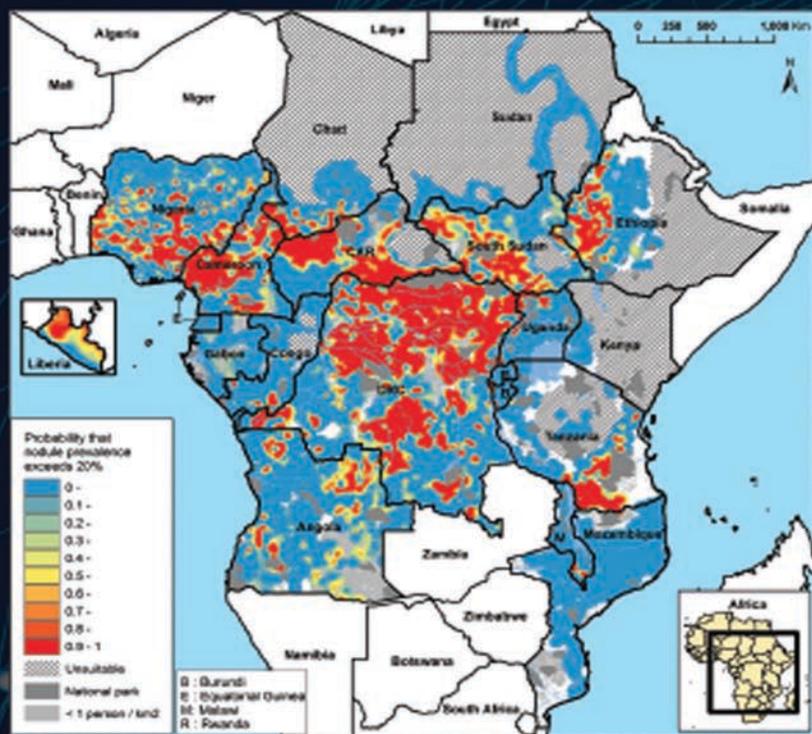


CHAPMAN & HALL/CRC
INTERDISCIPLINARY STATISTICS SERIES

MODEL-BASED GEOSTATISTICS FOR GLOBAL PUBLIC HEALTH

Methods and Applications



PETER J. DIGGLE
EMANUELE GIORGI



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Model-based Geostatistics for Global Public Health

Methods and Applications

CHAPMAN & HALL/CRC

Interdisciplinary y Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Recently Published Titles

MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION

S.Burgess and S.G.Thompson

POWER ANALYSIS OF TRIALS WITH MULTILEVEL DATA

M. Moerbeek and S.Teerenstra

STATISTICAL ANALYSIS OF QUESTIONNAIRES

A UNIFIED APPROACH BASED ON R AND STATA

F. Bartolucci, S. Bacci, and M. Gnaldi

MISSING DATA ANALYSIS IN PRACTICE

T. Raghunathan

SPATIAL POINT PATTERNS

METHODOLOGY AND APPLICATIONS WITH R

A. Baddeley, E Rubak, and R.Turner

CLINICAL TRIALS IN ONCOLOGY,THIRD EDITION

S. Green, J. Benedetti, A. Smith, and J. Crowley

CORRESPONDENCE ANALYSIS IN PRACTICE, THIRD EDITION

M. Greenacre

STATISTICS OF MEDICAL IMAGING

T. Lei

CAPTURE-RECAPTURE METHODS FOR THE SOCIAL AND MEDICAL SCIENCES

D. Böhning, P. G. M. van der Heijden, and J. Bunge

THE DATA BOOK

COLLECTION AND MANAGEMENT OF RESEARCH DATA

Meredith Zozus

MODERN DIRECTIONAL STATISTICS

C. Ley and T. Verdebout

SURVIVAL ANALYSIS WITH INTERVAL-CENSORED DATA

A PRACTICAL APPROACH WITH EXAMPLES IN R, SAS, AND BUGS

K. Bogaerts, A. Komarek, E. Lesaffre

STATISTICAL METHODS IN PSYCHIATRY AND RELATED FIELD

LONGITUDINAL, CLUSTERED AND OTHER REPEAT MEASURES DATA

Ralitza Gueorguieva

FLEXIBLE IMPUTATION OF MISSING DATA, SECOND EDITION

Stef van Buuren

COMPOSITIONAL DATA ANALYSIS IN PRACTICE

Michael Greenacre

MODEL-BASED GEOSTATISTICS for GLOBAL PUBLIC HEALTH

METHODS and APPLICATIONS

Peter J. Diggle and Emanuele Giorgi

For more information about this series, please visit: <https://www.crcpress.com/go/ids>

Model-based Geostatistics for Global Public Health

Methods and Applications

Peter J. Diggle
Emanuele Giorgi



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2019 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper
Version Date: 20190125

International Standard Book Number-13: 978-1-138-73235-3 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

