

---

---

# *Model-based geostatistics for global public health using R*

*Emanuele Giorgi  
Claudio Fronterre*

---

# **Contents**

---

<b>Preface</b>	<b>v</b>
<b>Author</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objectives of this book . . . . .	1
1.2 Pre-requisites for using this book . . . . .	2
1.2.1 Topics in probability . . . . .	2
1.2.2 Topics in statistics . . . . .	2
1.2.3 Topics in R programming . . . . .	3
1.3 Obtaining and running the R packages . . . . .	3
1.4 Data-sets used in the book . . . . .	4
1.4.1 Lead concentration in Galicia . . . . .	4
1.4.2 River-blindness in Liberia . . . . .	5
1.4.3 Malaria in the Western Kenyan Highlands . . . . .	6
1.4.4 <i>Anopheles gambiae</i> mosquitoes in Southern Cameroon .	7
1.4.5 Simulated-dataset . . . . .	7
1.5 Geostatistical problems and geostatistical models . . . . .	9
1.5.1 The Matern family of correlation functions . . . . .	12
1.6 Workflow of a statistical analysis and structure of the book .	15
<b>2 Handling of spatial data in R</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.2 Spatial data handling in geostatistical analysis . . . . .	17
2.3 Chapter overview . . . . .	18
2.4 Accessing covariates for disease mapping . . . . .	19
2.4.1 Example: Downloading administrative boundaries . . .	20
2.4.2 Example: Download population data . . . . .	21
2.5 Importing and standardizing spatial data . . . . .	23
2.5.1 Importing vector data . . . . .	23
2.5.2 Importing raster data . . . . .	23
2.5.3 Understanding coordinate reference systems (CRS) .	25
2.5.4 EPSG codes . . . . .	26
2.5.5 Convert a data frame to an <code>sf</code> object . . . . .	27
2.5.6 Working with CRSs in R . . . . .	28
2.5.7 Common Issues and Considerations . . . . .	32

2.6	Extracting covariate data . . . . .	32
2.6.1	Extracting covariates from raster layers . . . . .	34
2.7	Creating a predictive grid . . . . .	37
2.8	Visualizing spatial data . . . . .	38
2.8.1	Visualizing point data . . . . .	38
2.8.2	Visualizing polygon data . . . . .	40
2.8.3	Visualizing raster data . . . . .	41
2.8.4	Combining Multiple Spatial Data Types . . . . .	42
<b>3</b>	<b>Model formulation and parameter estimation</b>	<b>45</b>
	List of the main functions used in the chapter . . . . .	45
3.1	Exploratory analysis . . . . .	45
3.1.1	Exploring associations with risk factors using count data	45
3.1.2	Exploring overdispersion in count data . . . . .	59
3.1.3	Exploring residual spatial correlation . . . . .	66
3.2	The linear geostatistical model . . . . .	78
3.2.1	Evaluating the inclusion of the measurement error term $U_i$ and the specification of the smoothness parameter $\kappa$	78
3.2.2	Modelling hierarchical geostatistical data using the <code>re()</code> function . . . . .	84
3.3	Generalized linear geostatistical models . . . . .	88
3.3.1	Example: river-blindness in Liberia . . . . .	90
3.3.2	Example: <i>Anopheles Gambiae</i> mosquitoes in Cameroon . . . . .	92
3.4	Theory . . . . .	94
3.4.1	Maximum likelihood estimation for non-Gaussian mixed models with independent random effects . . . . .	94
3.4.2	Maximum likelihood estimation for non-Gaussian generalized linear geostatistical models . . . . .	95
3.5	Exercises . . . . .	106
<b>4</b>	<b>Model validation</b>	<b>109</b>
4.1	How to simulate geostatistical data from a fitted model . . . . .	109
4.2	Validating the calibration of the model . . . . .	109
4.3	Validating the spatial correlation of the model . . . . .	109
<b>5</b>	<b>Geostatistical prediction</b>	<b>111</b>
5.1	Pixel-level predictive targets . . . . .	111
5.2	Area-level predictive targets . . . . .	111
5.3	Comparing the predictive performance of geostatistical models	111
<b>6</b>	<b>Case studies</b>	<b>113</b>
6.1	Mapping stunting risk in Ghana . . . . .	113
6.2	Mapping river blindness in Malawi . . . . .	113
6.3	Mapping mosquitoes abundance in Cameroon . . . . .	113
	<b>References</b>	<b>115</b>

<b>Appendix</b>	<b>119</b>
An appendix section . . . . .	119



---

## ***Preface***

---

Its companion book “Model-based geostatistical for global public health” by Peter J. Diggle (2019) is a strongly recommended complementary read, as you work your way through this book.



---

## ***Author***

---

An author bio.

# 1

---

## *Introduction*

---

The book shows how to carry out model-based geostatistical analysis of public health data using the **RiskMap** R package. In this introductory chapter, we explain what are the pre-requisites for using this book and its learning objectives. We also explain what software should be installed and how. Finally, we give a brief overview of the class of models covered in this book, and the examples that will be used to illustrate the methods and use of software.

---

### **1.1 Objectives of this book**

The overall aim of this book is to provide you with the skills to perform a geostatistical analysis of a data-set using the R software environment. As you work your way through the book, you will learn to:

- explore geostatistical data-sets using graphical procedures and summary statistics;
- formulate and fit geostatistical models using the maximum likelihood estimation method;
- carry out prediction of health outcomes at different spatial scales;
- visualize and interpret the results from geostatistical models;
- model the relationships between spatially referenced risk factors and the health outcome of interest;
- validate the assumptions of geostatistical models and assess their predictive performance.

Although the focus of this book is on public health, the statistical ideas, as well as the software used, can also be applied for the analysis of geostatistical data-sets arising from other scientific fields.

---

## 1.2 Pre-requisites for using this book

To effectively understand and use the material presented in this book, it is expected that you should possess prior knowledge of basic probability theory, foundational topics in statistical modelling and R programming. Below we provide a more detailed explanation of the pre-requisites for each of these three fields.

### 1.2.1 Topics in probability

Basics probability theory is important to fully understand the content of this book. In particular, you should have knowledge of: the general definition and properties of continuous and discrete distribution; how to describe the properties of probability distributions through their mean, variance and skewness; the concepts of stochastic dependence and correlation; the distinction between marginal and conditional distributions; the basic properties of the Gaussian, Binomial and Poisson distributions; the definition and properties of the multivariate Gaussian distribution. The reader can find an extensive explanation and illustrations with examples of all these topics in Ross (2013).

### 1.2.2 Topics in statistics

Likelihood-based inference (whether frequentist or Bayesian) provides the theoretical bedrock for the estimation of almost any statistical model. In this book will focus on maximum likelihood estimation methods of inference. Extensive use of the notions of point and interval estimates obtained using the maximum likelihood estimation methods will be made through the book. Recommended readings include chapters 1, 2 and 4 of Pawitan (2001).

Good prior knowledge of Generalized linear models (GLMs) is essential, as the geostatistical modelling framework builds on these as an extension. Before embarking on the use of this book, we thus encourage you to review the basic theory of GLMs and, in particular, how these are applied and interpreted. In this book, we will cover examples that will model continuously measured outcomes and counts. Hence, good understanding of linear regression modelling and modelling of counts data using Binomial and Poisson regression should be the main focus of the review. For comprehensive overview of GLMs and their implementation in R, we refer you to Dobson and Barnett (2008).

### 1.2.3 Topics in R programming

Although this book does not require to possess advanced skills in R programming, it is important you have good knowledge in the following topics: creation and manipulation of vectors and matrices; logical vectors; character vectors; handling of lists and data frame objects; reading data into R; graphical procedures. A very large amount of freely available material covering these topics can be found online. Our recommendation is to start from the manual “An introduction to R” of the Comprehensive R Archive Network available at this link, available at R manual.

---

## 1.3 Obtaining and running the R packages

It is advised that you obtain the latest 64-bit version of R in order to run the R code of this book. To install R, go to the R website, where you can download the installer packages for Windows and Mac, and find instructions for Linux, using binary files.

- Windows
- Mac
- Linux

The list of the R packages used in this book is provided in Table 1.1.

Table 1.1: List of the R packages that will be used in the book with a description of their use in the data analysis. The packages marked by (E) are essential for the geostatistical analysis. Those instead marked by (R) are recommended and can be helpful to overcome issues as described under the column “Used for”.

R packages	Used for
<code>RiskMap</code> (E)	Estimating of geostatistical models and spatial prediction
<code>sf</code> (E)	Handling of spatial data in R
<code>terra</code> (E)	Handling of raster files in R
<code>ggplot2</code> (E)	Creating maps and exploratory plots
<code>crsuggest</code> (R)	Guessing a coordinate reference systems when unknown

To install packages in R for the first time, you can use the command `install.packages` in the R console, as shown below for the `RiskMap` package.

```
install.packages("RiskMap")
```

## 1.4 Data-sets used in the book

The geostatistical data-sets described in this section will be used throughout the book to illustrate the use of the R packages mentioned in the previous sections.

Each of the examples data-sets can be loaded from the **RiskMap** package, using the command

```
data(NAME_OF_THE_DATASET)
```

where in place of **NAME\_OF\_THE\_DATASET** you should type of the name of one of the data-sets listed in Table 1.2.

Table 1.2: List of data-sets available from the **RiskMap** package. Data-sets listed as “Example” are used throughout the book to illustrate the use of R functions. Data-sets listed as “Case study” are analysed in Chapter 6.

Names of the data-set	Short description	Used in this book as
<b>galicia</b>	Lead concentration m from moss samples collected in Galicia, Northern Spain	Example
<b>liberia</b>	Prevalence data on river-blindness from Liberia	Example
<b>malkenya</b>	Malaria prevalence data from a community and school survey conducted in Western Kenya	Example
<b>italy_sim</b>	Simulated geostatistical data-set within the Italian national boundaries	Example
<b>malnutrition</b>	Data on stunting among children in Ghana	Case study

### 1.4.1 Lead concentration in Galicia

Lead is a heavy metal which, in high concentrations, can cause chronic damage to living organisms over a long period of time. For this reason its spread



Figure 1.1: Data on the measured lead concentration (in micrograms per gram dry weight) in moss samples collected in Galicia, North-West of Spain.

and source must be regularly monitored. To assess the extent of the contamination in an area, measurements of lead are often taken from plants. The data here considered (Figure 1.1) consist of 132 locations of moss samples collected in 2000, in and around Galicia, a region in the North-Western part of Spain. One of the objectives of this survey was to establish the spatial pattern of lead concentration in Galicia so as to better identify possible sources of contamination; for more information, see Fernández, Rey, and Carballeira (2000).

In this case, geostatistical modelling can be used to predict the lead concentration across Galicia and allows to disentangle variation which is purely random, possibly due to measurement error, and genuine spatial variation, which is our main object of interest.

This data-set will be used in this book to show how to carry out the spatial analysis of continuously measured variables using linear geostatistical models.

#### 1.4.2 River-blindness in Liberia

In low-resource settings, where disease registries are typically absent, cross-sectional surveys are an essential monitoring tool that enables the estimation of the disease burden in a population of interest. The data considered in this example (Figure 1.2) have been collected as part of an Africa-wide initiative called the Rapid Epidemiological Mapping of Onchocerciasis (REMO) carried out in 2011 in 20 African countries (Zouré et al. 2014). The goal of REMO

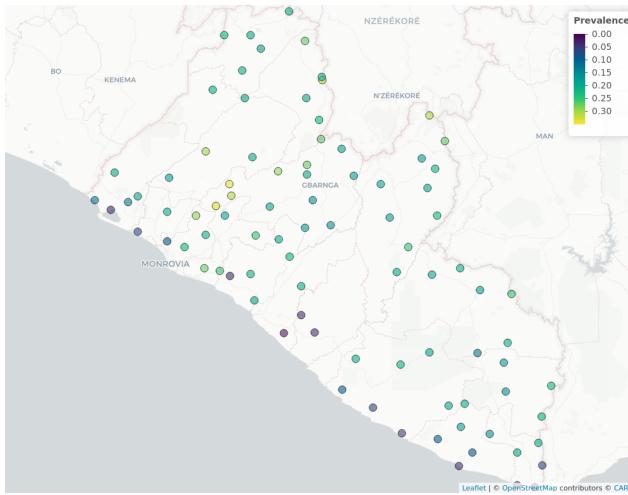


Figure 1.2: River-blindness data from a cross-sectional survey carried out in Liberia.

is to identify areas where river-blindness (or onchoceriasis), a disease transmitted by black flies who breed along fast flowing rivers, is still a public health problem. In this context, it is especially of interest to identify communities with a prevalence above 20% and for treatment is urgently needed.

In this book, we will use data collected from Liberia to model nodule prevalence, which is based on an alternative and cheaper diagnostic technique for river-blindness. In the analysis of this data-set, we will illustrate how to formulate and fit Binomial geostatistical models, and how these can be used to predict prevalence within a region of interest.

#### 1.4.3 Malaria in the Western Kenyan Highlands

Malaria is one of deadliest diseases that affects populations living in tropical and subtropical countries. It is caused by a parasite of the genus *Plasmodium* which is transmitted through the infectious bite of female *Anopheles* mosquitoes. In the following chapters, we shall analyse a data-set from a cross-sectional community survey carried out in July 2010 in Nyanza Province, in the Western Highlands of Kenya (Stevenson 2013).

What distinguishes this from the other examples data-sets is that the data contain both individual-level and household-level information. The outcome of interest is the result from a rapid diagnostic test for malaria which. In the book, we will illustrate how to account for the the hierarchical structure of the data and the binary nature of the outcome at each of the stages of the

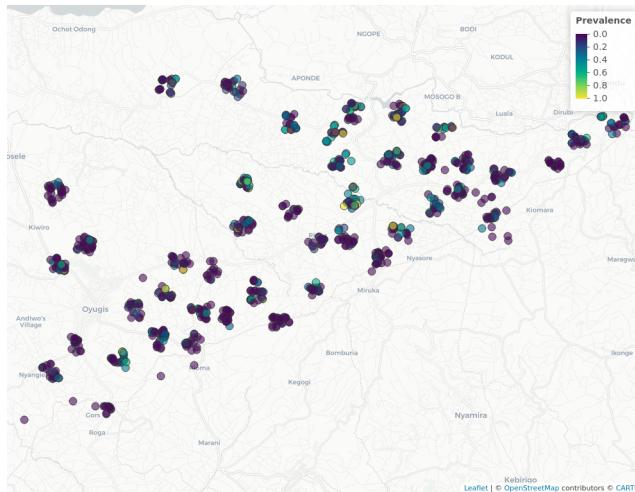


Figure 1.3: Malaria prevalence data from a cross-sectional survey carried out in Nyanza Province, in the Western Highlands of Kenya.

geostatistical analysis.

#### 1.4.4 *Anopheles gambiae* mosquitoes in Southern Cameroon

In studies of vector-borne and zoonotic diseases, understanding of the vector distribution can help to better guide the decision-making process for the implementation, monitoring and evaluation of control programmes. *Anopheles gambiae* mosquitoes are one of the main vectors for malaria transmission in sub-Saharan Africa. Their distribution over space is affected by several environmental and climatic factors, including temperature, humidity and vegetation.

The data-set on mosquitoes (Figure 1.4) that will use in the book consists of a sub-set taken from a large database (Tene Fossog et al. 2015). This was assembled in order to understand how the environment affects the distribution of different species of *Anopheles* mosquitoes in sub-Saharan Africa. This example data-set will be used to illustrate the application of Poisson geostatistical models for mapping mosquitoes abundance.

#### 1.4.5 Simulated-dataset

This data-set was generated using a geostatistical model, with the addition of unstructured random effects at provincial and regional level. More details on how this data-set was generated will be provided in Section 3.2. Whilst this

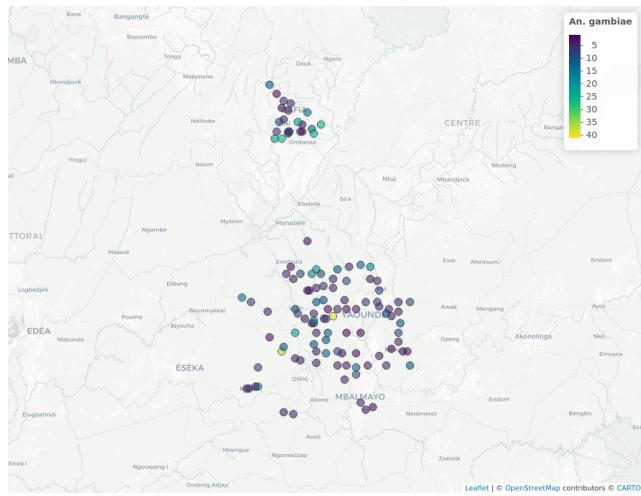


Figure 1.4: Map of the collected number of *Anopheles gambiae* mosquitoes in an area of Southern Cameroon.



Figure 1.5: Map of the locations of the simulated data-set generated over Italy for a continuous outcome.

data-set does not have any scientific relevance like the other data-sets used in this book, it will serve us to illustrate some of the more advanced features of the package that enable the inclusion of random effects, in addition to the latent Gaussian process that is common to all geostatistical models. The skills you will acquire through the analysis of this data-set will be useful for the analysis of data-sets presented as case studies in Chapter 6.

---

## 1.5 Geostatistical problems and geostatistical models

What the examples of the previous section have in common is that, in each case, the goal of statistical analysis is to draw inferences on an unobserved spatially continuous surface using data collected from a finite set of locations. The lead concentration in Galicia, the prevalence for river-blindness in Liberia and the abundance of *A. gambiae* mosquitoes in Cameroon can all be represented as spatially continuous processes that originate from the combined effects of environmental factors. We denote this class of inferential problems as *geostatistical problems* for which a solution can be found through the development and application of suitable *geostatistical models*, which are the subject of this book.

As one can soon realize, geostatistical problems are not unique to global health but arise in many other fields of science, including economics, physics, biology, geology and others. It thus comes to no surprise that geostatistics was initially developed in the South African mining industry in the 1950s (Krig 1951). This was then further developed as a self-contained discipline by Georges Matheron and other researchers at Fontainebleau, in France (Matheron 1963; Chilès and Delfiner 2016). In Watson (1971) and Watson (1972) a first connection is drawn between geostatistics and the prediction of stochastic processes. However, it is only with Ripley (1981) and then N. A. C. Cressie (1991) that geostatistics is explicitly brought into a classical statistical framework for the analysis of spatially referenced data. P. J. Diggle, Tawn, and Moyeed (1998) coined the term *model-based geostatistics* and introduced this as belonging to the general class of generalized linear mixed models (Breslow and Clayton 1993), while emphasizing the use of likelihood-based methods of inference. As in P. J. Diggle, Tawn, and Moyeed (1998), also in this book, we advocate the application of model-based geostatistical models as a class of parametric statistical models on which inference can be carried out using either maximum likelihood estimation or Bayesian methods.

More precisely, our attention will be directed at the class of *generalized linear geostatistical models*, or GLGM. To formally specify this, we first define the random variables  $S$ , a spatial stochastic process, and the random vari-

able  $Y = (Y_1, \dots, Y_n)$  which correspond to the outcome observed at a set of locations  $X = (x_1, \dots, x_n)$ . Let us use  $[A]$  to denote “the distribution of the random variable  $A$ ”. To formulate a GLGM, we should then specify the joint distribution of  $S$  and  $Y$ , which we write as

$$[Y, S] = [S][Y|S]. \quad (1.1)$$

On the right-hand side of the equation above, we have factorized the joint distribution of  $Y$  and  $S$ , as the product between the marginal distribution of  $S$  and the conditional distribution of  $Y$  given  $S$ . Hence, the formulation of a GLGM can be break down into the tasks of formulating  $[S]$  and  $[Y|S]$ .

In defining  $[S]$ , throughout the book, we shall assume that this is a zero-mean stationary and isotropic Gaussian process. In other words, these assumptions impose that the joint distribution of  $S(X) = (S(x_1), \dots, S(x_n))$ , i.e. the process  $S$  at the sampled locations  $x_1, \dots, x_n$ , is invariant with respect to rotations and translations of the locations  $X$ . In practical terms, the main implication of this is that, for any pair of locations  $x_i$  and  $x_j$  the correlation function  $\rho(\cdot)$  between  $S(x_i)$  and  $S(x_j)$  is purely a function of the Euclidean distance,  $u_{ij}$ , between  $x_i$  and  $x_j$ , i.e.

$$\text{cov}\{S(x_i), S(x_j)\} = \sigma^2 \rho(u_{ij}), \quad (1.2)$$

where  $\sigma^2$  is the variance of  $S(x)$  for all  $x$ . Below we provide more details on the type of covoriance functions that we consider in this book. Furthermore, the fact that we assume the process  $S$  to have mean zero is because this is process acts as a residual term in our modelling of  $Y$ . This aspect will be reiterated several times in the following chapters, as it as important implications for the interpretation of the other components of a geostatistical model, as well understanding the results of the analysis.

Finally, we model  $[Y|S]$ , i.e. the distribution of  $Y$  given  $S$ , is modeled as a set of mutually independent distributions which belong the exponential family, as defined in classical generalized linear modelling framework (Nelder and Wedderburn 1972). It then follows that, we can write  $[Y|S]$  as

$$[Y|S] = \prod_{i=1}^n [Y_i|S(x_i)]. \quad (1.3)$$

The final step then consists of specifying a distribution for  $[Y_i|S(x_i)]$ . Table 1.3 gives the range, mean and variance the three specifications for  $\$\[Y\_i | S(x\_i)]\$\$$  which we will consider in this book. In Table 1.3, the *canonical function*, say  $g(\cdot)$ , denotes the natural transformation of the mean component  $\mu_i$  that allows us to introduce both covariates and the spatial process  $S(x_i)$  into the model so as to explain the variation in  $\mu_i$  as

$$g(\mu_i) = d(x_i)^\top \beta + S(x_i). \quad (1.4)$$

where  $d(x_i)$  is a vector of spatially referenced covariates with associated regression coefficients  $\beta$ . Finally, the quantity  $m_i$ , which appears in the formulation of the Binomial and Poisson distributions, is an offset quantity and is used to account for the number of *tests* or the population size at a given location  $x_i$ .

Table 1.3: Type of outcomes  $Y_i$  considered in this book.

Distribution	Range of $Y_i$	Mean of $[Y_i S(x_i)]$	Variance of $[Y_i S(x_i)]$	Canonical link
Gaussian	$(-\infty, +\infty)$	$\mu_i$	$\tau^2$	$g(\mu_i) = \mu_i$
Binomial	$1, \dots, m_i$	$m_i\mu_i$	$m_i\mu_i(1 - \mu_i)$	$g(\mu_i) = \log\{\mu_i/(1 - \mu_i)\}$
Poisson	$1, 2, \dots, \infty$	$m_i\mu_i$	$m_i\mu_i$	$g(\mu_i) = \log\{\mu_i\}$

Based on the formulation in (1.4), we can see that  $S(x_i)$  quantifies residual spatial effects on  $\mu_i$  that have not been accounted for by the covariates  $d(x_i)$ . In an ideal scenario, the covariates  $d(x_i)$  should explain all the spatial variation without the need for  $S(x_i)$ . Although this unrealistic, in practice we may be able to most of the variation in  $\mu_i$  through  $d(x_i)$  and, hence, reduce  $S(x_i)$  to a negligible component. In Chapter 2, we will show how a thorough exploratory analysis can help to understand whether we have come close to that ideal scenario or, if instead, we need the use of GLGM to model the data.

The model described in (1.4) can be seen as the most basic GLGM that can be used for a geostatistical analysis. As we will see in the analysis of some of the examples and, in Chapter 6, for the case studies, extensions of this model will be required to accommodate the intrinsic non-spatial random variation of the data which is not captured by the covariates.

The types of problems that statistical models are applied to can be distinguished into three main categories: prediction problems; explanatory problems; problems of hypothesis testing. Most of the times, geostatistical problems tend to fall under the first category, where the goal is make predictive inferences on the process  $S(x)$  at location  $x$ , which is usually outside of the set of sampled locations. However, as will illustrate in the later chapters, geostatistical models play an important also in the other two types of problems. In particular, we will show that spatial correlation can have a substantial impact on the point estimates and standard errors for  $\beta$ . Hence, if the goal of the analysis is explain the relationship between a covariate  $d(x)$  with the mean component  $\mu$ .

### 1.5.1 The Matern family of correlation functions

Throughout the book, we shall consider the Matern (2013) family of correlation functions to model the spatial correlation of the Gaussian process  $S(x)$ . This defined as

$$\rho(u; \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa K_\kappa(u/\phi), \quad (1.5)$$

where  $\phi > 0$  and  $\kappa > 0$  are parameters and  $K_\kappa(\cdot)$  is the modified Bessel function of the third kind of order  $\kappa$ . The parameters  $\phi$  and  $\kappa$  regulate how fast the spatial correlation decays to zero for increasing distance and the smoothness of the process, respectively. A special case of Matern family of correlation functions, which is obtained when  $\kappa = 0.5$ , will be of particular relevance to the application considered in this book. This is the exponential correlation function which we write as

$$\rho(u; \phi) = \exp\{-u/\phi\}. \quad (1.6)$$

Another special case, which we do not consider in this book but has often been used in machine learning applications, is the Gaussian correlation function obtain as a limiting case for  $\kappa \rightarrow +\infty$  the possible smoothest process arising from the Matern family.

To better understand how  $\phi$  and  $\kappa$  affect the spatial correlation and the pattern of the spatial of the spatial surface, we now consider some examples.

Figure 1.6 shows six different Matern correlation functions. In panel (a), we have kept  $\kappa$  fixed to 0.5 and varied  $\phi$  over the values 0.05, 0.1 and 0.2. As expected, for larger values of  $\phi$  the correlation function has a slower decay to zero. Panels (a) to (c), in Figure 1.7, show three realizations of a Gaussian process from each of these correlation functions. The mean of the Gaussian process was set to zero and variance to 1. We can observe that spatial correlations with larger scales are associated with longer spatial trends, whilst smaller scales exhibit a patchier pattern. This is because, as  $\phi$  takes values that are closer to zero, the spatial surface will tend to show a less structured pattern and will revert towards its zero mean more rapidly. In our examples in the book, we will often use the so called *practical range* to interpret our estimates for  $\phi$ . The practical range is defined as the distance at which spatial correlation reaches 0.05, hence one can interpret this as the distance beyond which observations can be considered approximately independent. In the case of the exponential correlation function, the practical range is  $\log(20) \times \phi \approx 3\phi$ .

