Model-based geostatistics for global public health using R

Emanuele Giorgi
Claudio Fronterre

Table of contents

References **73**

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.

1

Introduction

The book provides 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 the describe the properties of probability distributions through their mean, variance and skeweness; the concepts of stochastic dependence and correaltion; 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 redear 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 frequensist 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
`RiskMap` (E)	Estimating of geostatistical models and spatial prediction
`sf` (E)	Handling of spatial data in R
`terra` (E)	Handling of raster files in R
`ggplot2` (E)	Creating maps and exploratory plots
`crsuggest` (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
`galicia`	Lead concentration m from moss samples collected in Galicia, Northern Spain	Example
`liberia`	Prevalence data on river-blindness from Liberia	Example
`malkenya`	Malaria prevalence data from a community and school survey conducted in Western Kenya	Example
`italy_sim`	Simulated geostatistical data-set within the Italian national boundaries	Example
`malnutrition`	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 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

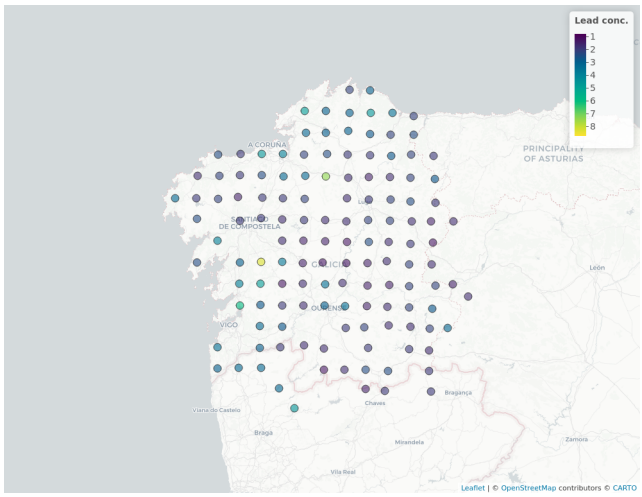


Figure 1.1: Data on the meausred lead concentration (in micrograms per gram dry weight) in moss samples collected in Galicia, North-West of Spain.

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; fore 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 Onchocerchiasis (REMO) carried out in 2011 in 20 African countries (Zouré et al. 2014). The goal of REMO is to identify areas where river-blindness (or onchocerchiasis), 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 a alternative and cheaper diagnostic technique for

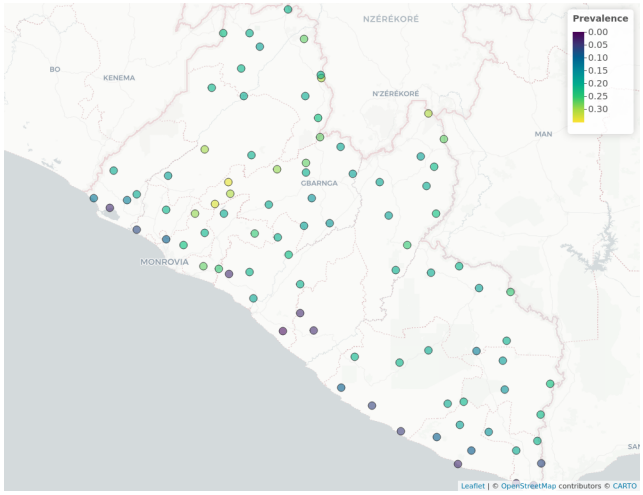


Figure 1.2: River-blindness data from a cross-sectional survey carried out in Liberia.

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 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. *Anophe-*

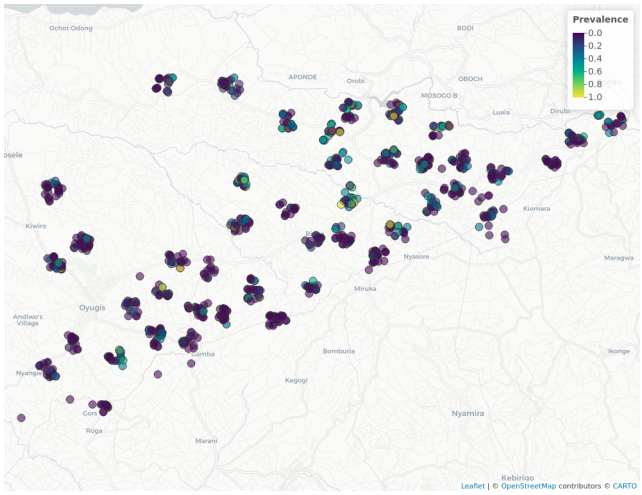


Figure 1.3: Malaria prevalence data from a cross-sectional survey carried out in Nyanza Province, in the Western Highlands of Kenya.

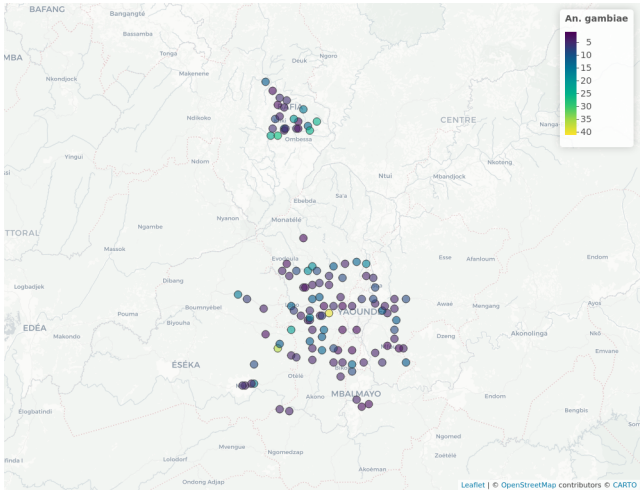


Figure 1.4: Map of the collected number of *Anopheles gambiae* mosquitoes in an area of Southern Cameroon.

les 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

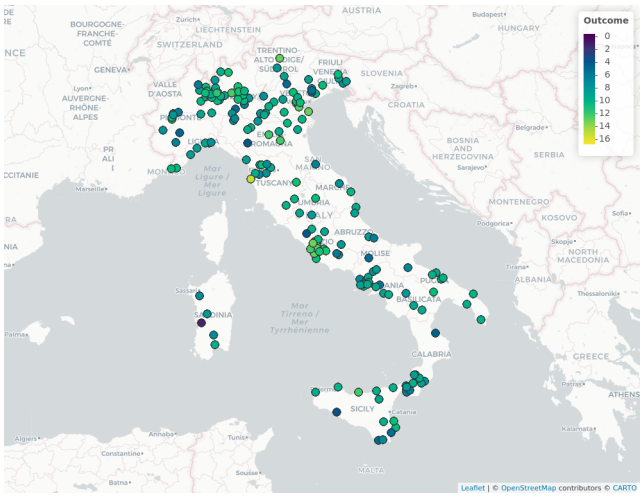


Figure 1.5: Map of the locations of the simulated data-set generated over Italy for a continuous outcome.