To Mandy and Iulia



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

Preface	xi
List of Figures	xv
List of Tables	xxv
1 Introduction	1
1.1 Motivating example: mapping river-blindness in Africa	1
1.2 Empirical or mechanistic models	5
1.3 What is in this book?	7
2 Regression modelling for spatially referenced data	11
2.1 Linear regression models	11
2.1.1 Malnutrition in Ghana	13
2.2 Generalised linear models	16
2.2.1 Logistic Binomial regression: river-blindness in Liberia	16
2.2.2 Log-linear Poisson regression: abundance of <i>Anopheles</i> <i>Gambiae</i> mosquitoes in Southern Cameroon	20
2.3 Questioning the assumption of independence	21
2.3.1 Testing for residual spatial correlation: the empirical variogram	24
3 Theory	29
3.1 Gaussian processes	29
3.2 Families of spatial correlation functions	31
3.2.1 The exponential family	31
3.2.2 The Matérn family	32
3.2.3 The spherical family	34
3.2.4 The theoretical variogram and the nugget variance . .	35
3.3 Statistical inference	37
3.3.1 Likelihood-based inference	38
3.4 Bayesian Inference	42
3.5 Predictive inference	43
3.6 Approximations to Gaussian processes	44
3.6.1 Low-rank approximations	45
3.6.2 Gaussian Markov random field approximations via stochastic partial differential equations	48

4	The linear geostatistical model	55
4.1	Model formulation	55
4.2	Inference	57
4.2.1	Likelihood-based inference	57
4.2.1.1	Maximum likelihood estimation	58
4.2.2	Bayesian inference	59
4.2.3	Trans-Gaussian models	61
4.3	Model validation	62
4.3.1	Scenario 1: omission of the nugget effect	63
4.3.2	Scenario 2: miss-specification of the smoothness parameter	64
4.3.3	Scenario 3: non-Gaussian data	64
4.4	Spatial prediction	66
4.5	Applications	70
4.5.1	Heavy metal monitoring in Galicia	70
4.5.2	Malnutrition in Ghana (continued)	75
4.5.2.1	Spatial predictions for the target population	78
5	Generalised linear geostatistical models	83
5.1	Model formulation	84
5.1.1	Binomial sampling	85
5.1.2	Poisson sampling	87
5.1.3	Negative binomial sampling?	88
5.2	Inference	89
5.2.1	Likelihood-based inference	89
5.2.1.1	Laplace approximation	89
5.2.1.2	Monte Carlo maximum likelihood	90
5.2.2	Bayesian inference	91
5.3	Model validation	93
5.4	Spatial prediction	94
5.5	Applications	95
5.5.1	River-blindness in Liberia (continued)	95
5.5.2	Abundance of <i>Anopheles Gambiae</i> mosquitoes in Southern Cameroon (continued)	98
5.6	A link between geostatistical models and point processes	99
5.7	A link between geostatistical models and spatially discrete processes	102
6	Geostatistical design	105
6.1	Introduction	105
6.2	Definitions	107
6.3	Non-adaptive designs	107
6.3.1	Two extremes: completely random and completely regular designs	108
6.3.2	Inhibitory designs	109

6.3.3	Inhibitory-plus-close-pairs designs	109
6.3.3.1	Comparing designs: a simple example	112
6.3.4	Modified regular lattice designs	114
6.3.5	Application: rolling malaria indicator survey sampling in the Majete perimeter, southern Malawi	115
6.4	Adaptive designs	117
6.4.1	An adaptive design algorithm	118
6.5	Application: sampling for malaria prevalence in the Majete perimeter (continued)	120
6.6	Discussion	122
7	Preferential sampling	123
7.1	Definitions	123
7.2	Preferential sampling methodology	125
7.2.1	Non-uniform designs need not be preferential	126
7.2.2	Adaptive designs need not be strongly preferential	126
7.2.3	The Diggle, Menezes and Su model	127
7.2.4	The Pati, Reich and Dunson model	127
7.2.4.1	Monte Carlo maximum likelihood using stochas- tic partial differential equations	128
7.3	Lead pollution in Galicia	130
7.4	Mapping ozone concentration in Eastern United States	134
7.5	Discussion	138
8	Zero-inflation	141
8.1	Models with zero-inflation	141
8.2	Inference	144
8.3	Spatial prediction	145
8.4	Applications	146
8.4.1	River blindness mapping in Sudan and South Sudan	146
8.4.2	Loa loa: mapping prevalence and intensity of infection	150
9	Spatio-temporal geostatistical analysis	157
9.1	Setting the context	158
9.2	Is the sampling design preferential?	160
9.3	Geostatistical methods for spatio-temporal analysis	163
9.3.1	Exploratory analysis: the spatio-temporal variogram	164
9.3.2	Diagnostics and novel extensions	166
9.3.2.1	Example: a model for disease prevalence with temporally varying variance	167
9.3.3	Defining targets for prediction	168
9.3.4	Accounting for parameter uncertainty using classical methods of inference	168
9.3.5	Visualization	170

