1 Overview of This Book

1.1 Volumes I and II

This book, *Beginner's Guide to Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA*, consists of two volumes. You are reading Volume I, *Using GLM and GLMM*. Volume II is entitled *Using GAM and Zero-Inflated Models*.

1.1.1 Volume I

In Volume I we explain how to apply linear regression models, generalised linear models (GLM), and generalised linear mixed-effects models (GLMM) to spatial, temporal, and spatial-temporal data. The models that will be employed use the Gaussian and gamma distributions for continuous data, the Poisson and negative binomial distributions for count data, the Bernoulli distribution for absence–presence data, and the binomial distribution for proportional data.

1.1.2 Volume II

In Volume II we apply zero-inflated models and generalised additive (mixed-effects) models to spatial and spatial-temporal data. We also discuss models with more exotic distributions like the generalised Poisson distribution to deal with underdispersion and the beta distribution to analyse proportional data.

1.2 What type of spatial data do we analyse in this book?

The short answer to this question is 'geostatistical' data. The long answer explains the three types of spatial data and then comes to the same answer. The spatial statistical literature distinguishes three types of spatial data, namely (i) lattice and areal data, (ii) geostatistical data, and (iii) spatial point pattern data (Cressie 1991; Schabenberger and Pierce 2002; Haining 2003). We briefly discuss each of these types of data.

1.2.1 Areal and lattice data

Suppose that we sample the number of tuberculosis cases in cattle in each county in England, or the total air pollution for each European country, or the total number of fish catches per European country, or the number of deaths per health clinic in a large city, or the number of babies born per country, or the crime rates per city block.

In a general notation we can state that we sample the realisations $y(\boldsymbol{s})$ of a stochastic process $Y(\boldsymbol{s})$, where $\boldsymbol{s}$ is part of a study area D. The study area in spatial studies is typically two-dimensional, which means that we can write $D \in \mathbb{R}^2$. If we have N samples in our study area D, then we can write

the observed values as $y(\boldsymbol{s}_1)$, …, $y(\boldsymbol{s}_N)$. This set of observations is one realisation of $Y(\boldsymbol{s}_1)$, …, $Y(\boldsymbol{s}_N)$. The latter is also called a random field.

In all examples that we gave in the first paragraph of this subsection the $\boldsymbol{s}$ represents an areal unit (country, county, or health clinic) and $y(\boldsymbol{s})$ is the aggregated or averaged value for that areal unit. In all these examples D is a countable collection of spatial units. If the areas are regular placed then we call it lattice data, otherwise it is areal data.

Areal data are common in medical science because quite often only aggregated data (per hospital or county) are available due to patient confidentiality. In ecology, one has to search much harder to find areal data. Two examples are the aggregation of numbers of species per area (e.g. fish landings per area of the sea; see www.ices.dk for examples) and average air pollution levels per country (see for examples www.eea.europa.eu/). In this book we will not focus on areal or lattice data.

1.2.2 Geostatistical data

In geostatistical data analysis we do not work with a value $y(\boldsymbol{s})$ for a specific areal unit $\boldsymbol{s}$. Instead $\boldsymbol{s}$ represents a spatial index like latitude and longitude. In Chapters 2, 4, and 12 we will use a field study in which pH is sampled at 253 locations in Ireland (see Figure 1.1); hence we sample $pH(\boldsymbol{s}_1)$, $pH(\boldsymbol{s}_2)$ to $pH(\boldsymbol{s}_{253})$ in the study area D, and the $\boldsymbol{s}_i$ contain the spatial coordinates.