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 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 aquire 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 (Krige 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 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 geostastics* 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 geostistical 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 variable $Y = (Y_1, \dots, Y_n)$ 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

$$[Y, S] = [S][Y|S]. \tag{1.1}$$

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 $S(X) = (S(x_1), ..., S(x_n))$, i.e. the process S at the sampled locations $x_1, ..., x_n$, is invariant with respect to rations 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.

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

where σ^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

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

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 μ_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 μ_i as

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

where $d(x_i)$ is a vector of spatially referenced covariates with associated regression coefficients β. 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	$(-\infty, +\infty)$	μ_i	τ^2	$g(\mu_i) = \mu_i$
Binomial	$1, ..., m_i$	$m_i\mu_i$	$m_i\mu_i(1-\mu_i)$	$g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$
Poisson	$1, 2, ..., \infty$	$m_i\mu_i$	$m_i\mu_i$	$g(\mu_i) = \log\{\mu_i\}$

Based on the formulation in (1.4), we can see that $S(x_i)$ quantifies residual spatial effects on μ_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 μ_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 β. Hence, if the goal of the analysis is explain the relationship between a covariate $d(x)$ with the mean component μ.

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

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

where $\phi > 0$ and $\kappa > 0$ are parameters and $K_\kappa(\cdot)$ is the modified Bessel function of the third kind of order κ. The parameters ϕ and κ 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 expeonential correlation function which we write as

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

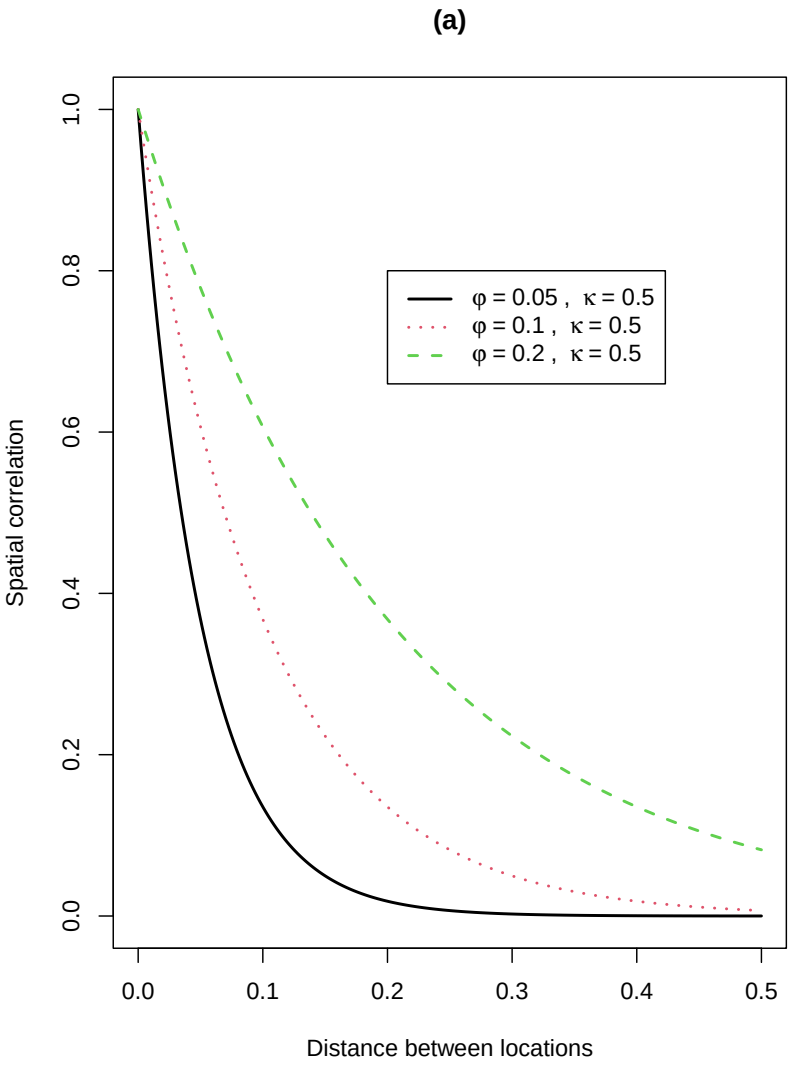
Another special case, which we dot consider in this book but has often been used in machine learning applications, is the Gaussian correlation function obtain as a limiting case for $\kappa \to +\infty$ the possible smoothest process arising from the Matern family.

To better understand how ϕ and κ 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 κ fixed to 0.5 and varied ϕ over the values 0.05, 0.1 and 0.2. As expected, for larger values of ϕ 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 ϕ 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.

Finally, let us consider the correlation functions, shown in the panel (b) of Figure 1.6. Here, we have varied κ over the values 0.5, 1.5 and 2.5, whilst ϕ 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 κ on the spatial surface for process that have approximately the same 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 the Guassian process, which determines its local behaviour. If you are interested in delving these theoretical aspects, we suggest reading Chapter 2 of Stein (1999).

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.