9.4	Historical mapping of malaria prevalence in Senegal from 1905 to 2014	171
9.5	Discussion	180
10	Further topics in model-based geostatistics	183
10.1	Combining data from multiple surveys	183
10.1.1	Using school and community surveys to estimate malaria prevalence in Nyanza province, Kenya	184
10.2	Combining multiple instruments	188
10.2.1	Case I: Predicting prevalence for a gold-standard diagnostic	189
10.2.2	Case II: Joint prediction of prevalence from two complementary diagnostics	190
10.3	Incomplete data	191
10.3.1	Positional error	191
10.3.2	Missing locations	195
10.3.2.1	Modelling of the sampling design	196
	Appendices	199
A	Background statistical theory	201
A.1	Probability distributions	201
A.1.1	The Binomial distribution	202
A.1.2	The Poisson distribution	202
A.1.3	The Normal distribution	203
A.1.4	Independent and dependent random variables	204
A.2	Statistical models: responses, covariates, parameters and random effects	206
A.3	Statistical inference	208
A.3.1	The likelihood and log-likelihood functions	208
A.3.2	Estimation, testing and prediction	210
A.3.3	Classical inference	211
A.3.4	Bayesian inference	215
A.3.5	Prediction	216
A.4	Monte Carlo methods	217
A.4.1	Direct simulation	218
A.4.2	Markov chain Monte Carlo	218
A.4.3	Monte Carlo maximum likelihood	220
B	Spatial data handling	223
B.1	Handling vector data in R	223
B.2	Handling raster data in R	227
	References	231
	Index	243

Preface

This book provides an introductory account of model-based geostatistics and its application in public health research.

The term *geostatistics* is a short-hand for the collection of statistical methods relevant to the analysis of geolocated data, in which the aim is to study geographical variation throughout a region of interest but the available data are limited to observations from a finite number of sampled locations. This scenario is typical of applications in low-resource settings where comprehensive disease registries do not exist. Accordingly, most of the examples in the book relate to public health research in low-to-middle-income countries, drawing on our experience of collaborative work in Africa, Asia and South America.

Geostatistical methods originated in the South African mining industry in the early 1950s (Krige, 1951). They were subsequently developed, by Georges Matheron and colleagues at Fontainebleau, France, into a self-contained methodology for addressing problems of spatial prediction; see Matheron (1963) or, for a book-length account, Chilès & Delfiner (2016). This methodology has subsequently been applied in many different fields, spanning the social, physical and health sciences. Watson (1971, 1972) first pointed out the connection between geostatistics and classical stochastic process prediction. The books by Ripley (1981) and Cressie (1991) subsequently placed geostatistics within the more general setting of statistical methods for spatially referenced data. Diggle et al. (1998) coined the term *model-based geostatistics* to mean the application of general principles of statistical modelling and inference to the analysis of geostatistical data. In particular, they emphasised the use of likelihood-based inference within an explicitly declared parametric model, typically a generalized linear mixed model (Breslow & Clayton, 1993) with a latent spatial process included in the linear predictor.

The R software environment (www.r-project.org) has become the standard vehicle for disseminating new statistical methodology as open-source software through the provision of R *packages* as add-ons to the basic R language. Packages are made available through the CRAN repository, which is accessible via the R project web-page, <https://cran.r-project.org/>. All of the analyses reported in this book can be reproduced using the R package `PrevMap` and its predecessor `geoR`. R scripts are provided on the book's web-site, <https://sites.google.com/site/mbgglobalhealth/>.

Many of the public health applications described in the book fall under the general heading of *disease mapping* problems. A basic scenario is the following. How can we best use data on empirical prevalences of a disease of interest at a

set of sampled locations within a designated region A to construct a map of the spatial variation in prevalence throughout A ? Many variations on this basic scenario arise according to the practical focus of particular applications. What is the relationship between disease risk and exposure to one or more spatially varying risk-factors? Where do unexplained “hot-spots” occur within A ? How is the spatial distribution of prevalence changing over time? What would be an efficient spatial sampling design for monitoring changes in prevalence over time?

The remainder of our applications concern *exposure mapping*, i.e. constructing spatially continuous maps of potential exposure to risk-factors, such as air pollutant concentrations, from a spatially discrete network of measurement sites. Similar questions are relevant in this context, and can again be answered using geostatistical methods.

Our aim has been to write a book that is accessible not only to statisticians but also to students and researchers in the public health sciences. Those in the latter category may initially struggle with some of the mathematical formalism that we use in describing the various statistical models and methods. However, we believe that the effort involved in becoming comfortable with mathematical notation, and with some basic concepts in probability and statistical inference, is well worthwhile for at least three reasons. Firstly, expressing a statistical model in mathematical terms forces precision of thought and explicit declaration of underlying assumptions, both of which can be masked by vague statements of the kind, “we fitted a regression model of disease risk on age, gender and socio-economic status.” Secondly, understanding the differences amongst statistical testing, estimation and prediction helps to ensure that the analysis of a set of data focuses on the correct scientific question. Finally, by embedding geostatistical methods within a general inferential paradigm we greatly reduce reliance on *ad hoc* methods and thereby ensure that our analyses are statistically efficient, i.e. within the declared model our inferences are as precise as they can be.

To help this second category of reader negotiate any initial technical difficulties, we have included a brief account of the underlying statistical theory and methods in Appendix A. Also, at the end of Chapter 1 we signpost those parts of the book that less mathematically inclined readers may wish to skip on a first reading. We emphasise that the reader needs only to understand the statements of the various results, not how they are derived.

Conversely, statisticians may be less familiar than public health scientists with software tools such as geographical information systems (GIS) for drawing the high-quality maps that are an essential part of communicating the results of a geostatistical analysis to users. We have therefore included Appendix B, which describes how to use R packages to do this. We could have used an open-source GIS instead. For example, we sometimes use the Quantum GIS (QGIS) system (<https://www.qgis.org/en/site/>) in our own work. But we think it is more helpful to the reader that we keep all of our analysis tools within a single software environment.

