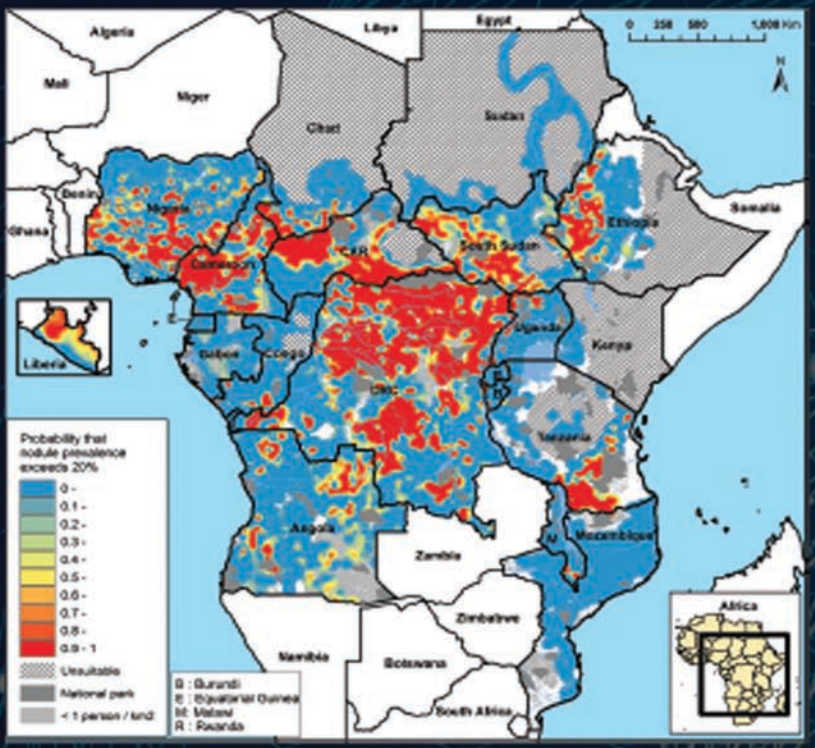
CHAPMAN & HALL/CRC
INTERDISCIPLINARY STATISTICS SERIES

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

Interdisciplinar 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 FORTHE SOCIALAND 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

FLEXBILE 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

Printed on acid-free paper
Version Date: 20190125

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

To Mandy and Iulia

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

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 decribed 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 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 heath 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

List of Tables

1

Introduction

CONTENTS

1.1 Motivating example: mapping river-blindness in Africa

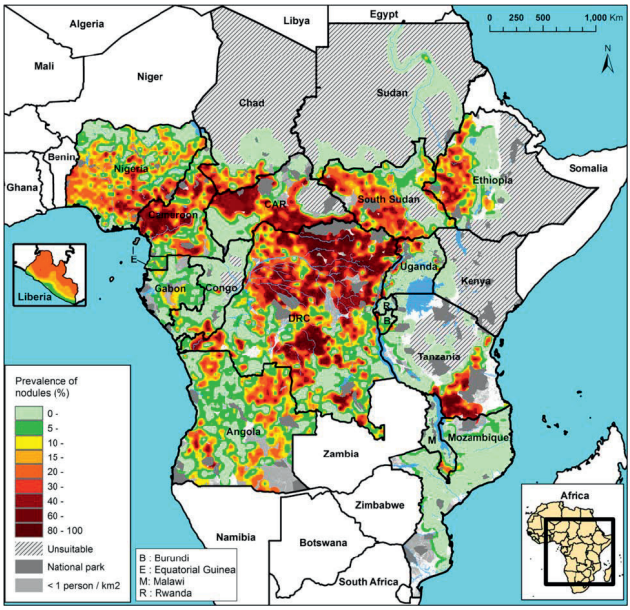


FIGURE 1.1
Map of estimated pre-control prevalence of onchocericasis infection Africa-wide.