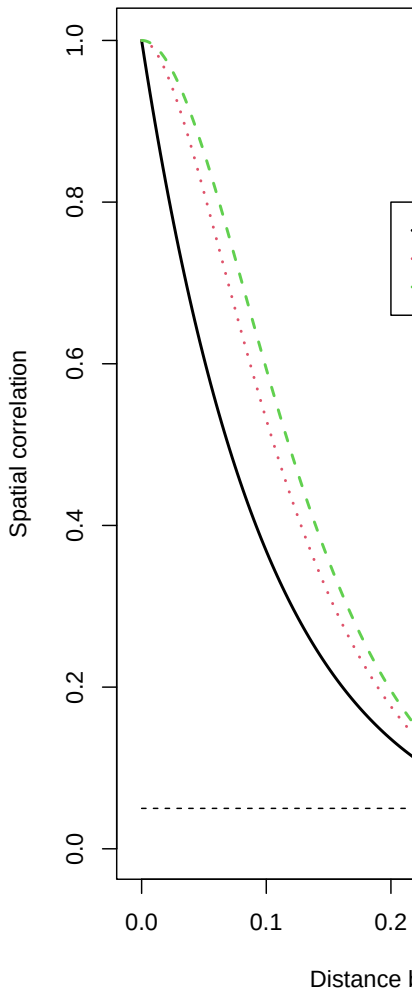


Figure 1.6: Examples of stationary and isotropic Matern correlation functions. Panel (a) shows three different correlations functions that have the same smothness parameter of $\kappa = 0.5$, while varying the scale parameter ϕ over $0.05, 0.1, 0.2$. In panel (b) the scale of the spatial correlation ϕ 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).

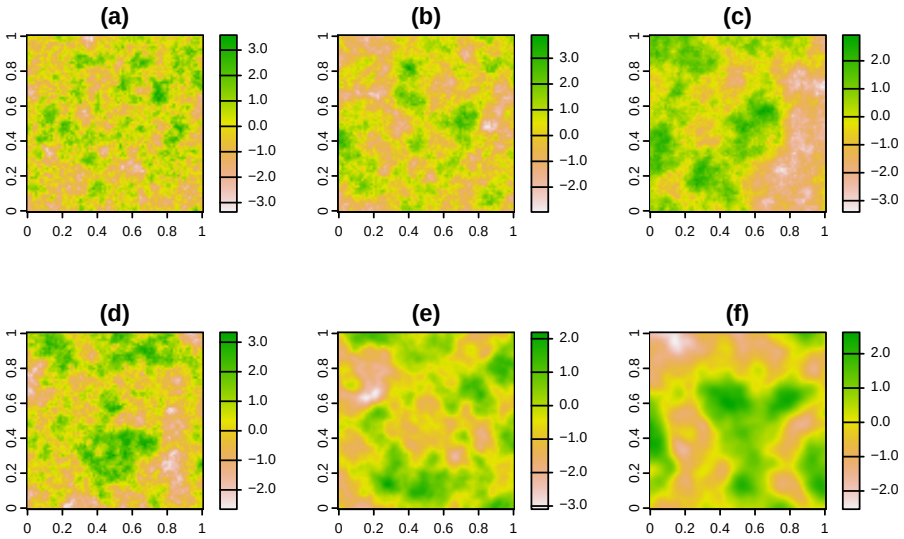


Figure 1.7: Simulated spatial surface using the three correltion functions shown in Figure 1.6. Panels (a), (b) and (c) correspond to the correlation functions from panel (a) in Figure 1.6 and in order these are: $\phi = 0.05$ and $\kappa = 0.5$; (b) $\phi = 0.1$ and $\kappa = 0.5$; (c) $\phi = 0.2$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

1.6 Workflow of a statistical analysis and structure of the book

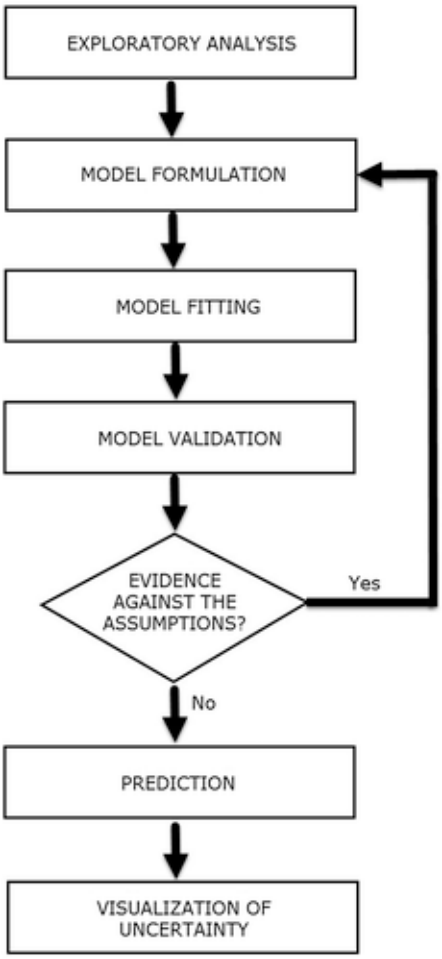


Figure 1.8: Stages of a statistical analysis

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 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 linear 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 sptatial 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 saptial data in R, in particular raster and shape files. 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

This is a book created from markdown and executable code.

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

```
1 + 1
```

```
[1] 2
```

2.1 Importing and processing spatial data in R

2.2 Visualizing geostatistical data

2.3

3

Model formulation and parameter estimation

List of the main functions used in the chapter

Function	R Package	Used for
`lmer`	`lme4`	Fitting linear mixed models
`glmer`	`lme4`	Fitting generalized linear mixed models
`glgm`	`RiskMap`	Fitting generalized linear mixed models
`s_variogram`	`RiskMap`	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$prev <- liberia$npos/liberia$ntest

