIRast))
278   return(x)
279 }
280
281 for (i in 1:12){
282   NDVIStack <- stack(NDVIStack, NDVIRasterFunction(paste0("NDVI", i, 20)))
283 }
284
285 for (i in 1 : 12){
286   meanTempStack <- stack(meanTempStack, crop(raster(climFilelist[grep(paste0("
          tmean", i, "_36"), climFilelist)])/10, extent(NDVIRast)))
287
288   minTempStack <- stack(minTempStack, crop(raster(climFilelist[grep(paste0("
          tmin", i, "_36"), climFilelist)])/10, extent(NDVIRast)))
289
290   maxTempStack <- stack(maxTempStack, crop(raster(climFilelist[grep(paste0("
          tmax", i, "_36"), climFilelist)])/10, extent(NDVIRast)))
291
292   rainStack <- stack(rainStack, crop(raster(climFilelist[grep(paste0("prec", i
          , "_36"), climFilelist)]), extent(NDVIRast)))
293 }
294
295 for (i in 1 : 19){
296   if (i < 12){
```

```
297     bioStack <- stack(bioStack, crop(raster(climFilelist[grep(paste0("bio", i,
            "_36"), climFilelist)])/10, extent(NDVIRast)))
298   }
299   else{
300     bioStack <- stack(bioStack, crop(raster(climFilelist[grep(paste0("bio", i,
            "_36"), climFilelist)]), extent(NDVIRast)))
301   }
302 }
303 altitudeLayer <-  crop(raster("/Volumes/JUSTJUBBA/Spatial/WorldClim_Botswana/
        alt_36.tif"), extent(NDVIRast))
304
305
306 # resample WolrdClim layers to that of model NDVI layers
307 meanTempStack <- resample(meanTempStack, NDVIStack)
308 minTempStack <- resample(minTempStack, NDVIStack)
309 maxTempStack <- resample(maxTempStack, NDVIStack)
310 rainStack <- resample(rainStack, NDVIStack)
311 bioStack  <- resample(bioStack, NDVIStack)
312 altitudeLayer <- resample(altitudeLayer, NDVIStack)
313
314 #Â caclulate min annual temp
315 minAnnualTempLayer <- (minTempStack[[1]] + minTempStack[[2]] + minTempStack
        [[3]] + minTempStack[[4]] + minTempStack[[5]] + minTempStack[[6]] +
        minTempStack[[7]] + minTempStack[[8]] + minTempStack[[9]] + minTempStack
        [[10]] + minTempStack[[11]] + minTempStack[[12]])/12
316
317 # name each layer in stack indicating appropriate month
318 names(NDVIStack)          <- rep(paste0("NDVI", 1:12))
319 names(rainStack)          <- rep(paste0("rain", 1:12))
320 names(meanTempStack)      <- rep(paste0("meanTemp", 1:12))
321 names(maxTempStack)       <- rep(paste0("maxTemp", 1:12))
322 names(minTempStack)       <- rep(paste0("minTemp", 1:12))
323 names(bioStack)           <- rep(paste0("bio", 1:19))
324 names(minAnnualTempLayer) <- "minAnnualTemp"
325 names(altitudeLayer)      <- "altitude"
326
327
328 # create multilayered spatial database - last step #
329
330 # write SpatialPointsDataFrame values to polygon that are inside Botswana
331 writeOGR(sitesSpdf, dsn = "getwd()", layer = 'sitesBots', driver = 'ESRI
        Shapefile', overwrite = T)
332
333 # read polygon as a new SPDF object
334 sitesPolyPoints = readOGR("getwd()", "sitesBots")
335
```

```
# add coords onto @data component of SPDF
sitesPolyPoints@data = cbind(sitesPolyPoints@data, sitesPolyPoints@coords)

# extract raster values at matching coords and add to @data component for each
    climate layer in stack
# then give layers column names and append them to the SPDF

rainLayers           <- raster:::extract(rainStack, as(sitesPolyPoints, "
    SpatialPoints"))
colnames(rainLayers) <- rep(paste0("rain", 1:12))
sitesPolyPoints@data <- cbind(sitesPolyPoints@data, rainLayers)

meanTempLayers           <- raster:::extract(meanTempStack, as(sitesPolyPoints
    , "SpatialPoints"))
colnames(meanTempLayers) <- rep(paste0("meanTemp", 1:12))
sitesPolyPoints@data     <- cbind(sitesPolyPoints@data, meanTempLayers)

maxTempLayers           <- raster:::extract(maxTempStack, as(sitesPolyPoints,
    "SpatialPoints"))
colnames(maxTempLayers) <- rep(paste0("maxTemp", 1:12))
sitesPolyPoints@data     <- cbind(sitesPolyPoints@data, maxTempLayers)

minTempLayers           <- raster:::extract(minTempStack, as(sitesPolyPoints,
    "SpatialPoints"))
colnames(minTempLayers) <- rep(paste0("minTemp", 1:12))
sitesPolyPoints@data     <- cbind(sitesPolyPoints@data, minTempLayers)

