Disease Mapping with WinBUGS and MLwiN

Andrew B. Lawson  
Department of Epidemiology and Biostatistics  
University of South Carolina, USA

William J. Browne  
School of Mathematical Sciences  
University of Nottingham, UK

Carmen L. Vidal Rodeiro  
Department of Epidemiology and Biostatistics  
University of South Carolina, USA
Disease Mapping with WinBUGS and MLwiN
Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title’s special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.
Disease Mapping with WinBUGS and MLwiN

Andrew B. Lawson
Department of Epidemiology and Biostatistics
University of South Carolina, USA

William J. Browne
School of Mathematical Sciences
University of Nottingham, UK

Carmen L. Vidal Rodeiro
Department of Epidemiology and Biostatistics
University of South Carolina, USA
## Contents

- **Preface** ix
- **Notation** xi
  - 0.1 Standard notation for multilevel modelling xi
  - 0.2 Spatial multiple-membership models and the MMMC notation xii
  - 0.3 Standard notation for WinBUGS models xii

1 **Disease mapping basics** 1
  - 1.1 Disease mapping and map reconstruction 2
  - 1.2 Disease map restoration 3

2 **Bayesian hierarchical modelling** 17
  - 2.1 Likelihood and posterior distributions 17
  - 2.2 Hierarchical models 18
  - 2.3 Posterior inference 19
  - 2.4 Markov chain Monte Carlo methods 20
  - 2.5 Metropolis and Metropolis–Hastings algorithms 21
  - 2.6 Residuals and goodness of fit 26

3 **Multilevel modelling** 29
  - 3.1 Continuous response models 30
  - 3.2 Estimation procedures for multilevel models 35
  - 3.3 Poisson response models 38
  - 3.4 Incorporating spatial information 42
  - 3.5 Discussion 43

4 **WinBUGS basics** 45
  - 4.1 About WinBUGS 45
  - 4.2 Start using WinBUGS 47
  - 4.3 Specification of the model 50
  - 4.4 Model fitting 59
  - 4.5 Scripts 64
  - 4.6 Checking convergence 65
Contents

4.7 Spatial modelling: GeoBUGS 67
4.8 Conclusions 72

5 MLwiN basics 75
5.1 About MLwiN 75
5.2 Getting started 77
5.3 Fitting statistical models 84
5.4 MCMC estimation in MLwiN 94
5.5 Spatial modelling 104
5.6 Conclusions 113

6 Relative risk estimation 115
6.1 Relative risk estimation using WinBUGS 115
6.2 Spatial prediction 137
6.3 An analysis of the Ohio dataset using MLwiN 139

7 Focused clustering: the analysis of putative health hazards 155
7.1 Introduction 155
7.2 Study design 156
7.3 Problems of inference 158
7.4 Modelling the hazard exposure risk 160
7.5 Models for count data 164
7.6 Bayesian models 166
7.7 Focused clustering in WinBUGS 167
7.8 Focused clustering in MLwiN 190

8 Ecological analysis 197
8.1 Introduction 197
8.2 Statistical models 198
8.3 WinBUGS analyses of ecological datasets 199
8.4 MLwiN analyses of ecological datasets 219

9 Spatially-correlated survival analysis 235
9.1 Survival analysis in WinBUGS 235
9.2 Survival analysis in MLwiN 237

10 Epilogue 251

Appendix 1: WinBUGS code for focused clustering models 253
A.1 Falkirk example 253
A.2 Ohio example 256
Appendix 2: S-Plus function for conversion to GeoBUGS format 263

Bibliography 267

Index 275
Statistics in Practice

Human and Biological Sciences

Brown and Prescott – Applied Mixed Models in Medicine
Ellenberg, Fleming and DeMets – Data Monitoring in Clinical Trials: A Practical Perspective
Lawson, Browne and Vidal Rodeiro – Disease Mapping with WinBUGS and MLwiN
Marubini and Valsecchi – Analysing Survival Data from Clinical Trials and Observation Studies
Parmigiani – Modeling in Medical Decision Making: A Bayesian Approach
Senn – Cross-over Trials in Clinical Research, Second Edition
Senn – Statistical Issues in Drug Development
Whitehead – Design and Analysis of Sequential Clinical Trials, Revised Second Edition
Whitehead – Meta-analysis of Controlled Clinical Trials

Earth and Environmental Sciences

Buck, Cavanagh and Litton – Bayesian Approach to Intrepreting Archaeological Data
Glasbey and Horam – Image Analysis for the Biological Sciences
Webster and Oliver – Geostatistics for Environmental Scientists

