MODELS AND METHODS FOR SPATIAL DATA: DETECTING OUTLIERS AND HANDLING ZERO-INFLATED COUNTS

by

Laurie Ainsworth

B.A.(Honors, Psychology), Simon Fraser University, 1990
M.Sc. (Mathematics and Statistics), Queen’s University, 1992

A THESIS submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Actuarial Science

© Laurie Ainsworth 2007
SIMON FRASER UNIVERSITY
Summer 2007

All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.
APPROVAL

Name: Laurie Ainsworth
Degree: Doctor of Philosophy
Title of Thesis: Models and Methods for Spatial Data: Detecting Outliers and Handling Zero-Inflated Counts

Examining Committee:
Dr. Richard Lockhart
Chair

Dr. Charmaine Dean
Senior Supervisor

Dr. Rick Routledge
Supervisory Committee

Dr. Steve Thompson
Supervisory Committee

Dr. Paramjit Gill
Internal External Examiner

Dr. Sylvia Esterby
External Examiner,
University of British Columbia, Okanagan

Date Approved:
Abstract

Hierarchical spatial modelling is useful for modelling complex spatially correlated data in a variety of settings. Due to the complexity of spatial analyses, hierarchical spatial models for disease mapping studies have not generally found application at Vital Statistics agencies. Chapter 2 compares penalized quasi-likelihood relative risk estimates to target values based on Bayesian Markov Chain Monte Carlo methods. Results show penalized quasi-likelihood to be a simple, reasonably accurate method of inference for exploratory studies of small-area relative risks and ranks of risks.

Often the identification of extreme risk areas is of interest. Isolated ‘hot spots’/‘low spots’ which are distinct from those of neighbouring sites are not accommodated by standard hierarchical spatial models. In Chapter 3, spatial methods are developed which allow extreme risk areas to arise in proximity to one another in a smooth spatial surface, or in isolated ‘hot spots’/‘low spots’. The former is modelled by a spatially smooth surface using a conditional autoregressive model; the latter is addressed with the addition of a discrete clustering component, which accommodates extreme isolated risks and is not limited by spatial smoothness. A Bayesian approach is employed, graphical techniques for isolating extremes are illustrated, and model assessment is conducted via cross-validation posterior predictive checks.

Zero-inflated data are not uncommon yet they are not handled well by standard models. These values may be of particular interest in species abundance studies where such zeros may provide clues to physical characteristics associated with habitat suitability or individual immunity. In Chapter 4 we review the overdispersion and zero-inflation literature and
develop a series of zero-inflated spatial models. Each model highlights different features of white pine weevil infestation data. The spatial models use a variety of structures for the probability of belonging to the zero component, thus allowing the probability of ‘resistance’ to differ across models. One model focuses on individually resistant trees which are located among infested trees while another focuses on clusters of resistant trees which are likely located within protective habitats.

The final chapter discusses future research ideas which have been motivated by this thesis.
Acknowledgments

I would like to extend a grateful thank you to my senior supervisor, Dr. Charmaine Dean, for her patience, guidance and support. I would also like to thank the Department of Statistics and Actuarial Science for providing a balanced, supportive environment in which to pursue graduate studies.

I feel fortunate to have shared my graduate experience with so many wonderful fellow students: Ruth Joy, Suman Jiwani, Farouk Nathoo, Laura Cowen, Jason Neilsen, Darby Thompson, Simon Bonner, Carolyn Huston, Elizabeth Juarez and so many more. Their camaraderie has been invaluable and I look forward to continued friendships and collaborations.

For their financial support, I would like to thank the British Columbia Health Research Foundation, Geomatics for Informed Decisions, a Canadian Network of Centres of Excellence, and Simon Fraser University.

Last, but certainly not least, is my family. They have provided endless support in so many ways. Thank you for your patience and understanding; I could not have done it without you!
Contents

Approval ii

Abstract iii

Acknowledgments v

Contents vi

List of Tables ix

List of Figures xi

1 Introduction 1

1.1 Overview ................................................................. 1

1.1.1 Hierarchical Spatial Models .................................. 2

1.1.2 The CAR model .................................................. 3

1.1.3 Zero-inflation .................................................... 4

1.2 Overview of Thesis .................................................. 6

1.2.1 Approximate Inference for Disease Mapping ................. 6

1.2.2 Detection of Outliers from Smooth Maps .................... 7

1.2.3 Zero-inflated Spatial Models .................................. 8

1.2.4 Discussion .......................................................... 8
## 2 Approximate Inference for Disease Mapping

2.1 Introduction ......................................................... 9
2.2 The Spatial Model .................................................. 11
2.3 Penalized Quasi-Likelihood Inference for Relative Risk Estimates .......... 13
2.4 MCMC Estimation .................................................... 14
2.5 Small-Sample Properties of the Estimators .................................. 15
  2.5.1 Infant Mortality Analysis .................................... 16
  2.5.2 Simulation Study .............................................. 20
2.6 Discussion .......................................................... 25

## 3 Detection of Outliers in Mapping Studies

3.1 Introduction ........................................................ 27
3.2 A Spatial Model with Discrete Components for Accommodating Local Outliers 29
  3.2.1 Priors ......................................................... 31
3.3 Application to Codling Moth Data .................................. 32
3.4 Application to Weevil Infestation Data ................................ 37
3.5 Application to Infant Mortality Data ................................ 39
  3.5.1 Hotspot identification ....................................... 40
3.6 Case Study .......................................................... 42
3.7 Sensitivity to Prior Specification ................................... 48
3.8 Discussion .......................................................... 49