NDVILayers           <- raster:::extract(NDVIStack, as(sitesPolyPoints, "
    SpatialPoints"))
colnames(NDVILayers) <- rep(paste0("NDVI", 1:12))
sitesPolyPoints@data <- cbind(sitesPolyPoints@data, NDVILayers)

bioLayers           <- raster:::extract(bioStack, as(sitesPolyPoints, "
    SpatialPoints"))
colnames(bioLayers)  <- rep(paste0("bio", 1:19))
sitesPolyPoints@data <- cbind(sitesPolyPoints@data, bioLayers)

minAnnualTemp        <- raster:::extract(minAnnualTempLayer, as(
    sitesPolyPoints, "SpatialPoints"))
sitesPolyPoints@data <- cbind(sitesPolyPoints@data, minAnnualTemp)

altitude            <- raster:::extract(altitudeLayer,as(sitesPolyPoints, "
    SpatialPoints") )
sitesPolyPoints@data <- cbind(sitesPolyPoints@data, altitude)


```

```
# assign SPDF to DF for caclulation convenience
s.p <- as.data.frame(sitesPolyPoints)

# subset meanTemp variables
s.pTemp <- subset(s.p, select = rep(paste0("meanTemp", 1:12)))
s.pRain <- subset(s.p, select = rep(paste0("rain", 1:12)))
s.pNDVI <- subset(s.p, select = rep(paste0("NDVI", 1:12)))

# caclulate annual std deviation of monthly variables
s.p$SDTemp <- apply(s.pTemp, 1, sd)
s.p$SDRain <- apply(s.pRain, 1, sd)
s.p$SDNDVI <- apply(s.pNDVI, 1, sd)

s.p$totTemp <- apply(s.pTemp, 1, sum)
s.p$totRain <- apply(s.pRain, 1, sum)
s.p$totNDVI <- apply(s.pNDVI, 1, sum)

# summer (months 12, 1, 2, 3) and winter caclulations (4,5,6,7,8,9,10)

s.p$summerTemp <- apply(s.p[c("meanTemp12", "meanTemp1", "meanTemp2", "
    meanTemp3")], 1, mean)
s.p$winterTemp <- apply(s.p[c("meanTemp4", "meanTemp5", "meanTemp6", "
    meanTemp7", "meanTemp8", "meanTemp9","meanTemp10")], 1, mean)
s.p$summerRain <- apply(s.p[c("rain12", "rain1", "rain2", "rain3")], 1, mean)
s.p$winterRain <- apply(s.p[c("rain4", "rain5", "rain6", "rain7", "meanTemp8",
     "rain9","rain10")], 1, mean)
s.p$summerNDVI <- apply(s.p[c("NDVI12", "NDVI1", "NDVI2", "NDVI3")], 1, mean)
s.p$winterNDVI <- apply(s.p[c("NDVI4", "NDVI5", "NDVI6", "NDVI7", "NDVI8", "
    NDVI9","NDVI10")], 1, mean)

meanTempCols  <- rep(paste0("meanTemp", 1:12))
rainCols <- rep(paste0("rain", 1:12))
NDVICols <- rep(paste0("NDVI", 1:12))

# rainfall concentration index and mean peak month around which rainfall is
    concentrated calculations #

angle <- rep(paste0("angle", 1:12))
s.pRain[angle] <- NA

for(i in 1:12){
  s.pRain[angle][i] <- (i*2*pi)/12
}

r1 <- rep(paste0("r1", 1:12))
r2 <- rep(paste0("r2", 1:12))
```

```
414 s.pRain[r1] <- NA
415 s.pRain[r2] <- NA
416
417 for(i in 1:12){
418   s.pRain[r1][i] <- s.pRain[rainCols][i]*cos(s.pRain[angle][i])
419   s.pRain[r2][i] <- s.pRain[rainCols][i]*sin(s.pRain[angle][i])
420 }
421
422 s.pRain$r <- sqrt((apply(s.pRain[r1], 1, sum))^2 + (apply(s.pRain[r2], 1, sum)
      )^2)
423
424 # concentration index
425 s.p$rCIndex <- (100*s.pRain$r)/s.p$totRain
426
427 # mean peak month around which rainfall is concentrated
428 s.p$q <- atan(apply(s.pRain[r2], 1, sum)/apply(s.pRain[r1], 1, sum))
429
430 # save workspace and write DF containing spatial database to table for later
      use
431 save.image('spatial_Bots.RData')
432 write.table(s.p, "/Volumes/JUSTJUBBA/Spatial/s.pBots")
433
434 # create prediction grid covering Botswana
435
436 # Notes: The same code principles apply as with the compilation of the spatial
       database above except at a lower resolution.
437 #        Instead of extracting values at sample points all of the raster data
      is used so
438 #        that an environmental or climatic value is present for each grid cell
      . All raster images
439 #        are stored in stacks and the same process as above is used to get a
      corresponding attribute for each cell.
440 #        Therefore this code is omitted.
441
442 # a prediction grid prepared using similar raster calculations as above
443 # covering Botswana at a 20km resolution was written to a table called
      gridPredBots20km:
444
445 write.table(grid20, "/Volumes/JUSTJUBBA/Spatial/gridPredBots20km")
446
447
448
449 #
450 # Model Building: Non-spatial analysis #
451 #
452
```