Finally, let us consider the correlation functions, shown in the panel (b) of Figure 1.6. Here, we have varied  $\kappa$  over the values 0.5, 1.5 and 2.5, whilst  $\phi$  has been fixed in order to force all three correlation functions to reach 0.05 for distance 0.3. In this way, we can better observe the effect of different values  $\kappa$  on the spatial surface for process that have approximately the same

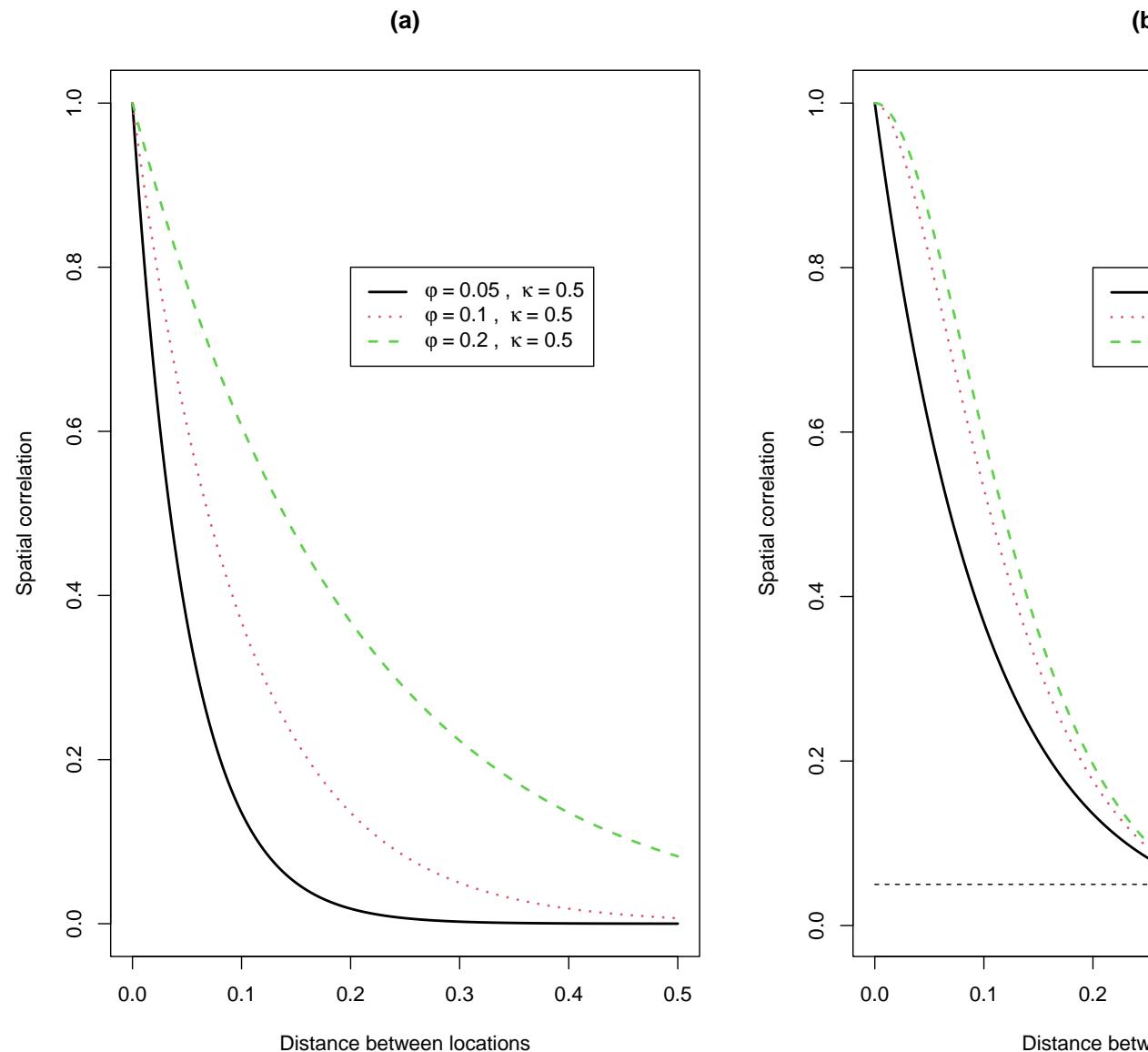


Figure 1.6: Examples of stationary and isotropic Matern correlation functions. Panel (a) shows three different correlations functions that have the same smoothness parameter of  $\kappa = 0.5$ , while varying the scale parameter  $\phi$  over 0.05, 0.1, 0.2. In panel (b) the scale of the spatial correlation  $\phi$  is chosen so that each of the three functions reaches 0.05 at distance 0.3 (as also shown by the horizontal and vertical black dashed segments).

range for the spatial correlation. In Figure 1.7, we observe three realizations from these correlation functions. We observe that the differences between the different surface are determined by the small spatial scale behaviour;  $\kappa = 0.5$  correspond to a rougher and less regular spatial pattern, whilst  $\kappa = 2.5$  shows a smoothest surface of the three processes considered. These properties of the spatial surface are related to the so called differentiability of the Gaussian process, which determines its local behaviour. If you are interested in delving these theoretical aspects, we suggest reading Chapter 2 of Stein (1999).

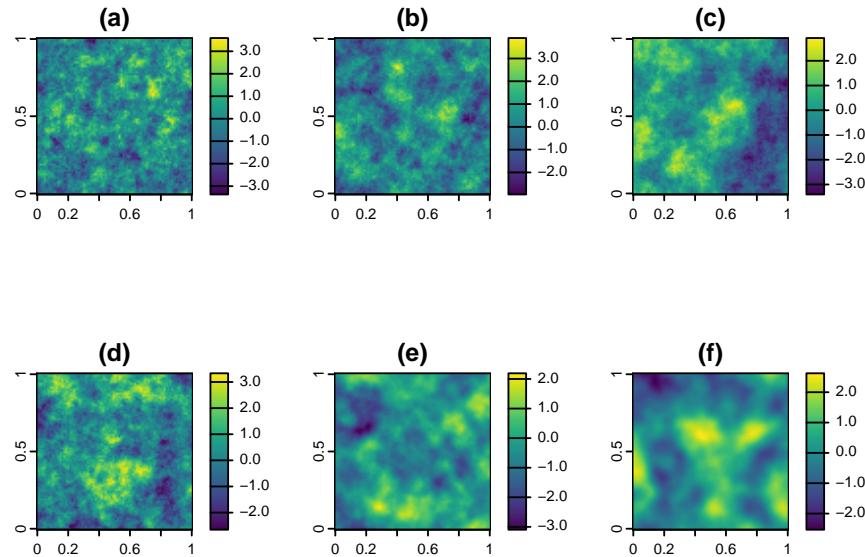


Figure 1.7: Simulated spatial surface using the three correlation functions shown in Figure 1.6. Panels (a), (b) and (c) correspond to the correlation functions from panel (a) in Figure 1.6 and in order these are:  $\phi = 0.05$  and  $\kappa = 0.5$ ; (b)  $\phi = 0.1$  and  $\kappa = 0.5$ ; (c)  $\phi = 0.2$  and  $\kappa = 0.5$ . Panels (c), (d) and (e) correspond to the correlation functions from panel (b) in Figure 1.6 and in order these are: (a)  $\phi = 0.1$  and  $\kappa = 0.5$ ; (b)  $\phi = 0.063$  and  $\kappa = 1.5$ ; (c)  $\phi = 0.051$  and  $\kappa = 2.5$ .

The flexibility provided by the Matern correlation function in capturing different forms of spatial correlations has made one of, if not the most widely used correlation function in model-based geostatistics (Stein 1999). For this reason, in this book we will consider the Matern correlation function. We will consider estimation issues related to the Matern correlation in Chapter 3.

## 1.6 Workflow of a statistical analysis and structure of the book

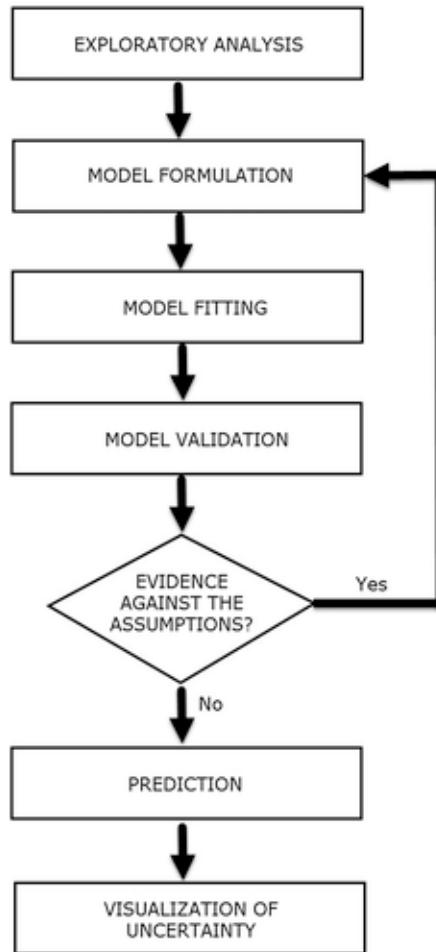


Figure 1.8: Stages of a statistical analysis

Figure 1.8 shows the different stages that will follow in carrying the geostatistical analysis of the examples introduced in Section 1.4. The exploratory analysis of the data is an essential first step that is used to understand the empirical associations between risk factors and the health outcome of interest. In our case, this first stage is also used to justify the use of geostatistical models by questioning the underlying assumptions of standard generalized lin-

ear models. Based on the results obtained from the exploration of the data, we then formulate a suitable statistical model and estimate its parameters using likelihood based methods of inference. These also allows us to obtain uncertainty measures about the strength of associations of regression relationships and the other model parameters that define the shape of the spatial correlation in the data. Following the estimation of the model, we then proceed to validate its underlying assumptions using suitable diagnostics that assess whether the model can later be sufficiently trusted to represent the observed variation in the modelled outcome. At this stage, if the diagnostics checks yield results that indicate the incompatibility of the model with the data, we then back to the stage of model formulation and address the issues arisen from the validation stage. If instead, we do not find any evidence against the fitted model we can proceed to carry out spatial prediction. At this stage, it is important to define suitable predictive targets that can help us to better answer the original research question and better assist the decision making process. The final step of visualization of uncertainty plays an important role in geostatistical analysis in order to convey the main findings of the study in an effective and easy-to-understand way for a wider audience which also consists of non-experts.

In the remainder of this book, each chapter focuses on a specific stage as shown in Figure 1.8. We treat visualization of uncertainty together with spatial prediction in Chapter 5.

Chapter 2 will provide an overview of how to handle spatial data in R, in particular raster and vector data (both points and polygons). The skills learned in this chapter will be applied throughout the book, and will especially be useful in Chapter 5 and Chapter 6 for generating predictive maps of the modelled outcome.

Chapter 3 focuses on the model building process and estimation of geostatistical models. This chapter will show how to carry out initial exploratory analyses of the data to inform the formulation of suitable geostatistical models and how these can be fitted using maximum likelihood estimation methods.

Chapter 4 illustrated the use of methods that can be used to validate the assumptions and calibration of statistical models.

Chapter 5 shows how geostatistical models can be used to carry out spatial prediction of a health outcome of interest both on a spatially continuous and spatially aggregated scales.

Finally, Chapter 6 presents the application of all the methods illustrated in the previous chapters to three additional data-sets. This chapter offers a summary of the content of book by putting together all the stages in the geostatistical analyses for each of the three case studies, and illustrates additional functionalities of the **RiskMap** R package not covered in the previous chapters.

# 2

---

## *Handling of spatial data in R*

---

### 2.1 Introduction

Geostatistical analysis primarily deals with **point-referenced data**. However, geostatistical modeling often requires more than just point data—**raster** and **areal (polygon)** data are frequently used to build essential covariates, enhance spatial understanding, or provide environmental context. For instance, population density, climate data, land use, and elevation, often key covariates in many spatial analyses, usually come in raster or polygon formats. Efficient handling of these different data types is critical for successful model-building. This chapter provides a comprehensive guide to the various stages of managing spatial data in R, using the **sf** and **terra** packages, to support a complete geostatistical workflow.

---

### 2.2 Spatial data handling in geostatistical analysis

Throughout a geostatistical analysis, handling spatial data takes place at multiple key stages:

1. **Retrieving External Spatial Data Sources:** Before starting the modeling process, one often needs to acquire spatial datasets from external sources to improve model accuracy. These could include satellite-derived environmental variables, population data from WorldPop, or climate data from WorldClim. Acquiring these datasets and ensuring they are in compatible formats and resolutions is an essential early step in spatial analysis.
2. **Importing and Standardizing Spatial Data:** Once external data is obtained, it must be imported into R. This involves handling various file formats such as shapefiles or geopackages for vector data and Geotiffs for raster data. Moreover, different datasets might

use different **Coordinate Reference Systems (CRS)**, requiring CRS standardization to ensure that all datasets align correctly. Failure to standardize CRS can result in spatial misalignments and incorrect results.

3. **Extracting Covariate Information for Modeling:** Covariate extraction is an essential step in geostatistical modeling. After importing spatial data, it is necessary to extract relevant covariate information both for the **sampled locations** (where we have observed data) and the **prediction locations** (where we wish to make predictions). This step involves linking raster or polygonal covariates—such as climate, population, or land cover data—to the geostatistical data points.
4. **Prediction and Creation of a Spatial Grid:** For predictive geostatistical models, a regular grid of points is often created over the study region. Covariates must then be assigned to each grid point to make predictions across the entire region of interest. Creating predictive grids and linking them to the necessary covariates is key to generating continuous spatial predictions from geostatistical models.
5. **Visualizing Spatial Data:** Visualization is crucial for exploring spatial data, interpreting model results, and communicating findings. Whether working with point-referenced data, polygons, or rasters, clear and effective visualization helps reveal patterns that inform the modeling process. Effective visualization can also help highlight covariate trends, spatial clusters, and uncertainty in predictions.

---

### 2.3 Chapter overview

This chapter introduces methods for handling spatial data at each stage of the geostatistical analysis workflow. You'll learn how to:

- Retrieve covariates from external spatial data sources;
- Import and standardize spatial data;
- Extract covariates for geostatistical analysis;
- Create predictive grids and link them with covariates;
- Visualize spatial data effectively to aid in analysis and presentation.

By mastering these steps, you'll be equipped to handle the complexity of spatial data in geostatistical modeling, enhancing your analyses and improving your predictions.

---

## 2.4 Accessing covariates for disease mapping

Covariates can play a crucial role in understanding and modeling the spatial variation in disease risk. In geostatistical analysis, incorporating environmental, demographic, and climatic variables might improve the predictive power of models or at least reduce the level of uncertainty in the predictions. These covariates can influence factors such as the spread of infectious diseases, the distribution of disease vectors, and the socio-economic conditions that impact health outcomes.

A wide range of open-access spatial datasets provide covariates for disease mapping. These include population density, climate, land cover, and human infrastructure data, which often come in raster or polygon formats. Alongside these environmental covariates, administrative boundaries are also crucial for public health analysis, as they help organize and aggregate data at different levels (e.g., country, region, or district). For example, aggregating health outcomes or covariate data by administrative units allows researchers to identify geographic disparities and allocate resources accordingly. Fortunately, open-source platforms such as geoBoundaries provide easily accessible administrative boundary data for many countries at various levels of granularity. This data is available in formats compatible with geospatial analysis tools in R, making it easy to integrate into geostatistical workflows.

Table 2.1 below summarizes key sources of covariates useful in disease mapping. Each dataset offers specific types of data, from satellite-derived environmental variables to gridded population estimates and administrative boundaries, which can be accessed through various R packages or APIs.

MENTION RSPATIAL DATA WEBSITE?? <https://rspatialdata.github.io/>

Table 2.1: Some key sources of covariates useful in disease mapping and R packages to retrieve them.

Source	Data Type	Description	R Package
WorldClim	Climate (temperature, rainfall)	Global climate data	geodata

Source	Data Type	Description	R Package
<b>MODIS</b>	Remote Sensing	Satellite imagery (e.g., vegetation indices)	MODISTsp
<b>OpenStreetMap (OSM)</b>	Human settlements, roads	Global geographic features	osmdata
<b>Google Earth Engine</b>	Satellite imagery	Large-scale environmental data analysis	rgee
<b>WorldPop</b>	Population data	Gridded population density estimates	worldpop (via API)

#### 2.4.1 Example: Downloading administrative boundaries

Administrative boundaries provide an essential spatial structure for many types of geostatistical analyses, particularly in disease mapping where data is often aggregated by administrative units such as regions, districts or provinces. The `rgeoboundaries` package, which are accessible via the `rgeoboundaries` package, provides openly available administrative boundary data for nearly every country, allowing researchers to integrate these boundaries into their analyses seamlessly.

```
# Load the rgeoboundaries package
library(rgeoboundaries)

# Download administrative boundaries for Liberia (level 0:
#   ↳ country)
liberia_admin0 <- gb_adm0("Liberia")

# Do the same for level 1: regions
liberia_admin1 <- gb_adm1("Liberia")

# Plot the administrative boundaries
plot(liberia_admin0$geometry)
plot(liberia_admin1$geometry)
```



Figure 2.1: Administrative boundaries for Liberia retrieved using the `rgeoboundaries` package.

#### 2.4.2 Example: Download population data

Population density is a key covariate in geostatistical models for public health research. In this section, we demonstrate how to retrieve high-resolution population data for Liberia from WorldPop using the `wpgpDownloadR` package. This package provides easy access to the WorldPop datasets, which offers gridded population estimates at various spatial resolutions.

Before downloading the data, we will search for available datasets for Liberia. The function `wpgpListCountryDatasets()` helps in retrieving a list of all available datasets for a specified country.

```
library(wpgpDownloadR)

# Search for datasets available for Liberia
# usign the ISO3 country code
lbr_datasets <- wpgpListCountryDatasets(ISO3 = "LBR")
```

```
if(knitr:::is_html_output()){
  lbr_datasets
} else {
  head(lbr_datasets)
}
```

ISO	ISO3	Country	Covariate
125	430	LBR	Liberia
374	430	LBR	Liberia
623	430	LBR	Liberia
872	430	LBR	Liberia

```
ISO ISO3 Country Covariate
125  430  LBR Liberia  ppp_2000
374  430  LBR Liberia  ppp_2001
623  430  LBR Liberia  ppp_2002
872  430  LBR Liberia  ppp_2003
```

```
1121 430 LBR Liberia ppp_2004
1370 430 LBR Liberia ppp_2005
```

125 Estimated total number of people per grid-cell 2000 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)  
 374 Estimated total number of people per grid-cell 2001 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)  
 623 Estimated total number of people per grid-cell 2002 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)  
 872 Estimated total number of people per grid-cell 2003 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)  
 1121 Estimated total number of people per grid-cell 2004 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)  
 1370 Estimated total number of people per grid-cell 2005 The dataset is available to download in Geotiff format at a resolution of 3 arc (approximately 100m at the equator)

We can check the description column to see what datasets are available. Let's download the population data for Liberia for the year 2014 at a 100m resolution. The `wpgpGetCountryDataset` function will then download a raster dataset based on ISO3 code and covariate name.

```
lbr_pop_url <- wpgpGetCountryDataset(ISO3 = "LBR", covariate =
  ↴ "ppp_2014")
```

This will download a raster file locally in a temporary directory. The path to the downloaded file is contained in the `lbr_pop_url` variable and when we introduce the `terra` package in the next sections we will show how to upload

the population raster into R. It is also possible to specify the directory where we want the raster to be downloaded using the `destDir` argument.

---

## 2.5 Importing and standardizing spatial data

In geostatistical analysis, importing and standardizing spatial data is a critical step to ensure that data from different sources align and can be used effectively. Spatial data, whether it's vector data (points, lines, polygons) or raster data (grids), can come in various formats and may use different **Coordinate Reference Systems (CRS)**. To perform accurate spatial analyses, it's essential to import data correctly and ensure consistency in terms of projection and format. This section will cover how to import vector and raster data into R, explain the concept of CRS and ensure that different datasets align properly for subsequent geostatistical analysis.

### 2.5.1 Importing vector data

Vector data are typically stored in formats such as **shapefiles** (.shp) or **geopackages** (.gpkg). The `sf` (simple features) package in R is the most common tool for handling vector data.

```
# Load the sf package
library(sf)

# Import a shapefile (e.g., administrative boundaries)
admin <- st_read("path_to_your_shapefile/Admin_Boundaries.shp")

# Inspect the data
print(admin)
plot(admin) # Basic plot of the shapefile
```

The `st_read()` function reads various spatial data formats, automatically recognizing file types.

### 2.5.2 Importing raster data

Raster data consists of a grid of cells, where each cell holds a value representing a spatial attribute such as elevation or temperature. The `terra` package in

R is designed to work with raster data and has superseded the older `raster` package due to better performance and greater functionality. Let's see now how to import a GeoTIFF file using `terra`. We can upload the population raster for Liberia that we have downloaded in Section 2.4.2.

```
# Load the terra package
library(terra)

# Import a raster file
lbr_pop_100 <- rast(lbr_pop_url)

# Inspect the raster data
print(lbr_pop_100)

class      : SpatRaster
dimensions : 5040, 4945, 1  (nrow, ncol, nlyr)
resolution : 0.0008333333, 0.0008333333 (x, y)
extent     : -11.48625, -7.365417, 4.352084, 8.552084  (xmin,
xmax, ymin, ymax)
coord. ref. : lon/lat WGS 84 (EPSG:4326)
source     : lbr_ppp_2014.tif
name       : lbr_ppp_2014
min value  : 0.006426936
max value  : 92.716874581

# Basic plot of the raster
plot(lbr_pop_100)
```

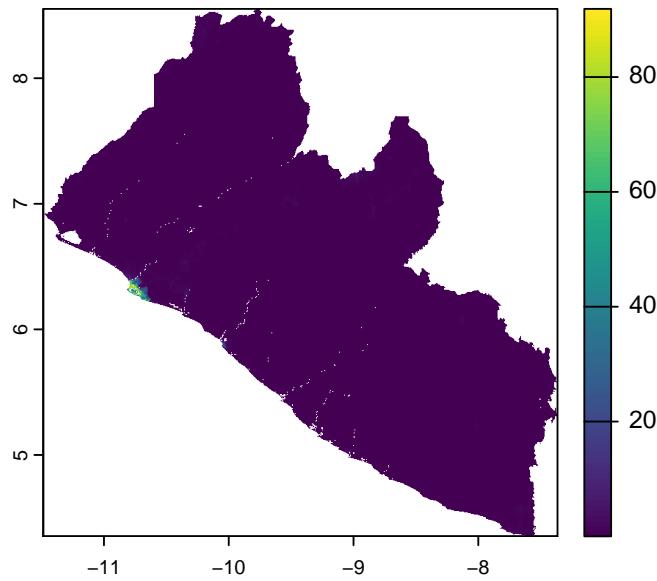


Figure 2.2: Liberia population count for 2014 at 100m resolution

The output of `print()` provides a detailed summary of the raster's properties. The `rast()` function reads raster files in formats like GeoTIFF, ASCII grid, or other common raster formats and then we used `plot()` as a quick way to visualize raster data.

### 2.5.3 Understanding coordinate reference systems (CRS)

When working with spatial data, especially from different sources, one of the most critical tasks is ensuring that all datasets share the same **Coordinate Reference System (CRS)**. A CRS defines how the two-dimensional map data corresponds to locations on the three-dimensional Earth. If different layers (e.g., raster and vector data) have different CRSs, they may not align correctly when plotted or analyzed together, leading to inaccurate analyses or visualizations.

CRSs can either be:

- **Geographic:** These use latitude and longitude coordinates to represent locations on the Earth's surface. The most common example is **WGS84** (EPSG:4326), the default CRS used by GPS and global datasets.

- **Projected:** These convert the Earth's curved surface to a flat map and preserve certain properties like area, distance, or direction. Examples include **Universal Transverse Mercator (UTM)** or **Albers Equal Area** projections.

In many cases, using a projected rather than a geographic CRS is preferred, especially when any summary statistic or parameter is distance-related. In a geographic CRS, distances between two points are calculated using angular coordinates (degrees), which do not translate easily into linear units like meters or kilometers. This makes interpreting distances challenging, as the length of a degree of latitude differs from the length of a degree of longitude. In contrast, a projected CRS uses a linear coordinate system (usually meters), ensuring that distances are accurately represented on a flat surface. This is important when computing spatial variograms, covariance functions, or some parameters in geostatistical models, where distance between sampling locations is a key factor. In this context, we assume that the default distance metric is Euclidean distance. However, an important exception to our recommendation arises when analyzing data across large, global-scale regions. In such cases, it is more appropriate to use a geographic CRS along with spherical distances, as these better reflect the curved nature of the Earth's surface.

#### 2.5.4 EPSG codes

An **EPSG code** is a unique identifier that defines a CRS. These codes, managed by the **European Petroleum Survey Group (EPSG)**, are widely used in geographic information systems (GIS) to simplify the use of specific projections, ensuring that spatial data is correctly aligned and interpreted. Each EPSG code corresponds to a unique CRS or map projection, making it easier to standardize and manage spatial data from different sources.

Some key and often used EPSG codes are:

- **EPSG:4326:** This code represents **WGS84**, the most commonly used geographic CRS, which uses latitude and longitude to describe locations on the Earth's surface. It is the default CRS for global datasets and GPS systems.
- **EPSG:326XX:** These codes represent the **UTM (Universal Transverse Mercator)** projection, which divides the world into zones. Each zone is optimized to preserve local distances and areas. For example (e.g. EPSG:32629: UTM Zone 29N, covering parts of Western Africa, including Liberia)
- **EPSG:3857:** This code is for the **Web Mercator projection**, which is widely used for web mapping services, including **Google Maps**, **OpenStreetMap**, and **Bing Maps**. This projected CRS uses meters as the unit of distance and is optimized for visualizing maps on a 2D plane, though it

distorts area and distance, especially at high latitudes. It is well-suited for interactive online mapping but not ideal for precise distance-based geostatistical analyses.

### 2.5.5 Convert a data frame to an `sf` object

In geospatial analysis, data is often provided in tabular formats like CSV files that contain spatial coordinates (e.g., latitude and longitude). To use these data effectively in R, it is necessary to convert the data frame into an `sf` object, which is the standard format for working with spatial data in R. Here we show how to achieve this, we can use the Liberia data available in the `RiskMap` package as it is a data frame.

```
# Load the RiskMap package
library(RiskMap)

# Load the Liberia data set
data("liberia")

# Convert the data frame to an sf object
liberia_sf <- st_as_sf(liberia,
                       coords = c("long", "lat"),
                       crs = 4326)

# Inspect the new sf object
liberia_sf
```

Simple feature collection with 90 features and 4 fields

Geometry type: POINT

Dimension: XY

Bounding box: xmin: -11.40335 ymin: 4.42552 xmax: -7.4928 ymax: 8.46911

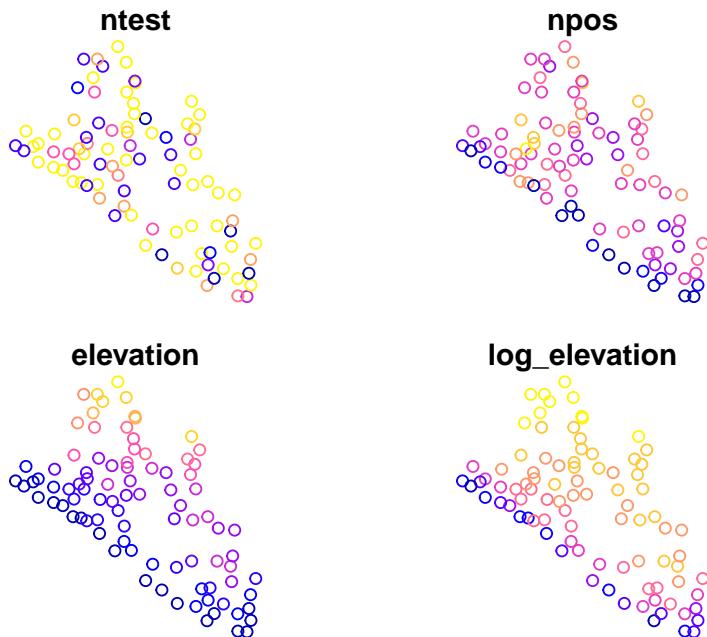
Geodetic CRS: WGS 84

First 10 features:

	ntest	npos	elevation	log_elevation	geometry
1	50	14	17.82621	2.880670	POINT (-10.4671 6.2878)
2	46	10	104.85070	4.652537	POINT (-10.45111 6.5725)
3	43	11	119.09543	4.779925	POINT (-10.02615 6.56667)
4	50	10	144.10921	4.970571	POINT (-10.28775 6.73333)
5	48	9	19.03508	2.946283	POINT (-10.03595 6.01667)
6	50	13	16.99954	2.833186	POINT (-10.335 6.266666)
7	43	9	137.75360	4.925467	POINT (-9.63333 6.13541)
8	50	11	101.75000	4.622519	POINT (-10.06875 6.2375)

```
9      47    11 147.12867      4.991308  POINT (-9.73333 6.3869)
10     43     0 26.61054      3.281308  POINT (-9.7856 5.7354)
```

```
plot(liberia_sf)
```



The `st_as_sf()` function converts the data frame into an `sf` object. The `coords` argument specifies which columns contain the spatial coordinates and the `crs` argument assigns the CRS that in this case we know being WGS84 (EPSG:4326). The `sf` object can now be used for operations such as spatial joins, distance calculations, and mapping with other spatial layers. Note that the columns containing the spatial coordinates have been replaced by a `geometry` column, which now stores this information. If you would like to retain the original coordinate columns in the output, you can set the `remove` argument to `FALSE` when converting the data frame to an `sf` object.

