Model-based geostatistics for global public health using R

Emanuele Giorgi
Claudio Fronterre

Table of contents

References **45**

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 Example 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(galicia)
```

for the lead concentration data from Galicia,

```
data(liberia)
```

for the river-blindness data-set,

```
data(malkenya)
```

for the malaria data in the Western Highlands of Kenya, and

```
data(anopheles)
```

for the Anopheles mosquitoes data-set.

In the final chapter of this book, we will consider the analysis of additional data-sets to review the main statistical concepts presented in this book.

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 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,

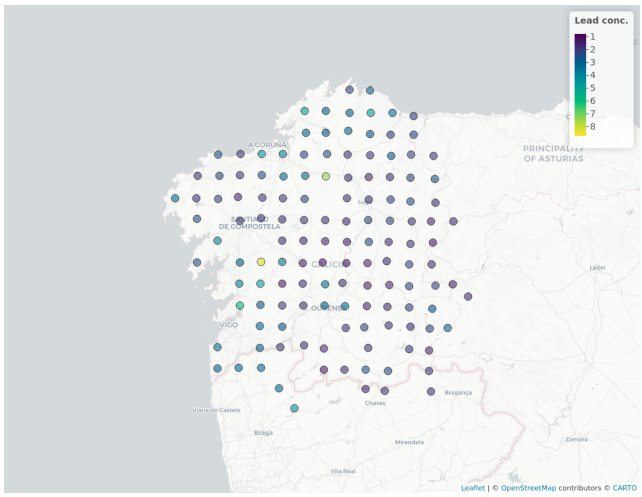

Figure 1.1: Data on the meausred lead concentration (in micrograms per gram dry weight) in moss samples collected in Galicia, North-West of Spain.

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 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.

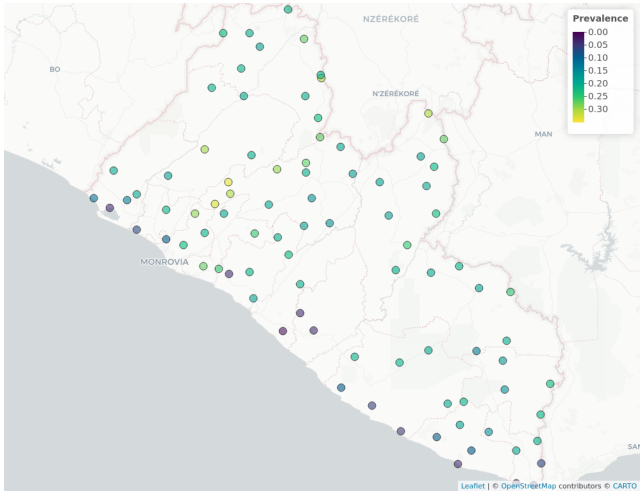

Figure 1.2: River-blindness data from a cross-sectional survey carried out in Liberia.

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. *Anopheles gambiae* mosquitoes are one of the main vectors for malaria transmission in sub-Saharan Africa. Their distribution over space is affected by several environmental and climatic factors, including temperature, humidity and vegetation.

The data-set on mosquitoes (Figure 1.4) that will use in the book consists of a

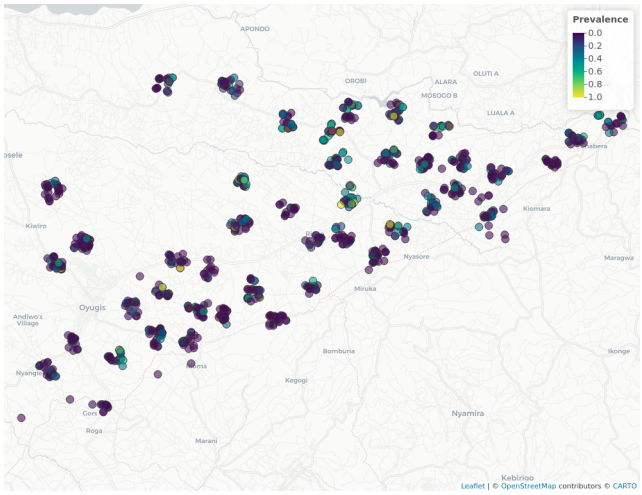

Figure 1.3: Malaria prevalence data from a cross-sectional survey carried out in Nyanza Province, in the Western Highlands of Kenya.

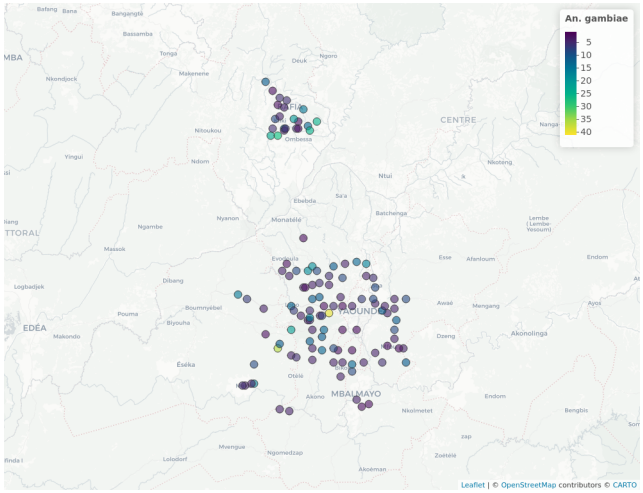

Figure 1.4: Map of the collected number of *Anopheles gambiae* mosquitoes in an area of Southern Cameroon.

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.