```
453 # create a df with only the variables of interest
454 spatialVars = s.p[which(colnames(s.p) %in% c(c("Pos", "Examined",  "DstTClW",
        "NDVI", "altitude", "SDTemp", "SDNDVI", "SDRain", "q", "rCIndex", "
        summerTemp", "winterTemp", "summerNDVI", "winterNDVI", "summerRain", "
        winterRain", "totRain"), rep(paste0("bio", 1:19))))]
455
456 # randomly partition data keeping 85% for derivation data set
457 derivIndex <- createDataPartition(spatialVars$Pos, p = .85, list = F, times =
        1)
458
459 # create validation and derivation
460 spatialVarsDeriv <- spatialVars[ derivIndex,]
461 spatialVarsTest  <- spatialVars[-derivIndex,]
462
463 #Â standardise variables
464 spatialVars[,3:36] <- scale(spatialVars[,3:36])
465
466 # 34 explanatory variables at start
467
468 # Univariate Logistic Regression - Model Building
469
470 AICList <-lapply(c(c( "DstTClW", "NDVI", "altitude", "SDTemp", "SDNDVI", "
        SDRain", "q", "rCIndex", "summerTemp", "winterTemp", "summerNDVI", "
        winterNDVI", "summerRain", "winterRain", "totRain"  ), rep(paste0("bio",
        1:19))),
471   function(var){
472    formula           <- as.formula(paste("Pos/Examined ~", var))
473    nonSpatialUniVar <- glm(formula, data = spatialVarsDeriv, weights =
           Examined, family = binomial)
474    cbind(summary(nonSpatialUniVar)$aic, exp(summary(nonSpatialUniVar)$coef[, "
           Estimate"]), summary(nonSpatialUniVar)$coef[, "Pr(>|z|)"])
475 })
476 # http://rstudio-pubs-static.s3.amazonaws.com/2989_
        ceae90d128554c728d5388439adf0661.html access: 28 Feb
477
478
479 # put list into matrix
480 m <- matrix(unlist(AICList), ncol=6, byrow=TRUE)
481
482 # delete 1st, 3rd, 5th column (intercept details not required)
483 m <- m[,-c(1  , 3, 5)]
484
485 # make col and row names nameable
486 dimnames(m) <- list(rownames(m, do.NULL = FALSE, prefix = "row"), colnames(m,
        do.NULL = FALSE, prefix = "col"))
487
```

```
# name columns
colnames(m) <- c("AIC", "OR", "Pr(>|z|)")

# name rows
rownames(m) <- c(c("DstTClW", "NDVI", "altitude", "SDTemp", "SDNDVI", "SDRain"
    , "q", "rCIndex", "summerTemp", "winterTemp", "summerNDVI", "winterNDVI",
    "summerRain", "winterRain", "totRain"  ), rep(paste0("bio", 1:19)))

# rank matrix from lowest AIC to highest
AICMatrix <- m[order(m[,1]),]

# preserve the order of univariate  AIC rankings in columns of df
colTestCols             <- paste0(rownames(AICMatrix)[1:nrow(AICMatrix)])
spatialVarsDerivColTest <- as.data.frame(spatialVarsDeriv[colTestCols])

# get column names for temp and rain related vars
tempThemeCols <- paste0(c("bio1", "bio2", "bio3", "bio4", "bio5", "bio6", "
    bio7", "bio8", "bio9", "bio10", "bio11", "summerTemp", "winterTemp", "
    SDTemp"))
rainThemeCols <- paste0(c("bio12", "bio13", "bio14", "bio15", "bio16", "bio17"
    , "bio18", "bio19", "totRain", "summerRain", "winterRain", "SDRain", "q",
    "rCIndex"))
NDVIThemCols <- paste0(c("NDVI", "summerNDVI", "winterNDVI", "SDNDVI")

# to preserve the order of columns by AIC rank remove all non-theme related
    vars from original df leaving only AIC ordered themed vars
tempTheme <- spatialVarsDerivColTest[-which(colnames(spatialVarsDerivColTest)
    %in%c(c(rainThemeCols), c("NDVI", "summerNDVI", "winterNDVI", "SDNDVI", "
    DstTClW", "altitude")))]
rainTheme <- spatialVarsDerivColTest[-which(colnames(spatialVarsDerivColTest)
    %in% c(c(tempThemeCols), c("NDVI", "summerNDVI", "winterNDVI", "SDNDVI", "
    DstTClW", "altitude")))]
NDVITheme <- spatialVarsDerivColTest[-which(colnames(spatialVarsDerivColTest)
    %in% c(c(tempThemeCols), c(rainThemeCols), c("DstTClW", "altitude")))]

### STAGE 2 ###

#Â function checks for multicolinearity in each theme
#Â set exact = F for appropriate asmyptotic methods to handle presence of ties
corrFunctionTheme <- function(varX, varDf){

 # create matrix to store multicolinearity test results per variable
 tst <- matrix(data = NA, nrow = ncol(varDf), ncol = 4)