### 2.5.6 Working with CRSs in R

When working with data from multiple sources, such as environmental layers, population data, or administrative boundaries, ensuring that all datasets share the same CRS is essential for accurate spatial analysis. This section covers the core tasks involved in managing CRSs in R: checking the CRS of spatial

data to ensure datasets are compatible and Reprojecting spatial data into a common CRS when necessary.

#### 2.5.6.1 Checking the CRS of Spatial Data

Before performing any spatial operation, it's crucial to check the CRS of your spatial datasets. Knowing whether your data uses geographic coordinates (e.g., WGS84) or a projected coordinate system (e.g., UTM) helps ensure that they are aligned and ready for analysis. Both `sf` and `terra` provide functions to retrieve and inspect the CRS, ensuring datasets are spatially aligned before analysis. The `st_crs()` function retrieves the CRS information for vector data.

```
# Check the CRS of the vector data
st_crs(liberia_sf)
```

```
Coordinate Reference System:
  User input: EPSG:4326
  wkt:
  GEOGCRS["WGS 84",
    ENSEMBLE["World Geodetic System 1984 ensemble",
      MEMBER["World Geodetic System 1984 (Transit)"],
      MEMBER["World Geodetic System 1984 (G730)"],
      MEMBER["World Geodetic System 1984 (G873)"],
      MEMBER["World Geodetic System 1984 (G1150)"],
      MEMBER["World Geodetic System 1984 (G1674)"],
      MEMBER["World Geodetic System 1984 (G1762)"],
      MEMBER["World Geodetic System 1984 (G2139)"],
      ELLIPSOID["WGS 84",6378137,298.257223563,
        LENGTHUNIT["metre",1]],
      ENSEMBLEACCURACY[2.0]],
    PRIMEM["Greenwich",0,
      ANGLEUNIT["degree",0.0174532925199433]],
    CS[ellipsoidal,2,
      AXIS["geodetic latitude (Lat)",north,
        ORDER[1],
        ANGLEUNIT["degree",0.0174532925199433]],
      AXIS["geodetic longitude (Lon)",east,
        ORDER[2],
        ANGLEUNIT["degree",0.0174532925199433]],
    USAGE[
      SCOPE["Horizontal component of 3D system."],
      AREA["World."],
```

```
BBOX[-90,-180,90,180],  
ID["EPSG",4326]]
```

The same can be achieved for raster data with the `crs()` function from the `terra` package.

```
# Check the CRS of the raster data  
crs(lbr_pop_100, proj = TRUE, describe = TRUE)
```

	name authority code area	extent
proj		
1 WGS 84	EPSG 4326 <NA> NA, NA, NA, NA +proj=longlat +datum=WGS84 +no_defs	

### 2.5.6.2 Reprojecting spatial data to a common CRS

If your datasets have different CRSs or if you want to change CRS (e.g. from geographical to projected) you will need to reproject one or more datasets so they can be spatially aligned. This ensures that they can be overlaid and analyzed together. For vector data, `st_transform()` reprojects the data into a specified CRS. This example transforms the liberia point data from WGS84 into UTM. To know what's the correct UTM zone and hence EPSG code for Liberia we can use the `get_epsg_utm` function from the `RiskMap` package.

```
# MOVE THIS INTO RISKMAP

get_epsg_utm <- function(sf_object) {  
  # Ensure the input is an sf object  
  if (!inherits(sf_object, "sf")) {  
    stop("Input must be an sf object.")  
  }  
  
  # Check if the CRS is WGS 84 (EPSG: 4326)  
  crs <- st_crs(sf_object)  
  if (is.null(crs) || crs$epsg != 4326) {  
    stop("Input sf object must have a CRS of WGS84 (EPSG:  
        4326).")  
  }  
  
  # Get the coordinates of the centroid of the sf object  
  centroid <- st_centroid(st_union(sf_object))  
  coords <- st_coordinates(centroid)
```

```

# Function to calculate UTM zone based on longitude
get_utm_zone <- function(lon) {
  return((floor((lon + 180) / 6) %% 60) + 1)
}

# Calculate UTM zone from longitude
utm_zone <- get_utm_zone(coords[1])

# Determine the EPSG code based on the latitude (north/south)
epsg_code <- if (coords[2] >= 0) {
  32600 + utm_zone # Northern Hemisphere
} else {
  32700 + utm_zone # Southern Hemisphere
}

# Return the calculated EPSG code
return(epsg_code)
}

# Obtain EPSG code for UTM for Liberia
get_epsg_utm(liberia_sf)

```

[1] 32629

```

# Reproject the vector data
liberia_sf_utm <- st_transform(liberia_sf, crs =
  ↪ get_epsg_utm(liberia_sf))

```

Reprojecting raster data is more complex than reprojecting vector data due to the continuous nature of raster grids. The process involves recalculating cell values to fit a new grid based on the new CRS, which can lead to challenges like resampling, distortion, and data loss. When reprojecting a raster, the grid must adjust to the new CRS, often requiring resampling of cell values. The method you choose depends on the data type: **nearest neighbor** is best for categorical data like land use while **bilinear or cubic interpolation** is good for continuous data like temperature, where smooth transitions are needed.

The function `project()` from the `terra` package can be used to reproject a raster.

```

lbr_pop_100_utm <- project(lbr_pop_100,
                           crs(liberia_sf_utm),
                           method = "bilinear")

```

Reprojecting a raster may alter its resolution. For example, reprojecting from geographic (degrees) to projected (meters) CRS can result in a mismatch between the original and new cell sizes. Moreover, distortion can occur when converting between projections, especially at high latitudes. Some cells may be stretched or compressed, leading to potential loss of information or edge artifacts. These distortions arise because the Earth is not flat, and projecting the curved surface of the Earth onto a flat plane (or vice versa) leads to trade-offs. For example, the Mercator projection preserves angles and shapes but distorts area, particularly near the poles.

For these reasons, it's often better to reproject vectors rather than rasters when both data types are used together and avoid as much as possible to change the CRS of raster. One way to achieve this is to work with WGS84 when performing all spatial operations like extraction of covariates from rasters and then transform only the point data to a projected CRS before fitting the model.

### 2.5.7 Common Issues and Considerations

#NOT SURE ABOUT THIS, MAYBE EXPAND OR REMOVE

- **CRS Mismatch:** One of the most common issues in spatial analysis is the mismatch of CRSs between datasets. Always check and reproject datasets if necessary before performing spatial operations like overlays or extractions.
- **Datum Transformations:** In some cases, especially when working with datasets from different regions, a simple projection transformation may not be enough. You might need to handle datum shifts (e.g., from NAD83 to WGS84). R can handle these transformations, but it's important to be aware of them.
- **Choosing the Right CRS:** The choice of CRS depends on your study area and the type of analysis. For global studies, WGS84 is often used. For regional studies, **UTM** or other local projections may be more appropriate to preserve accuracy in distances and areas.

---

## 2.6 Extracting covariate data

In geostatistical models, the inclusion of relevant covariates (environmental, demographic, or climatic) can potentially enhance predictive accuracy. Covariate data often come from raster or polygon sources, and extracting these

values for point locations is essential to link spatial context to point-referenced data. These covariates could include variables like temperature, elevation, land cover, or population density, which influence the spatial distribution of diseases. In this section, we will cover how to extract covariates at point locations from both polygon layers raster layers.

#### 2.6.0.1 Extracting covariates from polygon layers

Polygon layers contain discrete spatial entities, such as administrative boundaries or land use areas, with associated attributes. Extracting covariates from polygon layers involves associating point data with the attributes of the polygon in which they fall. Here is an example with.....NEED SOME DATA FOR THIS

```
# Perform a spatial join to transfer polygon attributes to the
# points
points_with_admin <- st_join(liberia_sf,
  liberia_admin1[["shapeName"]])

# View the results, points now include covariates from the
# polygon layer
head(points_with_admin)
```

```
Simple feature collection with 6 features and 5 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -10.4671 ymin: 6.016667 xmax: -10.02615
ymax: 6.73333
Geodetic CRS:  WGS 84
  ntest npos elevation log_elevation shapeName
  geometry
1    50   14  17.82621      2.880670 Margibi    POINT
  (-10.4671 6.2878)
2    46   10 104.85070      4.652537 Montserrado  POINT
  (-10.45111 6.5725)
3    43   11 119.09543      4.779925 Margibi    POINT
  (-10.02615 6.56667)
4    50   10 144.10921      4.970571 Margibi    POINT
  (-10.28775 6.73333)
5    48     9 19.03508      2.946283 Grand Bassa POINT
  (-10.03595 6.016667)
6    50   13 16.99954      2.833186 Grand Bassa    POINT
  (-10.335 6.266666)
```

This spatial join operation adds attributes (in this case the admin1 names) from the polygons to the points. These attributes can now be used in geostatistical models to explain spatial variation in disease risk based on the regions they fall into.

### 2.6.1 Extracting covariates from raster layers

Raster data provides continuous spatial information, such as elevation, climate data, or population density. Covariate values from raster layers can be extracted for specific points using the `extract()` function from the `terra` package. Each point will receive the value of the raster cell it overlaps.

```
# Extract raster values at the point locations
covariate_values <- extract(lbr_pop_100, liberia_sf)

# Combine the extracted values with the point data
liberia_sf$pop_total <- covariate_values[, 2]

# View the updated dataset
head(liberia_sf)
```

```
Simple feature collection with 6 features and 5 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -10.4671 ymin: 6.016667 xmax: -10.02615
ymax: 6.73333
Geodetic CRS:  WGS 84
  ntest npos elevation log_elevation           geometry
  pop_total
1  50   14  17.82621      2.880670  POINT (-10.4671 6.2878)
0.2914019
2  46   10 104.85070      4.652537  POINT (-10.45111 6.5725)
0.6087928
3  43   11 119.09543      4.779925  POINT (-10.02615 6.56667)
0.2044306
4  50    10 144.10921      4.970571  POINT (-10.28775 6.73333)
0.5078438
5  48     9 19.03508      2.946283  POINT (-10.03595 6.016667)
0.1767686
6  50   13  16.99954      2.833186  POINT (-10.335 6.266666)
1.6693381
```

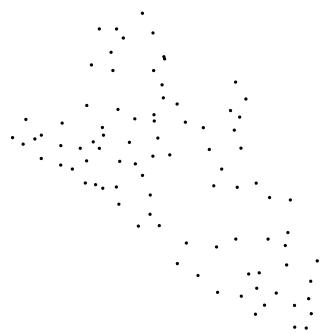
In this example, the `extract()` function assigns the raster value from the

population density layer to each point in the dataset. This allows the point data to include population density as a covariate in the analysis.

Instead of extracting values for exact point locations, it can sometimes be useful to aggregate covariate values within a defined area around each point. This is often done by creating a buffer around each point and calculating summary statistics (e.g., mean, sum) of the raster values within that buffer. For instance, you might want to calculate the average population density or temperature within a 2 km radius around each point to smooth out fine-scale variation.

```
# Create buffers around each point (e.g., 2 km radius)
buffered_points <- st_buffer(liberia_sf_utm, dist = 1000)

# Plot the buffers for visualization
plot(st_geometry(buffered_points), col = "blue", border =
  "black")
```



```
# Extract raster values within the buffer areas and calculate
  the
# mean or sum. Note that since we used the utm data to work on
  the
# meter scale we need to convert them back to WGS84
```

```
mean_pop_density <- extract(lbr_pop_100,
                           st_transform(buffered_points, crs =
                                         4326),
                           fun = mean, na.rm = TRUE)

# Add the averaged values to the points dataset
liberia_sf$pop_mean2km <- mean_pop_density[,2]

# View the updated dataset
head(liberia_sf)
```

Simple feature collection with 6 features and 6 fields

Geometry type: POINT

Dimension: XY

Bounding box: xmin: -10.4671 ymin: 6.016667 xmax: -10.02615  
ymax: 6.73333

Geodetic CRS: WGS 84

	ntest	npos	elevation	log_elevation	geometry
1	50	14	17.82621	2.880670	POINT (-10.4671 6.2878) 0.2914019
2	46	10	104.85070	4.652537	POINT (-10.45111 6.5725) 0.6087928
3	43	11	119.09543	4.779925	POINT (-10.02615 6.56667) 0.2044306
4	50	10	144.10921	4.970571	POINT (-10.28775 6.73333) 0.5078438
5	48	9	19.03508	2.946283	POINT (-10.03595 6.016667) 0.1767686
6	50	13	16.99954	2.833186	POINT (-10.335 6.266666) 1.6693381
			pop_mean2km		
1			0.3239253		
2			0.5534695		
3			0.2254926		
4			0.4708411		
5			0.2419380		
6			3.7534250		

You can modify the `fun` argument to calculate other summary statistics, such as the sum, min or max of the raster values within the buffer. This approach is particularly useful when the phenomenon being modeled (e.g., disease transmission) is influenced by broader spatial factors around the observation point, rather than just the value at the exact point location.

## 2.7 Creating a predictive grid

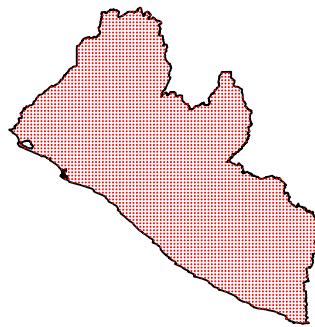
A predictive grid is a regularly spaced set of points or cells that spans the study region. This grid serves as the basis for predictions made by your model. The density of the grid (i.e., the distance between grid points) affects both the resolution of the prediction and the computational cost. For point-based predictions, we can generate a grid of points over a polygon (e.g., administrative boundary) using the `sf` package and the `st_make_grid` function.

```
# First we convert the Liberia boundaries to the UTM CRS
# because we want our grid in meters
liberia_admin0_utm <- liberia_admin0 |>
  st_transform(crs = get_epsg_utm(liberia_sf))

# Generate prediction grid at 5km resolution
pred_locations <- st_make_grid(liberia_admin0_utm,
                                cellsize = 5000,
                                what = "centers")

# Exclude locations that fall outside the study area
pred_locations <- st_intersection(pred_locations,
                                   liberia_admin0_utm)

# Visualize the result
plot(liberia_admin0_utm$geometry, col = "white")
plot(pred_locations, cex = .01, col = "red", pch = 19, add = T)
```



---

## 2.8 Visualizing spatial data

Visualization is a key part of spatial data analysis, as it allows you to explore and communicate spatial patterns and relationships effectively. R provides a lot of functionalities to visualize spatial data and create very beautiful maps. Until now we have used basic plotting functions. Here we introduce the `ggplot2` package that allows to combine different types of geographic data in a map. The `ggplot2` package in R provides a flexible and powerful framework for creating both simple and complex visualizations, including maps of point data, polygons, and rasters. With the help of extensions like `geom_sf()` and `geom_raster()`, `ggplot2` makes it easy to visualize spatial data, whether you're working with point locations, polygons, or continuous raster data.

### 2.8.1 Visualizing point data

Point data often represents the locations of observations (e.g., disease cases, sampling sites). `ggplot2` allows you to plot these points and optionally color them by a covariate (e.g., disease prevalence or population density). Here is

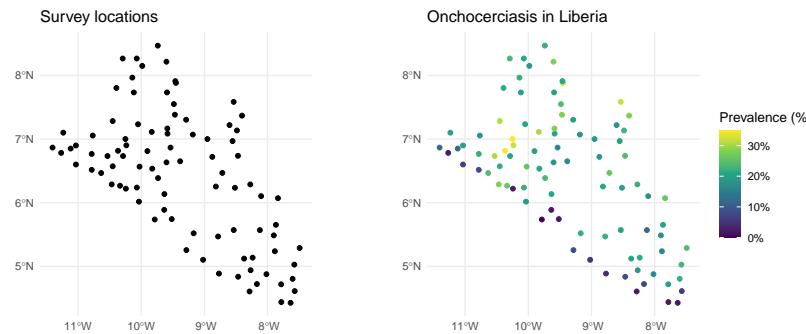
an example that uses the Liberia data.

```
# Load necessary libraries
library(ggplot2)

# Create a new variable with prevalence in the dataset
liberia_sf$prevalence <- liberia$npos / liberia$ntest

# Plot only the locations
ggplot(data = liberia_sf) +
  geom_sf(col = "black") +
  theme_minimal() +
  labs(title = "Survey locations")

# Color the points according to prevalence
ggplot(data = liberia_sf) +
  geom_sf(aes(color = prevalence)) +
  scale_color_viridis_c(labels = scales::label_percent()) +
  theme_minimal() +
  labs(title = "Onchocerciasis in Liberia",
       color = "Prevalence (%)")
```



Here `geom_sf()` is used to plot the spatial points. `aes(color = prevalence)` specifies that the points should be colored based on the `prevalence` covariate, providing a visual representation of spatial variation in disease risk. The `scale_color_viridis_c()` function applies the Viridis color scale, which is well-suited for continuous data and is friendly for those with color blindness. The `labels = scales::label_percent()` argument ensures that the color scale's labels are displayed as percentages (e.g., 5, 10%) rather than raw decimal values. To make the plot visually clean and minimal, `theme_minimal()` is applied, stripping away unnecessary background elements and keeping the focus on the data. Finally, the `labs()` defines the plot title and the color legend label.

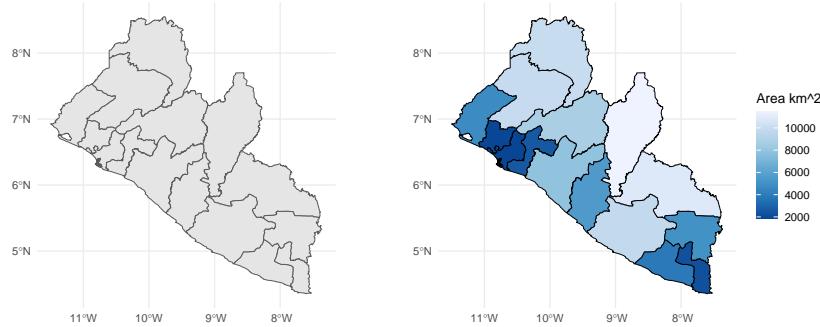
## 2.8.2 Visualizing polygon data

Polygon data typically represents administrative boundaries, land use, or other regional divisions. We can still use `geom_sf()` to create maps of polygons, optionally filling them by a covariate.

```
# Plot Liberia admin 1 level boundaries
ggplot(data = liberia_admin1) +
  geom_sf() +
  theme_minimal() +
  labs()

# We compute the area of each polygon
liberia_admin1$area <- as.numeric(st_area(liberia_admin1) / 1000
                                ^ 2)

# Color the polygons according tis new variable
ggplot(data = liberia_admin1) +
  geom_sf(aes(fill = area), color = "black") +
  scale_fill_distiller(direction = -1) +
  theme_minimal() +
  labs(fill = "Area km^2")
```



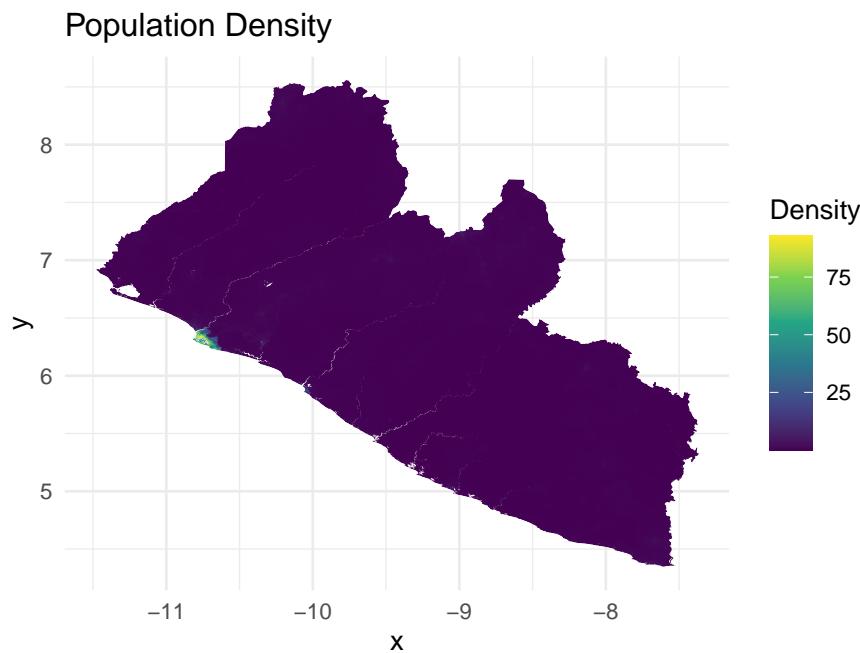
In this code, `aes(fill = area)` is used to fill each polygon with colors corresponding to its area. The `color = "black"` argument outlines the polygons in black, and you could set `fill = NA` to make the polygons transparent while still displaying the borders. The `scale_fill_distiller(direction = -1)` function applies a color gradient from ColorBrewer, with the `direction = -1` argument reversing the gradient (e.g., darker colors for larger areas).

### 2.8.3 Visualizing raster data

In `ggplot2`, you can visualize raster data by converting it into a data frame of coordinates and values. You can convert raster data into a format that `ggplot2` can handle by using the `as.data.frame()` function from `terra`.

```
# Convert the raster to a data frame for ggplot2
raster_df <- as.data.frame(lbr_pop_100, xy = TRUE)

# Plot raster using ggplot2
ggplot(data = raster_df) +
  geom_raster(aes(x = x, y = y, fill = lbr_ppp_2014)) +
  scale_fill_viridis_c() +
  coord_cartesian() +
  theme_minimal() +
  labs(title = "Population Density",
       fill = "Density")
```



In this example `as.data.frame()` converts the raster into a data frame with x and y coordinates and their corresponding raster values and `geom_raster()` is used to plot the raster cells, coloring them based on the population density.

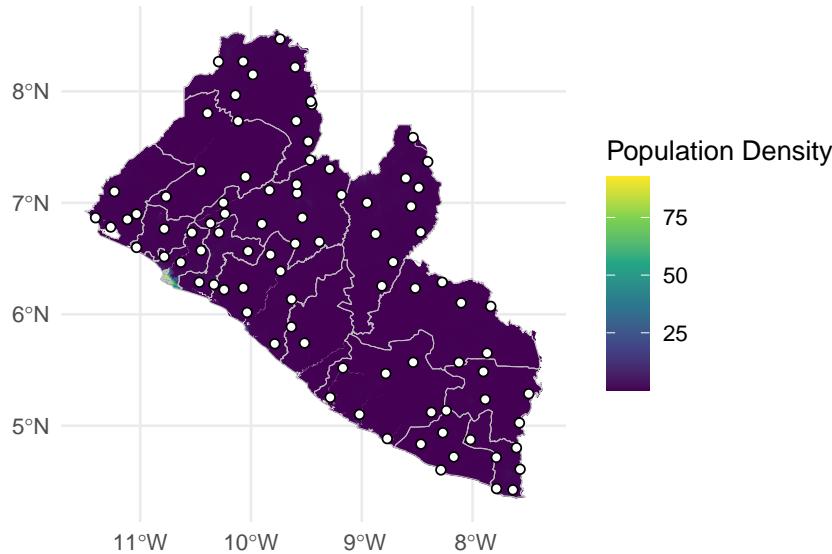
### 2.8.4 Combining Multiple Spatial Data Types

In many cases, it's useful to combine different spatial data types (points, polygons, and rasters) in a single visualization. `ggplot2` allows you to overlay these layers, providing a more comprehensive view of your spatial data.

```
# Combine points, polygons, and raster data in one plot
combined_map <- ggplot() +
  geom_raster(data = raster_df, aes(x = x, y = y, fill =
    ↪ lbr_ppp_2014)) +
  geom_sf(data = liberia_admin1, fill = NA, color = "grey") +
  geom_sf(data = liberia_sf, shape = 21, col = "black", fill =
    ↪ "white") +
  scale_fill_viridis_c() +
  theme_minimal() +
  labs(title = "Combined Spatial Data: Points, Polygons, and
    ↪ Raster",
    fill = "Population Density",
    x = "", y = "")

combined_map
```

Combined Spatial Data: Points, Polygons, and Raster



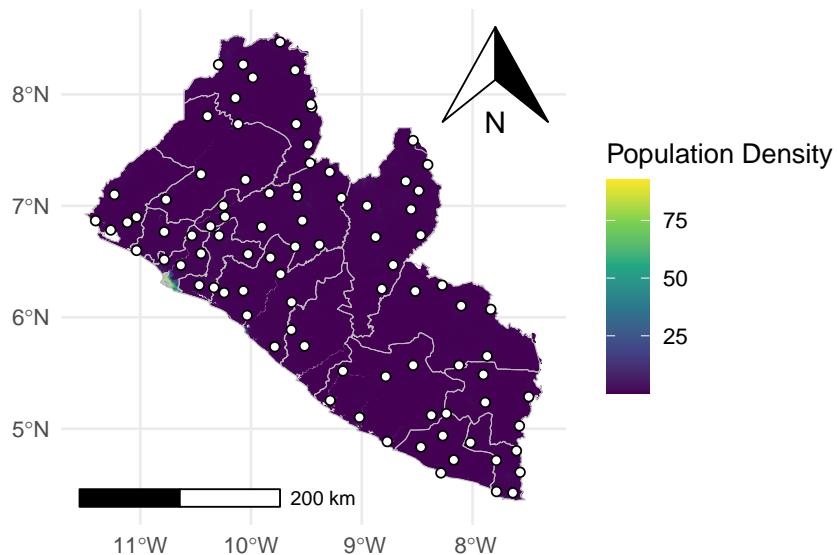
To enhance spatial visualizations in `ggplot2`, adding a scale bar and a north

arrow improves map readability and professionalism. The `ggspatial` package offers tools to easily integrate these elements into your maps. Below is an example that demonstrates how to use `ggspatial` to add a scale bar and north arrow to a map that includes raster, polygon, and point data.

```
# Load necessary libraries
library(ggspatial)

# Add scale bar and north arrow
combined_map +
  annotation_scale(location = "bl", width_hint = 0.5) +
  annotation_north_arrow(location = "tr", which_north = "true")
```

Combined Spatial Data: Points, Polygons, and Raster





# 3

---

## *Model formulation and parameter estimation*

---

### **List of the main functions used in the chapter**

Function	R Package	Used for
<code>lmer</code>	<code>lme4</code>	Fitting linear mixed models
<code>glmer</code>	<code>lme4</code>	Fitting generalized linear mixed models
<code>glgm</code>	<code>RiskMap</code>	Fitting generalized linear geostatistical models
<code>s_variogram</code>	<code>RiskMap</code>	Computing the empirical variogram and carrying out permutation test for spatial independence

---

### **3.1 Exploratory analysis**

As illustrated in Figure 1.8, exploratory analysis is the first step that should be carried out in a statistical analysis. This stage is essential to inform how covariates should be introduced in the model and, in our case, whether the variation unexplained by those covariates exhibits spatial correlation.