In writing this book we have benefited greatly from discussion and collaboration with many friends, colleagues and students without whom this book would never have been written. Special thanks are due to Madeleine Thomson (Columbia University), who introduced PJD to the world of global public health research having spotted the potential for geospatial statistical methods to contribute to this important area of work.

The opportunity to make a contribution, however small, to public health in some of the world's poorest countries has been both a humbling and an enormously rewarding experience for us both.

Peter J Diggle and Emanuele Giorgi, Lancaster, 17 November, 2018



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

List of Figures

1.1	Map of estimated pre-control prevalence of onchocerciasis infection Africa-wide.	1
1.2	Prevalence map for onchocerciasis in Liberia produced by WHO (methodology unspecified).	3
1.3	Prevalence map for onchocerciasis in Liberia produced by the application of model-based geostatistics.	4
1.4	Predictive probability map for exceedance of 20% onchocerciasis prevalence in Liberia, the WHO-defined threshold for “treatment priority areas.”	6
1.5	Residential locations (unit post-codes) of reported cases of non-specific gastrointestinal illness in Hampshire, UK.	8
1.6	Estimated relative risk of lung cancer mortality in Castile-La Mancha, Spain.	8
2.1	Gauss and the Gaussian distribution depicted on a 10 Deutschmark banknote.	12
2.2	Curves of child-growth based on length-for-age (0-2 years) and height-for-age (2-5 years). Each curve correspond to a specific length/height-for-age Z-score value (HAZ) of -3, -2, 0, 2 and 3, with 0 being the standard level of growth.	13
2.3	Plots for the data on malnutrition in Ghana: (a) point-map showing the geo-referenced locations of the sampled clusters; (b) scatter plot of height-for-age Z-scores against age; (c) box-plots of height-for-age Z-scores for each category of maternal education; (d) box-plots of height-for-age Z-scores for each category of household wealth score. The unit of distance between geographical locations in (a) is 1 km.	15
2.4	The estimated effect (solid line) of age on HAZ from model (2.2). This is formally expressed by the <i>broken stick</i> in (2.2). The dashed lines are the 95% confidence intervals.	17
2.5	Predictive probability map for exceedance of 20% skin nodule prevalence in Liberia from the Binomial regression model in (2.3).	18
2.6	Map of the locations of the <i>Anopheles gambiae</i> counts in Southern Cameroon.	19

2.7 Scatter plot of the log-counts of mosquitoes against elevation for the data on *Anopheles gambiae* counts in Southern Cameroon; the solid line is the fitted mean using the Poisson log-linear model (2.5). 20

2.8 Empirical variograms of residuals from GLMM fits to data on river blindness in Liberia (upper panels) and abundance of *Anopheles gambiae* in Cameroon (lower panels). The left-hand panels shows un-binned empirical variograms, the right-hand panels empirical variograms using a bin-width of 25 km (upper panel) and 10 km (lower panel). 26

2.9 Plots of the diagnostic check on the presence of residual spatial correlation for the data on river blindness in Liberia (upper panels) and abundance of *Anopheles gambiae* in Cameroon (lower panels). Left panels: empirical variograms of the estimate residual variation \hat{Z}_i (solid lines) and 95% tolerance bandwidth (grey area) generated under the hypothesis of spatial independence. Right panels: null distribution of the test statistic in 2.12 with the solid point on the abscissa corresponding to the observed value of the test statistic. 27

3.1 Exponential correlation functions, $\rho(u) = \exp(-u/\phi)$, for $\phi = 0.1$ and $\phi = 1.0$ (solid and dashed lines, respectively). The thin horizontal line intersects each $\rho(u)$ at its practical range, $u \approx 3 \times \phi$ 32

3.2 Realisations of stationary processes $S(x)$ with correlation function $\rho(u) = \exp(-u/\phi)$. The upper and lower panels correspond to $\phi = 0.1$ and $\phi = 1.0$, respectively. 33

3.3 Matérn correlation functions, $\rho(u; \phi, \kappa)$, for $\kappa = 0.5, 1.5, 2.5$ (solid, dashed and dotted lines, respectively) and values of ϕ chosen so that in each case the practical range is 3.0. The thin horizontal line at height 0.05 therefore intersects each correlation function at the same value, $u = 3$ 34

3.4 Realisations of stationary processes $S(x)$ with Matérn correlation function $\rho(u; \phi, \kappa)$ The upper and lower panels correspond to $\kappa = 1.5$ and $\kappa = 2.5$, respectively. 35

3.5 Spherical correlation function and Matérn correlation function with $\kappa = 1.5$ (solid and dashed lines, respectively) with values of ϕ chosen so that in each case the practical range is 3.0. . . 36

3.6 The generic form of the variogram, $V(u)$. The *nugget* is another name for the parameter τ^2 , the *sill* is likewise another name for the parameter σ^2 , whilst the *practical range* is the distance u at which the correlation function $\rho(u)$ decays to 0.05. 38