 # name columns
```

```
521  dimnames(tst) <- list(rownames(tst, do.NULL = FALSE, prefix = "row"),
         colnames(tst, do.NULL = FALSE, prefix = "col"))
522  colnames(tst) <- c("Upper p", "Lower p", "Two.sided p", "rho")
523
524    for(i in 1:ncol(varDf)){
525      tst[i,1] = cor.test(varX, varDf[,i], method = "spearm", alternative = "g",
           exact = F)$p.value
526      tst[i,2] = cor.test(varX, varDf[,i], method = "spearm", alternative = "l",
           exact = F)$p.value
527      tst[i,3] = cor.test(varX, varDf[,i], method = "spearm", alternative = "two
         .sided", exact = F)$p.value
528      tst[i,4] = cor.test(varX, varDf[,i], method = "spearm", alternative = "two
         .sided", exact = F)$estimate
529      rownames(tst)[i] = names(varDf)[i]
530    }
531    return(tst)
532  }
533
534  # test multicollinearity among all variables and rank from lowest to highest
       according to Spearman's r and check if r > 0.85
535  # criteria for excluding a variable: keep variable with lowest AIC from
       univariate analysis in the presence of collinearity
536  # starting with lowest ranked AIC variable in each theme
537  # so that everything correlated to it will have a higher AIC and can be
       removed
538  # at each round the variable tested, with lowest AIC, is put into a list
539
540  # initialize lists
541  tempThemeKept     <- list()
542  tempThemeRemoved  <- list()
543  rainThemeKept     <- list()
544  rainThemeRemoved  <- list()
545  NDVIThemeKept     <- list()
546  NDVIThemeRemoved  <- list()
547
548  # runs until one var remains in themed df
549
550  # Temperature theme
551
552  while (ncol(tempTheme) > 1){
553
554      # perform correlation test
555      corrTemp         <- corrFunctionTheme(varX = tempTheme[,1] , varDf =
           tempTheme )
556      corrTempRanked   <- corrTemp[order(abs(corrTemp[,4])),]
```

```
557     corrTempVars     <- corrTempRanked[which(abs(corrTempRanked [,4]) > 0.85),
            arr.ind = TRUE]
558
559     # whether or not correlated tested var always kept - lowest AIC
560     tempThemeKept[length(tempThemeKept)+1] <-  list(colnames(tempTheme)[1])
561
562     # add lowest ranking AIC var to list
563
564     # else if statement needed becasue row names disappear when only 1 row
            remains in matrix
565
566     if (length(nrow(corrTempVars)) < 1) {
567         tempThemeRemoved[length(tempThemeRemoved)+1] <- list(NA)
568         tempTheme                                    <- tempTheme[-which(
                colnames(tempTheme) %in% colnames(tempTheme)[1])]
569     } else{
570         if (nrow(corrTempVars) == 2) {
571             ind = which(rownames(corrTempVars) !=  colnames(tempTheme)[1])
572             tempThemeRemoved[length(tempThemeRemoved)+1] <- rownames(
                    corrTempVars)[ind]
573             tempTheme <- tempTheme[-which(colnames(tempTheme) %in% rownames(
                    corrTempVars))]
574         }  else{
575             tempThemeRemoved[length(tempThemeRemoved)+1] <- list(rownames(
                    corrTempVars[-which(rownames(corrTempVars) %in% colnames(
                    tempTheme)[1]) ,]))
576             tempTheme <- tempTheme[-which(colnames(tempTheme) %in% rownames(
                    corrTempVars))]
577         }
578     }
579
580 }
581
582
583 # Rain theme
584
585 while (ncol(rainTheme) > 1){
586
587     # perform correlation test
588     corrRain         <- corrFunctionTheme(varX = rainTheme[,1] , varDf =
            rainTheme )
589     corrRainRanked   <- corrRain[order(abs(corrRain[,4])),]
590     corrRainVars     <- corrRainRanked[which(abs(corrRainRanked [,4]) > 0.85),
            arr.ind = TRUE]
591
592     # whether or not correlated tested var always kept - lowest AIC
```

```
593     rainThemeKept[length(rainThemeKept)+1] <-  list(colnames(rainTheme)[1])
594
595     # add lowest ranking AIC var to list
596
597     # else if statement needed because row names disappear when only 1 row
            remains in matrix
598
599     if (length(nrow(corrRainVars)) < 1) {
600         rainThemeRemoved[length(rainThemeRemoved)+1] <- list(NA)
601         rainTheme                                    <- rainTheme[-which(
                colnames(rainTheme) %in% colnames(rainTheme)[1])]
602     } else{
603         if (nrow(corrRainVars) == 2) {
604             ind = which(rownames(corrRainVars) !=  colnames(rainTheme)[1])
605             rainThemeRemoved[length(rainThemeRemoved)+1] <- rownames(
                    corrRainVars)[ind]
606             rainTheme <- rainTheme[-which(colnames(rainTheme) %in% rownames(
                    corrRainVars))]
607         } else{
608             rainThemeRemoved[length(rainThemeRemoved)+1] <- list(rownames(
                    corrRainVars[-which(rownames(corrRainVars) %in% colnames(
                    rainTheme)[1]),]))
609             rainTheme <- rainTheme[-which(colnames(rainTheme) %in% rownames(
                    corrRainVars))]
610         }
611     }
612
613 }
614
615
616 # NDVI theme
617
618 while (ncol(NDVITheme) > 1){
619
620     # perform correlation test
621     corrNDVI         <- corrFunctionTheme(varX = NDVITheme[,1] , varDf =
            NDVITheme )
622     corrNDVIRanked   <- corrNDVI[order(abs(corrNDVI[,4])),]
623     corrNDVIVars     <- corrNDVIRanked[which(abs(corrNDVIRanked [,4]) > 0.85),
            arr.ind = TRUE]
624
625     # whether or not correlated tested var always kept - lowest AIC
626     NDVIThemeKept[length(NDVIThemeKept)+1] <-  list(colnames(NDVITheme)[1])
627
628     # add lowest ranking AIC var to list
629
```