In the exploratory analysis of count data, we will also look at how overdispersion, which is a necessary, though not sufficient, condition for residual spatial correlation.

#### **3.1.1 Exploring associations with risk factors using count data**

Assessment of the association between the health outcome of interest and non-categorical (i.e. continuous) risk factors can be carried using graphical

tools, such scatter plots. The graphical inspection of the empirical association between the outcome and the covariates is especially useful to identify non-linear patterns in the relationship which should then be accounted for in the model formulation.

In this section, we look more closely at the case when the observed outcome is a count which requires a different treatment from continuously measured outcomes, which are generally covered by most statistics textbooks (see, for example, Chapter 1 of Weisberg (2014)).

### 3.1.1.1 When the outcome is an aggregated count

Let us first consider the example of the river-blindness data in Liberia (Section 1.4.2), and examine the association between prevalence and elevation. We first generate a plot of the prevalence against the measured elevation at each of the sample locations

```
liberia$prev <- liberianpos/liberia$ntest  
  
ggplot(liberia, aes(x = elevation, y = prev)) + geom_point() +  
  labs(x="Elevation (meters)",y="Prevalence")
```

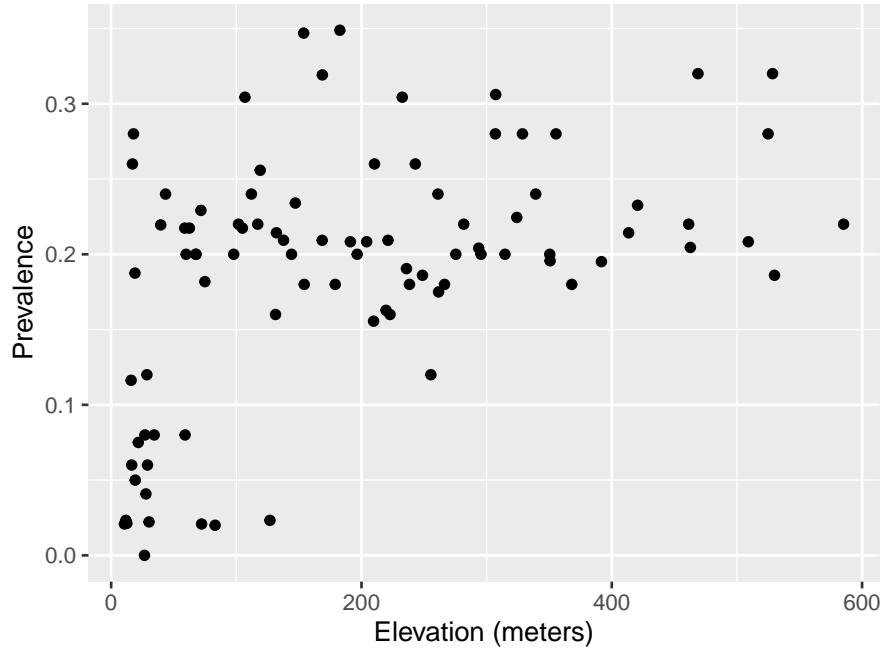


Figure 3.1: Scatter plot of the empirical prevalence for river-blindness against elevation, measured in meters.

The plot shown in Figure 3.1 shows that, as elevation increases from 0 to around 150 meters, prevalence rapidly increases to around 0.25 and, for larger values in elevation than 150 meters, the relationship levels off. This begs the question of how we can account for this in a regression model. To answer this question rigorously, however, the plot in Figure 3.1 cannot be used. This is because, when the modelled outcome is a bounded Binomial count, regression relationships are specified on the logit-transformed prevalence (log-odds) scale; see Table 1.3 in Section 1.5. To explore regression relationships in the case of prevalence data, it is convenient to use the so-called empirical logit in place of the empirical prevalence. The empirical logit is defined as

$$l_i = \log \left\{ \frac{y_i + 1/2}{n_i - y_i + 1/2} \right\} \quad (3.1)$$

where  $y_i$  are the number of individuals who tested positive for river-blindness and  $n_i$  is the total number of people tested at a location. The reason for using the empirical logit, rather than the standard logit transformation applied directly to the empirical prevalence, is that it allows to generate finite values for empirical prevalence values of 0 and 1, for which the standard logit transformation is not defined.

```
# The empirical logit
liberia$elogit <- log((liberianpos+0.5) /
                         (liberia$ntest-liberianpos+0.5))

ggplot(liberia, aes(x = elevation, y = elogit)) + geom_point() +

# Adding a smoothing spline
labs(x="Elevation (meters)",y="Empirical logit") +
stat_smooth(method = "gam", formula = y ~ s(x),se=FALSE) +

# Adding linear regression fit with log-transformed elevation
stat_smooth(method = "lm", formula = y ~ log(x),
            col="green",lty="dashed",se=FALSE) +

# Adding linear regression fit with change point in 150 meters
stat_smooth(method = "lm", formula = y ~ x + pmax(x-150, 0),
            col="red",lty="dashed",se=FALSE)
```

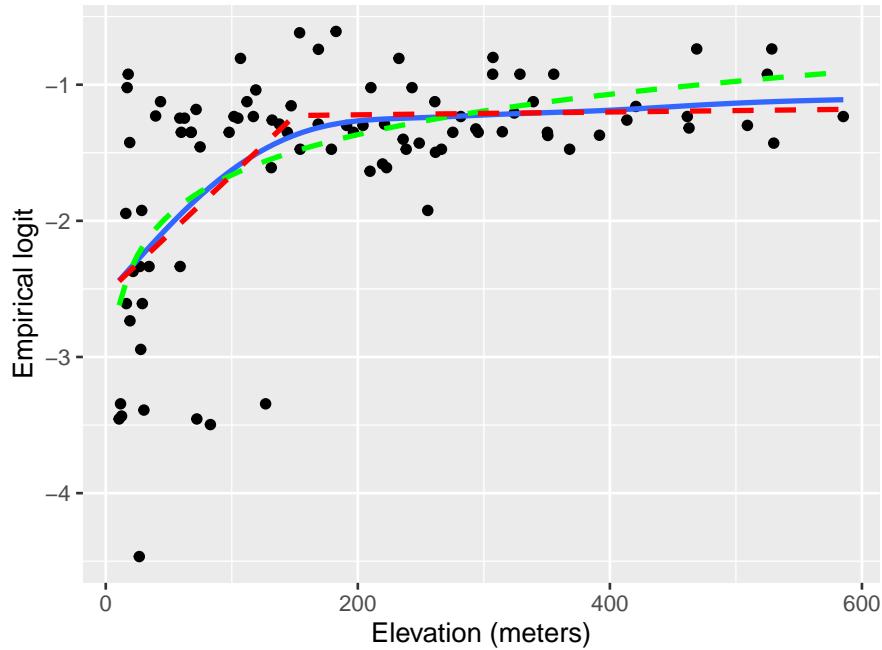


Figure 3.2: Scatter plot of the empirical prevalence for river-blindness against elevation, measured in meters.

Figure 3.2 shows the scatter plot of the empirical logit against elevation. In this plot, we have also added three lines through the `stat_smooth` from the `ggplot2` package. Using this function, we first pass the term `gam` to `method` to add a penalized smoothing spline (Hastie, Tibshirani, and Friedman 2001), represented by the blue solid line. The smoothing spline allows us to better discern how the type of relationship and how to best capture it using a standard regression approach. As we can see from Figure 3.2, the smoothing spline corroborates our initial observation of a positive relationship up to about 150 meters, followed by a plateau.

To capture this non-linear relationship, we can use the two following approaches. The first is based on a simple log-transformation of elevation and is represented in Figure 3.2 by the green line. If were to express this relationship using a standard Binomial regression model, this would take the form

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = \beta_0 + \beta_1 \log\{e(x_i)\} \quad (3.2)$$

where  $p(x_i)$  and  $e(x_i)$  are the river-blindness prevalence and elevation at sampled location  $x_i$ , respectively.

Alternatively, the non-linear effect of elevation on prevalence could be captured using a linear spline. Put in simple terms, we want to fit a linear regression model that allows for a change in slope above 150 meters. Formally, this is expressed in a Binomial regression model as

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = \beta_0 + \beta_1 e(x_i) + \beta_2 \max\{e(x_i) - 150, 0\}. \quad (3.3)$$

Based on the equation above, the effect of elevation below 150 meters is quantified by the parameter  $\beta_1$ . Above 150 meters, instead, the effect of elevation becomes  $\beta_1 + \beta_2$ . Note that the function `pmax` (and not the standard base function `max`) should be used in R when the computation of the maximum between a scalar value and each of the components of a numeric vector is required.

Before proceeding further, it is important to explain the differences between the use of the logarithmic transformation (Equation 3.2) and the linear spline (Equation 3.3). We observe that both curves provide a similar fit to the data, with larger differences observed for larger values in elevation, where the log-transformed elevation models yields larger values for the predicted prevalence. This also suggests that if we were to extrapolate the predictions beyond 600 meters in elevation the implied pattern by the model with the log-transformed elevation would predict an increasingly larger elevation, which is unrealistic, since the fly that transmits the diseases cannot breed at those altitudes. The linear spline model instead would generate predictions that would be very similar to those observed between 150 and 600 meters. From this point view, the linear spline model would thus have more scientific validity than the other

model. However, which of the two approaches should be chosen to model the effect of elevation is a question that closely depends on the research question to be addressed.

If the interest of the study was in better understanding the association between elevation and prevalence, the linear spline model does not only provide a more credible explanation but also its regression parameters can be more easily interpreted. In fact, for a unit increase in elevation, the multiplicative change in the odds for river-blindness is  $\exp\{\beta_1\}$ , if elevation is below 150 meters, and  $\exp\{\beta_1 + \beta_2\}$ , if elevation is above 150 meters. When instead we use the log-transformed elevation, the interpretation of  $\beta_1$  in Equation 3.2 is slightly more complicated, as it is based on the multiplicative increase in elevation by the same amount given by the base of the algorithm, which is about  $e \approx 2.718^1$ . To avoid this, one could rescale the regression coefficient as, for example,  $\beta_1/\log_2(e)$  which would be interpreted as the multiplicative change in the odds for river-blindness for a doubling in elevation. However, a doubling in elevation is less meaningful when considering larger values of elevation.

When the goal of statistical analysis is instead in developing a predictive model for the outcome of interest, the explanatory power and interpretability of the model may be of less concern. For this reason, the model with the log-transformed model could be preferred over the model with the linear spline, if it shown to yield more predictive power. We will come back to this point again in Chapter 5, where will show how to assess and compare the predictive performance of different geostatistical models.

The other type of aggregated count data that we consider are unbounded counts. The Anopheles mosquitoes data-set (Section 1.4.4) is an example of this, since there is no upper limit to the number of mosquitoes that can be trapped at a location. Let us consider the covariate represented by elevation. In this case, the simplest model that can be used to analyse the data is a Poisson regression, where the linear predictor is defined on the log of the mean number of mosquitoes (Table 1.3). Hence, exploratory plots for the association with covariates should be generated using the log transformed counts of mosquitoes. In this instance, to avoid taking the log of zero, we can add 1 to the reported counts, if required. The variable of the `An.gambiae` in the `anopheles` data-set does not contain any 0, hence we simply apply the log transformation without adding 1.

```
anopheles$log_counts <- log(anopheles$An.gambiae)
ggplot(anopheles, aes(x = elevation, y = log_counts)) +
  geom_point() +
  # Adding a smoothing spline
```

---

<sup>1</sup>The letter  $e$  stands for the so called Euler's number and represents the base of the natural logarithm. In the book, we write  $\log(\cdot)$  to mean the "natural logarithm of  $\cdot$ ".

```
labs(x="Elevation (meters)",y="Log number of An. gambiae
  ↵ mosquitoes") +
  stat_smooth(method = "lm", formula = y ~ x, se=FALSE)
```

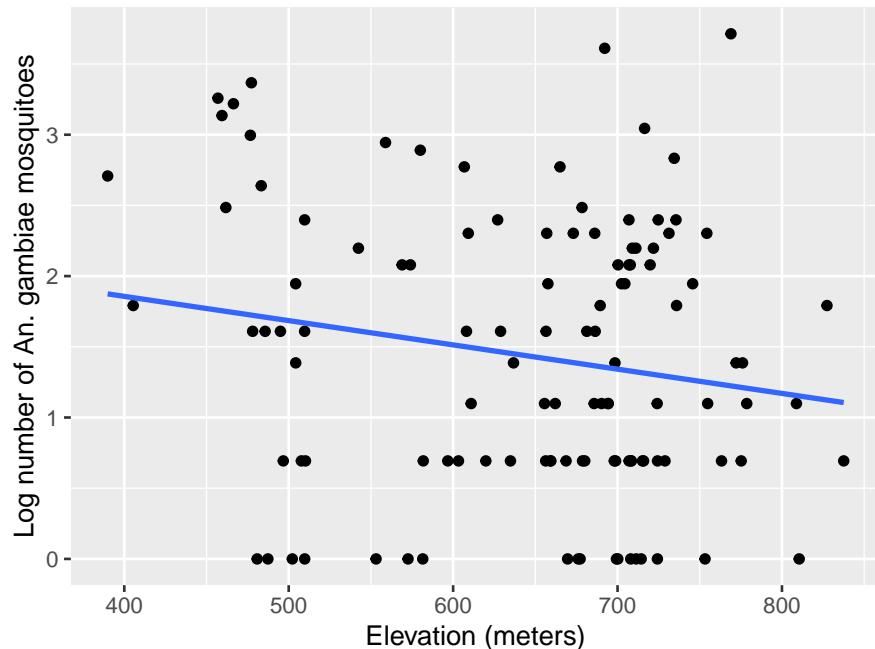


Figure 3.3: Scatter plot of the log transformed number of *Anopheles gambiae* mosquitoes against elevation, measured in meters. The blue line is generated using the least squares fit.

The scatter plot of Figure 3.3 shows that there is a negative, though weak, association, with the average number of mosquitoes decreasing for increasing elevation. In this instance, the assumption of a linear relationship with elevation would be a reasonable choice.

### 3.1.1.2 When the outcome is an individual-level binary indicator

We now consider the malaria data from Kenya (Section 1.4.3) where the main outcome is the result from a rapid diagnostic test (RDT) for malaria from individuals within households. In this case, because the outcome only takes two values, 1 for a positive RDT test result and 0 otherwise, the direct application of the empirical logit from Equation 3.1 would not help us to generate informative scatter plots. Throughout the book, we will consider the data from the

community survey only, hence we work with a subset of the data which we shall name `malkenya_comm`

```
malkenya_comm <- malkenya[malkenya$Survey=="community", ]
```

To show how this issue can be overcome, let us consider the variables age and gender. To generate a plot that can help us understand between the relationship with malaria prevalence and the two risk factors, we proceed as follows.

```
# Grouping of ages into classes defined through "breaks"
malkenya_comm$Age_class <- cut(malkenya_comm$Age,
                                breaks = c(0, 5, 10, 15, 30, 40, 50,
                                          ↪ 100),
                                include.lowest = TRUE)
```

Using the `cut` function, we first split age (in years) into classes through the argument `breaks`. The classification of age into [0, 5], (5, 10] and (10, 15] is common in many malaria epidemiology studies, as children are one of the groups at highest risk malaria. The choice of the other classes of age reflects instead the need to balance the number of observations falling in each of the classes.

```
# Computation of the empirical logit by age groups and gender
age_class_data <- aggregate(RDT ~ Age_class + Gender,
                             data = malkenya_comm,
                             FUN = function(y)
                               log((sum(y)+0.5)/(length(y) ↪ -sum(y)+0.5)))
```

We then compute the empirical logit, using the total number of cases within age group and by gender. For a given age group and gender, which we denote as  $\mathcal{C}$ , the empirical logit in Equation 3.1, now takes the form

$$l_{\mathcal{C}} = \log \left\{ \frac{\sum_{i \in \mathcal{C}} y_i + 0.5}{|\mathcal{C}| - \sum_{i \in \mathcal{C}} y_i + 0.5} \right\} \quad (3.4)$$

where  $y_i$  are the individual binary outcomes and  $i \in \mathcal{C}$  is used to indicate that the sum is carried out over all the individuals who belong the class  $\mathcal{C}$ , identified by a specific age group and gender. Finally,  $|\mathcal{C}|$  is the number of individuals who fall within  $\mathcal{C}$ . In the code above, the empirical logit in Equation 3.4 is computed using the `aggregate` function. An inspection of the object `age_class_data`, a data frame, shows that the empirical is found in the column named `RDT`.

```
# Computation of the average age within each age group
age_class_data$age_mean_point <- aggregate(Age ~ Age_class +
  Gender,
  data = malkenya_comm,
  FUN = mean)$Age

# Number of individuals within each age group, by gender
age_class_data$n_obs <- aggregate(Age ~ Age_class + Gender,
  data = malkenya_comm,
  FUN = length)$Age
```

In order to generate the scatter-plot, we compute the average age within each age group by gender, and use these as our values for the x-axis. Note that since we only need to obtain the average age from this output, we use \$Age to extract this only and allocate to the column age\_mean\_point. Finally, we also compute the number of observations within each of classes and place this in n\_obs.

```
ggplot(age_class_data, aes(x = age_mean_point, y = RDT,
  size = n_obs,
  colour = Gender)) +
  geom_point() +
  labs(x="Age (years)",y="Empirical logit")
```

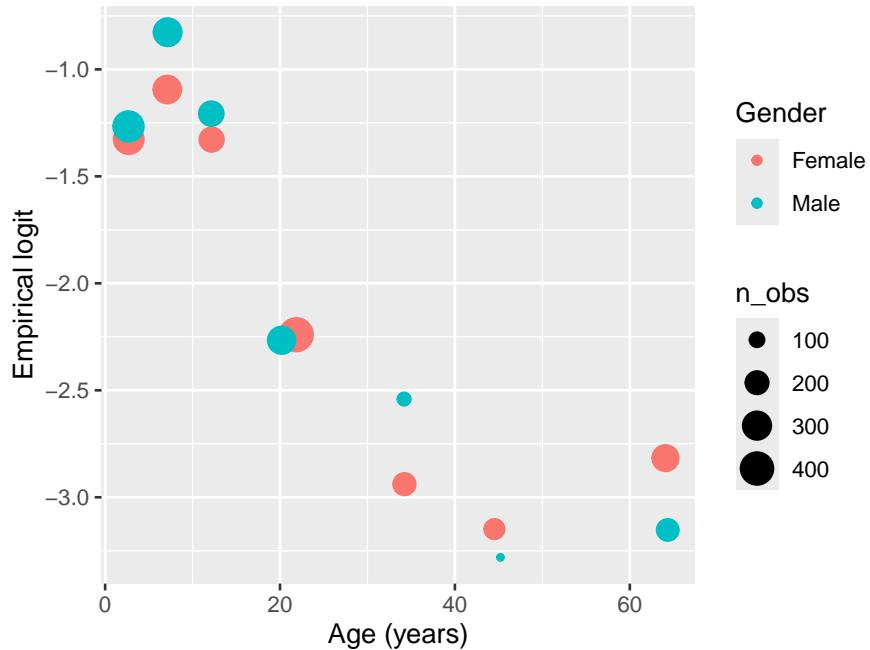


Figure 3.4: Plot of the empirical logit against age, for males and females. The size of each solid point is rendered proportional to the number of individuals within age group, as indicated in the legend.

The resulting plot in Figure 3.4 shows the empirical logit against age by gender, with the size of each of the points proportional to the number of observations falling within each class. The observed patterns are explained by the fact that young children, especially those under the age of five, are particularly vulnerable to severe malaria infections. This is primarily due to their immature immune systems and lack of acquired immunity. As individuals grow older, they generally develop partial immunity to malaria through repeated exposure to the disease. This acquired immunity can provide some level of protection against severe malaria. At the same time, gender roles and activities can influence exposure to malaria-carrying mosquitoes. For example, men may spend more time outdoors for work or other activities, increasing their exposure to mosquito bites and thus their risk of infection. In addition, there are also biological factors to consider. Hormonal and genetic differences between males and females may also contribute to variations in immune responses to malaria infection. The interaction between age and gender is complex and may vary depending on the specific context and population being studied. A 2020 report from the Bill & Melinda Gates foundation provides a detailed overview of this and other aspects related to gender and malaria (Katz and Bill & Melinda Gates Foundation 2020).

To account for age in a model for malaria prevalence, several approaches are possible, some of which have been developed using biological models (Smith et al. 2007). To model the patterns observed in Figure 3.4, we can follow the same approach used in the previous section to model the relationship between elevation and river-blindness prevalence. First, let us consider age without the effect of gender. Let  $p_j(x_i)$  denote the probability of a positive RDT for the  $j$ -th individual living in a household at location  $x_i$ . Assuming that malaria risk reaches its peak at 15 years of age, we can capture the non-linear relationship using a linear spline with two knots, one at 15 years and a second one at 40 years. This is expressed as

$$\log \left\{ \frac{p_j(x_i)}{1 - p_j(x_i)} \right\} = \beta_0 + \beta_1 a_{ij} + \beta_2 \times \max\{a_{ij} - 15, 0\} + \beta_3 \max\{a_{ij} - 40, 0\} \quad (3.5)$$

where  $a_{ij}$  is the age, in years, for the  $j$ -th individual at household  $i$ . Based on this model the effect of age on RDT prevalence is  $\beta_1$ , for  $a_{ij} < 15$ ,  $\beta_1 + \beta_2$ , for  $15 < a_{ij} < 40$ , and  $\beta_1 + \beta_2 + \beta_3$  for  $a_{ij} > 40$ .

Figure 3.4 indicates that age may interact with gender, meaning that the effect of gender on RDT prevalence changes across age, with larger differences observed between males and females for ages above 20 years. To assess such differences using a standard Binomial regression model, the linear predictor for RDT prevalence can be formulated as

$$\log \left\{ \frac{p_j(x_i)}{1 - p_j(x_i)} \right\} = \beta_0 + (\beta_1 + \beta_1^* g_{ij}) \times a_{ij} + (\beta_2 + \beta_2^* g_{ij}) \times \max\{a_{ij} - 15, 0\} + (\beta_3 + \beta_3^* g_{ij}) \times \max\{a_{ij} - 40, 0\} \quad (3.6)$$

where  $g_{ij}$  is the indicator for gender, with 1 corresponding to male and 0 to female. The coefficients  $\beta_1^*$ ,  $\beta_2^*$  and  $\beta_3^*$  thus quantify the differences in risk between the two genders for ages below 15 years, between 15 and 40 years, and above 40 years, respectively. If all of those coefficients were 0, the model in Equation 3.5 would be recovered.

```
glm_age_gender_interaction <- glm(RDT ~ Age + Gender:Age +
  pmax(Age-15, 0) +
  ↪   Gender:pmax(Age-15, 0) +
  pmax(Age-40, 0) +
  ↪   Gender:pmax(Age-40, 0),
  data = malkenya_comm, family =
  ↪   binomial)

summary(glm_age_gender_interaction)
## 
## Call:
```

```

## glm(formula = RDT ~ Age + Gender:Age + pmax(Age - 15, 0) +
##       Gender:pmax(Age -
##                     15, 0) + pmax(Age - 40, 0) + Gender:pmax(Age - 40, 0),
##       family = binomial,
##       data = malkenya_comm)
##
## Coefficients:
##                               Estimate Std. Error z value
## Pr(>|z|)
## (Intercept)           -1.05835   0.10245 -10.331  <
## 2e-16 ***
## Age                  -0.03384   0.01310  -2.584
## 0.00978 **
## pmax(Age - 15, 0)    -0.03975   0.02356  -1.687
## 0.09162 .
## pmax(Age - 40, 0)    0.09170   0.02482   3.695
## 0.00022 ***
## Age:GenderMale        0.01428   0.01221   1.170
## 0.24202
## GenderMale:pmax(Age - 15, 0) -0.03625   0.03145  -1.153
## 0.24908
## GenderMale:pmax(Age - 40, 0)  0.02451   0.04320   0.567
## 0.57052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2875.8 on 3351 degrees of freedom
## Residual deviance: 2673.8 on 3345 degrees of freedom
## AIC: 2687.8
##
## Number of Fisher Scoring iterations: 5

```

The code above shows how to fit the model specified in Equation 3.6. The terms `Age`, `pmax(Age-15, 0)` and `pmax(Age-40, 0)` respectively correspond to  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ , whilst the `Gender:Age`, `Gender:pmax(Age-15, 0)` and `Gender:pmax(Age-40, 0)` to  $\beta_1^*$ ,  $\beta_2^*$  and  $\beta_3^*$ , respectively. In the summary of the fitted model, we observe that the interaction coefficients are non-statistically significant. However, removing the interaction based on the fact that each of the coefficients have each p-values larger than the conventional level of 5% would be wrong. Instead we should carry out the likelihood ratio test, as shown below.

```

glm_age_gender_no_interaction <- glm(RDT ~ Age + pmax(Age-15,
  ↵  0) + pmax(Age-40, 0),
  ↵   data = malkenya_comm, family =
  ↵   binomial)

anova(glm_age_gender_no_interaction, glm_age_gender_interaction,
  ↵  test = "Chisq")
## Analysis of Deviance Table
##
## Model 1: RDT ~ Age + pmax(Age - 15, 0) + pmax(Age - 40, 0)
## Model 2: RDT ~ Age + Gender:Age + pmax(Age - 15, 0) +
  ↵  Gender:pmax(Age -
##           15, 0) + pmax(Age - 40, 0) + Gender:pmax(Age - 40, 0)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3348     2675.6
## 2      3345     2673.8  3    1.8051   0.6138

```

To carry out the likelihood ratio test to assess the null hypothesis that  $\beta_1^* = \beta_2^* = \beta_3^* = 0$ , we first fit the simplified nested model under this null hypothesis. The likelihood ratio test can then be carried out using the `anova` command as shown. The p-value indicates that we do not find evidence against the null hypothesis, hence in our analysis of the data we might favour the simplified model that does not assume an interaction between the two genders.

The approach just illustrated, can also be applied to explore the association with other continuous variables that are a property of the household and not of the individual. Let us, for example, consider the variable `elevation` from the `malkenya` data-set.

```

malkenya_comm$elevation_class <- cut(malkenya_comm$elevation,
  ↵   breaks = quantile(malkenya_comm$]
  ↵   elevation, seq(0, 1, by = 0.1)),
  ↵   include.lowest = TRUE)

```

Following the same approach used for age, we first split `elevation` into classes. To define these, we use the deciles of the empirical distribution of `elevation` which we calculate using the `quantile` function above. In this way we also ensure that the number of observations falling within each class of `elevation` is approximately the same.

```

# Computation of the empirical logit by classes of elevation
elev_class_data <- aggregate(RDT ~ elevation_class,
  ↵   data = malkenya_comm,
  ↵   FUN = function(y)

```

```

log((sum(y)+0.5)/(length(y) ]
      ↵ -sum(y)+0.5)))
# Computation of the average elevation within each class of
# elevation
elev_class_data$elevation_mean <- aggregate(elevation ~
      ↵ elevation_class,
              data = malkenya_comm,
              FUN = mean)$elevation

```

We then compute the empirical logit and the average elevation for each class of elevation. The empirical logit is computed as already defined in Equation 3.4, where now the definition of  $\mathcal{C}$  is given by a specific decile used to split the distribution of elevation.

```

ggplot(elev_class_data, aes(x = elevation_mean, y = RDT),
       size = n_obs) +
  geom_point() +
  labs(x="Elevation (meters)",y="Empirical logit")

```

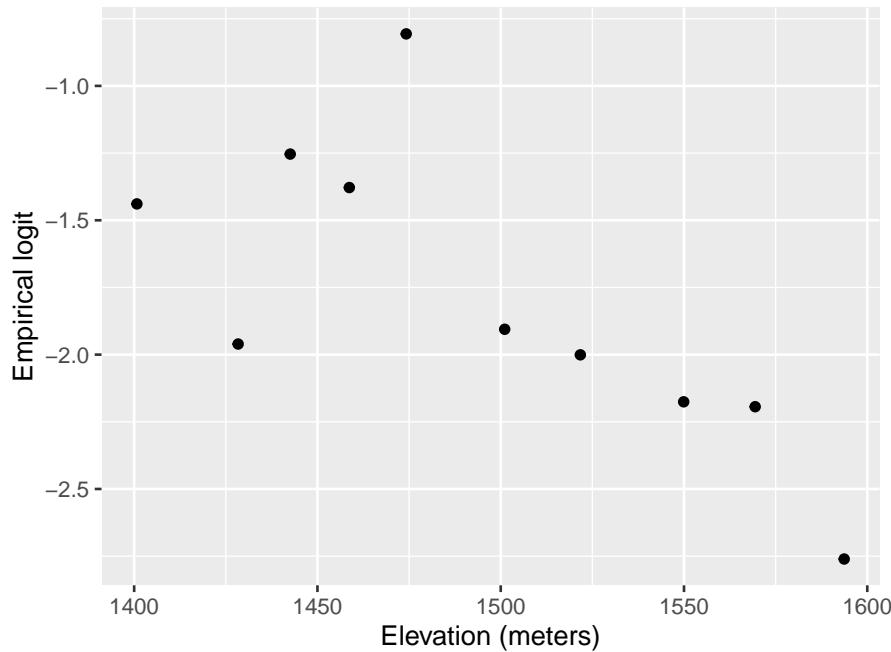


Figure 3.5: Plot of the empirical logit against elevation measured in meters.