Industry, Commerce and Finance

Aitken – Statistics and the Evaluation of Evidence for Forensic Scientists
Ohser and Mücklich – Statistical Analysis of Microstructures in Materials Science
Preface

The analysis of disease maps has seen considerable development over the last decade. This development has been reflected in a fast-increasing literature and has been matched by the development of software tools. The intersecting areas of spatial statistical methods development and geographical information systems (GIS) have both witnessed this growth. With increasing public health concerns about environmental risks and even bioterrorism, the need for good methods for analysing spatial health data is immediate. Two major software tools have now been developed, which allow the modelling of spatially-referenced small area health data. These tools, MLwiN and WinBUGS, both provide facilities for sophisticated modelling of realistically complex health data. WinBUGS was developed to allow the application of a wide range of hierarchical Bayesian models, exploiting modern computational advances, in particular Gibbs sampling. MLwiN was developed to allow the fitting of models to multilevel data where a natural parameter hierarchy exists. Originally, this was implemented using iterative likelihood and quasi-likelihood estimation methods. However, the most recent versions of the package have implemented Bayesian computational methodology and now have many parallel capabilities. Increasingly both packages are being used by researchers and also now there is a desire to be able to apply such methodology in practical public health applications. In response to this need, the authors have attempted to provide an introduction to the methods and types of applications where such modelling is feasible. We do not claim to provide a comprehensive text on disease mapping and have confined our attention to the main application of these methods to counted data, where numbers of cases are recorded within small areas.

This book is designed to be of interest to final-year undergraduate and graduate level statistics and biostatistics students but will also be of relevance to epidemiologists and public health workers both in higher education and beyond. The book provides in the introductory chapters (Chapters 1–5) general background to disease mapping, Bayesian hierarchical modelling and multilevel modelling approaches, and basic introductions to the use of WinBUGS and MLwiN. The latter part of the book is focused on application areas, and is divided between relative risk estimation (Chapter 6), focused clustering (Chapter 7), ecological analysis (Chapter 8), and finally spatial survival analysis (Chapter...
Throughout the book we provide clear descriptions of the model programming execution and analysis of and interpretation of results. We have adopted the philosophy that we would attempt to demonstrate how MLwiN and WinBUGS approach the same data example, but also have included examples where either one or the other packages have limitations. We cannot necessarily hope to provide definitive answers to how modelling is to be approached in every case. However, we would hope that we provide useful pointers to the issues and potential benefits of the approaches described. As both MLwiN and WinBUGS are evolving packages, it is to be expected that features described here may vary in the future. However, we have done our best to describe the current or soon-to-be current form of the packages which is relevant to the potential audience for this published work. All the material described here is available in WinBUGS 1.4 (see Section 4.8.2 for download information and website www.mrc-bsu.cam.ac.uk/bugs), and in MLwiN (see section 5.6.1 and website http://multilevel.ioe.ac.uk/index.html for more details). Most datasets used in this book are available to download (with associated WinBUGS code) from the site http://www.sph.sc.edu/alawson/.

We would like to acknowledge the help and contribution of a number of people during the development of this work. First, we would like to acknowledge the help of the MLwiN project team, in particular Jon Rasbash, Harvey Goldstein, Amy Burch, Lisa Brennan, Fiona Steele and Min Yang. We would like to thank Allan Clark, Robin Puett, Lance Waller, Tom Richards, James Hebert, Alastair Leyland, Sudipto Banerjee, Robert McKeown, Ken Kleinman, Peter Rogerson, Dan Wartenburg and Martin Kulldorff for support and encouragement in the project. In addition, we would like to acknowledge the data availability afforded by the sophisticated online public access GIS layer system developed by, amongst others, Guang Zhao of the South Carolina Department of Health and Environmental Control. Finally, the continuing support and encouragement of Sian Jones and Rob Calver at Wiley Europe must be acknowledged and is much appreciated.

Andrew Lawson (Columbia, SC, USA)
William Browne (Nottingham, UK)
Carmen Vidal Rodeiro (Columbia, SC, USA)

March 2003
Notation

In complex random effects models there is often a myriad of different ‘standard’ notations to represent a statistical model. This is generally because the models were first discovered by many different authors at roughly the same time and each author had their own particular notation and style.

0.1 STANDARD NOTATION FOR MULTILEVEL MODELLING

In this book, in the multilevel modelling sections, as we will be using the MLwiN software package, we will use the notation used by this software package.