```
    # else if statement needed because row names disappear when only 1 row
        remains in matrix

    if (length(nrow(corrNDVIVars)) < 1) {
        NDVIThemeRemoved[length(NDVIThemeRemoved)+1] <- list(NA)
        NDVITheme                                    <- NDVITheme[-which(
            colnames(NDVITheme) %in% colnames(NDVITheme)[1])]
    } else{
        if (nrow(corrNDVIVars) == 2) {
            ind = which(rownames(corrNDVIVars) != colnames(NDVITheme)[1])
            NDVIThemeRemoved[length(NDVIThemeRemoved)+1] <- rownames(
                corrNDVIVars)[ind]
            NDVITheme <- NDVITheme[-which(colnames(NDVITheme) %in% rownames(
                corrNDVIVars))]
        } else{
            NDVIThemeRemoved[length(NDVIThemeRemoved)+1] <- list(rownames(
                corrNDVIVars[-which(rownames(corrNDVIVars) %in% colnames(
                NDVITheme)[1]),]))
            NDVITheme <- NDVITheme[-which(colnames(NDVITheme) %in% rownames(
                corrNDVIVars))]
        }
    }

}

# left with tempThemeKept, rainThemeKept, NDVIThemeKept, DstTClW, altitude


### STAGE 3 ###

spatialVarsDerivDf = spatialVarsDeriv[which(colnames(spatialVarsDeriv) %in% c
    (c(tempThemeKept, rainThemeKept, NDVIThemeKept), c("DstTClW", "altitude",
    "Pos", "Examined")))]

fit.Stage3 <- glm(Pos/Examined ~ bio9+SDTemp+bio5+summerTemp+bio3+bio7+rCIndex
    +bio13+bio19+bio14+totRain+bio18+summerNDVI+DstTClW+altitude, weights =
    Examined, data = spatialVarsDerivDf, family="binomial")

# this will yield a basic candidate model (Candidate List: )
bootGLM.Stage3 <- boot.stepAIC(fit.Stage3, spatialVarsDerivDf, direction = "
    backward", alpha = 0.05, B = 1000)

fitTest.1     <- glm(Pos/Examined ~ bio9      , weights = Examined, data =
    spatialVarsDerivDf, family="binomial")
summary(fitTest.1)

```

```
fitTest.2      <- glm(Pos/Examined ~ bio9 + altitude           , weights =
    Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.2)


fitTest.3      <- glm(Pos/Examined ~ bio9 + altitude + bio5          , weights =
     Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.3)

fitTest.4      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7  , weights =
    Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.4)

fitTest.5      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+summerTemp,
    weights = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.5) # exclude summerTemp


fitTest.6      <- glm(Pos/Examined ~  bio9+altitude+bio5+bio7+summerNDVI,
    weights = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.6) # exclude summerNDVI


fitTest.7      <- glm(Pos/Examined ~  bio9+altitude+bio5+bio7+SDTemp , weights
    = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.7) # exclude SDTemp


fitTest.8      <- glm(Pos/Examined ~  bio9+altitude+bio5+bio7+DstTClW, weights
    = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.8)

fitTest.9      <- glm(Pos/Examined ~  bio9+altitude+bio5+bio7+DstTClW+bio18,
    weights = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.10)

fitTest.10      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
    rCIndex , weights = Examined, data = spatialVarsDerivDf, family="binomial"
    )
summary(fitTest.11) # exclude rCIndex

fitTest.11      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
    bio3, weights = Examined, data = spatialVarsDerivDf, family="binomial")
summary(fitTest.12) # exclude bio3

```

```
697 fitTest.12      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
        totRain, weights = Examined, data = spatialVarsDerivDf, family="binomial")
698 summary(fitTest.13)
699
700 fitTest.13      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
        totRain+bio13, weights = Examined, data = spatialVarsDerivDf, family="
        binomial")
701 summary(fitTest.14) # exclude bio13
702
703
704 fitTest.14      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
        totRain+bio19, weights = Examined, data = spatialVarsDerivDf, family="
        binomial")
705 summary(fitTest.15) # exclude bio19
706
707
708 fitTest.15      <- glm(Pos/Examined ~ bio9+altitude+bio5+bio7+DstTClW+bio18+
        totRain+bio14, weights = Examined, data = spatialVarsDerivDf, family="
        binomial")
709 summary(fitTest.16) # exclude bio14
710
711 # Stage 4: bio9+altitude+bio5+bio7+DstTClW+bio18+totRain
712
713
714 ### STAGE 5 ###
715
716 # further tests based on enviro themes of previously excluded variables
717 #Â criteria: bring variables that were exlcuded (high correlation + AIC
        ranking) into candidate model (consider bootGLM + manual stepwise) and re-
        assess based on
718 # frequency of selection in bootsrapped samples (and if variable was not
        excluded by stepAIC() algorithm)
719
720 # kept bio9 ahead of bio6+bio1+winterTemp+bio11
721
722 fit.1 <- glm(Pos/Examined ~bio9+altitude+bio5+bio7+DstTClW+bio18+totRain+bio6+
        bio1+winterTemp+bio11, weights = Examined, data = spatialVarsDeriv, family
        = "binomial")
723
724 bootGLM.1 <- boot.stepAIC(fit.1, spatialVarsDeriv, direction = "backward",
        alpha = 0.05, B = 1000) # keep winterTemp instead of bio9
725
726 # kept bio5 ahead of bio10
727 fit.2 <- glm(Pos/Examined ~winterTemp+altitude+bio5+bio7+DstTClW+bio18+totRain
        +bio10, weights = Examined, data = spatialVarsDeriv, family = "binomial")
728
```