The resulting plot in Figure 3.5 shows an approximately linear relationship with decreasing values of the empirical logit for increasing elevation. This is expected because the cooler environment at higher altitudes is less favourable to the development of the overall mosquito life cycle.

An alternative approach to generate a scatter plot for assessing the association between elevation and the empirical logit would be to aggregate the data at household level, rather than using classes of elevation. However, this approach does not work as the one illustrated above when only one individual is sampled for each location. In the case of the `malkenya` data, the great majority of the locations only include one individual making this second approach less useful than the one illustrated.

Other more sophisticated approaches for the exploration of the associations between covariates and binary outcomes are available. For example, the use of the empirical logit could be avoided by using non-parametric regression methods for Binomial outcomes (Bowman 1997), also implemented in `sm` package in R. Our view is that a careful exploratory analysis based on simpler methods, as those illustrated above, can be equally effective to inform the module formulation.

### 3.1.2 Exploring overdispersion in count data

One of the main advantages in the use of covariates is the ability to attribute part of the variation of the outcome to a set of measured variables and, hence, reduce the uncertainty of our inferences. However, it almost always the case that the finite number of covariates at our disposal is not enough to fully explain the variation in the outcome. In other words, the existence of unmeasured covariates that are related to the modelled outcome give rise to the so called residual variation. In a standard linear regression model the extent to which we are able to account for important covariates is directly linked to the size of the variance of the residuals. In the case of count data, instead, this link is less well defined and one of the main consequences of the omission of covariates, which we address in this chapter, is *overdispersion*.

Overdispersion occurs when the variability of the data is larger than that implied by the generalized linear model (GLM) fitted to them. For example, if we consider the Binomial distribution, the presence of overdispersion implies that  $V(Y_i) > n_i\mu_i(1-\mu_i)$ , where we recall that  $n_i$  is the Binomial denominator and  $\mu_i$  is the probability of “success” for each of the  $n_i$  Bernoulli trials; for a Poisson distribution with  $E(Y_i) = \mu_i$ , instead, overdispersion implies that  $V(Y_i) > \mu_i$ .

Assessment of the overdispersion for count data can be carried out in different ways depending on the goal of the statistical analysis. Since the focus of this book is to illustrate how to formulate and apply geostatistical models, the

most natural approach to assess overdispersion is through the use of generalized linear mixed models (GLMMs). The class of GLMMs that we consider in this and the next section are obtained by replacing the spatial Gaussian process  $S(x_i)$  introduced in Equation 1.4 with a set of mutually independent random effects, which we denote as  $Z_i$ , and thus write

$$g(\mu_i) = d(x_i)^\top \beta + Z_i. \quad (3.7)$$

The model above accounts for the overdispersion in the data through  $Z_i$  whose variance can be interpreted as an indicator of the amount of overdispersion. To show this, we carry out a small simulation as follows. For simplicity, we consider the Binomial mixed model with an intercept only, hence

$$\log \left\{ \frac{\mu_i}{1 - \mu_i} \right\} = \beta_0 + Z_i \quad (3.8)$$

and assume that the  $Z_i$  follow a set of mutually independent Gaussian variables with mean 0 and variance  $\tau^2$ . In our simulation we vary  $\beta_0$  over the set  $\{-3, -2, -1, 0, 1, 2, 3\}$  and set  $\tau^2 = 0.1$  and the binomial denominators to  $n_i = 100$ . For a given value of  $\beta_0$ , we then proceed through the following iterative steps.

- Simulate 10,000 values for  $Z_i$  from a Gaussian distribution with mean 0 and variance  $\tau^2$ .
- Compute the probabilities  $\mu_i$  based on Equation 3.8.
- Simulate 10,000 values from a Binomial model with probability of success  $\mu_i$  and denominator  $n_i$ .
- Compute the empirical variance of the counts  $y_i$  simulated in the previous step.
- Change the value of  $\beta_0$  and repeat the previous steps, for all the values of  $\beta_0$ .

The code below shows the implementation of the above steps in R.

```
# Number of simulations
n_sim <- 10000

# Variance of the Z_i
tau2 <- 0.1

# Binomial denominator
bin_denom <- 100
```

```
# Intercept values
beta0 <- c(-3, -2, -1, 0, 1, 2, 3)

# Vector where we store the computed variance from
# the simulated counts from the Binomial mixed model
var_data <- rep(NA, length(beta0))

for(j in 1:length(beta0)) {
  # Simulation of the random effects Z_i
  Z_i_sim <- rnorm(n_sim, sd = sqrt(tau2))

  # Linear predictor of the Binomial mixed model
  lp <- beta0[j] + Z_i_sim

  # Probabilities of the Binomial distribution conditional on
  # Z_i
  prob_sim <- exp(lp)/(1+exp(lp))

  # Simulation of the counts from the Binomial mixed model
  y_i_sim <- rbinom(n_sim, size = bin_denom, prob = prob_sim)

  # Empirical variance from the simulated counts
  var_data[j] <- var(y_i_sim)
}

# Probabilities from the standard Binomial model (Z_i = 0)
probs_binomial <- exp(beta0)/(1+exp(beta0))

# Variance from the standard Binomial model
var_bimomial <- bin_denom*probs_binomial*(1-probs_binomial)

matplot(beta0, cbind(var_data, var_bimomial), type = "b", pch =
  20,
  lty = "solid", ylab = "Variance", xlab =
  expression(beta[0]))
legend(-3, 80, c("Binomial mixed model", "Standard Binomial
model"),
  col=1:2, lty = "solid", cex = 0.75)
```

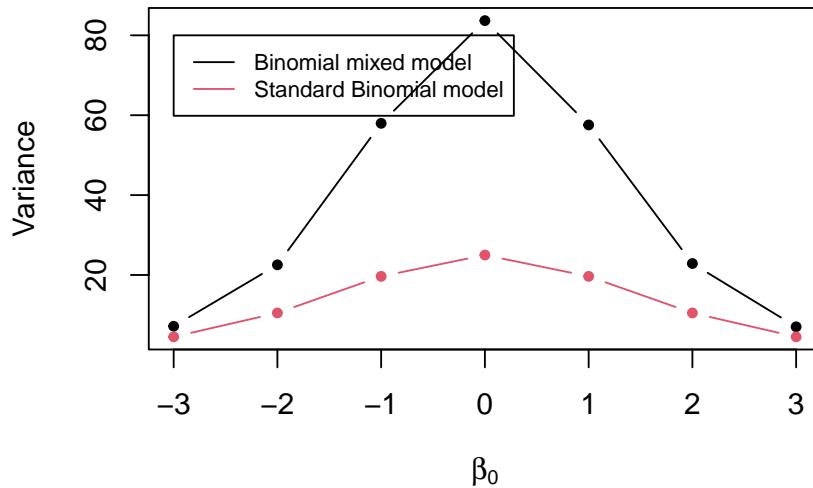


Figure 3.6: Plot of the variances of the standard Binomial model and the Binomial mixed model (see Equation 3.8) against  $\beta_0$

Figure 3.6 shows the results of the simulation. In this figure, the red line corresponds to the variance of a standard Binomial model, obtained by setting  $Z_i = 0$  and computed as  $n_i \mu_i(1 - \mu_i)$  with  $\mu_i = \exp\{\beta_0\}/(1 + \exp\{\beta_0\})$ . As expected, this plot shows that the variance of the simulated counts from the mixed model in Equation 3.8 exhibit a larger variance than would be expected under the standard Binomial model. It also indicates that the chosen value for the variance of  $Z_i$  of  $\tau^2 = 0.1$  corresponds to a significant amount of dispersion. One way to relate  $\tau^2$  to the amount of overdispersion is by considering that, following from the properties of a univariate Gaussian distribution, *a priori* the  $Z_i$  will take values between  $-1.96\sqrt{\tau^2}$  and  $+1.96\sqrt{\tau^2}$  with approximately 95% probability. That implies that  $\exp\{Z_i\}$ , which expresses the effect of the random effects on the odds ratios, will be with 95% probability between  $\exp\{-1.96\sqrt{\tau^2}\}$  and  $\exp\{+1.96\sqrt{\tau^2}\}$ . By replacing  $\tau^2$  with the chosen values for the simulation, those two becomes about 0.54 and 1.86, meaning that with the  $Z_i$  with 95% probability will have a multiplicative effect on the odds ratios between 0.54 and 1.86.

We encourage you to do Exercise 1 and Exercise 2 at the end of this chapter, to further explore how generalized linear mixed models can be used as a tool to account for overdispersion.

### 3.1.2.1 Maximum likelihood estimation of generalized linear mixed models for count data

We now illustrate how to fit a generalized linear mixed, using the `anopheles` data-set as an example. We consider two models: an intercept-only model and one that uses elevation as a covariate. Let  $\mu(x_i)$  be the number of mosquitoes captured at a location  $x_i$ ; then the linear predictor with elevation as a covariate takes the form

$$\log\{\mu_i\} = \beta_0 + \beta_1 d(x_i) + Z_i \quad (3.9)$$

where  $d(x_i)$  indicates the elevation in meters at location  $x_i$  and the  $Z_i$  are independent and identically distributed Gaussian variables with mean 0 and variance  $\tau^2$ . The model with an intercept only is simply obtained by setting  $\beta_1 = 0$ .

We carry out the estimation in R using the `glmer` function from the `lme4` package (see Bates et al. (2015) for a detailed tutorial). The `glmer` function implements the maximum likelihood estimation for generalized linear mixed models. The code below shows how the `glmer` is used to carry out this step for the model in Equation 3.9 and the one without covariates.

```
# Create the ID of the location
anopheles$ID_loc <- 1:nrow(anopheles)

# Poisson mixed model with elevation
fit_glmer_elev <- glmer(An.gambiae ~ scale(elevation) +
  (1|ID_loc), family = poisson,
  data = anopheles, nAGQ = 25)

# Poisson mixed model with intercept only
fit_glmer_int <- glmer(An.gambiae ~ (1|ID_loc), family =
  poisson,
  data = anopheles, nAGQ = 25)
```

To fit the model with `glmer`, we first must create a variable in our data-set that allows us to identify the location associated with each count. In this case, since every row corresponds to a different location, we simply use the row number to identify the locations and save this in the `ID_loc` variable. The random effects  $Z_i$  are then included in the model by adding `(1 | ID_loc)` in the formula of the `glmer` function.

When introducing the variable elevation, we standardize the variable so that its mean is 0 and its variance is 1. This is done to aid the convergence of the algorithm used to fit the model and it is generally considered good practice, especially when many variables with different scales are used as covariates. However, we emphasize that standardizing a variable does not affect the fit

of the model to the data. This is because the model with the standardized variable is a reparametrization of the model with the unstandardized variable. In other words, a model that uses standardized covariates only attaches a different interpretation to its regression coefficients while maintaining the same goodness of fit of the model with that uses the covariates on their original scale.

The argument `nAGQ` is used to define the precision of the approximation of the maximum likelihood estimation algorithm. By default `nAGQ = 1`, which corresponds to the Laplace approximation. Values for `nAGQ` larger than 1 are used to define the number of points of the adaptive Gaussian-Hermite quadrature. The general principle is that the larger `nAGQ` the better, but at the expense of an increased computing time. Based on the guidelines and help pages of the `lme4` package, it is stated that a reasonable value for `nAGQ` is 25. For more technical details on this aspect, we refer the reader to Bates et al. (2015).

We can now look at the summary of the fitted models to the mosquitoes data-set.

```
### Summary of the model with elevation
summary(fit_glmer_elev)
## Generalized linear mixed model fit by maximum likelihood
#>   (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]
## Family: poisson ( log )
## Formula: An.gambiae ~ scale(elevation) + (1 | ID_loc)
##   Data: anopheles
##
##       AIC      BIC  logLik deviance df.resid
##   291.8    300.1   -142.9     285.8      113
##
## Scaled residuals:
##       Min     1Q Median     3Q    Max
## -0.89574 -0.42469 -0.09483  0.29445  0.53352
##
## Random effects:
## Groups Name        Variance Std.Dev.
## ID_loc (Intercept) 0.7146   0.8453
## Number of obs: 116, groups: ID_loc, 116
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.53042   0.09365 16.342 <2e-16 ***
## scale(elevation) -0.19794   0.08950 -2.212   0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
#> 1
```

```

## 
## Correlation of Fixed Effects:
##          (Intr)
## scale(lvtn) 0.036

### Summary of the model with the intercept only
summary(fit_glmer_int)
## Generalized linear mixed model fit by maximum likelihood
##   (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]
## Family: poisson  ( log )
## Formula: An.gambiae ~ (1 | ID_loc)
## Data: anopheles
##
##      AIC      BIC  logLik deviance df.resid
## 294.6    300.1   -145.3     290.6     114
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -0.73816 -0.42718 -0.06941  0.26564  0.45022
##
## Random effects:
## Groups Name        Variance Std.Dev.
## ID_loc (Intercept) 0.761    0.8724
## Number of obs: 116, groups: ID_loc, 116
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.52849   0.09584 15.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##                1

```

From the summary of the model that uses elevation, we observe that the estimated regression coefficient  $\beta_1$  is statistically significant different from 0. The interpretation of the estimated regression coefficient is the following: for an increase of about 100 meters in elevation, all other things being equal, the average number of mosquitoes decreases by about  $100\% \times [1 - \exp\{-0.19794\}] \approx 18\%$ . Note that when using a standardized variable, a unit increase for this corresponds to an increase in the original unstandardized variable equal to its standard deviation, which for the `elevation` variable is about 100 meters.

From the summaries of the two models, under `Random effects:`, we obtain the estimates associates with the random effects introduced in the model. In this case, since we only have introduced  $Z_i$ , this part of summary provides the

maximum likelihood estimate for  $\tau^2$ , the variance of  $Z_i$ , which found on the line where `ID_loc` is printed. We then observe that the estimates for  $\tau^2$  for the intercept-only model is 0.761, whilst for the model with elevation this is 0.7146. Note that the figures reported under `Std.Dev.` are simply the square root of the value reported under `Variance`. As expected, the introduction of elevation contributes to the explanation of the residual variation captured by  $Z_i$ , though by a very small amount. The estimated values of  $\tau^2$  thus suggest that there is extra-Binomial variation in the data that is not account for by elevation.

In the next section, we will illustrate how to assess the presence of residual correlation for continuous measurements and overdispersed count data.

### 3.1.3 Exploring residual spatial correlation

In its most basic form, the concept of spatial correlation can be succinctly encapsulated by Tobler (1970) first law of geography, which posits that “everything is interconnected, but objects in close proximity exhibit stronger relationships than those situated farther apart.” After we have identified the key variables to introduce as covariates in the model (Section 3.1.1) and, in the case of count data, assessed the presence of overdispersion (Section 3.1.2), our final exploratory step consists of assessing whether the residuals of the non spatial model show evidence of spatial correlation. Hence, in geostatistical modelling, the interest is not in the spatial correlation of the data, but rather on understanding whether the variation in the outcome unexplained by the covariates exhibits spatial correlation. We call this *residual spatial correlation*, to emphasize that spatial correlation is a concept relative to the covariates that we have introduced in the model.

In the context of geostatistical analysis, the tool that is generally used to assess the residual spatial correlation is the so called *empirical variogram*. Before looking at the mathematical definition of the empirical variogram, let us consider a generalized linear mixed model as expressed in Equation 3.7. Our goal is then to question the assumption of independently distributed random effects  $Z_i$  by asking whether the  $Z_i$  show evidence of spatial correlation. Let  $Z_i$  and  $Z_j$  be two random effects that are associate with two different locations  $x_i$  and  $x_j$ , respectively, and let us take the squared difference between the two

$$V_{ij} = (Z_i - Z_j)^2. \quad (3.10)$$

How does the spatial correlation affect the value of  $V_{ij}$ ? To answer this question, we can refer to the aforementioned Tobler’s law of geography. When  $x_i$  and  $x_j$  will be closer to each other, then  $Z_i$  and  $Z_j$  will also tend to be more similar to each other, thus making  $V_{ij}$  smaller, on average. On the contrary, when  $x_i$  and  $x_j$  will be further apart, then  $V_{ij}$  will become larger, on average. We can then construct the empirical variogram by considering all possible

pairs of locations  $x_i$  and  $x_j$ , for which we then compute  $V_{ij}$  and plot this against the distance between  $x_i$  and  $x_j$ , which we denote as  $u_{ij}$ . If there is spatial correlation in the random effects  $Z_i$ , then this will manifest as an average increase in the  $V_{ij}$  as  $u_{ij}$  increases. However, there are still two issues that we have to address before we can generate and plot the empirical variogram.

The first issue is that we do not observe  $Z_i$  as, by definition, this is a latent variable. Hence, we require an estimate for  $Z_i$  which we can then feed into  $V_{ij}$ . To emphasize this point, from now on, we shall replace Equation 3.10 with

$$\hat{V}_{ij} = (\hat{Z}_i - \hat{Z}_j)^2. \quad (3.11)$$

Several options are available for estimating  $Z_i$ . Our choice is to use the model of the predictive distribution of  $Z_i$ , that is the distribution of  $Z_i$  conditioned to the data  $y_i$ . This estimator for  $Z_i$  is also readily available from the output of the `lmer` and `glmer` functions of the `lme4` package, as we will illustrate later in our example in this section.

The second issue is that if simply plot  $\hat{V}_{ij}$  against the distances  $u_{ij}$  (also known as *cloud variogram*), due to the high noiseness in the  $\hat{V}_{ij}$ , it may be quite difficult to assess the presence of an increasing trend in the  $\hat{V}_{ij}$  and thus detect spatial correlation. Hence, it is general practice to group the distances  $u_{ij}$  into classes, say  $\mathcal{U}$ , and then take average of all the  $\hat{V}_{ij}$  that fall within  $\mathcal{U}$ .

We can now write the formal definition of the empirical variogram as

$$\hat{V}(\mathcal{U}) = \frac{1}{2|\mathcal{U}|} \sum_{(i,j):(u_i, u_j) \in \mathcal{U}} \hat{V}_{ij} \quad (3.12)$$

where  $|\mathcal{U}|$  denotes the number of pairs of locations that fall within the distance class  $\mathcal{U}$ . The rationale behind dividing by 2 in  $1/2|\mathcal{U}|$  from the above equation, will be elucidated in Section 3.2, and there is no need for us to delve into this matter at this juncture. When creating the empirical variogram plot, we select the midpoint values of the distance classes  $\mathcal{U}$  to represent the x-axis values.

Before we can evaluate residual spatial correlation, there remains one crucial concern: relying solely on a visual inspection of the empirical variogram is susceptible to human subjectivity. Furthermore, it is worth noting that even a seemingly upward trend observed in the empirical variogram might be merely a product of random fluctuations, rather than a reliable indication of actual residual spatial correlation. To address these concerns and enhance the objectivity of the use of the empirical variogram, one approach would involve comparing the observed empirical variogram pattern with those generated in the absence of spatial correlation. Following this principle, we then use a permutation test that allows us to generate empirical variograms under the assumption of absence of spatial correlation through the following iterative steps.

1. Permute the order of the locations in the data-set while keeping everything else fixed.
2. Compute the empirical variogram  $\hat{V}(\mathcal{U})$  for the permuted data-set.
3. Repeat 1 and 2 a large number of times, say 10,000.
4. Use the resulting 10,000 empirical varigograms to compute 95% confidence intervals, by taking the 0.025 and 0.975 quantiles of these for each distance class  $\hat{V}(\mathcal{U})$ .
5. If the observed empirical variogram falls fully within the envelope generated in the previous point, we then conclude that the data do not exhibit residual spatial correlation. If, instead, the observed empirical variogram partly falls outside the envelope we conclude that the data do exhibit residual spatial correlation.

We now show an application of all the concepts introduced in this section to the Liberia data on river-blindness.

### 3.1.3.1 Example: assessing spatial correlation for the Liberia data

We consider the Binomial mixed model that uses the log-transformed elevation as a covariate to model river blindness prevalence, hence

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = \beta_0 + \beta_1 \log\{e(x_i)\} + Z_i \quad (3.13)$$

where  $e(x_i)$  is the elevation in meters at location  $x_i$  and the  $Z_i$  are i.i.d. Gaussian variables with mean 0 and variance  $\tau^2$ . We first fit the model above using the `glmer` function.

```
# Convert the data-set into an sf object
liberia <- st_as_sf(liberia, coords = c("lat", "long"), crs =
  ↵ 4326)

# Create the ID of the location
liberia$ID_loc <- 1:nrow(liberia)

# Binomial mixed model with log-elevation
fit_glmer_lib <- glmer(cbind(npos, ntest) ~ log(elevation) +
  ↵ (1|ID_loc), family = binomial,
  data = liberia, nAGQ = 25)

summary(fit_glmer_lib)
```

Generalized linear mixed model fit by maximum likelihood  
 (Adaptive Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]  
 Family: binomial ( logit )  
 Formula: cbind(npos, ntest) ~ log(elevation) + (1 | ID\_loc)  
 Data: liberia

AIC	BIC	logLik	deviance	df.resid
127.9	135.4	-61.0	121.9	87

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.46033	-0.63341	-0.07633	0.61995	3.12732

Random effects:

Groups	Name	Variance	Std.Dev.
ID_loc	(Intercept)	0.003097	0.05565

Number of obs: 90, groups: ID\_loc, 90

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.96292	0.21184	-13.987	< 2e-16 ***
log(elevation)	0.26143	0.04071	6.422	1.35e-10 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' '

Correlation of Fixed Effects:

(Intr)
log(elevtn) -0.981

From the output, we observe that the estimate for  $\tau^2$  is about 0.003, indicating a moderate level of overdispersion in the data.

```
liberia$Z_hat <- ranef(fit_glmer_lib)$ID_loc[,1]
```

Through the function `rانef` when extract the estimates of the random effects  $Z_i$  and save these in the data set. We then use the function `s_variogram` from the `RiskMap` package to compute the empirical varigogram for the estimated  $\hat{Z}_i$ .

```
scale_to_km = TRUE,
n_permutation = 10000)
```

Through the argument `bins` we can specify the the classes of distance, previously denoted by  $\mathcal{U}$ ; check the help page of `s_variogram` to see how this is defined by default. The value passed to `bins` in the code above correspond to define the following classes of distance  $\mathcal{U}$ : [15, 30], (30, 40] and so forth, with the last class being [350,  $+\infty$ ), i.e. all pairs of locations whose distances are above 350km. The argument `n_permutation` allows the user to specify the number of permutations that are performed the generate the envelope for absence of spatial correlation previously described.

```
dist_summaries(data = liberia,
               scale_to_km = TRUE)
## $min
## [1] 3.34536
##
## $max
## [1] 533.0733
##
## $mean
## [1] 206.7424
##
## $median
## [1] 192.6496
```

The `dist_summaries` function within the `RiskMap` package can be used for gauging the extent of the area covered by your dataset, aiding in the selection of appropriate values to be passed to the `bins` argument. In the provided output above, we can observe that for the Liberia dataset, the minimum and maximum distances span approximately 3km and 533km, respectively. While there is not a one-size-fits-all recommendation for setting `bins`, two fundamental principles should inform your decision-making. Firstly, it is advisable to avoid choosing overly large distance intervals, as the uncertainty associated with the empirical variogram tends to increase with distance due to fewer available pairs of observations for estimation. Secondly, especially when spatial correlation is not strong, it is crucial to carefully explore the behavior of the variogram at smaller distances. Consequently, it is generally advisable to experiment with different `bins` configurations and observe how they impact the pattern of the empirical variogram.

```
plot_s_variogram(liberia_variog,
                  plot_envelope = TRUE)
```

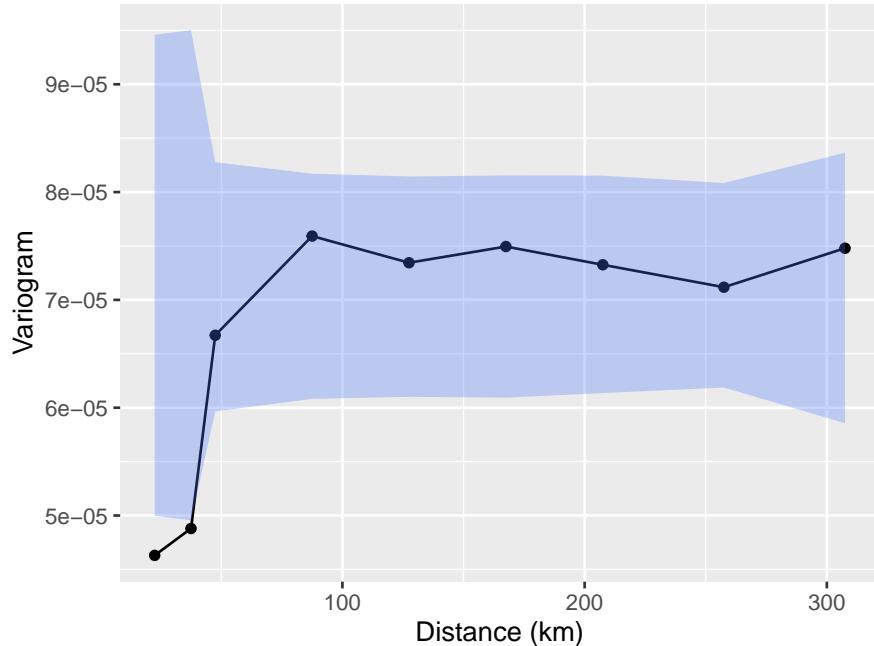


Figure 3.7: Plot of the empirical variogram (solid line) computed using the estimated random effects from the model in Equation 3.13. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

Finally, the `plot_s_variogram` function enables us to visualize the empirical variogram and, through the `plot_envelope` argument, include the envelope generated by the permutation procedure. As illustrated in Figure 3.7, we observe that the empirical variogram falls outside the envelope at relatively short distances, typically below 30km. However, for distances exceeding 30km, the behavior of the empirical variogram does not significantly differ from variograms generated under the assumption of spatial independence. In summary, we interpret the evidence presented in Figure 3.7 as indicative of residual spatial correlation within the data. Nevertheless, it is essential to exercise caution when attempting to ascertain the scale of the spatial correlation using the empirical variogram. As we will emphasize throughout this book, the empirical variogram's sensitivity to the choice of bins values renders it an unreliable tool for drawing statistical inferences. In other words, we advocate employing the empirical variogram primarily to assess the presence of residual correlation.