3.7	A process model expresses a scientist's assumptions about nature, a data model expresses a statistician's assumptions about how nature generates observations, statistical inference links the two.	40
3.8	Kernel regression estimates of a function $s(x)$ calculated from data Y_i observed at 100 equally spaced points x_i in one spatial dimension. Black, red and blue lines correspond to values of $\phi = 0.2, 0.4$ and 0.8 , respectively.	46
3.9	The correlation function corresponding to (3.23) with kernel function $f(u) = 2\pi^{-1}(1 - \ u\ ^2) : \ u\ \leq 1$ and $\phi = 1, 0.5$ (solid and dashed lines, respectively).	48
3.10	A Matérn correlation function (solid line) and the limiting form of the correlation function of an approximating low-rank model (3.23) with kernel function $f(u) = 2\pi^{-1}(1 - \ u\ ^2) : \ u\ \leq 1$ and $\phi = 0.9$ (dashed line).	49
3.11	The two plots show the neighbouring cells to a single grid location (red cell) whose entries in the corresponding conditional autoregressive process are different from zero. Each colour identifies entries with the same reported value. Blank cells correspond to an entry value of zero. The upper configuration is used to approximate a Matérn field with $\kappa = 1$ and the lower with $\kappa = 2$	51
3.12	The Matérn correlation function with scale parameter $\phi = 0.1$ and smoothness parameter $\kappa = 1$ (solid line). The points correspond to the correlation of a CAR process defined over a regular grid with spacing $1/64$ and precision matrix given by the upper panel of Figure 3.11.	52
3.13	Left panel: triangulations within the unit square under different constraints. Right panel: resulting piece-wise linear approximation of a continuous surface. The solid blue line is the domain polygon.	53
4.1	Log-Gaussian, Gamma and Uniform density functions, $f(\theta)$. In all three cases, $\int_1^{10} f(\theta)d\theta = 0.95$	61
4.2	Results from the graphical test based on the algorithm described in Section 4.3 for each of the models of Section 4.3.1 to 4.3.3. In each panel, the solid line is the empirical variogram of the residuals from a standard linear regression model. The shaded area is a 95% pointwise tolerance band generated under the fitted geostatistical model.	65
4.3	Point-map of the Galicia lead concentration data. Each circle has center at a sampled spatial location and radius proportional to the measured lead concentration. The continuous line shows the administrative boundary of Galicia.	71

4.4	Histograms of the Galicia lead concentration data(left panel) and log-transformed data (right panel).	72
4.5	Galicia lead concentration data. Results from variogram diagnostic check for the presence of residual spatial correlation (left-hand panel) and for compatibility of the data with the fitted geostatistical model (4.19, right-hand panel). The solid line is the empirical variogram of the data. The shaded areas are 95% tolerance bands under the hypothesis of spatial independence (left-hand panel) and under the fitted model (4.19, right-hand panel).	73
4.6	Spatial prediction for Galicia lead concentration data. Point predictions (left panel) and standard errors (right panel) of the log-transformed lead concentration were computed on a 20 by 20 km regular grid over Galicia. In both plots, the vertical axis corresponds to results from the likelihood-based method, and the horizontal axis to results from the Bayesian method. . . .	75
4.7	Spatial prediction of lead concentrations in moss samples, based on the model (4.19). Predictive mean (left-hand panel) and predictive probability of exceeding 4ppm dry weight (right-hand panel).	76
4.8	Results from the diagnostic check on the compatibility of the adopted spatial correlation function for the malnutrition data. The solid line corresponds to the empirical variogram of the residuals from the standard linear regression. The shaded area represents the 95% tolerance bandwidth obtained using the algorithm described in Section 4.3.	78
4.9	Results from the spatial prediction of the malnutrition data. Upper panels: predicted surfaces of height-for-age Z-scores (HAZ). Lower panels: maps of the predictive probability that HAZ lies below a threshold of -2. Left panels: age is fixed at 1 year, maternal education and wealth index at their lowest score. Right panels: results for the general population of children in Ghana aged less than 5 years; see Section 4.5.2.1 for more details on how these are obtained. In each of the maps, the results are reported on a 10 by 10 km regular grid covering the whole of Ghana.	79
4.10	Histogram of the age distribution in the data on malnutrition.	81
5.1	Plot of the marginal probability that $Y_i = 0$ from model (5.7) against $1/\text{Var}[e_i^U]$, when e^{U_i} is log-Gaussian (black line) or Gamma (red line) distributed.	88

5.2	Diagnostic results for the analysis on river-blindness in Liberia. The solid line is the empirical variogram based on estimated random effects from a non-spatial binomial mixed model. The shaded area is a 95% tolerance bandwidth generated by the algorithm of Section 5.3.	97
5.3	Maps of the predicted nodule prevalence (left panel) and exceedance probability of 20% prevalence (right panel) in Liberia. Contours of 20% prevalence and of 25% and 75% exceedance probabilities are shown in the left and right panels, respectively.	97
5.4	Schematic representation of individual events (solid dots) occurring in a geographical region. In the left-hand panel, events are coloured red or black according to whether they do or do not fall within one of a set of delineated sampling areas. In the right-hand panel, the number of events in each sampled area is shown as a circle at the centroid of each sampled area, with radius size-coded to correspond to counts of 0, 1, 2, 3.	100
6.1	Completely random (left-hand panel) and square lattice (right-hand panel) designs, each consisting of $n = 100$ sampling locations on the unit square.	108
6.2	Three inhibitory designs with $n = 100$ points and packing densities 0.1, 0.2 and 0.4 (left, centre and right panels, respectively).	110
6.3	A hypothetical empirical variogram (open circles) and two theoretical variograms (solid and dashed lines) that give equally good fits.	111
6.4	Two inhibitory plus close pairs designs with $n = 100$ points on the unit square. The design in the left-hand panel has 95 points in its inhibitory component and 5 paired points, each within a distance 0.05 of their companion. The design in the right-hand panel has 75 points in its inhibitory component and 25 paired points, each within a distance 0.02 of their companion. In both designs, the packing density of the inhibitory component is 0.4. Inhibitory and paired points are shown as closed and open circles, respectively.	112
6.5	Three sampling designs with $n = 100$ points on the unit square: completely random (left-hand panel); inhibitory (centre panel); inhibitory plus close pairs (right-hand panel). See text for detailed specifications.	113
6.6	Realisations of two stationary process each with mean $\mu = 0$, variance $\sigma^2 = 1$ and exponential correlation function, $\rho(u) = \exp(-u/\phi)$. In the left-hand panel, $\phi = 0.1$. In the right-hand panel, $\phi = 0.2$	114