If we consider a three-level nested Normal model, then the standard multilevel model will be written as

\[ y_{ijk} = \beta X + v_k + u_{jk} + e_{ijk}, \quad v_k \sim N(0, \sigma_v^2), \quad u_{jk} \sim N(0, \sigma_u^2), \quad e_{ijk} \sim N(0, \sigma_e^2) \]

Here the fixed effects are represented by \( \beta \), \( X \) is a design matrix, and the random effects at levels 1, 2 and 3 are represented by \( e, u, \) and \( v \) respectively. Level 1 units are indexed \( i \), level 2 units \( j \) and level 3 units \( k \). There is a rather unfortunate notational clash as in disease mapping \( e \) is typically used to represent the expected counts. However, the level 1 residuals disappear from the equation in the Poisson response multilevel model which minimizes confusion. A three-level Poisson response model is typically written in MLwiN as

\[ y_{ijk} \sim \text{Poisson}(\pi_{ijk}), \]
\[ \log (\pi_{ijk}) = \log (e_{ijk}) + X\beta + v_k + u_{jk}, \]
\[ v_k \sim N(0, \sigma_v^2), \quad u_{jk} \sim N(0, \sigma_u^2). \]

In standard disease mapping \( \theta \) is often used rather than \( \pi \), and the \( e_{ijk} \) is often put on the right-hand side of the equation.
0.2 SPATIAL MULTIPLE-MEMBERSHIP MODELS AND THE MMMC NOTATION

The disadvantage of the standard multilevel notation is that it relies on the nested structure of the model. Browne et al. (2001) consider more general random effect structures including crossed random effects and multiple-membership structures. Rather than give an index for each classification (level in a nested structure) they instead use mapping functions to define the unit in the classification that a particular observation belongs. For example let us consider the three-level Poisson model and assume that the levels are counties within regions within nations. Then in the notation of Browne et al. (2001) we can write Equation (1) as follows:

\[ y_i \sim \text{Poisson}(\pi_i), \]
\[ \log (\pi_i) = \log (e_i) + X\beta + u^{(3)}_{\text{nation}[i]} + u^{(2)}_{\text{region}[i]}, \]
\[ u^{(3)}_{\text{nation}[i]} \sim N(0, \sigma^2_{u^{(3)}}), \quad u^{(2)}_{\text{region}[i]} \sim N(0, \sigma^2_{u^{(2)}}). \]

So here we define all terms with respect to the lowest (observation) level which is labelled \( i \). The functions \( \text{nation}[i] \) and \( \text{region}[i] \) are mapping functions that return the nation and region respectively that observation \( i \) belongs to. As the random part of the model consists of a set of classifications which need not now be ordered in terms of nesting (and if the model contained crossed effects could not) we simply define each set of random effects with the letter \( u \) but include a superscript that gives the classification a number. We start numbering from 2 as 1 is reserved for the observation level.

The spatial multiple-membership models that we will consider later can be easily written in this notation as follows:

\[ y_i \sim \text{Poisson}(\pi_i), \]
\[ \log (\pi_i) = \log (e_i) + X\beta + \sum_{j \in \text{neigh}(i)} w_{ij} u^{(3)}_j + u^{(2)}_{\text{region}[i]}, \]
\[ u^{(3)}_j \sim N(0, \sigma^2_{u^{(3)}}), \quad u^{(2)}_{\text{region}[i]} \sim N(0, \sigma^2_{u^{(2)}}). \]

Here we have a set of region effects indexed by 2 and a set of neighbour effects that are indexed by 3.

0.3 STANDARD NOTATION FOR WinBUGS MODELS

In hierarchical models for disease maps, the notation commonly used is slightly different from that used in multilevel models. The basic Poisson likelihood model is defined as
where \( e_i \) is the expected count and \( \theta_i \) is the relative risk in the \( i \)th small area. Note that in the notation of multilevel models, \( \pi_i = e_i \theta_i \). It is also common to use \( \hat{\lambda}_i = e_i \theta_i \), and this form is used in Chapter 7.

Modelling focuses on \( \theta_i \). Here we assume this notation for all the standard analysis within WinBUGS. In addition to region specific notation we also introduce space–time notation with a second subscript denoting the time period:

\[
y_{ik} \sim \text{Poisson}(e_{ik} \theta_{ik}).
\]

Here, \( k \) denotes the relevant time period and the expected count and relative risk are allowed to vary over time periods.

When random effects are introduced into models it is usual to denote region-specific uncorrelated heterogeneity as \( v_i \), and correlated heterogeneity for the same unit as \( u_i \). This differs slightly from the convention in multilevel models.

In each section the relevant notation for that section is introduced and it is hoped that any differences between sections will not create difficulties for the reader.