1.5 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. In Chapter 3, we will look more closely at what type of correlation functions can be used for $\rho(\cdot)$ and how these affect our predictive inferences. Furthermore, the fact that 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 1.2 gives the range, mean and variance the three specifications for $[Y_i | S(x_i)]$ which we will consider in this book. In Table 1.2, 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 1.2: 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	$1, \ldots, m_i$	$m_i\mu_i$	$m_i\mu_i(1-\mu_i)$	$g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$
Poisson	$1, 2, \ldots, \infty$	$m_i\mu_i$	$m_i\mu_i$	$g(\mu_i) = \log\{\mu_i\}$

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.

1.6 Workflow of a statistical analysis and structure of the book

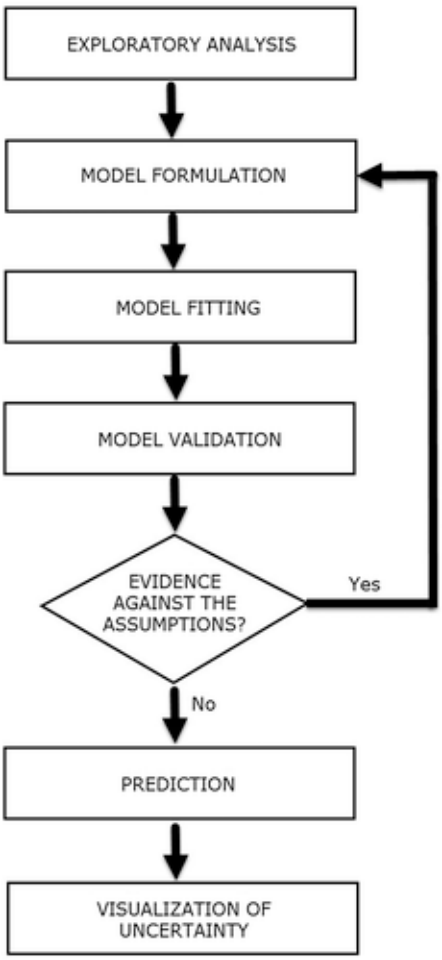

Figure 1.5: Stages of a statistical analysis

Figure 1.5 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 1.5. We treat visualization of uncertainty together with spatial prediction in Chapter 7.

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 7 and Chapter 8 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 6 illustrated the use of methods that can be used to validate the assumptions and calibration of statistical models.

Chapter 7 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 8 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

3.1 Exploratory analysis

As illustrated in Figure 1.5, 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

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 of Binomial outcomes which a require a different treatment from other type of outcomes. If you want to

read more on how to carry out exploratory analysis of the data, we refer you to Chapter 1 of Weisberg (2014).

3.1.1.1 When the outcome is an aggregated Binomial 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")
```