```
729 bootGLM.2 <- boot.stepAIC(fit.2, spatialVarsDeriv, direction = "backward",
        alpha = 0.05, B = 1000) # keep bio5
730
731 # kept bio7 ahead of bio2
732 fit.3 <- glm(Pos/Examined ~winterTemp+altitude+bio5+bio7+DstTCIW+bio18+totRain
        +bio2, weights = Examined, data = spatialVarsDeriv, family = "binomial")
733
734 bootGLM.3 <- boot.stepAIC(fit.3, spatialVarsDeriv, direction = "backward",
        alpha = 0.05, B = 1000) # keep bio7
735
736 # kept totRain ahead of bio12
737 fit.4 <- glm(Pos/Examined ~winterTemp+altitude+bio5+bio7+DstTCIW+bio18+totRain
        +bio12, weights = Examined, data = spatialVarsDeriv, family = "binomial")
738
739 bootGLM.4 <- boot.stepAIC(fit.4, spatialVarsDeriv, direction = "backward",
        alpha = 0.05, B = 1000) # keep totRain
740
741 fit.Stage5 <- glm(Pos/Examined ~winterTemp+altitude+bio5+bio7+DstTCIW+bio18+
        totRain, weights = Examined, data = spatialVarsDeriv, family = "binomial")
742
743 bootGLM.Stage5 <- boot.stepAIC(fit.Stage5, spatialVarsDeriv, direction = "
        backward", alpha = 0.05, B = 1000)
744
745 # End Stage 5 model:  winterTemp+altitude+bio5+bio7+DstTCIW+bio18+totRain
746
747 ## end non-spatial model building
748
749
750 #
751 # Spatial Analysis
752 #
753
754 # using full dataset keep only Stage 5 variables
755 dfFull = s.p[which(colnames(s.p) %in% c(c("Pos", "Examined", "Lon", "Lat", "
        winterTemp","altitude","bio5","bio7","DstTCIW","bio18","totRain")))]
756
757 # using derivation dataset
758 dfFull = s.p[derivIndex,][which(colnames(s.p[derivIndex,]) %in% c(c("Pos", "
        Examined", "Lon", "Lat", "winterTemp","altitude","bio5","bio7","DstTCIW",
        "bio18","totRain")))]
759
760 # the following code demonstrates the spatial model for the full dataset the
        same workings apply when using only the derivation data
761
762 # transform DF to a SPDF
763 coordinates(dfFull) <- cbind("Lon", "Lat")
```

```
764
765 # standardise data
766 dfFull@data[,3:9] <-  scale(dfFull@data[,3:9])
767
768 # run a non-spatial GLM to obtain starting values for the MH step
769 dfFull = s.p[derivIndex,][which(colnames(s.p[derivIndex,]) %in% c(c("Pos", "
    Examined", "Lon", "Lat", "winterTemp","altitude","bio5","bio7","DstTCldW",
    "bio18","totRain")))]
770
771 #starting values from non-spatial analysis for beta
772 fit <- glm(Pos/Examined ~ winterTemp+altitude+bio5+bio7+DstTCldW+bio18+totRain,
     weights = Examined, data = dfFull , family=binomial("logit"))
773 beta.starting <- coefficients(fit)
774
775 # use the variance covariance matrix as the proposal (tuning) distribution for
     the MH step
776 beta.tuning <- t(chol(vcov(fit)))
777
778 # get maximum Euclidean distance between sites
779 d.max <- max(iDist(dfFull@coords))
780
781 # this defines the number of simulations to be run in batches each of certain
    length as well the burn in period
782 n.batch <- 3500
783 batch.length <- 100
784 n.samples <- n.batch*batch.length
785 burn.in <- 0.8*n.samples
786
787 # 3 spatial GLMMs are run as follows
788 # notes: effective range of the spatial weights is controlled by phi and is
    roughly 3/phi
789 #        posterior inference is based on three MCMC chains each of length 350
    000
790
791
792 m1.Full <- spGLM(Pos ~ winterTemp+altitude+bio5+bio7+DstTCldW+bio18+totRain,
793                  data = dfFull,
794                  weights= dfFull@data$Examined, family = "binomial", coords =
                       dfFull@coords,
795                  starting = list("beta" = beta.starting, "phi" = 3/(0.5*d.max
                      ), "sigma.sq" = 1,   "w" = 0),
796                  tuning = list("beta" = beta.tuning,   "phi" = 0.06,  "sigma.
                      sq" = 0.5, "w" = 0.5),
797                  priors   = list(phi.Unif = c(3/d.max, 3/(0.1*d.max)), "sigma
                      .sq.IG" = c(2,1)),
```