ggplot(liberia, aes(x = elevation, y = prev)) + geom_point() +
  labs(x="Elevation (meters)",y="Prevalence")
```

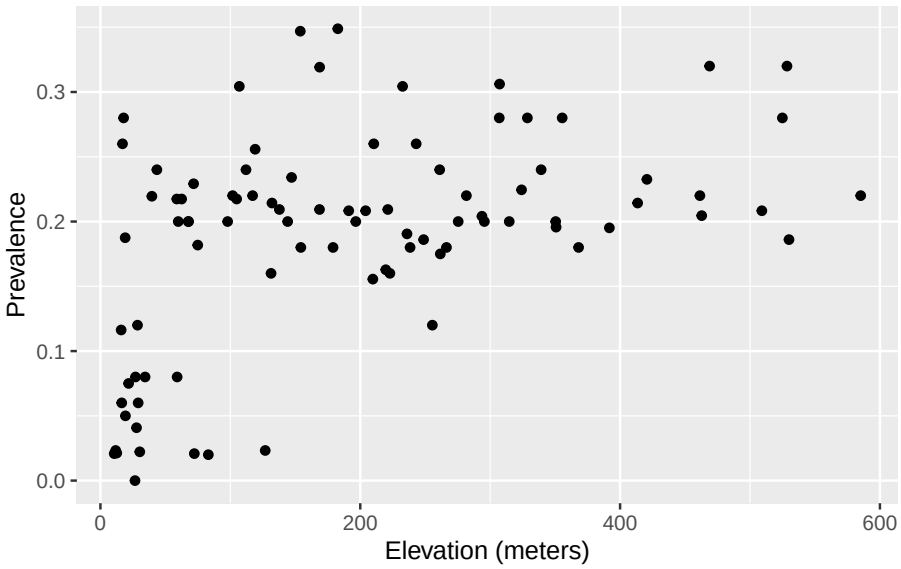


Figure 3.1: Scatter plot of the empirical prevalence for river-blindess against elevation, measured in meters.

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 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

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

where y_i are the number of individuals who tested positive for riverblindness 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$elogit <- log((liberia$npos+0.5)/
                      (liberia$ntest-liberia$npos+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)
```

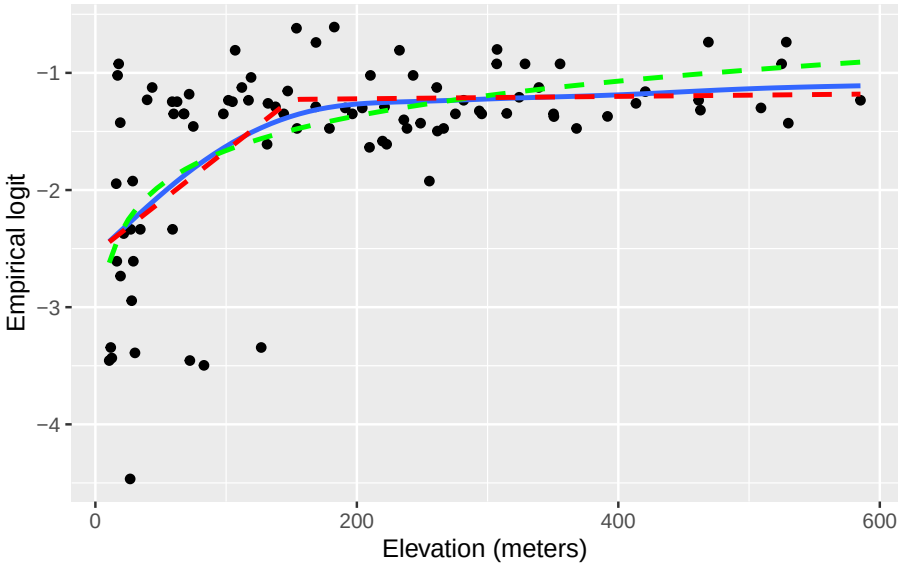


Figure 3.2: Scatter plot of the empirical prevalence for river-blindess against elevation, measured in meters.

Figure 3.2 shows the scatter plot of the empirical logit against elevation. In this plot, we have also added three lines though 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 er 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

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

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

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

Based on the equation above, the effect of elevation below 150 meters is quantified by the parameter β_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 β_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

[1]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 $\cdot$".

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 tranformation without adding 1.

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

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

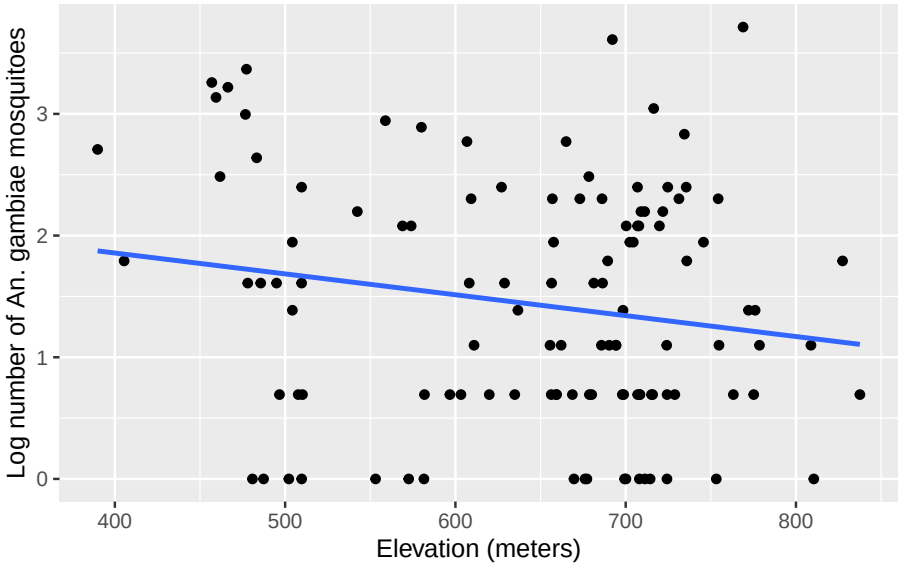


Figure 3.3: Scatter plot of the log tranformed number of *Anopheles gambiae* mosquitoes against elevation, measured in meters. The blue line is generated using the least squares fit.

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 invidual-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