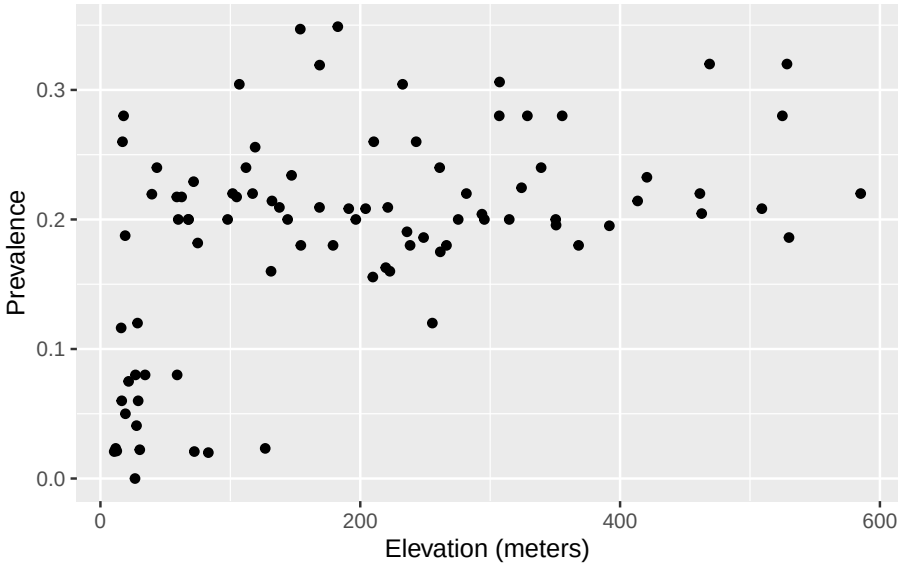

Figure 3.1: Scatter plot of the empirical prevalence for river-blindess against elevation, measured in meters.