```
                    amcmc = list(n.batch = n.batch, batch.length = batch.length,
                         accept.rate = 0.43),
                    cov.model = "exponential", verbose = T, n.report = 1)

m2.Full <- spGLM(Pos ~ winterTemp+altitude+bio5+bio7+DstTClW+bio18+totRain,
                    data = dfFull,
                    weights= dfFull@data$Examined, family = "binomial", coords =
                         dfFull@coords,
                    starting = list("beta" = beta.starting, "phi" = 3/(0.5*d.max
                        ), "sigma.sq" = 1,   "w" = 0),
                    tuning = list("beta" = beta.tuning,   "phi" = 0.06,  "sigma.
                        sq" = 0.5, "w" = 0.5),
                    priors   = list(phi.Unif = c(3/d.max, 3/(0.1*d.max)), "sigma
                        .sq.IG" = c(2,1)),
                    amcmc = list(n.batch = n.batch, batch.length = batch.length,
                         accept.rate = 0.43),
                    cov.model = "exponential", verbose = T, n.report = 1)

m3.Full <- spGLM(Pos ~ winterTemp+altitude+bio5+bio7+DstTClW+bio18+totRain,
                    data = dfFull,
                    weights= dfFull@data$Examined, family = "binomial", coords =
                         dfFull@coords,
                    starting = list("beta" = beta.starting, "phi" = 3/(0.5*d.max
                        ), "sigma.sq" = 1,   "w" = 0),
                    tuning = list("beta" = beta.tuning,   "phi" = 0.06,  "sigma.
                        sq" = 0.5, "w" = 0.5),
                    priors   = list(phi.Unif = c(3/d.max, 3/(0.1*d.max)), "sigma
                        .sq.IG" = c(2,1)),
                    amcmc = list(n.batch = n.batch, batch.length = batch.length,
                         accept.rate = 0.43),
                    cov.model = "exponential", verbose = T, n.report = 1)

# consolodate posteriors to perform convergence diagnostics on the fit of the
    spGLM MCMC chains
posteriors <- as.mcmc.list(list(m1.Full$p.beta.theta.samples,m2.Full$p.beta.
    theta.samples, m3.Full$p.beta.theta.samples))

# compute Gelman diagnostics to assess convergence.  These compare within-
    chain to
# between-chain variability. Values near 1 suggest full convergence.
print(gelman.diag(posteriors))


#Â define samples taken after burn in
sub.samps <- burn.in:n.samples

```

```
# credibility intervals for each simulated regression and variance parameter estimated
quantile(m1.Full$p.beta.theta.samples[sub.samps,1], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,2], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,3], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,4], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,5], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,6], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,7], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,8], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,9], prob=c(0.025, 0.975))
quantile(m1.Full$p.beta.theta.samples[sub.samps,10], prob=c(0.025, 0.975))

# odds ratio calculated for each simulated paramater's mean and credibility intervals
# calculated on odds ratio scale via exponentiation of paramater's coefficent

# Intercept odds ratio
intercept <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,1])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,1], prob=c(0.025, 0.975))))

# winterTemp odds ratio
winterTemp <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,2])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,2], prob=c(0.025, 0.975))))

# altitude odds ratio
altitude <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,3])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,3], prob=c(0.025, 0.975))))

# bio5 odds ratio
bio5 <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,4])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,4], prob=c(0.025, 0.975))))

# bio7 odds ratio
bio7 <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,5])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,5], prob=c(0.025, 0.975))))

# DstTClW odds ratio
DstTClW <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,6])), exp(quantile(m1.Full$p.beta.theta.samples[sub.samps,6], prob=c(0.025, 0.975))))

# bio18 odds ratio
```

```
bio18 <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,7])), exp(quantile(
    m1.Full$p.beta.theta.samples[sub.samps,7], prob=c(0.025, 0.975))))

# totRain odds ratio
totRain <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,8])), exp(
    quantile(m1.Full$p.beta.theta.samples[sub.samps,8], prob=c(0.025, 0.975)))
    )

# sigma.sq odds ratio
sigma.sq <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,9])), exp(
    quantile(m1.Full$p.beta.theta.samples[sub.samps,9], prob=c(0.025, 0.975)))
    )

# phi odds ratio
phi <- c(exp(mean(m1.Full$p.beta.theta.samples[sub.samps,10])), exp(quantile(
    m1.Full$p.beta.theta.samples[sub.samps,10], prob=c(0.025, 0.975))))


# spatial prediction #

# prepare grid for prediction of gridded sites across Botswana
# 20 km resolution

# read in prediction grid
grid20 <- read.table("Volumes/JUSTJUBBA/Spatial/gridPredBots20km.txt")

# keep only relevant explanatory variables
pred.grid20 <- grid20[which(colnames(grid20) %in% c("Lon", "Lat", "winterTemp"
    ,"altitude","bio5","bio7","DstTClW","bio18","totRain"))]

# convert DF to gridded SPDF and set correct projection
coordinates(pred.grid20) <- cbind("Lon", "Lat")
gridded(pred.grid20) <- TRUE
proj4string(pred.grid20) <- "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84
    +towgs84=0,0,0"

pred.coords20 <- pred.grid20@coords
pred.covars20 <- scale(pred.grid20@data)

#Â run prediction command spPredict in order to get a prediction of malaria
    risk at each cell for all MCMC samples using one spGLM chain
m1.pred.Full.20 <- spPredict(m1.Full, pred.coords = pred.coords20, pred.covars
    = as.matrix(cbind(1,pred.covars20)),start=burn.in)

# mean and standard deviation of prediction probability at each cell
y.pred.grid.prob.mu <- apply(m1.pred.Full.20$p.y.predictive.samples,1,mean)
```

```
y.pred.grid.prob.sd <- apply(m1.pred.Full.20$p.y.predictive.samples,1, sd)

# plot predicted mean probability of malaria risk at 20km resolution
# requires mba.surf to yield interpolation surfaces at 100x100 resolution on the x and y axis
res <- 100
surf <- mba.surf(cbind(pred.grid20@coords, y.pred.grid.prob.mu), res, res, extend=TRUE, sp=TRUE)$xyz.est
plot(botswana)
image.plot(as.image.SpatialGridDataFrame(surf), asp=1.25, add = T)
plot(dfFull, add = T, pch = 20, cex = 0.3, col = 'black')
plot(botswana, add= T)
title(main="Predicted mean probability of malaria risk- 20km grid")