6.7	Two modified square lattice-based designs: lattice-plus-close-pairs (left-hand panel); lattice-plus-in-fill (right-hand panel). completely random (left-hand panel) and square lattice (right-hand panel) designs, each consists of $n = 100$ sampling locations on the unit square.	115
6.8	The map of Malawi, showing Majete Wildlife Reserve highlighted (left) and its perimeter with focal areas A, B and C highlighted (right).	116
6.9	The optimised inhibitory plus close pairs design in sub-area A. Sampled and unsampled households are indicated by blue and black dots, respectively.	118
6.10	An efficiency comparison between non-adaptive (NAGD) and adaptive (AGD) designs with respect to spatially averaged prediction variance. All designs are inhibitory with minimum permissible distance between sampled locations $d_0 = 0.03$. For the adaptive designs, initial designs were of size $n_0 = 30, 40, \dots, 90$, batch sizes were $b = 1, 5, 10$. Calculations assumed a Gaussian model with Matérn correlation and parameters $\phi = 0.05$, $\kappa = 1.5$. Adapted from Chipeta et al. (2016).	120
6.11	Initial inhibitory sampling design (red dots) and first wave of adaptive sampling locations (blue dots) in sub-area B. Inset shows an expanded view of a sub-set of sampled locations. . .	121
7.1	Sampling locations for the two surveys of lead concentrations in moss samples. The two maps correspond to surveys conducted in 1997 (left panel) and 2000 (right panel). Unit of distance is 100 km. Each point is represented by a symbol corresponding to a quintile class of the observed lead-concentration values as indicated by the legend.	124
7.2	Kernel density estimate, $\hat{f}(x)$, of the sampling density for the 1997 lead concentration data (left-hand panel), scatter plot of log-transformed lead concentrations, Y_i , against $\log \hat{f}(x_i)$ (right-hand panel, dashed line is the least squares fit.	130
7.3	Triangulated mesh used for the SPDE representation of the field \mathcal{S}_1	131
7.4	Trace plot (central panel), correlogram (central panel) and empirical cumulative density function (right panel) of the first and second 5000 samples of the region-wide average of $\tilde{\mathcal{S}}_1$, obtained from the independence sampler algorithm.	132
7.5	Plug-in predictions for the spatial variation in lead concentrations in 1997 (upper left panel) and 2000 (upper panel), and the ratio of the latter over the former (lower panel)	133
7.6	Locations of the air pollution monitoring stations across the Eastern United States.	135

7.7 Scatter plots of the sampling intensity against the log-transformed population density (upper panel), the ozone concentration against the log-transformed population density (middle panel) and the ozone concentration against the sampling intensity. The sampling intensity is estimated using a Gaussian kernel density estimate. The red solid lines are least squares fit with corresponding 95% confidence intervals as dashed lines. The *broken sticks* in the upper and lower panels are constructed using a knot at 5 for the log-transformed population density. 136

7.8 Predicted surfaces for the sampling intensity (left panel) and ozone concentration in ppb (right panel). 137

8.1 Plot of the three different disease prevalence patterns at the boundary of endemic areas: continuous and decreasing trajectory that never reaches exactly zero (black line); continuous and decreasing trajectory that exactly reaches zero (red line); and discontinuous trajectory that exactly reaches zero (green line). 142

8.2 Map of the 900 sampled locations in Sudan and South Sudan. The background is a physical map of the region with the solid light blue line representing the Nile river. 147

8.3 Map of the estimated probability of suitability for river-blindness, $\pi(x)$, from zero-inflated model of Section 8.4.1. . . 148

8.4 Surfaces of the estimated river-blindness nodule prevalence (left panels) and its probability of exceeding a threshold of 20% (right panels) from the standard (upper panels) and zero-inflated (middle panels) geostatistical models. The lower panels show the difference between the surfaces from the two models. 149

8.5 Expected fraction of individuals with more than 8000 MF counts per ml of blood, in each of the sampled viallges in the study sites in Cameroon (upper panel), the Republic of Congo (central panel) and the Democratic Republic of Congo (lower panel). 153