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

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")
```

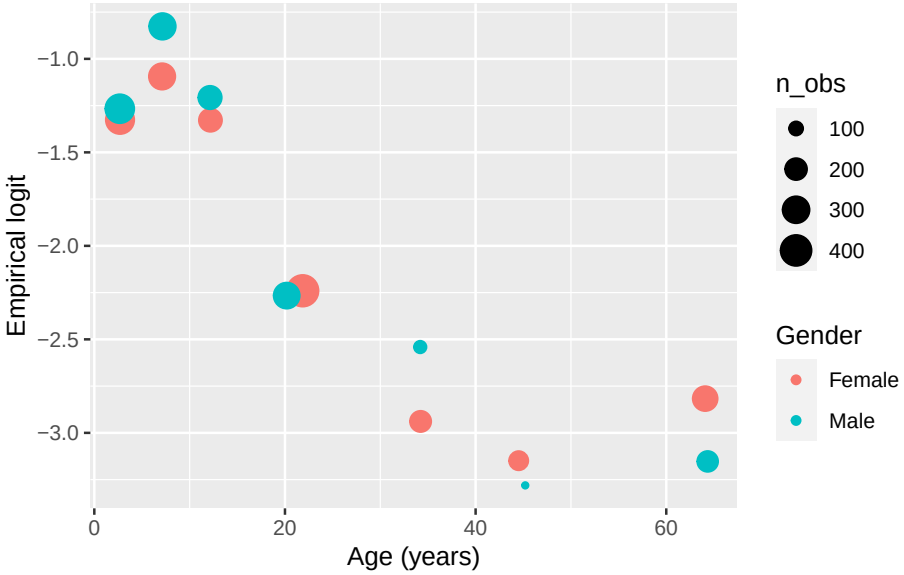


Figure 3.4: Plot of the empirical logit against age, for males and females. The size of each solid point is rendered proportional to the number of individuals within age group, as indicated in the legend.

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

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

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 β_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

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

where g_{ij} is the indicator for gender, with 1 corresponding to male and 0 to female. The coefficients β_1^*, β_2^* and β_3^* thus quantify the differences in risk between the two genders for ages below 15 years, betwee 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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7681  -0.7051  -0.4940  -0.2734   2.7294
##
## 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 β_1, β_2 and β_3, whilst the `Gender:Age`, `Gender:pmax(Age-15, 0)` and `Gender:pmax(Age-40, 0)` to β_1^*, β_2^* and β_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 the 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 assumes 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
                                     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")
```

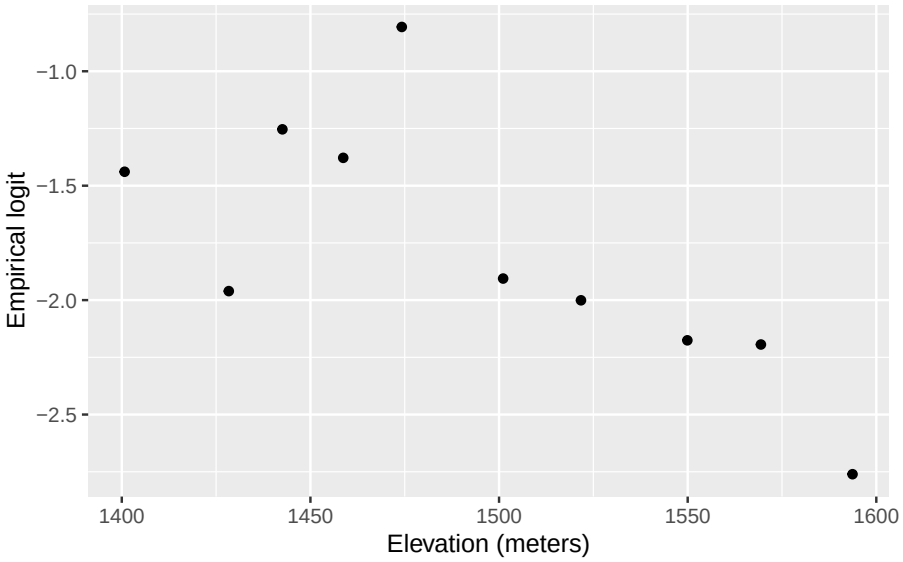


Figure 3.5: Plot of the empirical logit against elevation measured in meters.

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 $S(x_i)$ in introduced in Equation 1.4 with a set of mutually independent random effects, which we denote as Z_i, and thus write

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

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

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

and assume that the Z_i follow a set of mutually independent Gaussian variables with mean 0 and variance τ^2. In our simulation we vary β_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 β_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 τ^2.
- Compute the probabilities μ_i based on Equation 3.8.
- Simulate 10,000 values from a Binomial model with probability of success μ_i and denominator n_i.
- Compute the empirical variance of the counts y_i simulated in the previous step.
- Change the value of β_0 and repeat the previous steps, for all the values of β_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)
```

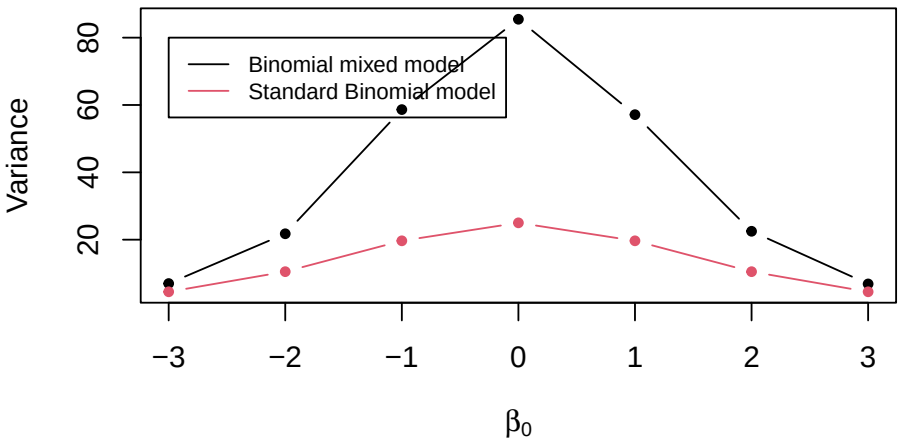


Figure 3.6: Plot of the variances of the standard Binomial model and the Binomial mixed model (see Equation 3.8) against β_0

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 τ^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 τ^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 generalize 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

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

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 τ^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 withuot 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 the that the estimated regression coefficient β_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 τ^2, the variance of Z_i, which found on the line where `ID_loc` is printed. We then observe that the estimates for τ^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 τ^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 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

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

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

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

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 knwon 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

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

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 $\widehat{V}(\mathcal{U})$ 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 varigorams to compute 95% confidence intervals, by taking the 0.025 and 0.975 quantiles of these for each distance class $\widehat{V}(\mathcal{U})$.
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

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

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 τ^2. We first fit the model above using the `glmer` function.

```r
# 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 = binom
                       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 ' ' 1

Correlation of Fixed Effects:
            (Intr)
log(elevtn) -0.981
```

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

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

Throuhg the function `ranef` 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 varigoram for the estimated $\hat{Z}_i$.