# plot predicted standard deviation of malaria risk at 20km resolution
res <- 100
surf <- mba.surf(cbind(pred.grid20@coords, y.pred.grid.prob.sd), res, res, extend=TRUE, sp=TRUE)$xyz.est
plot(botswana)
image.plot(as.image.SpatialGridDataFrame(surf), asp=1.25, add = T)
plot(dfFull, add = T, pch = 20, cex = 0.3, col = 'black')
plot(botswana, add= T)
title(main="Predicted standard deviation of malaria risk- 20km grid")


# Cross-validation calculations #

# Predicted vs observed prevelance at validation sites
# Notes: For this section of code the derivation dataset is used, i.e. dfDeriv. The same spatial code as above applies but inseatd of using the full dataset
#       dfDeriv was used yielding 3 spatial chains given by: m1.Deriv, m2.Deriv, m3.Deriv.
#       The code showing the spatial modelling for the derivation data is ommitted.

# non-spatial prediction accuracy

# derivation and validation subsets of the data
dfDeriv = s.p[derivIndex,][which(colnames(s.p[derivIndex,]) %in% c(c("Pos", "Examined", "Lon", "Lat", "winterTemp","altitude","bio5","bio7","DstTClW","bio18","totRain")))]
dfValid = s.p[-derivIndex,][which(colnames(s.p[derivIndex,]) %in% c(c("Pos", "Examined", "Lon", "Lat", "winterTemp","altitude","bio5","bio7","DstTClW","bio18","totRain")))]

```

```
# non-spatial glm fit
fit.Deriv <- glm(Pos/Examined ~ winterTemp+altitude+bio5+bio7+DstTClW+bio18+
    totRain, weights = Examined, data = dfDeriv, family = "binomial")

#Â test prediction ability of non-spatial model by applying fitted model (
    deriv data) to validation sample (predicted probabilities)
predValid <- predict(fit.Deriv, dfValid, type="response", se = T)

# calculate mean absolute error (MAE) and mean error (ME) between observed and
     predicted using non-spatial model at validation sites on probability
    scale
obs <- dfValid$Pos/dfValid$Examined
pred <- exp(predValid$fit)/(1+exp(predValid$fit))
ME.nonSpatial <- mean(obs-pred)
MAE.nonSpatial <- mean(abs(obs-pred))

# spatial prediction accuracy

# n x 2 matrix of n prediction location coordinates
pred.coords <- dfValid@coords

# An n x q design matrix or data frame containing the covariates associated
    with pred.coords
pred.covars <- dfValid@data[3:9]

# holds the posterior predictive samples given model output m1.Deriv from
    spGLM function
m1.pred.valid <- spPredict(m1.Deriv, pred.coords = pred.coords, pred.covars =
    as.matrix(cbind(1,pred.covars)),start=burn.in)

# mean predicted probability of malaria at each site in validation subset
y.pred.valid.prob.mu <- apply(m1.pred.valid$p.y.predictive.samples,1,mean)

# calculate ME and MAE between observed and predicted using spatial model at
    validation sites
obs <- dfValid$Pos/dfValid$Examined
ME.spatial <- mean(obs-y.pred.valid.prob.mu)
MAE.spatial <- mean(abs(obs-y.pred.valid.prob.mu))
```

thesisCode.R

Appendix B: Variable Calculations

Standard Deviation (SD)

The SD of an explanatory variable was calculated as follows:

$$\text{SD} = \sqrt{\sum_{m=1}^{12} (\hat{\text{y}} - \mathbf{y}_m)^2}$$

where $\mathbf{y}_m$ = the monthly value and $\hat{\mathbf{y}}$ = mean of all $\mathbf{y}_m$.

Mean Peak Month Around Which Rainfall is Concentrated (q):

Monthly rainfall is expressed as a vector $(\mathbf{r}_m, \boldsymbol{\theta}_m)$ where $\mathbf{r}_m$ is the magnitute of rainfall and and $\boldsymbol{\theta}_m$ represents its the angle expressed in arc units for months $m = \{1, \ldots, 12\}$:

$$\boldsymbol{\theta}_m = \frac{m2\pi}{12}.$$

The twelve monthly vectors are added the total vector $(\mathbf{r}_t, \boldsymbol{\theta}_t)$:

$$\mathbf{r}_t = \sqrt{(\sum_{m=1}^{12} \mathbf{r}_m \text{cos}\boldsymbol{\theta}_m)^2 + (\sum_{m=1}^{12} \mathbf{r}_m \text{sin}\boldsymbol{\theta}_m)^2}. \tag{1}$$

The mean peak month around which rainfall is concentrated (q) is then given by:

$$\boldsymbol{\theta}_t = \tan^{-1}\left(\frac{\sum_{m=1}^{12} \mathbf{r}_m \text{sin}\boldsymbol{\theta}_m}{\sum_{m=1}^{12} \mathbf{r}_m \text{cos}\boldsymbol{\theta}_m}\right).$$

Rainfall Concentration (rCIndex)

Using Equation 1 the rainfall concentration index (rCIndex) is calculated as:

$$\text{rCIndex} = \frac{100\mathbf{r}_t}{\text{annual total rainfall}}.$$