8.6 Empirical standardized variograms based on the predictive mean of the random effects associated with prevalence (a) and intensity (b), and their standardized cross-variogram (c), from the non-spatial model of Schlüter et al. (2016). The dashed lines represent the theoretical standardized variograms and cross-variogram from fitted geostatistical model to the Loa loa data, given by (8.8) and (8.9), respectively. 154

8.7 Scatter plot of the point estimates (a) and length of the 95% predictive intervals (b) for the prediction target, defined in (8.10) with $c = 8000$, from the non-spatial model of Schlüter et al. (2016) and the spatial model of Giorgi et al. (2017). . . . 154

9.1	Diagram of the different stages of a statistical analysis. . . .	163
9.2	User interface of a Shiny application for visualization of results. The underlying data are described in Section 9.4.	169
9.3	Locations of the sampled communities in each of the time-blocks indicated by Table 9.1.	171
9.4	The plots show the results from the Monte Carlo methods used to test the hypotheses of spatio-temporal independence (upper panels) and of compatibility of the adopted covariance model with the data (lower panels). The shaded areas represent the 95% tolerance region under each of the two hypotheses. The solid lines correspond to the empirical variogram for $\tilde{Z}(x_i, t_i)$, as defined in Section 9.3.1. In the lower panels, the theoretical variograms obtained from the least squares (dotted lines) and maximum likelihood (dashed lines) methods are shown.	174
9.5	Density functions of the maximum likelihood estimator for each of the model parameters based on parametric bootstrap (PB), as black lines, and the Gaussian approximation (GA), as orange lines; the blue lines correspond to the posterior density from the Bayesian fit.	175
9.6	Profile deviance (solid line) for the parameter of spatio-temporal interaction ξ of the Gneiting (2002) family given by (9.11). The dashed line is the 0.95 quantile of a χ^2 distribution with one degree of freedom.	177
9.7	Scatter plots of the point estimates (upper panels) and standard errors (lower panels) of <i>Plasmodium falciparum</i> prevalence for children between 2 and 10 years of age, using plugin, parametric bootstrap and Bayesian methods. The dashed red lines in each panel is the identity line.	178
9.8	(a) Predictive mean (solid line) of the country-wide average prevalence with 95% predictive intervals. (b) Predictive probability of the country-wide average prevalence exceeding a 50% threshold.	178
9.9	(a) Predictive mean surface of prevalence for children between 2 and 10 ($PfPR_{2-10}$); (b) Exceedance probability surface for a threshold of 5% $PfPR_{2-10}$. Both maps are for the year 2014. The contour lines correspond to 5% $PfPR_{2-10}$, in the left panel, and to 25%, 50% and 75% exceedance probability, in the right panel.	179
9.10	Prevalence estimates (left panel) and standard errors (right panel) based on the Demographic and Health Survey conducted in Senegal in 2014. Those are obtained from a model using a spatial indicator for urban and rural communities (x-axis) and excluding this explanatory variable (y-axis). The dashed line in both graphs is the identity line.	181

10.1	Geographical coordinates of the sampled compounds in the community (black points) and school (grey triangles) surveys.	185
10.2	Maps of: (a) the point predictions of $B^*(x)$, the bias surface associated with the school-based sample of Kenyan children; (b) $r(x)$, the predictive probability that $B^*(x)$ lies outside the interval from 0.9 to 1.1 ().	187
10.3	(a) scatter plot of the standard errors for $S(x)$ using models fitted to the data from the community survey against the community and school surveys, with dashed identity line; (b) plot of the prediction locations where black (grey) points correspond to locations where a reduction (an increase) in the standard errors for $S(x)$ is estimated when using the data from both the community and school surveys.	188
10.4	Representation of the Giorgi et al. (2015) model as a directed acyclic graph.	189
A.1	Three examples of the binomial distribution, $p(y)$, with $m = 10$ and $\theta = 0.2, 0.5, 0.8$ (solid, dashed and dotted lines, respectively). The first and third are mirror images of each other. The second is symmetric about the point $y = 5$. In all three, the maximum value of $p(y)$ is at $y = m\theta$, the expectation of Y	203
A.2	Two examples of the Poisson distribution, $p(y)$, with $\mu = 2, 5$ (solid and dashed lines, respectively), and a binomial distribution with $m = 100$ and $p = 0.05$, hence expectation $\mu = 5$ (dotted line). The second and third are almost identical.	204
A.3	The probability density functions, $f(y)$ of three Normal distributions, $N(5, 1)$, $N(10, 1)$, $N(10, 4)$ (solid, dashed and dotted lines, respectively). Changing μ shifts $f(y)$ along the y -axis; changing σ^2 increases or decreases the spread of $f(y)$	205
A.4	The Poisson log-likelihood function for a single observation $y = 10$	209
A.5	The Poisson deviance function for a single observation $y = 10$ (solid line) and for a set of five independent observations with sample mean $\bar{y} = 10$ (dashed line).	210
A.6	Graphical calculation of a maximum likelihood estimate and a likelihood interval. The plotted function is the Poisson log-likelihood (A.10) for sample size $n = 100$ and sample mean $\bar{y} = 10$	213
B.1	The left panel shows a map of the second level subdivision of Liberia. The central and right panels add a 5 by 5 regular grid and the area of each district, respectively.	226
B.2	Interactive map of the area of each district in Liberia obtained using the <code>tmap</code> package.	226
B.3	Maps of the elevation, in meters, in Liberia.	228