```
liberia_variog <- s_variogram(data = liberia,
                              variable = "Z_hat",
                              bins = c(15, 30, 40, 80, 120,
                                       160, 200, 250, 300, 350),
                              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 funda-

mental 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)
```

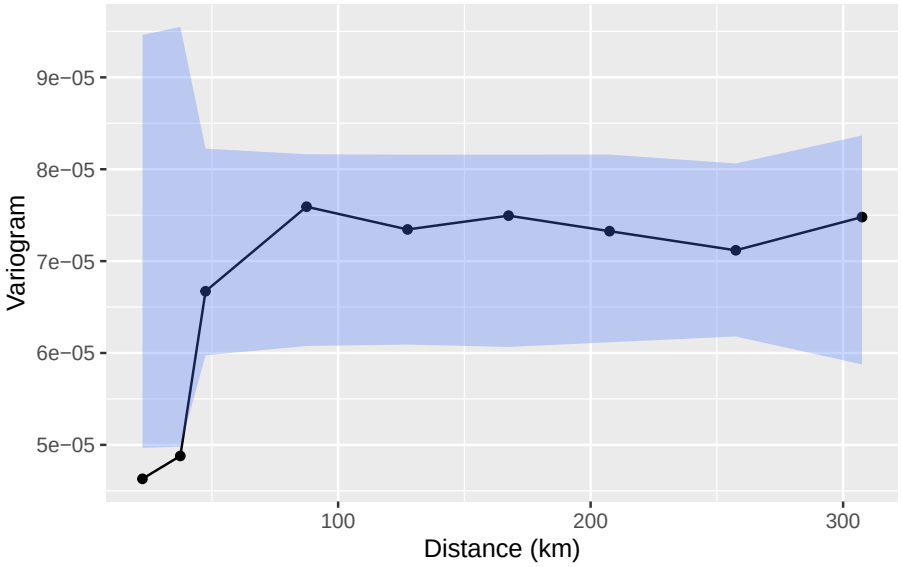


Figure 3.7: Plot of the empirical variogram (solid line) computed using the estimated random effects from the model in Equation 3.13. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

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 spa-

tial 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 Guassian 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.

3.1.3.2.1 One observation per location

To illustrate the use of the variogram under this scenario, we shall use the Gailicia 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

$$Y_i = \beta_0 + U_i \tag{3.14}$$

where the U_i are i.i.d. Gaussian variables with mean zero and variance ω^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 is 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 τ^2 and fit the model

$$Y_i = \beta_0 + Z_i + U_i, \tag{3.15}$$

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, ω (recall that ω^2 is the variance of U_i); b) or, if we do not know ω, 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)
```

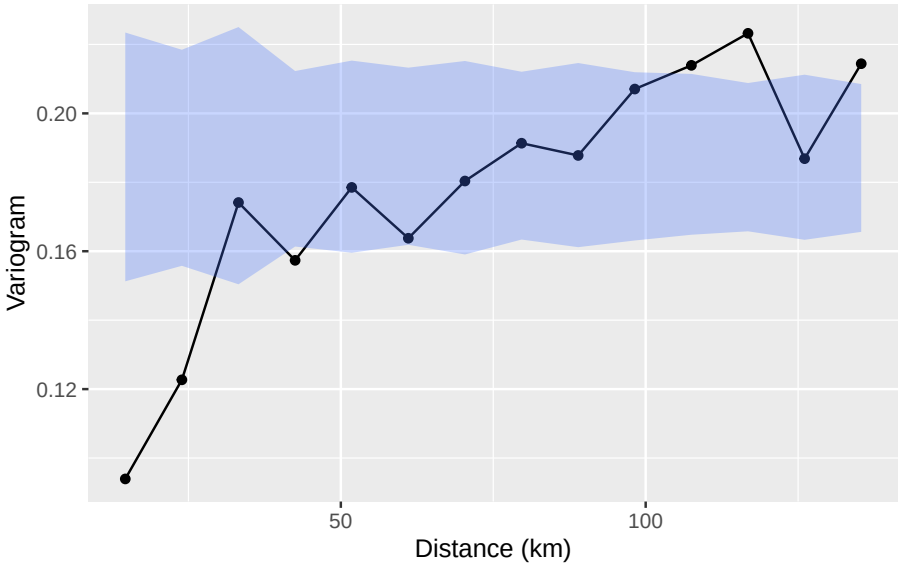


Figure 3.8: Plot of the empirical variogram (solid line) computed using the estimated residuals from the model in Equation 3.14. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

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.

3.1.3.2.2 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 belong 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 ans as U_{ij} the random variation due to the measurement error inherent to each observation. Hence, we write

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

where $d(x_i)$ is the population density at location x_i; as before, we use τ^2 and ω^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)
```

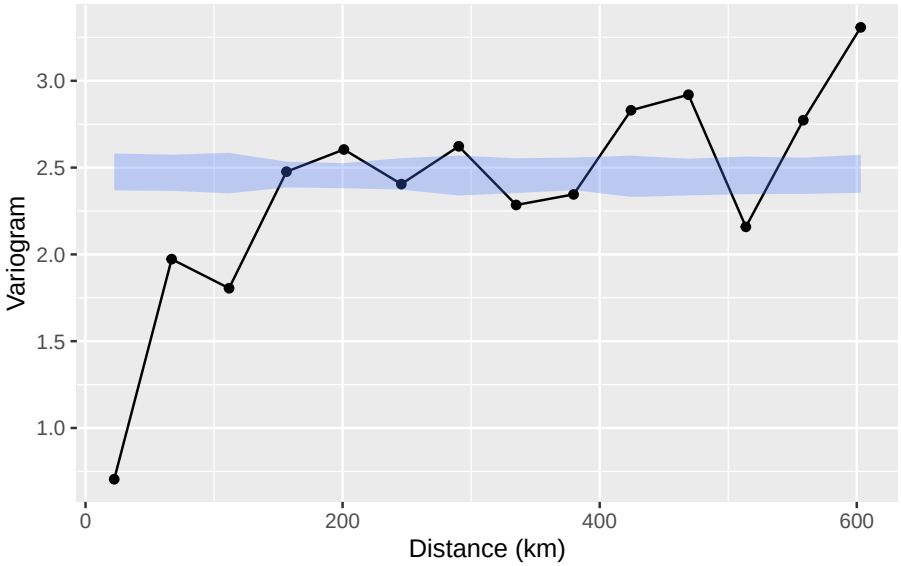


Figure 3.9: Plot of the empirical variogram (solid line) computed using the estimated random effects from the model in Equation 3.16. The blue shaded area is the 95% confidence level envelope generated using the permutation procedure described in Section 3.1.3.

In the code above, the introduction of the Z_i random effect is specified in the `lmer` function though `(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 β_0 and β_1, as well for the variances τ^2 listed under `Random effects:`. The estimate of σ^2 is found on the line of printed output starting with `ID_loc`, whilst that for `$ ^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 Parameter estimation: 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

$$\mu_i = d(x_i)^\top \beta + S(x_i). \tag{3.17}$$

Hence, in this case, we interpret β as the effect on μ_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 ω^2. Thanks to the linear properties of Gaussian random variables, we can also express the linear model in a compact expression, as

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

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 μ_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 μ_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 κ.

3.2.1 Evaluating the inclusion of the measurement error term U_i and the specification of the smoothness parameter κ

In this section we analyse the Galicia data using a linear geostatistical geostatitsical 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

$$Y_i = \mu + S(x_i) + U_i \tag{3.19}$$

where were $S(x)$ is a Matern process with variance σ^2, scale parameter ϕ and and smothness parameter κ; the U_i correspond to the measurement error term and denote with ω^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 κ at prespecified value by the user. The object passed to the argument `data` in `glpm` 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 3.1.3.2.2. 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 ϕ.