### 3.1.3.2 Exploring residual spatial correlation with linear Gaussian models

When using a linear model to assess spatial correlation, it is important to distinguish two cases: 1) when the data contain only one location per location; 2) when more than one observation per location is available. We now consider each of these two scenarios separately.

#### *One observation per location*

To illustrate the use of the variogram under this scenario, we shall use the Galicia data on lead concentration in moss samples. The simplest possible model for the data is a standard linear model without covariates which assumes independence among the observations, hence

$$Y_i = \beta_0 + U_i \quad (3.14)$$

where the  $U_i$  are i.i.d. Gaussian variables with mean zero and variance  $\omega^2$ . At this stage, our goal is then to assess whether the assumption of independence for the  $Z_i$  is supported by the data or whether there is evidence of spatial correlation. However, the measurement error of the device may be present as part of the natural random variation in  $Z_i$  which might mask the detection of spatial correlation using the residuals  $Z_i$  challenging, especially if the measurement error dominates the spatial variation of the data. One would be tempted to introduce an additional, location-specific random effect, say  $Z_i$  with mean zero and variance  $\tau^2$  and fit the model

$$Y_i = \beta_0 + Z_i + U_i, \quad (3.15)$$

where we interpret  $U_i$  as random variation due to the measurement device and  $Z_i$  as a random effect accounting for unmeasured covariates that contribute to the variation between locations in lead concentration. However, the model in Equation 3.15 is not identifiable because we cannot disentangle the separate contributions of  $Z_i$  and  $U_i$  from the variation of the data, unless: a) we know the precision of the measurement device,  $\omega$  (recall that  $\omega^2$  is the variance of  $U_i$ ); b) or, if we do not know  $\omega$ , we can then separate the two sources of variation only if we have multiple observations per location (the scenario which we shall consider in the next section).

For the Galicia data, for we do not know the measurement device precision and we only have one observation per location. However, this does not prevent us from using the variogram based on the residuals from Equation 3.14, while keeping in mind the limitations and uncertainty that are inherent to this exploratory tool as remarked at the end of the last paragraph.

```
# Fitting of the linear model and extraction of the residuals
lm_fit <- lm(log(lead) ~ 1, data = galicia)
galicia$residuals <- lm_fit$residuals

# Convert the galicia data frame into an "sf" object
galicia_sf <- st_as_sf(galicia, coords = c("x", "y"), crs =
  ↵ 32629)

# Compute the variogram, using the residuals from the linear
# model fit,
# and the 95% confidence level envelope for spatial independence
galicia_variog <- s_variogram(galicia_sf, variable =
  ↵ "residuals",
  scale_to_km = TRUE,
  bins = seq(10, 140, length = 15),
  n_permutation = 10000)

# Plotting the results
plot_s_variogram(galicia_variog, plot_envelope = TRUE)
```

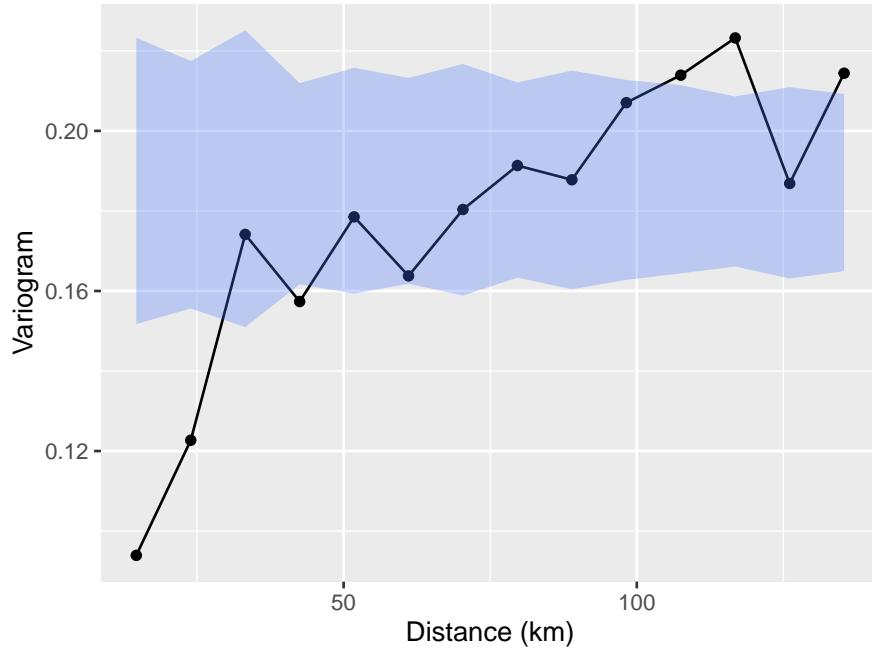


Figure 3.8: Plot of the empirical variogram (solid line) computed using the estimated residuals from the model in Equation 3.14. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

In the code above, we first fit the linear model in Equation 3.14 and then extract the residuals from this. Note that for this simple model, the residuals of the model are obtained by simply centering the outcome to zero by subtracting its mean. However, for more complex linear models that use covariates the computation of the residuals is more involved and can be carried out through a simple extension of the code above by specifying an appropriate `formula` in the `lm` function.

Figure 3.8 shows the empirical variogram and the 95% envelope for spatial independence. This clearly shows that the measurement of lead concentration are spatially correlated.

#### *More than one observation per location*

For the case of more than one observation per location, we shall consider the `italy_sim` data-set. This data-set contains 10 observations per location, for a

total of 200 locations. The variable `ID_loc` is a numeric indicator that can be used to identify the location each observation belongs to. For this data-set, we use the population density, named `pop_dens`, as a log-transformed covariate; we leave you as an exercise to assess that is a reasonable modelling choice.

To specify a non-spatial mixed model for the outcome, we then use two subscripts:  $i$  to identify a given location;  $j$  to identify the  $j$ -th observation for a given location  $i$ . We denote as  $Z_i$  the location-specific random effect and as  $U_{ij}$  the random variation due to the measurement error inherent to each observation. Hence, we write

$$Y_{ij} = \beta_0 + \beta_1 \log\{d(x_i)\} + Z_i + U_{ij}, \quad (3.16)$$

where  $d(x_i)$  is the population density at location  $x_i$ ; as before, we use  $\tau^2$  and  $\omega^2$  to denote the variances of  $Z_i$  and  $U_{ij}$ , respectively. To fit this model to the data we use the `lmer` function from the `lme4` package.

```
# Fitting a linear mixed model to the italy_sim data-set
# See main text for model specification
lmer_fit <- lmer(y ~ log(pop_dens) + (1 | ID_loc), data =
  italy_sim)

summary(lmer_fit)
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ log(pop_dens) + (1 | ID_loc)
##   Data: italy_sim
##
## REML criterion at convergence: 3390.9
##
## Scaled residuals:
##       Min     1Q   Median      3Q     Max
## -2.85209 -0.64198 -0.01832  0.66014  3.01870
##
## Random effects:
##   Groups   Name        Variance Std.Dev.
##   ID_loc   (Intercept) 2.5006   1.5813
##   Residual           0.1959   0.4426
##   Number of obs: 2000, groups: ID_loc, 200
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -0.40448   0.57613 -0.702
## log(pop_dens) 1.33798   0.07643 17.506
##
## Correlation of Fixed Effects:
##              (Intr)
```

```
## log(pp_dns) -0.981

# Incorporating the estimated random effects into the data
italy_sim$rand_eff <- ranef(lmer_fit)$ID_loc[italy_sim$ID_loc,1]

# Converting the italy_sim data frame into an "sf" object
italy_sim_sf <- st_as_sf(italy_sim, coords=c("x1", "x2"), crs =
  ↪ 32634)

# Compute the variogram, using the random effects from the
# linear mixed model fit,
# and the 95% confidence level envelope for spatial independence
italy_sim_variog <- s_variogram(italy_sim_sf, variable =
  ↪ "rand_eff",
  scale_to_km = TRUE,
  n_permutation = 200)

# Plotting the results
plot_s_variogram(italy_sim_variog, plot_envelope = TRUE)
```

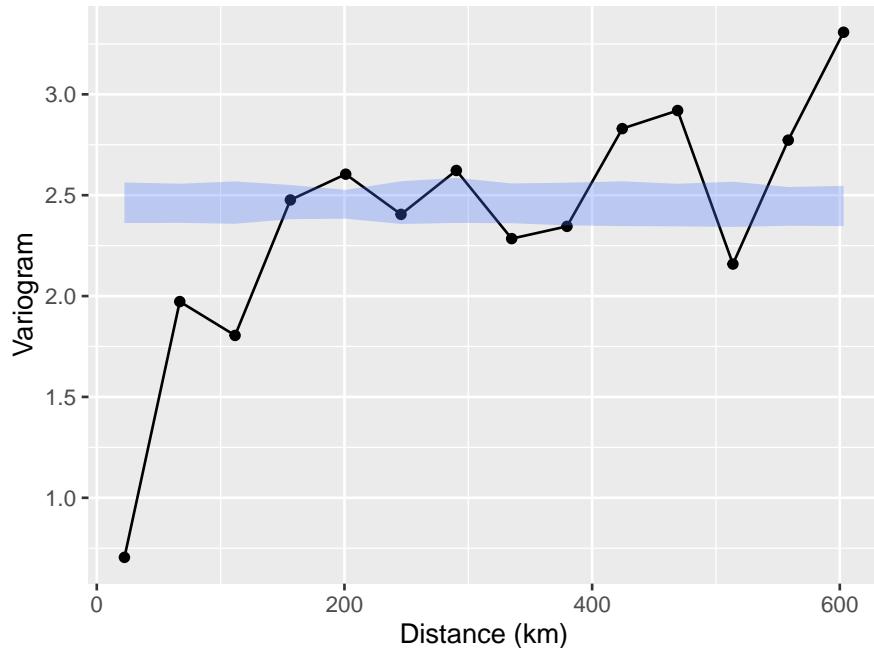


Figure 3.9: Plot of the empirical variogram (solid line) computed using the estimated random effects from the model in Equation 3.16. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

In the code above, the introduction of the  $Z_i$  random effect is specified in the `lmer` function through `(1|ID_loc)` in the formula; recall that `ID_loc` is the numerical indicator that identifies each location. From the summary of the model, we obtain both the estimates of the regression coefficients  $\beta_0$  and  $\beta_1$ , as well for the variances  $\tau^2$  listed under `Random effects::`. The estimate of  $\sigma^2$  is found on the line of printed output starting with `ID_loc`, whilst that for  $\$ \hat{\sim}^2$  is next to `Residual`.

The application of the empirical variogram, whose results are shown in Figure 3.9, indicate the presence of residual correlation. This is because we observe that the solid line representing the empirical variogram falls outside of the 95% envelope for spatial independence.

### 3.2 The linear geostatistical model

In this section, we consider spatially referenced outcomes  $Y_i$  that are continuous. We first consider the simpler case of a single measurement  $Y_i$  per location  $x_i$ . Recalling the class of generalized linear models introduced in Section 1.5, the linear predictor takes the form

$$\mu_i = d(x_i)^\top \beta + S(x_i). \quad (3.17)$$

Hence, in this case, we interpret  $\beta$  as the effect on  $\mu_i$  for a unit increase in  $d(x_i)$ . Let  $U_i$  denote i.i.d. random variables representing the measurement error associated with  $Y_i$ , each having mean zero and variance  $\omega^2$ . Thanks to the linear properties of Gaussian random variables, we can also express the linear model in a compact expression, as

$$Y_i = \mu_i + U_i = d(x_i)^\top \beta + S(x_i) + U_i. \quad (3.18)$$

To fully specify a geostatistical model for our dataset, we must address two critical aspects.

1. Defining the relationship between each covariate and the mean value  $\mu_i$ .
2. Selecting an appropriate correlation function for  $S(x)$ .

As illustrated in the previous sections, the initial step of exploratory analysis allows us to handle the first aspect, where we determine the regression relationship between covariates and the mean value  $\mu_i$ . However, based on existing methods of exploratory analysis, it is difficult to understand what is a suitable correlation function at this stage. A commonly recommended starting point is the Matern correlation function (see Equation 1.5), which offers considerable flexibility in capturing a wide range of correlation structures, under the assumption stationarity and isotropy. As we shall illustrate in the next example, even estimating a Matern correlation function is a task that poses many inferential challenges due to the poor identifiability, especially, of its smoothness parameter  $\kappa$ .

#### 3.2.1 Evaluating the inclusion of the measurement error term $U_i$ and the specification of the smoothness parameter $\kappa$

In this section we analyse the Galicia data using a linear geostatistical geostatistical model for the log-transformed lead concentration, which we denote as  $Y_i$ . Since we do not use covariates, we then write the model as

$$Y_i = \mu + S(x_i) + U_i \quad (3.19)$$

where were  $S(x)$  is a Matern process with variance  $\sigma^2$ , scale parameter  $\phi$  and smoothness parameter  $\kappa$ ; the  $U_i$  correspond to the measurement error term and denote with  $\omega^2$  their variance.

We carry out the parameter estimation of the model using the `glgpm` function from the `RiskMap` package. This function implements maximum likelihood estimation for generalized linear mixed models using a Matern correlation function, while fixing the smoothness parameter  $\kappa$  at prespecified value by the user. The object passed to the argument `data` in `glpmp` can either be a `data.frame` object or an `sf` object. Below we illustrate the use of `glgpm` while distinguishing between these two cases.

```
# Parameter estimation when the argument passed to `data` is a
# data-frame
fit_galicia <-
  glgpm(log(lead) ~ gp(x, y, kappa = 1.5), data=galicia, family =
    "gaussian",
    crs = 32629, scale_to_km = TRUE, messages = FALSE)
```

The code above shows the use of `glgpm` by passing `galicia` as a `data.frame` object to `data`. The specification of the Guassian process  $S(x)$  is done through the addition of the term `gp()` in the formula. The function `gp` allows you to specify the columns of the coordinates in the data, in this case `x` and `y`, the smoothness parameter through the argument `kappa`, set to 1.5 in this example. In the help page of `gp`, you can see that by default the nugget term (denoted in this book by the random variable  $Z_i$ ) is excluded from the model by default; to include and estimate the variance parameter of the nugget, you should set `nugget=NULL` in the `gp()` function. However, doing so for a linear model that only has one observation per location will generate error message as this is a non-identifiable model for the same reasons given in Section 3.1.3.2. However, if the measurement error variance is known this can be fixed by the user using the argument `fix_var_me` and the inclusion of the nugget term is then possible (to better understand this point, try Exercise 7 at the end of this chapter).

The argument `crs` is used to specify the coordinate reference system (CRS) of the data. For the Galicia data, as well as for every other data-set used in this book, the CRS is reported in the help page description of the data-set. If `crs` is not specified, the function will assume that the coordinates are in longitude/latitude format and will use these without applying any transformation. Finally, the argument `scale_to_km`, is used to specify whether the distances between locations should be scaled to kilometers or maintained in meter; this argument will not affect the scale, and thus the interpretation of the spatial correlation parameter  $\phi$ .

```
# Parameter estimation when the argument passed to `data` is an
# sf object
galicia_sf <- st_as_sf(galicia, coords = c("x", "y"), crs =
# 32629)

fit_galicia_sf <-
glgpm(log(lead) ~ gp(kappa = 1.5), data=galicia_sf, family =
# "gaussian",
      scale_to_km = TRUE, messages = FALSE)
```

The code above shows the alternative approach to estimate the model, when the argument passed to `data` is an `sf` object. In this case, the data-set `galicia` is converted into an `sf` object before the use of the `glgpm` function using `st_as_sf`. When when then fit the linear geostatistical model with `galicia_sf`, the only differences with the previous chunk of code that used `galicia` instead, is that the coordinates names in `gp()` and the `crs` argument in `glgpm` do not need to be specified as they are both directly obtained from `galicia_sf`.

```
summary(fit_galicia)
## Linear geostatsitical model
## 'Lower limit' and 'Upper limit' are the limits of the 95%
# confidence level intervals
##
## Regression coefficients
## Estimate Lower limit Upper limit StdErr z.value
# p.value
## (Intercept) 0.707418 0.552762 0.862075 0.078908 8.9651
# < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
# 1
## Offset included into the linear predictor
##
## Estimate Lower limit Upper limit
## Measurement error var. 0.0154636 0.0051797 0.0462
##
## Spatial Guassian process
## Matern covariance parameters (kappa=1.5)
## Estimate Lower limit Upper limit
## Spatial process var. 0.17127 0.13303 0.2205
## Spatial corr. scale 9.02085 7.59521 10.7141
## Variance of the nugget effect fixed at 0
##
```

```
## Log-likelihood: 69.03029
##
## AIC: -138.0606
##
```

We can then inspect the point and interval estimates for the model through the `summary` function of the model as shown in the code chunk above. This outputs is presented in three sections: in the first section, we have the results for the regression coefficients; in the second section, we have the estimate for the variance of the measurement error component,  $\omega^2$  whose point estimate is about 0.015; in the final section, we have the estimates for the parameters of the spatial covariance function,  $\sigma^2$  and  $\phi$  which are about 0.171 and 9.021 (km), respectively. The message `Variance of the nugget effect fixed at 0` indicates that the nugget has not been included in the mode; try Exercise 7 to see how this summary changes in the presence of the nugget term.

At this point, you may be wondering, why we have used 1.5 for the value of  $\kappa$  and whether there is a statistical approach to find the most suitable value for this. We show such an approach in the code below. However, before examining the code, we would like to point an important aspect that relates to how the value of  $\kappa$  can affect the estimate of the measurement error variance  $\omega^2$ . As we have shown, in Section 1.5.1, values of  $\kappa$  that are closer to zero will give a rougher and less regular surface for  $S(x)$ . In the case of a single observation, when  $\kappa$  is closer to zero, it will thus become increasingly difficult to estimate  $\omega^2$  since the likelihood function may attribute most of the noisiness to  $S(x)$  and rely less on  $U_i$  to explain the unstructured random variation found in the data. As a consequence of this, we can expect that for value of  $\kappa$  closer to zero, the estimates of  $\omega^2$  will also be smaller and, on the contrary, when  $\kappa$  is larger, the estimates of  $\omega^2$  will also increase. The results generated in the below clearly illustrate this point.

```
# Number of the values chosen for kappa
n_kappa <- 10

# Set of values for kappa
kappa_values <- seq(0.5, 3.5, length = n_kappa)

# Vector that will store the values of the likelihood function
# evaluated at the maximum likelihood estimate
llik_values <- rep(NA, length = n_kappa)

# Vector that will store the maximum likelihood estimates
```

```

# of the variance of the measurement error
sigma2_me_hat <- rep(NA, length = n_kappa)

# List that will contain all the geostatistical models fitted
# for
# the different values of kappa specified in kappa_values
fit_galicia_list <- list()

for(i in 1:n_kappa) {
  fit_galicia_list[[i]] <- glgpm(log(lead) ~ gp(x, y, kappa =
  kappa_values[i]),
  data=galicia, family =
  "gaussian",
  crs = 32629, scale_to_km =
  TRUE, messages = FALSE)
  llik_values[i] <- fit_galicia_list[[i]]$log.lik
  sigma2_me_hat[i] <- coef(fit_galicia_list[[i]])["sigma2_me"]
}

```

By examining the results shown in panel (a) of Figure 3.10, we observe that, as expected, smaller values for  $\kappa$  leads to smaller point estimates for  $\omega^2$  and viceversa. This begs the question, what should be our chosen value for  $\kappa$ ?

To answer this question, a natural approach is to estimate the model for different values of  $\kappa$  and see which one give the best fit to the data, according to the likelihood function. In panel (b) of Figure 3.10, we show the results of this approach where we considered 10 values for  $\kappa$  within the range 0.5 to 3.5 (see code above). In this plot, we have also added a horizontal line to help us to approximate a range of the most plausible value for  $\kappa$ . More precisely, the horizontal dashed line is computed by taking the maximum observed values of the likelihoods computed for the different values of  $\kappa$ , say  $\hat{M}$  and we subtract the quantile 0.95 of a  $\chi^2$  distribution with 1 degree of freedom. In R this horizontal line is obtained as

```
max(llik_values)-pchisq(0.95, df = 2)/2
```

where `llik_values` is as defined in the previous chunk of code. Note that this approach is essentially constructing the profile likelihood for  $\kappa$  which one could use to derive a confidence interval for  $\kappa$  with a finer segmentation for `kappa_values`. However, our current objective is not to derive the confidence interval for  $\kappa$ , but rather to gain a broad understanding of the  $\kappa$  values supported by the dataset. The values of  $\kappa$  that corresponds to likelihood values above the horizontal line are approximately between 0.75 and 2.75. Hence, selecting  $\kappa = 1.5$  seems to be a reasonable one in this case. Now, you may

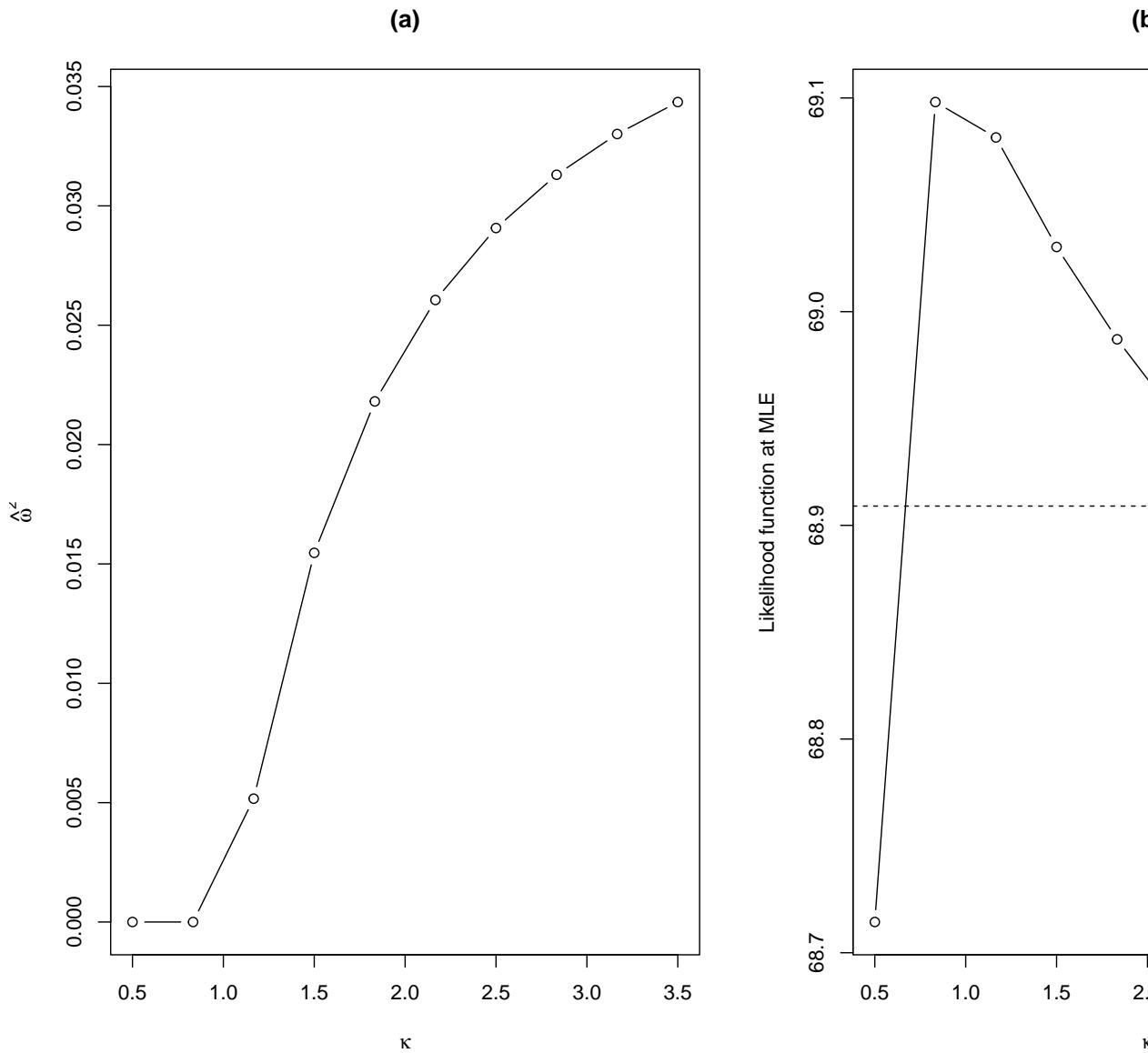


Figure 3.10: (a) plot of the maximum likelihood estimate for the variance of the measurement error  $U_i$ , denoted in the text as  $\omega^2$ , against the chosen fixed value for  $\kappa$ ; (b) profile likelihood for  $\kappa$  with the horizontal dashed line corresponding to the (approximate) threshold for constructing a 95% confidence interval based on a  $\chi^2$  with one degree of freedom.

be pondering: are there value other than  $\kappa = 1.5$  that could fit the data even better? Our answer is that it is not worth the effort to try estimate  $\kappa$  more precisely because it is empirically very difficult and, under some scenarios, even impossible. Estimating  $\kappa$  poses a well-documented challenge in geostatistics (Zhang (2004)), which justifies our adoption of a pragmatic approach that sets it at a predefined value. This issue is also further exacerbated when analyzing count data, which tend to be less informative about the correlation structure than continuously measured data.

### 3.2.2 Modelling hierarchical geostatistical data using the `re()` function

We now consider the analysis of geostatistical data with a hierarchical structure and show how to formulate and fit a geostatistical model that accounts for the effects of the different layers of the data. For this purpose, we use the `italy_sim` where each of the sampled locations can be grouped according to two administrative subdivisions of Italy, regions (Admin level 2) and provinces (Admin level 3), as shown in Figure 3.11.

```
library(rgeoboundaries)
library(mapview)

italy_regions <- geoboundaries(country = "italy", adm_lvl =
  ↵ "adm2")

italy_provinces <- geoboundaries(country = "italy", adm_lvl =
  ↵ "adm3")

par(mfrow = c(1,2))

# Map of the data with the region boundaries
map_regions <- ggplot() +
  geom_sf(data = italy_sim_sf, pch = 4, color = "red") +
  geom_sf(data = italy_regions, fill = NA) +
  theme_void() +
  labs(title = "Regions") +
  theme(plot.title = element_text(hjust = 1/2))

# Map of the data with the province boundaries
map_provinces <- ggplot() +
  geom_sf(data = italy_sim_sf, pch = 4, color = "red") +
  geom_sf(data = italy_provinces, fill = NA) +
  theme_void() +
  labs(title = "Provinces") +
  theme(plot.title = element_text(hjust = 1/2))
```

```
library(gridExtra)
grid.arrange(map_regions, map_provinces, ncol = 2)
```

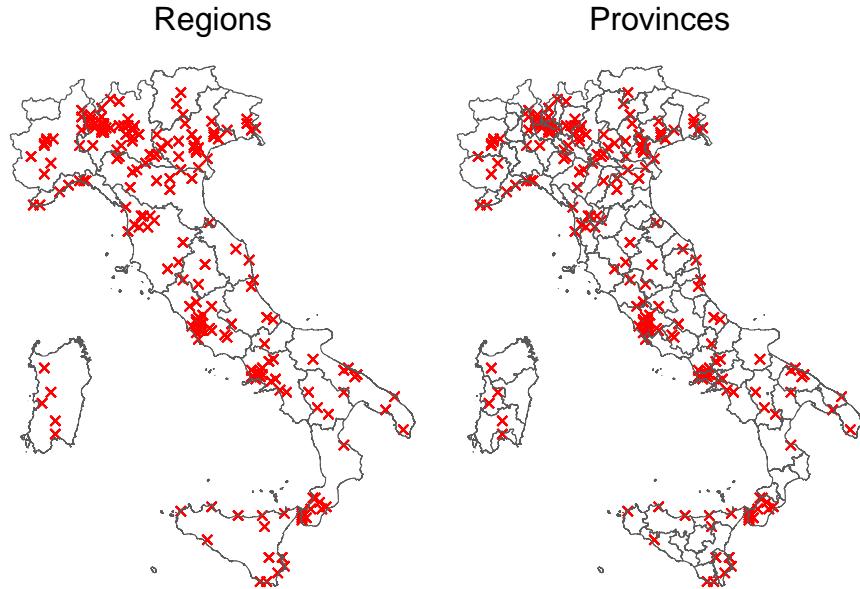


Figure 3.11: The plot shows the location of the data-set `italy_sim` (red crosses), with the boundaries of the regions (left) and provinces (right) of Italy.