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 1.2 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)
```

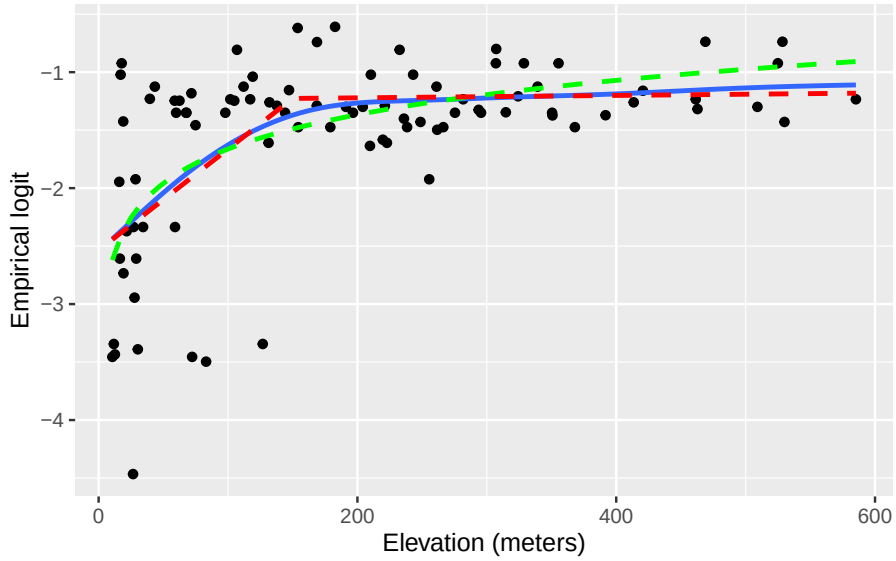

Figure 3.2: Scatter plot of the empirical prevalence for river-blindess against elevation, measured in meters.

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

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

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 we move on, it is important to briefly discuss the differences between using 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 $e \approx 2.718$ (Euler's number). To avoid this, one could rescale the regression coefficient as, for example, $\beta_1 / \log_2(e)$ which would be interpreted as the multiplicative change in the odds for river-blindness for a doubling in elevation. However, a doubling in elevation is less meaningful when considering larger values of elevation.

When the goal of statistical analysis is instead in developing a predictive model for the outcome of interest, the explanatory power and interpretability of the model may be of less concern. For this reason, the model with the log-transformed model could be preferred over the model with the linear spline,

if it shown to yield more predictive power. We will come back to this point again in Chapter 7, where will show how to assess and compare the predictive performance of different geostatistical models.

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.

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$Age_class <- cut(malkenya$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,
                                    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$age_mean_point <- aggregate(Age ~ Age_class
 ↪  + Gender,
                                   data = malkenya,
                                   FUN = mean)$Age


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

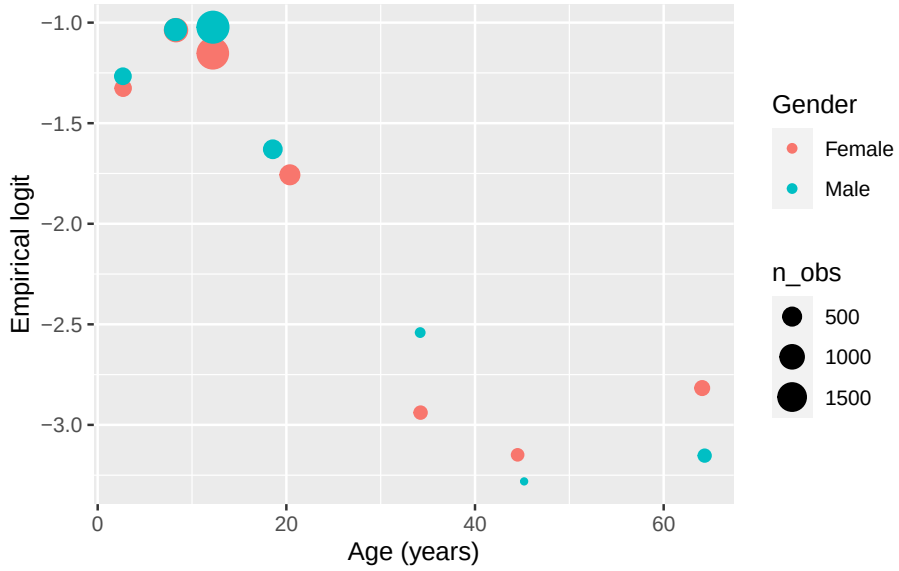

Figure 3.3: Plot of the empirical logit against age, for males and females. The size of each solid point is rendered proportional to the number of individuals within age group, as indicated in the legend.

The resulting plot in Figure 3.3 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 3.3, 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 3.3 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, 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)
##
## Coefficients:
##                                Estimate Std. Error z
 ↪  value Pr(>|z|)
## (Intercept)                   -1.135655   0.089314
 ↪  -12.715  < 2e-16 ***
## Age                           -0.001674   0.008502
 ↪  -0.197    0.844
## pmax(Age - 15, 0)             -0.095720   0.016123
 ↪  -5.937 2.91e-09 ***
## pmax(Age - 40, 0)              0.121917   0.022680
 ↪  5.376 7.64e-08 ***
## Age:GenderMale                 0.007733   0.005080
 ↪  1.522    0.128
## GenderMale:pmax(Age - 15, 0) -0.026943   0.022766
 ↪  -1.183    0.237
## GenderMale:pmax(Age - 40, 0)  0.022514   0.040543
 ↪  0.555    0.579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
 ↪  0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be
 ↪  1)
##
##     Null deviance: 8538.2  on 8203  degrees of freedom
## Residual deviance: 8224.6  on 8197  degrees of freedom
## AIC: 8238.6
##
## 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, 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      8200     8227.4
## 2      8197     8224.6  3   2.8706    0.412
```

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 illustrated in this section to explore the association with gender, can also be applied to explora the association with other continuous variables which could also be a property of the household and not of the individual (e.g. elevation).

3.2 Exploring overdispersion in count data

3.3 Exploring residual spatial correlation

4

Linear Gaussian model

5

Generalized linear geostatistical models

6

Model validation

This is a book created from markdown and executable code.

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

```
1 + 1
```

```
[1] 2
```

6.1 How to simulate geostatistical data from a fitted model

6.2 Validating the calibration of the model

6.3 Validating the spatial correlation of the model

7

Geostatistical prediction

This is a book created from markdown and executable code.

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

```
1 + 1
```

```
[1] 2
```

7.1 Pixel-level predictive targets

7.2 Area-level predictive targets

7.3 Comparing the predictive performance of geostatistical models

8

Case studies

This is a book created from markdown and executable code.

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

```
1 + 1
```

```
[1] 2
```

8.1 Mapping stunting risk in Ghan

8.2 Mapping river blindness in Malawi

8.3 Mapping mosquitoes abundance in Cameroon

References

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.

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.

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.

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.

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.