```
# 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)
## Geostatistical linear model
## 'Lower limit' and 'Upper limit' are the limits of the95% 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
##
##                           Estimate Lower limit Upper limit
## Measuremment 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, ω^2 whose point estimate is about 0.015; in the final section, we have the estimates for the parameters of the spatial covariance function, σ^2 and ϕ 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 κ 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 κ can affect the estimate of the measurement error variance ω^2. As we have shown, in Section 1.5.1, values of κ that are closer to zero will give a rougher and less regular surface for $S(x)$. In the case of a single observation, when κ is closer to zero, it will thus become increasingly difficult to estimate ω^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 κ closer to zero, the estimates of ω^2 will also be smaller and, on the contrary, when κ is larger, the estimates of ω^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 κ leads to smaller point estimates for ω^2 and viceversa. This begs the question, what should be our chosen value for κ?

To answer this question, a natural approach is to estimate the model for different values of κ 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 κ 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 κ. More precisely, the horizontal dashed line is computed by taking the maximum observed values of the likelihoods computed for the different values of κ, say $\hat{M}$ and we subtract the quantile 0.95 of a χ^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 κ which one could use to derive a confidence interval for κ with a finer segmentation for `kappa_values`. However, our current objective is not to derive the confidence interval for κ, but rather to gain a broad understanding of the κ values supported by the dataset. The values of κ 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 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 κ more precisely because it is empirically very difficult and, under some scenarios, even impossible. Estimating κ 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.

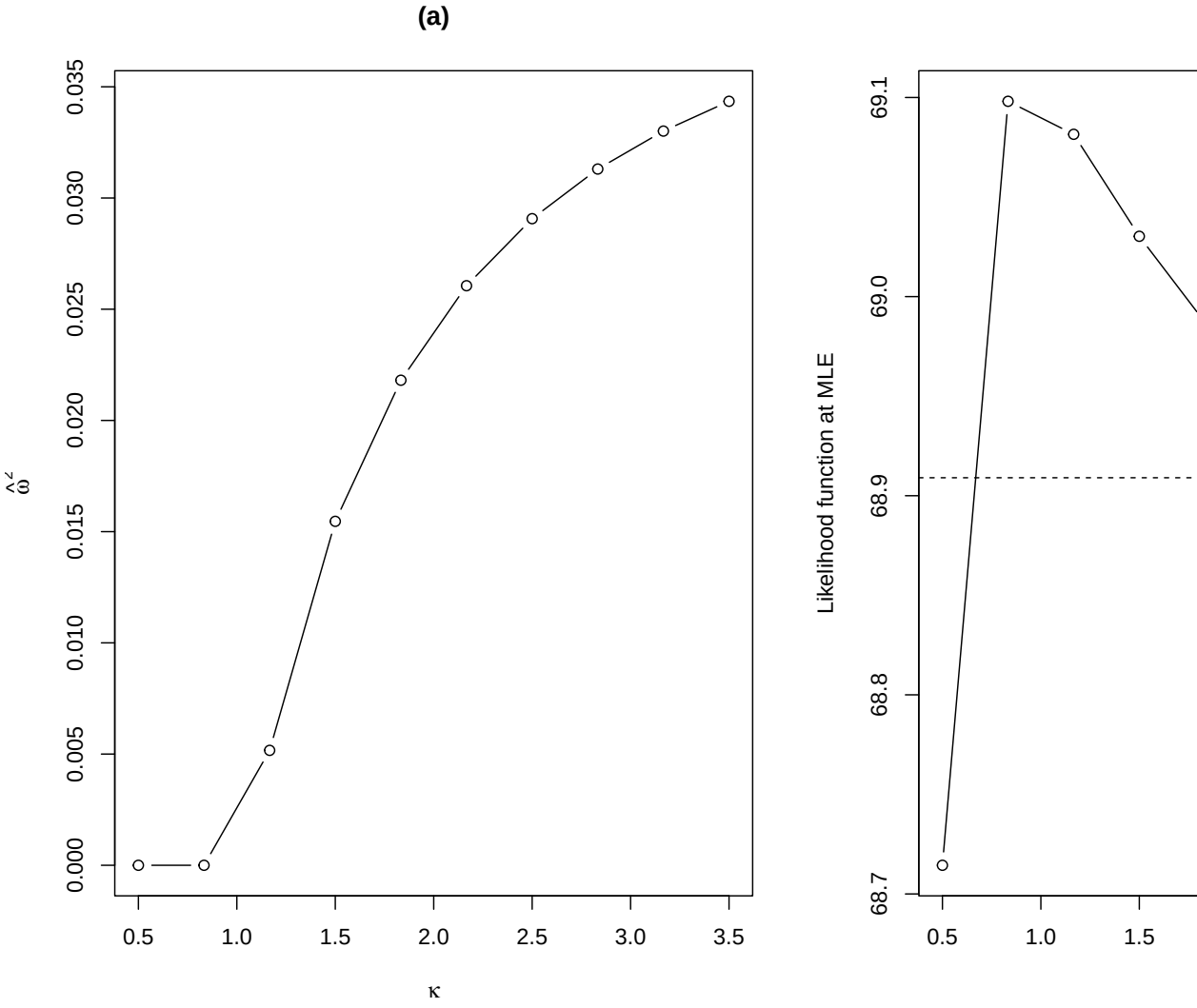


Figure 3.10: (a) plot of the maximum likelihood estimate for the variance of the measurement error U_i, denoted in the text as ω^2, against the chosen fixed value for κ; (b) profile likeihood for κ with the horizontal dashed line corresponding to the (approximate) threshold for constructing a 95% confidence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