Let  $V_h$ , for  $h = 1, \dots, n_V$  and  $Z_k$ , for  $k = 1, \dots, n_Z$ , be a set of mutually independent Gaussian variables with zero means and variances  $\sigma_V^2$  and  $\sigma_Z^2$ , respectively. Finally let  $\mathcal{A}_h$ , for  $h = 1, \dots, n_V$  and  $\mathcal{B}_k$ , for  $k = 1, \dots, n_Z$ , denote the areas encompassed by the boundaries of the region and provinces, respectively. Since we have 10 repeated observations at each locations, we shall use  $Y_{ij}$  to denote the  $j$ -th outcome (found in the data under the column  $y$ ) at the  $i$ -th location  $x_i$ . The model from which we generated data  $y_{ij}$  is

$$Y_{ij} = \beta_0 + \underbrace{\overbrace{\beta_1 d(x_i) + S(x_i)}^{\text{Location-level effect}}}_{\text{Region effect}} + \sum_{h=1}^{n_V} I(x_i \in \mathcal{A}_h) \times V_h + \underbrace{\sum_{k=1}^{n_Z} I(x_i \in \mathcal{B}_k) \times Z_k}_{\text{Province effect}} + \underbrace{U_{ij}}_{\text{Measurement error}}, \quad (3.20)$$

for  $j = 1, \dots, 10$  and  $i = 1, \dots, 200$ , where,  $d(x_i)$  is the log-transformed population density,  $I(x \in \mathcal{R})$  is an indicator function that takes value 1 if the location  $x$  falls within the boundaries of the area denoted by  $\mathcal{R}$  and 0 otherwise. The covariance function chosen in the generation of the data was an exponential correlation function hence  $\text{cov}\{S(x), S(x')\} = \sigma^2 \exp\{-||x - x'||/\phi\}$ .

The inclusion of the random effects  $V_h$ , for the region, and  $Z_k$ , for the province, can be included by using the `re()` function into the `formula` passed to `glgpm` as shown below.

```

# Location-level effect
italy_fit <- glgpm( y ~ log(pop_dens) + gp(kappa = 0.5, nugget =
  0) +
  # Region and province effects
  re(region, province),
  data = italy_sim_sf, scale_to_km = TRUE,
  family = "gaussian")
##   0:    441.00731: -0.404481  1.33798  0.00000  4.62406
##   0.00000  0.00000  0.00000
##   1:   -130.05497: -0.408825  1.35003 -0.0387110  4.66059
##   -0.910409 -0.00265292  0.0107965
##   2:   -268.13436: -0.865126  1.39525 -0.0305269  5.92133
##   -1.96079 -0.666581  0.487537
##   3:   -329.24560: -1.44091  1.38139 -0.546306  4.98616
##   -1.67936 -1.15202  0.383821
##   4:   -331.20861: -1.38498  1.37256  0.165463  5.61366
##   -1.63033 -2.11435  0.371263
##   5:   -331.28418: -0.874796  1.37322 -0.271990  5.18071
##   -1.62887 -0.970704  0.371407
##   6:   -331.42968: -1.04087  1.37294 -0.119652  5.33634
##   -1.62883 -1.33321  0.375031
##   7:   -331.43375: -1.07312  1.37288 -0.0878211  5.36725
##   -1.62883 -1.42011  0.376251
##   8:   -331.43375: -1.07445  1.37288 -0.0866254  5.36840
##   -1.62883 -1.42489  0.376323
##   9:   -331.43375: -1.07445  1.37288 -0.0866228  5.36840
##   -1.62883 -1.42491  0.376323
##  10:   -331.43375: -1.07445  1.37288 -0.0866227  5.36840
##   -1.62883 -1.42491  0.376323
##  11:   -331.43375: -1.07445  1.37288 -0.0866227  5.36840
##   -1.62883 -1.42491  0.376323

summary(italy_fit)
## Linear geostatistical model

```

```

## 'Lower limit' and 'Upper limit' are the limits of the 95%
## confidence level intervals
##
## Regression coefficients
##             Estimate Lower limit Upper limit     StdErr
## z.value p.value
## (Intercept) -1.074454   -2.031694   -0.117213  0.488397
##                -2.2 0.02781 *
## log(pop_dens) 1.372877    1.352018    1.393735  0.010642
##                129.0 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##                1
## Offset included into the linear predictor
##
##             Estimate Lower limit Upper limit
## Measurement error var. 0.19616      0.19361      0.1987
##
## Spatial Guassian process
## Matern covariance parameters (kappa=0.5)
##             Estimate Lower limit Upper limit
## Spatial process var. 0.91702      0.59470      1.414
## Spatial corr. scale 214.52032    153.60797    299.587
## Variance of the nugget effect fixed at 0
##
## Unstructured random effects
##             Estimate Lower limit Upper limit
## region (random eff. var.) 0.24053      0.14553      0.3976
## province (random eff. var.) 1.45692      1.37851      1.5398
##
## Log-likelihood: 331.4338
##
## AIC: -662.8675
##

```

In the code above, the factor variables `region` and `province` found in `italy_sim` are passed to `re()` and these are estimated as unstructured random effects as defined by Equation 3.20. From the summary of model, we then obtain the parameter and interval estimates as reported in Table 3.2.

Table 3.2: Maximum likelihood estimates and, lower and upper limits of the 95% confidence interval for the parameters of the model in Equation 3.16.

Parameter	Point estimate	Lower limit	Upper limit
$\beta_0$	-1.074	-2.032	-0.117
$\beta_1$	1.373	1.352	1.394
$\sigma^2$	0.917	0.595	1.414
$\phi$	214.520	153.608	299.587
$\sigma_V^2$	0.241	0.146	0.398
$\sigma_Z^2$	1.457	1.379	1.540
$\omega^2$	0.196	0.194	0.199

The function `to_table` from the `RiskMap` package can be used to obtain the Latex or HTML code directly from a fit of the model. Here is an example.

```
to_table(italy_fit, digits = 3)
## % latex table generated in R 4.4.1 by xtable 1.8-4 package
## % Sun Nov 3 02:15:02 2024
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrr}
## \hline
## & Estimate & Lower limit & Upper limit \\
## \hline
## (Intercept) & -1.074 & -2.032 & -0.117 \\
## log(pop\_dens) & 1.373 & 1.352 & 1.394 \\
## Spatial process var. & 0.917 & 0.595 & 1.414 \\
## Spatial corr. scale & 214.520 & 153.608 & 299.587 \\
## region (random eff. var.) & 0.241 & 0.146 & 0.398 \\
## province (random eff. var.) & 1.457 & 1.379 & 1.540 \\
## Measurement error var. & 0.196 & 0.194 & 0.199 \\
## \hline
## \end{tabular}
## \end{table}
```

### 3.3 Generalized linear geostatistical models

We now consider the modelling of count outcomes, focusing our attention to the Binomial and Poisson geostatistical models.

The use of Binomial geostatistical model is appropriate whenever the reported counts have a known upper bound. For example, when analyzing aggregated disease counts from cross-sectional survey data, the total number of people sampled at a location defines the largest number of possible reported disease cases. Recalling the model specification of Table 1.3, we specify the Binomial geostatistical model by stating that the reported disease counts  $Y_i$  out of  $n_i$  (the maximum value that  $Y_i$  can take) at location  $x_i$ , conditionally on the spatial random effects  $S(x_i)$ , follow a Binomial distribution with probability  $p(x_i)$  and logit-linear predictor

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = d(x_i)^\top \beta + S(x_i). \quad (3.21)$$

In this model,  $p(x_i)$  is commonly understood as representing the prevalence of the disease under consideration, defined as the proportion of individuals with the disease at a specific moment. While  $p(x_i)$  is not strictly a proportion, its interpretation as disease prevalence is justified by the fact that if we were to sample a sufficiently large population at location  $x_i$ , the proportion of individuals with the disease would converge closely to the value of  $p(x_i)$ .

A Binomial geostatistical model is also used when analyzing data at individual level where the outcome is binary and usually indicates the positive or negative result of a test for a disease under investigation. In this analysis, a mix of individual-level variables and location-level variables could be introduced to model the probability of a positive test. To formally write the model, we now use  $Y_{ij}$  to indicate the binary outcome taking value 1, if the test was positive, and 0, if negative, for the  $j$ -th individual sampled at location  $x_i$ . The linear predictor now takes the form

$$\log \left\{ \frac{p_{ij}}{1 - p_{ij}} \right\} = e_{ij}^\top \gamma + d(x_i)^\top \beta + S(x_i). \quad (3.22)$$

In the model above we have distinguished between covariates  $e_{ij}$ , that express the effect on  $p_{ij}$  due to individual traits (such as age and gender), and covariates  $d(x_i)$ , that instead quantify the effect on  $p_{ij}$  resulting from the environmental attributes of location  $x_i$  (such as elevation, temperature, and precipitation). In this model  $p_{ij}$  – the probability of a positive test – once again can be interpreted as the disease prevalence for a population characterized by the individual traits  $e_{ij}$  and residing at location  $x_i$ .

Finally, when the counts arise from a population whose size is unknown or when the occurrence of a disease is very small relative to the size of the target population, one should consider the use of a Poisson geostatistical model. We then use  $Y_i$  to denote the reported counts at location  $x_i$  and assume that this, conditionally on a Gaussian process  $S(x_i)$ , follows a Poisson distribution with mean  $\lambda(x_i)$  and log-linear predictor

$$\log \{\lambda(x_i)\} = d(x_i)^\top \beta + S(x_i). \quad (3.23)$$

If for example,  $Y_i$  corresponds to the number of new cases reported at a health facility at location  $x_i$  over a certain, the  $\lambda(x_i)$  is interpreted as the average number of new cases reported over that period. If the information on the size of the population at risk for the diseases were available, this can be incorporated into the model so as to enable the interpretation of  $\lambda(x_i)$  as the disease incidence, defined as the total number of new cases reported over a given period of time divided by the population at risk. If we denote the latter as  $m_i$ , then we would specify the mean of  $Y_i$  as  $m_i\lambda(x_i)$  an model  $\lambda(x_i)$  as already specified in Equation 3.23. Another prevalent application of the Poisson geostatistical model in epidemiology involves mapping the abundance of disease vectors. Here,  $\lambda(x_i)$  corresponds to the mean number of vectors per unit area. Also, note that in this last scenario  $m_i$  cannot be defined as it is the goal of the analysis to infer the vector population size at location  $x_i$ .

In all the three models, namely Equation 3.21, Equation 3.22 and Equation 3.23, one might consider the inclusion of an additional random effect  $Z_i$ , previously used for the linear geostatistical model, to account for small spatial variation or overdispersion that stems from the omission of covariates that may not necessarily exhibit spatial structure, such as behavioral or genetic traits. We shall further elaborate and provide guidance on this point in the following examples of this section.

### 3.3.1 Example: river-blindness in Liberia

In section Section 3.1.1.1, we carried out the exploratory analysis of the river-blindness prevalence data to assess how the covariate of elevation could be introduced into the model. In Section 3.1.3.1, we then tested for residual spatial correlation, when the log-transformed elevation is used as a covariate, using the empirical variogram. Based on these results, we now formulate a fit a Binomial geostatistical model to data. To achieve this, we extend the model presented in Equation 3.13 through the inclusion of a spatial Gaussian process,  $S(x)$ , hence we write

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = \beta_0 + \beta_1 \log\{e(x_i)\} + S(x_i) + Z_i. \quad (3.24)$$

We shall assume that  $S(x_i)$  is a stationary isotropic Gaussian process with an exponential correlation function function  $\rho(u) = \exp\{-u/\phi\}$ , with scale parameter  $\phi$ .

We apply the Monte Carlo Maximum Likelihood (MCML) method for parameter estimation, which is implemented via the `glgpm` function. In this section, we set aside the technical details of the estimation process to focus exclusively on interpreting the results. A more detailed explanation of the MCML theory can be found in Section 3.4.2, with a concise summary of the practical aspects provided in Section 3.4.2.3.

```
set.seed(1)
fit_liberia <-
  glgpm(npos ~ log(elevation) + gp(long, lat, nugget = NULL),
         distr_offset = ntest, data = liberia,
         crs = 4326,
         convert_to_crs = 32629,
         family = "binomial", messages = FALSE)

summary(fit_liberia)
```

```
Binomial geostatistical linear model
'Lower limit' and 'Upper limit' are the limits of the 95%
confidence level intervals

Regression coefficients
Estimate Lower limit Upper limit StdErr
z.value p.value
(Intercept) -2.948777 -3.376292 -2.521263 0.218123
-13.519 < 2.2e-16
log(elevation) 0.298443 0.257020 0.339867 0.021135
14.121 < 2.2e-16

(Intercept) ***
log(elevation) ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Offset included into the linear predictor

Spatial Guassian process
Matern covariance parameters (kappa=0.5)
Estimate Lower limit Upper limit
Spatial process var. 0.246161 0.191336 0.3167
Spatial corr. scale 60.068879 43.265064 83.3992
Variance of the nugget 0.101348 0.071311 0.1440

Log-likelihood: 39.78642
```

When using the `glgpm` function to fit a Binomial model, we need to specify the Binomial denominators, i.e. the number of tested at a location in this case, through the argument `distr_offset`. Other arguments of `glgpm` that are used for carrying out MCML are `control_mcmc`, that defines the setting parameters of the Monte Carlo Markov Chain algorithm, and `par0`, which represents our best guess for the unknown parameters of the Binomial geostatistical model. For more information on how the default values of these two arguments, we

refer the reader the to the help page of the `glgpm` function. To understand how to change the default parameters in order to improve the MCML estimation, we advise you to read Section 3.4.2.1. Also, remember that to enable the estimation of the nugget effect,  $Z_i$ , you should set `nugget=NULL` in the `gp` function, since by default the `glgpm` function excludes  $Z_i$  when the out is non-Gaussian.

In the above summary of the model, the interpretation of the regression coefficients is the same as in a standard logistic regression model. Hence, by taking the exponential of the estimate for the regression coefficient  $\beta_1$ , we could then express the effect for a unit increase in the covariate – in this case the log-transformed elevation – on the odds of disease prevalence. Finally, the interpretation of estimates of the covariance parameters is not affected by the type of distribution used to model the main outcome. From the above summary, we note the larger estimates of  $\sigma^2$  than  $\tau^2$  which suggest the dominance of spatial variation over the unstructured variation in the unexplained variation of disease prevalence. Finally, the estimate of  $\phi$  indicates a relatively large spatial correlation with a *practical range* close to  $(58.67 \times 3 \approx) 180\text{km}$ .

### 3.3.2 Example: *Anopheles Gambiae* mosquitoes in Cameroon

In Section 3.1.1.1, we explored the association between elevation and the number of mosquitoes and found that the data showed that on average, the number of “*Anopheles Gambiae*\* (*A. Gambiae*) decreases for increasing elevation. We now fit a log-linear Poisson geostatistical model, with linear predictor

$$\lambda(x_i) = \beta_0 + \beta_1 d(x_i) + S(x_i) \quad (3.25)$$

where  $d(x_i)$  correspond to the elevation in meters at location  $x_i$ . This time, we first convert the data frame `anopheles` into an `sf` object. This step is optional but recommended as it allows us to simply the code syntax.

```
anopheles <- st_as_sf(anopheles, coords = c("web_x", "web_y"),
                      crs = 3857)
```

We can then use the following script to fit the model in RiskMap.

```
set.seed(1)
an_fit <-
glgpm(An.gambiae ~ elevation + gp(),
      data = anopheles,
      family = "poisson", messages = FALSE)

summary(an_fit)
```

```
Poisson geostatistical linear model
'Lower limit' and 'Upper limit' are the limits of the 95%
confidence level intervals

Regression coefficients
            Estimate Lower limit Upper limit      StdErr
            z.value   p.value
(Intercept) 3.69384399  3.00784990  4.37983808  0.35000342
10.554 < 2.2e-16
elevation    -0.00283997 -0.00336073 -0.00231922  0.00026569
-10.689 < 2.2e-16

(Intercept) ***
elevation   ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Offset included into the linear predictor

Spatial Guassian process
Matern covariance parameters (kappa=0.5)
            Estimate Lower limit Upper limit
Spatial process var.  0.67165     0.59987     0.7520
Spatial corr. scale   4.38649     3.84756     5.0009
Variance of the nugget effect fixed at 0

Log-likelihood: 9.453628
```

The estimate,  $\hat{\beta}_1$ , for the regression coefficient  $\beta_1$  suggests that for a 100 meters drop in elevation, the average number of mosquitoes decreases by about  $(100 \times \exp\{100 \times -0.018\} \approx) 17\%$ . The order of magnitude of the estimates suggest the presence of residual spatial correlation, albeit on a rather small scale.

```
dist_summaries(anopheles, scale_to_km = TRUE)
```

```
$min
[1] 0.3865243

$max
[1] 161.9637

$mean
[1] 55.12497
```

```
$median
[1] 44.14232
```

Using the `dist_summaries` function, we can see that the average distance is about 55km and the minimum is about 387 meters. The estimate of phi of 4km is thus relatively small compared to the the distances between locations in the data. In Chapter 5, we will further investigate to what extent this small-scale spatial correlation can affect the predictions of the *An. Gambiae* counts.

### 3.4 Theory

#### 3.4.1 Maximum likelihood estimation for non-Gaussian mixed models with independent random effects

In this section, we provide more technical details on the derivation of the likelihood function for the mixed model specified in Equation 3.7. While not essential for understanding the application and interpretation of geostatistical models, this section provides useful insights into computationally challenges inherent to the models presented in this book. It is also important to point out that in this section we consider the case when the outcome  $Y_i$  is a count.

Let  $\theta = (\beta, \tau^2)$  denote the vectore of unkown parameters, where  $\beta$  is the vector of regression coefficients and  $\tau^2$  is the variance of the random effects  $Z = (Z_1, \dots, Z_n)$ . The likelihood function for a set of observations  $Y = (Y_1, \dots, Y_n)$ , for  $i = 1, \dots, n$ , under the model in Equation 3.7, is obtained by deriving the marginal distribution of  $Y_i$ , hence

$$L(\theta) = [Y] = \int_{R^n} [Z][Y | Z] dZ$$

Since the elements  $Z_i$  in  $Z$  are mutually independent, we can write

$$[Z] = \prod_{i=1}^n [Z_i]$$

and, using the conditional independence among the  $Y$  given  $Z$ , we obtain

$$[Y | Z] = \prod_{i=1}^n [Y_i | Z_i].$$

From these, it then follows that

$$L(\theta) = \int_{R^n} \prod_{i=1}^n [Z_i] \prod_{i=1}^n [Y_i | Z_i] dZ = \prod_{i=1}^n \int_R [Z_i] [Y_i | Z_i] dZ_i. \quad (3.26)$$

This shows that under the model in Equation 3.7, whilst accounting for overdispersion, this model does not account for any source of correlation between the observations. This is because the likelihood function can be expressed, as shown in the equation above, as the product of  $n$  likelihoods associated to each of the observations.

Another important observation is that the integrals that express the likelihood in Equation 3.26, cannot be solved analytically and need to be approximated. Several solution are available that either based on maximum likelihood estimation or Bayesian methods of inference. The function `glmer` used in Section 3.1.3.2 provides both the Laplace and Gaussian quadrature approximations options. The former consists of finding the *peak* of the integrand - i.e. the function that is being integrate - and approximating that with a Gaussian distribution centered at that *peak*, using the curvature of the original distribution to determine the width of the Gaussian distribution. The Gaussian quadrature is another numerical integration technique used to approximate the integral of a function. It involves choosing specific points (nodes) and weights within the integration interval to approximate the integral. By strategically selecting these points and weights based on a Gaussian distribution, Gaussian quadrature can provide highly accurate results compared to other numerical integration methods, such as the Laplace approximation. For a comprehensive overview of these and other methods of approximations of intractable integrals, we find that Bishop (2006) is one of the most accessible textbooks on this topic.

### 3.4.2 Maximum likelihood estimation for non-Gaussian generalized linear geostatistical models

To derive the likelihood function of a generalized linear geostatistical model (GLGM), we consider the special case of a single observation per location without loss of generality. Let  $Y$  and  $Z$  denote the same vectors of random variables introduced in the previous section, with components  $Y_i$  and  $Z_i$  corresponding to the set of locations  $X = (x_1, \dots, x_n)$ . Additionally, we define  $S(X) = (S(x_1), \dots, S(x_n))$  to represent the spatial Gaussian process at the sampled locations  $X$ . In this case the vector of unknown parameters is  $\theta = (\beta, \sigma^2, \phi, \tau^2)$  and, as done in the previous section, we can obtain the likelihood function through the marginal distribution of the data, hence we write

$$L(\theta) = [Y] = \int_{R^n} \int_{R^n} [Z][S][Y | Z, S] dZ dS \quad (3.27)$$

To simplify the explanation and notation in this section and the following ones, we will first rewrite the likelihood function in terms of the random effect  $W_i = S(x_i) + Z_i$  and will denote this in its vector form as  $W = (W_1, \dots, W_n)$ .

Hence we write

$$L(\theta) = [Y] = \int_{R^n} [W][Y | W] dW \quad (3.28)$$

In this model formulation, we assume that  $Y$  conditionally on  $W$  is a set of mutually independent outcomes, hence

$$[Y|W] = \prod_{i=1}^n [Y_i|W_i].$$

However, unlike in the previous section, the distribution of  $[W]$  is a Multivariate Gaussian distribution with correlated components, hence we cannot split the integral into separate univariate ones as before. More specifically,  $W$  follows a Multivariate Gaussian distribution with mean vector 0 and covariance matrix  $\Omega$ , whose  $(i,j)$ -th entry is

$$[\Omega]_{ij} = \begin{cases} \sigma^2 \rho(u_{ij}; \phi, \kappa) & \text{if } i \neq j \\ \sigma^2 + \tau^2 & \text{if } i = j \end{cases}$$

where  $\rho(u; \phi, \kappa)$  is the Matern correlation function with scale parameter  $\phi$  and smoothness parameter  $\kappa$ ; see Equation 1.5.

In the next two sections we provide a brief overview of Monte Carlo maximum likelihood and the use of Markov chain Monte Carlo to approximate Equation 3.28.

### 3.4.2.1 Monte Carlo maximum likelihood

In maximum likelihood estimation, we aim to maximize the likelihood function to find the parameters that make our data the most likely to have been observed. Monte Carlo Maximum Likelihood (MCML) is a method introduced by Charles J. Geyer (1991), designed to handle situations where the likelihood function is difficult or impossible to compute directly. As shown above, non-Gaussian generalized linear geostatistical models yield an intractable likelihood function that involves a high-dimensional integral, hence we use MCML to approximate this.

MCML solves this problem using Monte Carlo methods, which involve random sampling to approximate the value of these integrals. One way to apply Monte Carlo methods in this context is through *importance sampling*. Importance sampling allows us to approximate the likelihood function by rewriting the likelihood in a form that can be estimated using samples from an easier distribution, known as the *importance sampling distribution*. To this end, we

then rewrite Equation 3.28 as

$$\begin{aligned}
 L(\theta) &= \int [W, Y; \theta] dW \\
 &= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W, Y; \theta_0] dW \\
 &\propto \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W|Y; \theta_0] dW \\
 &= E \left\{ \frac{[W, Y; \theta]}{[W, Y; \theta_0]} \right\},
 \end{aligned} \tag{3.29}$$

where  $\theta_0$  represents our best guess of  $\theta$  and the expectation in the equation above is taken with respect to  $[W|Y; \theta_0]$ , which represent our *importance sampling distribution*.

Let  $W_{(j)}$  represent the  $j$ -th sample out of  $B$  drawn from the distribution  $[W | Y; \theta_0]$ . To simulate samples from  $[W | Y; \theta_0]$ , Markov chain Monte Carlo (MCMC) algorithms are required and we explain more in detail in the next section.

We approximate the expression in Equation 3.29 as follows:

$$L_{MC}(\theta) = \frac{1}{B} \sum_{j=1}^B \frac{[W_{(j)}, Y; \theta]}{[W_{(j)}, Y; \theta_0]}. \tag{3.30}$$

As  $B \rightarrow \infty$ , the Monte Carlo likelihood  $L_{MC}$  converges to the exact likelihood function described in Equation 3.29. The maximization of this Monte Carlo likelihood is performed using an optimization algorithm that uses the analytical expressions of the first and second derivatives. This approach is especially useful for exploring the likelihood surface, which can be relatively flat, particularly in the space of covariance parameters. To improve the approximation of the likelihood function in equation Equation 3.30, this process can be repeated iteratively, replacing  $\theta_0$  after each step with the estimate obtained from equation Equation 3.30, until convergence is achieved. However, due to Monte Carlo error, the estimates after each iteration may still vary considerably, making convergence difficult to reach if very strict criteria are used.

For spatial prediction, we take a pragmatic approach by running an initial MCML iteration, followed by a second iteration with 10,000 samples, without further iterations. This is because spatial prediction is generally less affected by Monte Carlo error, as will be demonstrated later. On the other hand, when the goal is to draw inferences about model parameters, more iterations may be needed to reduce Monte Carlo error to acceptable levels.

### 3.4.2.2 Monte Carlo Markov Chain: sampling from the predictive distribution of the spatial Gaussian process

Building on the outline of the MCML method from the previous section, we now address the task of sampling from the distribution  $[W | Y]$  to generate the samples  $W_{(j)}$  used in equation Equation 3.30. To simplify the notation, we will use  $\$[W | Y]$$$  in place of  $[W | Y; \theta_0]$ , but it is important to emphasize that the parameters  $\theta_0$  are considered fixed when sampling from  $[W | Y]$ . Since direct simulation from  $[W | Y]$  is not feasible, we must rely on Markov chain Monte Carlo (MCMC) methods, which we briefly introduce next.

The *Metropolis-Hastings algorithm* is a foundational technique in the field of MCMC methods, which are widely used for sampling from complex probability distributions. The primary goal of MCMC methods is to generate a sequence of samples from a target distribution where direct sampling is challenging, such as the distribution of  $[W | Y]$ . The Metropolis-Hastings algorithm, first introduced by Metropolis et al. (1953) and later generalized by Hastings (1970), achieves this by constructing a Markov chain that has  $\$[W | Y]$$$  as its stationary distribution. In what follows, we shall use  $\pi(W)$  to denote the density function of  $[W | Y]$  up to a constant of proportionality.

Given a current realization  $W_{(k)}$  of the distribution  $[W | Y]$  in the Markov chain, the algorithm proceeds as follows.