B.4 Map of the distance from the closest river on a 5 by 5 km regular grid. The black lines represent the digitized rivers contained in the `water` object. 228

List of Tables

2.1	Ordinary least squares estimates with nominal standard errors (Std. Error) and 95% confidence intervals (CI) for the regression coefficients of the linear regression model in (2.2).	16
2.2	Maximum likelihood estimates with associated standard errors (Std. Error) and 95% confidence intervals (CI) for the regression coefficients of the Binomial regression model in (2.3).	18
2.3	Maximum likelihood estimates with associated standard errors (Std. Error) and 95% confidence intervals (CI) for the regression coefficients of the Poisson regression model in 2.5.	21
4.1	P-values based on the test statistic (4.13) for the misspecified and true model under each of the three scenarios in Section 4.3.1 to 4.3.3.	64
4.2	Galicia lead concentration data. Parameter estimation using likelihood-based and Bayesian methods. The table gives maximum likelihood estimates (MLE), standard errors (Std. Error) and 95% confidence interval for the likelihood-based method, posterior means, posterior standard errors and 95% credible intervals for the Bayesian method.	74
4.3	Maximum likelihood estimates with associated standard errors (Std. Error) and 95% confidence intervals (CI) for the regression coefficients of the linear geostatistical model in (4.20).	77
4.4	Empirical joint distribution of maternal education (from 1=“Poorly educated” to 3=“Highly educated”) and wealth index (from 1=“Poor” to 3=“Wealthy”).	80
5.1	Maximum likelihood estimates and associated 95% confidence intervals (CI) for the model in (5.19) using the Laplace (Section 5.2.1.1) and the Monte Carlo likelihood (Section 5.2.1.2) methods.	96
5.2	Monte Carlo maximum likelihood estimates and associated 95% confidence intervals for the model in (5.20).	98
6.1	Average squared prediction errors for three designs (D1, D2, D3) and realisations of two simulations models (SGP1, SGP2). See text for specification of designs and simulation models. . .	113

6.2	Monte Carlo maximum likelihood estimates and 95 % confidence intervals for the binomial logistic geostatistical model fitted to malaria prevalence data in Majete sub-area B.	117
6.3	Monte Carlo maximum likelihood estimates and 95 % confidence intervals for the model fitted to the initial Majete malaria data.	122
7.1	Maximised log-likelihoods for four models fitted by Diggle et al. (2010) to the Galicia lead pollution data.	132
7.2	Monte Carlo maximum likelihood parameter estimates and 95% confidence interval for the model fitted to the 1997 and 2000 surveys of lead concentrations in moss samples. The unit of measure for ϕ_1 and ϕ_2 is 100km.	133
7.3	Monte Carlo maximum likelihood estimates and associated 95% confidence intervals for the model on ozone concentration.	137
8.1	Monte Carlo maximum likelihood estimates and association 95% confidence intervals for the standard and zero-inflated geostatistical models of Section 8.4.1.	148
8.2	Monte Carlo maximum likelihood estimates and their 95% confidence intervals (CI) for the geostatistical model of Section 8.4.2.	152
9.1	Number of surveys and country-wide average <i>Plasmodium falciparum</i> prevalence, in each time-block.	173
9.2	Maximum likelihood estimates of the model parameters and their 95% confidence intervals (CI) based on the asymptotic Gaussian approximation (GA) and parametric bootstrap (PB).	176
9.3	Posterior mean and 95% credible intervals of the model parameters from the Bayesian fit.	176
10.1	Explanatory variables used in the analysis of malaria prevalence in Nyanza province, Kenya.	186
10.2	Monte Carlo maximum likelihood estimates and corresponding 95% confidence intervals.	186

1

Introduction

CONTENTS

1.1	Motivating example: mapping river-blindness in Africa	1
1.2	Empirical or mechanistic models	5
1.3	What is in this book?	7

1.1 Motivating example: mapping river-blindness in Africa

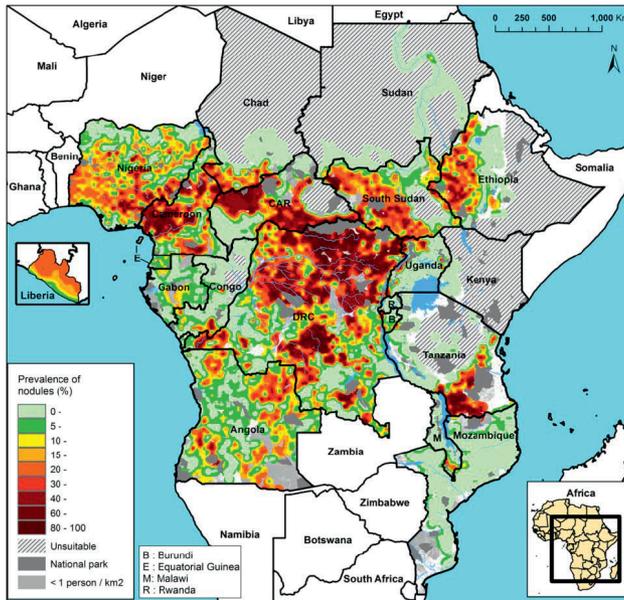


FIGURE 1.1

Map of estimated pre-control prevalence of onchocerciasis infection Africa-wide.