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)
```

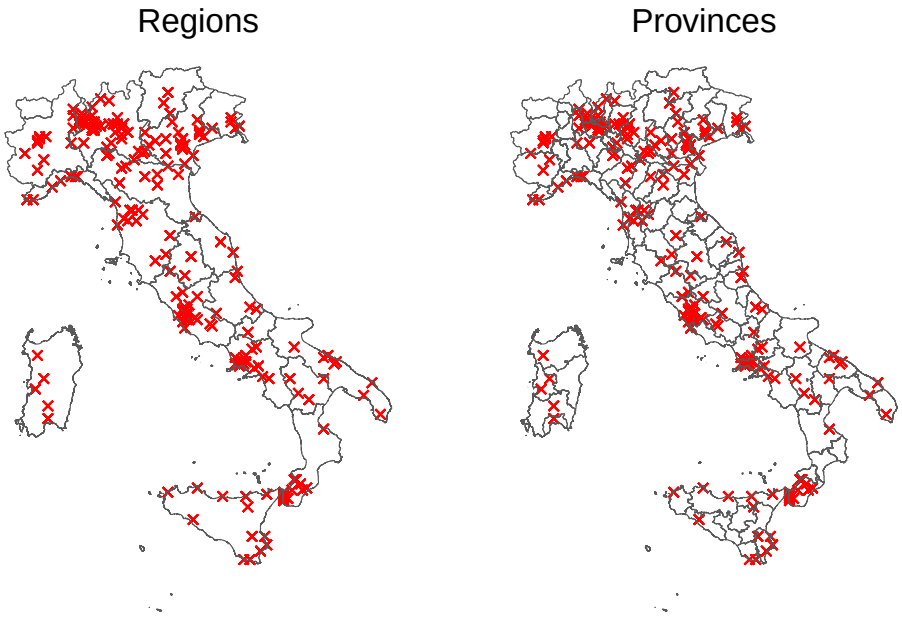


Figure 3.11: The plot shows the location of the data-set `italy_sim` (red crosses), with the boundaries of the regions (left) and provinces (right) of Italy.

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 σ^2_V and σ^2_Z, 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

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

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")
## The CRS used is EPSG:32634
## Distances between locations are computed in kilometers
##   0:     430.59466: -0.404481  1.33798  0.00000  5.30607  0.00000  0.00000  0.00000
##   1:    -139.11498: -0.407668  1.36548 -0.0189346  5.32468 -0.908799 -0.00256388 0.0
##   2:    -279.05702: -1.21582  1.37480 -0.538617  4.98928 -1.94843 -0.565719 0.365940
##   3:    -330.45019: -1.11959  1.37294 -0.103810  5.35702 -1.67443 -1.18312 0.367271
##   4:    -331.43272: -1.07102  1.37291 -0.100149  5.35516 -1.62984 -1.37447 0.375269
##   5:    -331.43375: -1.07439  1.37288 -0.0867550  5.36830 -1.62883 -1.42311 0.376285
##   6:    -331.43375: -1.07445  1.37288 -0.0866232  5.36840 -1.62883 -1.42491 0.376323
##   7:    -331.43375: -1.07445  1.37288 -0.0866232  5.36840 -1.62883 -1.42491 0.376323

summary(italy_fit)
## Geostatistical linear model
## 'Lower limit' and 'Upper limit' are the limits of the95% confidence level intervals
##
##  Regression coefficients
##                Estimate Lower limit Upper limit    StdErr z.value p.value
## (Intercept)   -1.074453   -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
##
##                          Estimate Lower limit Upper limit
## Measuremment 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  214.52024   153.60790     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
β_0	-1.074	-2.032	-0.117
β_1	1.373	1.352	1.394
σ^2	0.917	0.595	1.414
ϕ	214.520	153.608	299.587
σ^2_V	0.241	0.146	0.398
σ^2_Z	1.457	1.379	1.540
ω^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 4.1.2 by xtable 1.8-4 package
## % Fri Nov 24 14:43:05 2023
## \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 \\
##   Measuremment error var. & 0.196 & 0.194 & 0.199 \\
##    \hline
```

```
## \end{tabular}
## \end{table}
```

3.3 Generalized linear geostatistical models

3.4 Theory

3.4.1 The likelihood function of a genearlized linear mixed model

3.4.2 The likelihood function of a genearlized linear geostatistical model

3.4.3 Monte Carlo maximum likelihood

3.5 Exercises

2. 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?

3. 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 τ^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 τ^2 and summarize your findings. NOTE: In this simulation the offset n_i can be set to 1.

4. 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?

5. 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?

6. 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?

7. Consider a linear model for the Galicia data as in Section 3.2.1 but now also introduce a nugget term, hence

$$Y_i = \mu + S(x_i) + Z_i + U_i \tag{3.21}$$

where Z_i are i.i.d. Gaussian variables with mean 0 and variance τ^2. Fit the model by fixing the variance of U_i, ω^2 to 0.01 using the argument `fix_var_me` in `glgpm`. Using this model, reproduce the graph as shown in Figure 3.10. Repeat this, but now fix ω^2 to 0.02. How do the curves change and why?

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 Ghan

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.

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.

Chilès, J-P, and P. Delfiner. 2016. *Geostatistics (Second Edition)*. Hoboken: Wiley.

Cressie, N. A. C. 1991. *Statistics for Spatial Data*. New York: Wiley.

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.

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.

Katz, Elizabeth, and Bill & Melinda Gates Foundation. 2020. "Gender and Malaria Evidence Reivew." Bill & Melinda Gates Foundation. 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.

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.

Ross, Sheldon. 2013. *First Course in Probability, a*. 9th ed. Harlow: Pearson Education UK.

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