- Propose a new realization  $W_{(k+1)}$  from  $q(W_{(k+1)} | W_{(k)})$ , where  $q(\cdot | \cdot)$  is the density function of a chosen proposal distribution.
- Compute the acceptance probability:

$$\alpha = \min \left( 1, \frac{\pi(W_{(k+1)})q(W_{(k)} | W_{(k+1)})}{\pi(W_{(k)})q(W_{(k+1)} | W_{(k)})} \right).$$

- Accept the proposed realization  $W_{(k+1)}$  with probability  $\alpha$ ; otherwise, retain the current realization  $W_{(k)}$ .

The algorithm ensures that, under certain conditions, the chain converges to the target distribution  $\pi(W)$  as the number of steps increases Roberts and Tweedie (1996).

The functions of the **RiskMap** package for fitting Binomial and Poisson geostatistical models implement the *Metropolis-Adjusted Langevin Algorithm (MALA)*, which is an improvement over traditional Metropolis-Hastings MCMC methods. Introduced by Rossky, Doll, and Friedman (1978) and further developed by Roberts and Tweedie (1996), MALA incorporates gradient information from the target distribution to propose more informed moves, potentially leading to faster convergence and more efficient exploration of the target distribution.

MALA combines the Metropolis-Hastings framework with Langevin dynamics, where proposals  $q(\cdot|\cdot)$  are defined as:

$$W_{(k+1)} = W_{(k)} + \frac{\epsilon^2}{2} \nabla \log \pi(W_{(k)}) + \epsilon \eta,$$

where  $\epsilon$  is a step size parameter,  $\nabla \log \pi(W_{(k)})$  is the gradient of the log of the target density with respect to  $W_{(k)}$ , and  $\eta \sim \mathcal{N}(0, I)$  is a Gaussian random variable with mean 0 and independent components with unit variance.

By leveraging the gradient, MALA can propose moves that are more likely to be accepted, especially in high-dimensional spaces or regions of low probability density.

In the context of Laplace sampling for geostatistical models, the *Metropolis-Adjusted Langevin Algorithm (MALA)* is used to construct an efficient proposal distribution. The primary objective is to generate samples from a conditional distribution of latent Gaussian processes, which is challenging due to high dimensionality and complex dependence structures.

The Laplace sampling method, as implemented in the **RiskMap** package, utilizes a specific form of MALA. Here, the proposal distribution is derived from a second-order Taylor expansion around the mode of the target distribution  $[W | Y]$ . Let  $\hat{w}$  be the mode of this conditional distribution, and let  $\hat{\Omega}$  be the inverse of the negative Hessian matrix evaluated at  $\hat{w}$ , which approximates the local covariance structure of the distribution. We then define the random variable  $\tilde{W}$  as

$$\tilde{W} = \hat{\Omega}^{-1/2}(W - \hat{w})$$

This linear transformation is applied to decorrelate the components of  $W$ , making the MCMC updates more efficient (Charles J. Geyer 1994). The function `Laplace_sampling_MCMC` in **RiskMap** uses this approach where by at each iteration the variable  $\tilde{W}$  using the MALA algorithm, and samples for  $[W|Y]$  are then obtained by inverting the linear transformation from the above equation. The logic behind using this approach is that it allows the proposal distribution to adapt to the local geometry of the posterior, thus improving convergence properties and reducing the correlation between successive samples (see O. F. Christensen (2004)).

Once the MCMC algorithm has completed all iterations, it is important to ensure that it has converged to the target distribution. The diagnostic methods listed below can be used to assess convergence using the obtained MCMC samples.

1. **Trace Plots:** Plotting the sampled values against the iterations can help visually inspect if the chains have reached a stationary

distribution. A well-mixed chain should appear as a “random walk” around a stable mean value.

2. **Autocorrelation Plots:** These plots measure the correlation between samples as a function of the lag between them. Low autocorrelation across lags indicates good mixing, suggesting convergence.
3. **Effective Sample Size (ESS):** The ESS measures the number of independent samples that an MCMC chain is equivalent to, accounting for the autocorrelation in the chain. It is calculated as:

$$\text{ESS} = \frac{N}{1 + 2 \sum_{k=1}^{\infty} \rho_k},$$

where  $N$  is the total number of MCMC samples and  $\rho_k$  is the autocorrelation at lag  $k$ . A high ESS indicates that the chain has good mixing properties and the samples are nearly independent. The summation of autocorrelations up to a sufficiently large lag is used to estimate the reduction in effective sample size due to the dependence between samples. A small ESS suggests high autocorrelation and that more samples are needed to achieve the same precision as a set of independent samples.

A high ESS (Effective Sample Size) indicates that the chain has good mixing properties, meaning the samples are nearly independent. The summation of autocorrelations up to a sufficiently large lag is used to estimate the reduction in effective sample size due to the dependence between samples. A small ESS suggests high autocorrelation, meaning more samples are needed to achieve the same precision as a set of independent samples. In addition to those, other diagnostics can provide further assurance of convergence. For example, the Gelman-Rubin diagnostic compares the variance between multiple chains to the variance within each chain; a diagnostic value close to 1 indicates that the chains have converged to a common distribution (Gelman and Rubin 1992). Also, the Geweke diagnostic tests for convergence compares the means of the first part and last part of a single MCMC chain. If the means are not significantly different, it suggests that the chain has reached stationarity (Geweke 1992). Using a combination of these diagnostics can be used to assess whether the MCMC algorithm has adequately explored the target distribution.

It is important to point out that in MCMC algorithms, to the quality of samples we need to carefully consider these three control parameters of the MCMC.

- **Number of Simulations:** Increasing the number of simulations allows the algorithm to explore the posterior distribution more thoroughly, reducing

the variability in estimates. A higher number of simulations typically leads to more accurate results but also requires more computational time.

- **Burn-in:** The burn-in phase refers to the initial set of iterations where the Markov chain has not yet converged to its stationary distribution. These initial samples are discarded to ensure that the remaining samples represent the target distribution more accurately. Properly setting the burn-in period helps avoid bias due to initial conditions.
- **Thinning:** Thinning reduces autocorrelation by only keeping every  $n$ -th sample from the chain. While this decreases autocorrelation between retained samples, it also reduces the overall sample size. Thinning is used when chains exhibit high autocorrelation to ensure a more independent set of samples is analyzed. Initially, these values can be set based on prior experience or general rules of thumb (e.g., a burn-in of 10-20% of the total iterations). In the next summary section, we provide more details on specific implementation of these in RiskMap.

For a more thorough and complete overview of MCMC algorithms, chapters 10 to 13 of Gelman and Rubin (1992) are a recommended read.

### 3.4.2.3 Summary with example

In summary, here are the steps required to perform Monte Carlo Maximum Likelihood (MCML):

1. Start by selecting an initial guess for the model parameters, denoted as  $\theta_0$ .
2. Simulate Samples from  $[W | Y; \theta_0]$  using MCMC methods. Perform a check on the MCMC samples to assess convergence of the MCMC algorithm.
3. Compute and maximize the approximate likelihood function in Equation 3.30.
4. Set the newly estimated parameter as  $\theta_0$  for the next iteration.
5. Repeat 1 to 4 until convergence is achieved.
6. Monitor the estimates for convergence. If strict convergence criteria are set, Monte Carlo error may cause fluctuations, making it harder to achieve convergence.
7. Perform additional iterations, if necessary. For inference on model parameters, continue iterating to reduce Monte Carlo error. For spatial prediction, a pragmatic approach may involve stopping after a second iteration with 10,000 samples, as prediction is less affected by Monte Carlo error.

In **RiskMap**, steps 1 to 3 are automated (with the exception of the check on the convergence of the MCMC algorithm in step 2), while steps 4 to 7 require manual implementation. In Section 3.3.1, estimation is performed using a geostatistical model fitted to prevalence data. The `par0` argument in the `glgpm` function allows you to specify initial guesses for the model parameters  $\theta_0$ . If no values are provided for `par0` (as in Section 3.3.1), the following defaults are used:

- The initial guess for the regression coefficients  $\beta$  is set to the maximum likelihood estimates obtained from a standard generalized linear model (GLM), i.e., a GLM without random effects or overdispersion.
- The initial values for both  $\sigma^2$  (the variance of the spatial Gaussian process) and  $\tau^2$  (the variance of the nugget term, if included) are set to 1.
- The initial guess for the scale parameter of the spatial correlation function  $\phi$  is based on the 0.1 quantile of the distances between all pairs of locations.

In our experience, this set of initial guesses works reasonably well. However, more iterations of the MCMC algorithm, as suggested in the step 7 above, are generally recommended.

An alternative approach for setting initial guesses for the covariance parameters  $\sigma^2$ ,  $\tau^2$ , and  $\phi$  is to apply an interpolation technique to the variogram, as proposed in the seminal work by N. Cressie (1985). However, in our experience, this approach does not consistently provide better approximations of the likelihood function compared to the method mentioned above. This may be due to the instability in estimating the variogram based on residuals from mixed models, as well as the sensitivity of empirical variogram estimation to the chosen distance bins.

To better illustrate some of the points above, we continue the analysis of the Liberia data from Section 3.3.1. Diagnostic summaries on the convergence of the MCMC algorithm can be obtained with the `check_mcmc` function by passing the fitting model `fit_liberia` to the function.

```
check_mcmc(fit_liberia)
```

**Spatial random effect Effective sample size: 727.**

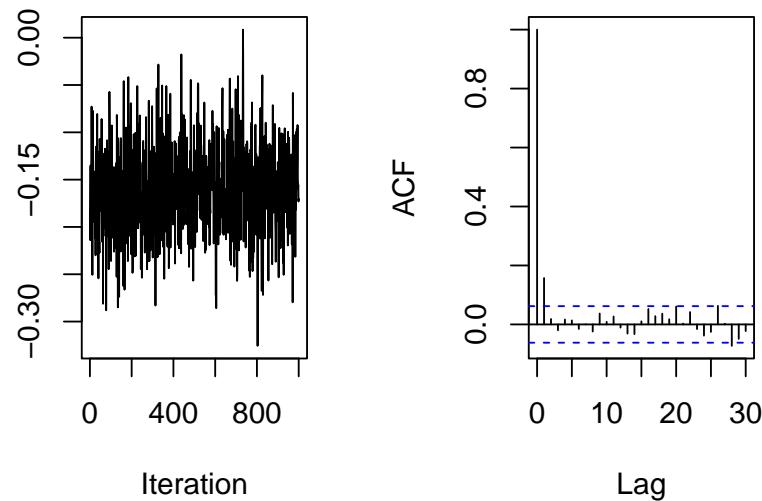


Figure 3.12: Trace plot (left panel) and autocorrelation function plot (right panel) for the averaged spatial random effects from the river-blindness data analysis. The title of the right panel also shows the effective sample size of the 1000 Markov chain Monte Carlo (MCMC) samples. The MCMC samples are obtained by running the MCMC algorithm for 12000 iterations, with a burn-in of 2000 and thinning of 10.

The plot shown in Figure 3.12 are for the MCMC samples of the averaged spatial random effect, that is  $\bar{S} = \sum_{i=1}^n S(x_i)/n$ . Performing the check on  $\bar{S}$  is done to avoid the need of checking the convergence for each individual component  $S(x_i)$ . However, one of the risks of doing this is that if fewer components of  $\bar{S}$  have not reached convergence this may not be shown through the inspection of the samples for  $\bar{S}$ . However, this is an unlikely scenario when using the type of MCMC algorithm implemented in **RiskMap** (see Section 3.4.2.2). Nonetheless, if one wants to check the convergence for individual components this can be done as follows.

```
check_mcmc(fit_liberia, check_mean = FALSE, component = 10)
```

**Spatial random effect Effective sample size: 727.**

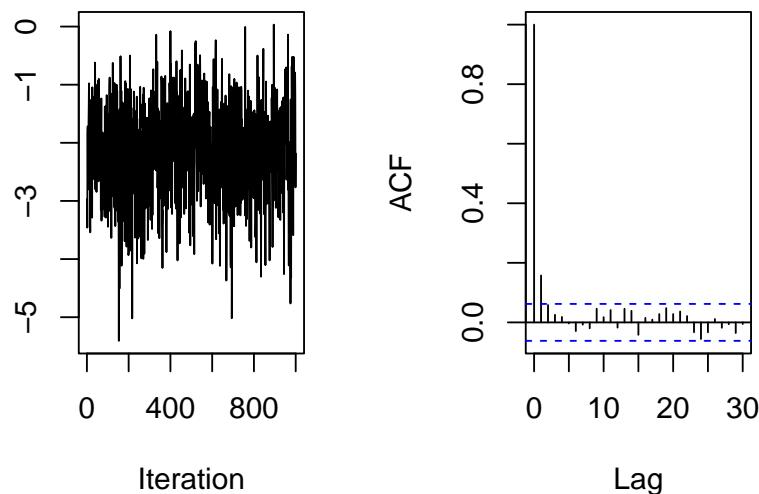


Figure 3.13: Trace plot (left panel) and autocorrelation function plot (right panel) for the tenth component of the spatial random effects from the river-blindness data analysis.

Both Figure 3.12 and Figure 3.13, shows a reasonably good mixing and relatively low correlation in the samples. We point out that the number of MCMC samples in `glgpm` are set through the function `set_control_sim` passed to the argument `control_mcmc`. By default `set_control_sim` uses 12000 iterations, a burn-in of 2000 and thinning of 10, thus giving 1000 MCMC samples. This configuration can be enough for the first run of the Monte Carlo maximum Likelihood (MCML) method, however it is advisable to increase the number of iterations to obtain a larger number of samples with reduced correlation.

```
# Update the guesses for the model parameters
par0_liberia <- coef(fit_liberia)

# Refit the model and increase the MCMC samples
fit_liberia2 <-
  glgpm(npos ~ log(elevation) + gp(long, lat, nugget = NULL),
        distr_offset = ntest, data = liberia,
        crs = 4326,
        convert_to_crs = 32629,
```

```

par0 = par0_liberia,
control_mcmc = set_control_sim(n_sim = 110000,
                                burnin = 10000,
                                thin = 10),
family = "binomial", messages = FALSE)

check_mcmc(fit_liberia2)

```

**Spatial random effect Effective sample size: 7679**

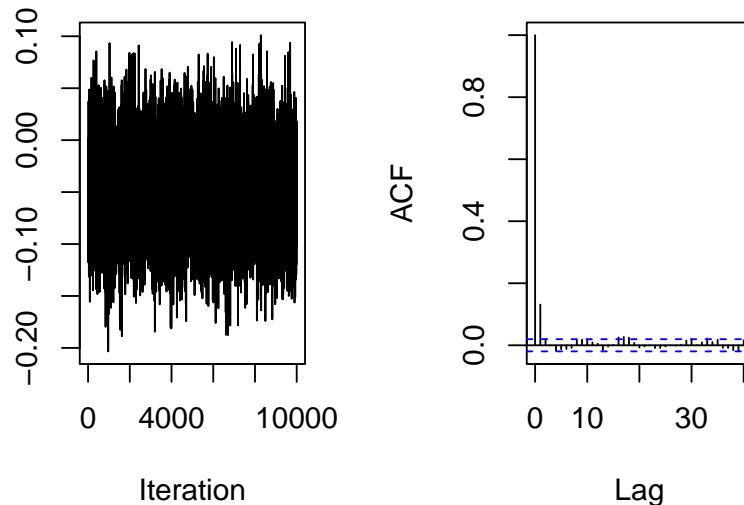


Figure 3.14: Trace plot (left panel) and autocorrelation function plot (right panel) for the tenth component of the spatial random effects from the river-blindness data analysis. The Markov chain Monte Carlo (MCMC) samples are obtained by running the MCMC algorithm for 110000 iterations, with a burn-in of 10000 and thinning of 10.

The code above performs an additional step in the MCML estimation, using 10,000 MCMC samples. As shown in Figure 3.14, these MCMC samples exhibit good mixing and low autocorrelation, which indicates that they adequately approximate the likelihood function of the Binomial geostatistical model. Although additional MCMC iterations could be run to obtain estimates closer to the true maximum likelihood estimate (since we are approximating the likelihood function and not obtaining the exact MLE), we choose

not to do so here. This decision is due to both the computational time required and the minimal impact of Monte Carlo error on spatial predictions in this case.

### **3.4.2.4 Alternative approaches to parameter estimation**

In the sections above, we focused specifically on the implementation of the Monte Carlo Maximum Likelihood (MCML) method in **RiskMap**. However, other approaches are available, some of which are computationally more efficient than the MCML method we have illustrated. Our preference for this estimation method is based on two main reasons: 1) it delivers robust performance across a wide range of data scenarios, and 2) it does not require the specification of arbitrary prior distributions when prior information on the model parameters is unavailable.

Alternative methods worth mentioning include the expectation-maximization algorithm for maximum likelihood estimation (Zhang 2002), integrated nested Laplace approximation (INLA) (Rue, Martino, and Chopin 2009), and implementations of the Metropolis-Adjusted Langevin Algorithm (MALA) for Bayesian inference (O. Christensen, Roberts, and Sköld 2006). While these methods are valuable, explaining and comparing them is not our focus here, as it would detract from the practical orientation of this book.

## **3.5 Exercises**

1. Consider the Binomial mixed model with linear predictor as defined in Equation 3.7. By editing the code for the simulation shown in Section 3.1.2, generate a graph as in Figure 3.6 under the two following scenarios: *i*)  $\tau^2 = 0.2$  and  $n_i = 100$ ; *ii*)  $\tau^2 = 0.1$  and  $n_i = 1$ . How does the variance of  $Y_i$  change under *i*) and *ii*) in comparison to Figure 3.6? How do you explain the differences?
2. Similarly to the previous exercise, consider a Poisson mixed model with linear predictor

$$\log \{\mu_i\} = \beta_0 + Z_i,$$

where  $Z_i$  are a set of mutually independent Gaussian variables with mean 0 and variance  $\tau^2$ . Using the code shown in Section 3.1.2, carry out a simulation study to compute the variance of  $Y_i$  and generate a graph similar to Figure 3.6 to compare the variance of the Poisson mixed model with that of a standard Poisson model.

Generate the graph for different values of  $\tau^2$  and summarize your findings. NOTE: In this simulation the offset  $n_i$  can be set to 1.

3. Create an R function that computes the *cloud variogram*. As explained in Section 3.1.3, the cloud variogram is obtained by plotting  $\hat{V}_{ij}$  (see Equation 3.11) against the distances  $u_{ij}$ . The function should take as input a data-set with three columns: the variable for which the cloud variogram is to be computed; and two columns corresponding to the location of the data. Then, use this function to create the cloud variogram for the model for river-blindness in Equation 3.13. How does this compare to empirical variogram that takes the averages within predefined distance classes, as shown in Figure 3.7?
4. Fit a Binomial mixed model to the Liberia data-set on river-blindness without any covariates, i.e.

$$\log \left\{ \frac{p(x_i)}{1 - p(x_i)} \right\} = \beta_0 + Z_i.$$

Making use of the R code presented in Section 3.1.3.1, use the function `s_variogram` to generate the empirical variogram for this model and compare this to the empirical variogram of Figure 3.7. What differences do you observe?

5. Fit a Poisson mixed model to the anopheles mosquito data using elevation as a covariate, i.e.

$$\log\{\mu(x_i)\} = \beta_0 + \beta_1 d(x_i) + Z_i$$

where  $d(x_i)$  is the measured elevation at location  $x_i$ . How strong is the overdispersion in the data? After fitting the model, extract the estimates of the random effects  $Z_i$  and compute the empirical variogram with the 95% envelope for spatial independence. Repeat this for different classes of distances by changing the input passed to `bins` in the `s_variogram` function. How do the different specifications of `bins` affect the results?



# 4

---

## *Model validation*

---

This is a book created from markdown and executable code.

See (**knuth84?**) for additional discussion of literate programming.

[1](#) [+](#) [1](#)

[1] 2

---

### 4.1 How to simulate geostatistical data from a fitted model

---

### 4.2 Validating the calibration of the model

---

### 4.3 Validating the spatial correlation of the model

---



# 5

---

## *Geostatistical prediction*

---

This is a book created from markdown and executable code.

See ([knuth84?](#)) for additional discussion of literate programming.

[1](#) [+](#) [1](#)

[1] 2

---

### 5.1 Pixel-level predictive targets

---

---

### 5.2 Area-level predictive targets

---

---

### 5.3 Comparing the predictive performance of geostatistical models



# 6

---

## *Case studies*

---

This is a book created from markdown and executable code.

See (**knuth84?**) for additional discussion of literate programming.

[1](#) [+](#) [1](#)

[1] 2

---

### 6.1 Mapping stunting risk in Ghana

---

---

### 6.2 Mapping river blindness in Malawi

---

---

### 6.3 Mapping mosquitoes abundance in Cameroon



---

## References

---

- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Bowman, A. W. 1997. *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with s-Plus Illustrations*. Oxford Statistical Science Series ; 18. Oxford : New York: Clarendon Press ; Oxford University Press.
- Breslow, N. E., and D. G. Clayton. 1993. “Approximate Inference in Generalized Linear Mixed Models.” *Journal of the American Statistical Association* 88: 9–25.
- Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press.
- Chiles, J-P, and P. Delfiner. 2016. *Geostatistics (Second Edition)*. Hoboken: Wiley.
- Christensen, OF, GO Roberts, and M Sköld. 2006. “Robust Markov Chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models.” *Journal of Computational and Graphical Statistics* 15 (1): 1–17.
- Christensen, Ole F. 2004. “Monte Carlo Maximum Likelihood in Model-Based Geostatistics.” *Journal of Computational and Graphical Statistics* 13 (3): 702–18.
- Cowles, Mary Kathryn, and Bradley P. Carlin. 1996. “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.” *Journal of the American Statistical Association* 91 (434): 883–904.
- Cressie, N. A. C. 1991. *Statistics for Spatial Data*. New York: Wiley.
- Cressie, Noel. 1985. “Fitting Variogram Models by Weighted Least Squares.” *Mathematical Geology* 17 (5): 563–86.
- Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. “Model-Based Geostatistics.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47 (3): 299–350. <https://doi.org/10.1111/1467-9876.00113>.
- Diggle, Peter J. 2019. *Model-Based Geostatistics for Global Public Health : Methods and Applications*. Chapman and Hall/CRC Interdisciplinary Statistics Ser. Milton: Chapman; Hall/CRC.
- Dobson, A. J., and A. Barnett. 2008. *An Introduction to Generalized Linear Models*. Third. Chapman; Hall/CRC.
- Fernández, J. A, A Rey, and A Carballeira. 2000. “An Extended Study of

- Heavy Metal Deposition in Galicia (NW Spain) Based on Moss Analysis." *Science of The Total Environment* 254 (1): 31–44. [https://doi.org/10.1016/S0048-9697\(00\)00431-9](https://doi.org/10.1016/S0048-9697(00)00431-9).
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. CRC Press.
- Gelman, Andrew, and Donald B Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72.
- Geweke, John. 1992. "Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments." Edited by Jose M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith, 169–93.
- Geyer, Charles J. 1991. "Markov Chain Monte Carlo Maximum Likelihood." *Journal of Computational and Graphical Statistics* 1 (4): 39–55.
- Geyer, Charles J. 1994. "Likelihood and Exponential Families." *Department of Statistics, University of Minnesota*.
- . 1996. "Markov Chain Monte Carlo Maximum Likelihood." *Department of Statistics, University of Minnesota*.
- . 2019. "Monte Carlo Methods in MCMC." In *Handbook of MCMC*, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, 3–48. CRC Press.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.
- Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57 (1): 97–109.
- Katz, Elizabeth, and Bill & Melinda Gates Foundation. 2020. "Gender and Malaria Evidence Review." Bill & Melinda Gates Foundation. [https://www.gatesgenderequalitytoolbox.org/wp-content/uploads/BMGF\\_Malaria-Review\\_FC.pdf](https://www.gatesgenderequalitytoolbox.org/wp-content/uploads/BMGF_Malaria-Review_FC.pdf).
- Krige, D. G. 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: 119–39.
- Matern, B. 2013. *Spatial Variation*. Lecture Notes in Statistics. Springer New York. <https://books.google.co.uk/books?id=HrbSBwAAQBAJ>.
- Matheron, G. 1963. "Principles of Geostatistics." *Economic Geology* 58: 1246–66.
- Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.
- Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society A* 135: 370–84.
- Pawitan, Yudi. 2001. *In All Likelihood : Statistical Modelling and Inference Using Likelihood*. Oxford ; New York: Clarendon Press : Oxford University Press.
- Ripley, B. D. 1981. *Spatial Statistics*. New York: Wiley.

- Roberts, Gareth O., and Richard L. Tweedie. 1996. "Exponential Convergence of Langevin Distributions and Their Discrete Approximations." *Bernoulli* 2 (4): 341–63.
- Ross, Sheldon. 2013. *First Course in Probability*, a. 9th ed. Harlow: Pearson Education UK.
- Rossky, Peter J., J. D. Doll, and Harold L. Friedman. 1978. "Brownian Dynamics as Smart Monte Carlo Simulation." *The Journal of Chemical Physics* 69 (10): 4628–33.
- Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–92. <https://doi.org/10.1111/j.1467-9868.2008.00700.x>.
- Smith, David L, Carlos A Guerra, Robert W Snow, and Simon I Hay. 2007. "Standardizing Estimates of the Plasmodium Falciparum Parasite Rate." *Malaria Journal* 6 (1): 131–31.
- Stein, Michael L. 1999. *Interpolation of Spatial Data Some Theory for Kriging*. 1st ed. 1999. Springer Series in Statistics. New York, NY: Springer New York : Imprint: Springer.
- Stevenson, Gillian H. AND Gitonga, Jennifer C. AND Stresman. 2013. "Reliability of School Surveys in Estimating Geographic Variation in Malaria Transmission in the Western Kenyan Highlands." *PLOS ONE* 8 (10). <https://doi.org/10.1371/journal.pone.0077641>.
- Tene Fossog, Billy, Diego Ayala, Pelayo Acevedo, Pierre Kengne, Ignacio Ngomo Abeso Mebuy, Boris Makanga, Julie Magnus, et al. 2015. "Habitat Segregation and Ecological Character Displacement in Cryptic African Malaria Mosquitoes." *Evolutionary Applications* 8 (4): 326–45. <https://doi.org/10.1111/eva.12242>.
- Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46: 234–40.
- Watson, G. S. 1971. "Trend -Surface Analysis." *Mathematical Geology* 3: 215–26.
- . 1972. "Trend Surface Analysis and Spatial Correlation." *Geological Society of America Special Paper* 146: 39–46.
- Weisberg, Sanford. 2014. *Applied Linear Regression*. Fourth. Hoboken NJ: Wiley. <http://z.umn.edu/alr4ed>.
- Zhang, Hao. 2002. "On Estimation and Prediction for Spatial Generalized Linear Mixed Models." *Biometrics* 58 (1): 129–36.
- . 2004. "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics." *Journal of the American Statistical Association* 99 (465): 250–61.
- Zouré, Honorat GM, Mounkaila Noma, Afework H Tekle, Uche V Amazigo, Peter J Diggle, Emanuele Giorgi, and Jan HF Remme. 2014. "Geographic Distribution of Onchocerciasis in the 20 Participating Countries of the African Programme for Onchocerciasis Control: (2) Pre-Control Endemic-

ity Levels and Estimated Number Infected." *Parasites & Vectors* 7 (1): 326–26.

---

## ***Appendix***

---

### **An appendix section**

Tempor justo turpis amet egestas inceptos. Conubia tristique mattis cras dui fames nisi curae. Dictumst convallis egestas, elementum viverra ex morbi. Sem libero est curae cursus, tempor rutrum etiam. Convallis scelerisque metus primis mus scelerisque vehicula potenti congue augue. Lacus taciti ullamcorper dapibus elit vulputate natoque erat. Nisi facilisi nisl urna ullamcorper imperdiet at nullam ac imperdiet.