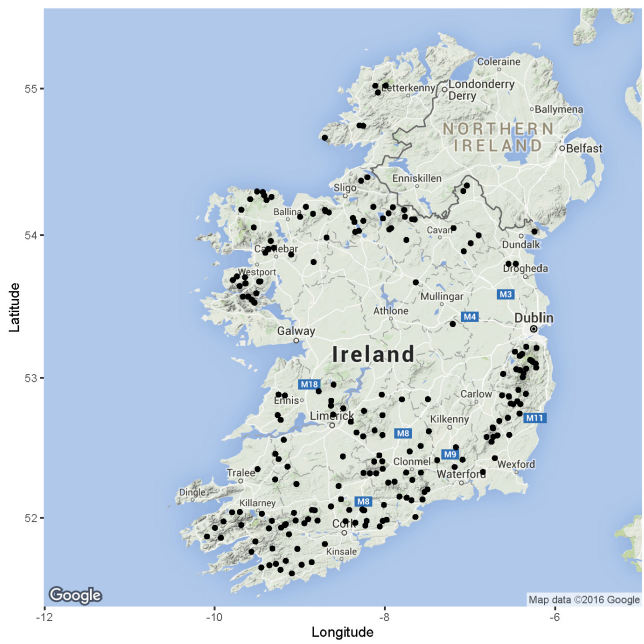

Figure 1.1. Spatial positions (black dots) of the 257 sampling locations in Ireland.

Other examples of geostatistical data that are analysed in Volume I are a plant species diversity index sampled at 890 sites on La Palma, Spain (Chapter 13), numbers of fledged bird chicks sampled at 181 nests on Santa Catalina Island, California (Chapter 15), and absence–presence of coral disease in 68 reef colonies in Haulover Bay on the northeast side of St. John, US Virgin Islands (Chapter 16).

1.2.3 Spatial point pattern data

Suppose that in a bird study we sample N areas. If a particular species is present in area i, then we set $y(\boldsymbol{s}_i) = 1$; otherwise it is 0. Or suppose we sample N areas for the presence of a disease in humans. Again we end up with $y(\boldsymbol{s}_1)$ to $y(\boldsymbol{s}_N)$ values that are either 0 (absent) or 1 (present).

The null hypothesis in these studies is complete spatial randomness of the areas where we have a 1. Interest is then whether there is any spatial clustering of areas where the disease occurs or where the bird is present. It also possible to include covariates in this type of analysis. For example, we can investigate whether the patterns where a certain disease appears differ in terms of the covariate effects. Or a pollution effect may cause birds to appear in only certain parts of the study area.

In spatial point pattern data analysis the sampling locations $\boldsymbol{s}$ are random, and we investigate the patterns in the positions of the points.

In this book we will not focus on spatial point data.

1.3 Outline of this book

In Chapter 2 we discuss an important topic: dependency. Ignoring this means that we have pseudoreplication. We present a series of examples and discuss how dependency can manifest itself.

We briefly discuss frequentist tools that are available for the analysis of temporal and spatial data in Chapters 3 and 4, and we will conclude that their application is rather limited, especially if non-Gaussian distributions are required. We will therefore consider alternative models, but these require Bayesian techniques.

In Chapter 5 we discuss linear mixed-effects models to analyse hierarchical (i.e. clustered or nested) data, and in Chapter 6 we outline how we add spatial and spatial-temporal dependency to regression models via spatial (and/or temporal) correlated random effects.

In Chapter 7 we introduce Bayesian analysis, Markov chain Monte Carlo techniques (MCMC), and Integrated Nested Laplace Approximation (INLA). INLA allows us to apply models to spatial, temporal, or spatial-temporal data.

In Chapters 8 through 16 we present a series of INLA examples. We start by applying linear regression and mixed-effects models in INLA (Chapters 8 and 9), followed by GLM examples in Chapter 10. In Chapters 11 through 13 we show how to apply GLM models on spatial data. In Chapter 14 we discuss time-series techniques and how to implement them in INLA. Finally, in Chapters 15 and 16 we analyse spatial-temporal models in INLA.

1.4 Prerequisites

We assume that readers are familiar with R (see for example Zuur et al. 2009b) and multiple linear regression. Working knowledge of Poisson, negative binomial, and Bernoulli GLM is also recommended though we do provide a short revision of these topics in Chapter 10.

1.5 Availability of the R code and data

All R code and data sets are available on the website for the book. In the preface we explain how to open the R files.