## 4 Zero-inflated Spatial Models

4.1 Introduction and Overview ......................................... 51
  4.1.1 Overdispersion .............................................. 52
  4.1.2 Zero-inflated Models ....................................... 53
  4.1.3 Zero-inflated Models for Correlated Data .................... 54
  4.1.4 Zero-inflated Models for Spatially Correlated Data .......... 55
4.2 Zero-inflated Spatial Models ....................................... 56
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>PQL and MCMC estimates for the infant mortality data</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>Mean estimated ranks of the highest and lowest relative risks</td>
<td>22</td>
</tr>
<tr>
<td>2.3</td>
<td>Mean square errors (MSE) of relative risks (RR) by percentile of true relative risk (PQL MSE= 0.025, MCMC MSE= 0.033)</td>
<td>23</td>
</tr>
<tr>
<td>3.1</td>
<td>Model specification: $\mu_i^b$ is the mean, conditional on the linear combination of random effects, $b$.</td>
<td>31</td>
</tr>
<tr>
<td>3.2</td>
<td>Analysis of moth counts: posterior mean estimates and 95% credible intervals for model parameters</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Analysis of infant mortality: posterior mean estimates and 95% credible intervals for model parameters</td>
<td>40</td>
</tr>
<tr>
<td>3.4</td>
<td>Estimates for selected LHAs from the analysis of infant mortality: $\hat{r}$ is the relative risk estimate (subscripts indicate the model), a superscript ‘nbr’ indicates the mean of estimates for neighbouring sites</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Estimates for the case studies: $r_{true}$ is the true relative risk, $\hat{r}$ is the relative risk estimate (subscripts indicate the model), a superscript ‘nbr’ indicates the mean of estimates for neighbouring sites</td>
<td>44</td>
</tr>
<tr>
<td>3.6</td>
<td>Case studies: estimated mean of posterior predictive distributions</td>
<td>48</td>
</tr>
<tr>
<td>4.1</td>
<td>Model specification</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td>Posterior mean estimates of proportion of years infested, proportion of resistant trees, and variability of random effects. CI is the credibility interval</td>
<td>63</td>
</tr>
</tbody>
</table>
Table 4.3  Classification of observed data and posterior median number of values generated for each category under each model...
List of Figures

Figure 1.1  Direct Acyclic Graph of the hierarchical Bayesian spatial model with a convolution Gaussian prior on the log relative risks .......................... 5

Figure 2.1  Infant mortality relative risks (RR) ............................................. 17
Figure 2.2  Relative risk of infant mortality for local health areas in British Columbia 18
Figure 2.3  Gelman-Rubin plots for assessing convergence of selected parameter estimates in the infant mortality analysis ........................................... 19
Figure 2.4  Population sizes and true relative risks for the simulation study .... 21
Figure 2.5  Coverage probabilities by population size quartiles and by relative risk size ................................................................. 24

Figure 3.1  Map of moth trap locations identifying active traps (counts ≤ 10, blue circles; counts >10, solid blue circles) empty traps with the largest (red) and smallest (yellow) posterior probabilities of membership in the zero component, and remaining empty traps (green) .............................................. 35
Figure 3.2  Mean of moth counts for neighbouring traps and posterior probabilities of component membership with the 10 largest posterior probabilities of membership in the zero component indicated in red ............................. 36
Figure 3.3  Map of spruce tree locations identifying uninfested trees (in green) highly resistant trees (in red) and infested trees (circles size represent the proportion of years infested) .................................................. 38
Figure 3.4  Analysis of infant mortality: \( \hat{r} \) is the relative risk estimate (subscripts indicate the model), a superscript ‘nbr’ indicates the mean of the estimates of neighbouring LHAs, circle radius equals the square root of the population size, and flagged LHAs are indicated in red.

Figure 3.5  Case studies: posterior probabilities of membership in the extreme components (inflated site in red); each sub-plot displays the posterior probability of non-membership in component 1 versus the posterior probability of membership in component 5.

Figure 3.6  Case studies for a site with moderate population size (Saanich): comparison of the LC model component weight (\( \hat{p}_{LC} \)), the mean of the estimated weights of neighbours (\( \bar{p}_{LC}^{nbr} \)) and model M1 posterior probabilities of membership in component 5, inflated site in red.

Figure 3.7  Case studies for a site with moderate population size (Saanich): identification of outliers using estimates from the LC model and M1; \( \hat{r} \) is the relative risk estimate (subscript indicates model), \( \hat{p}_{LC} \) is the estimated spatial component weight under the LC model, a superscript ‘nbr’ indicates the mean of the estimates of neighbours, and inflated values are indicated in red.

Figure 4.1  Correspondence of the posterior probability of resistance among the spatial models.

Figure 4.2  The relationship between the posterior estimate of resistance for trees with no infestations and measurements on neighbouring trees: mean proportion of infestations observed for neighbouring trees and the proportion of neighbouring trees which were never infested.
Figure 4.3  Largest and smallest posterior estimates of the probability of resistance (a and b) and the random effect associated with the non-resistant component (c and d). Infested trees are indicated in black with larger circles indicating a larger proportion of years infested. In a and b, trees with the largest probability of resistance are indicated in yellow; remaining non-infested trees are indicated in blue. The 100 trees with the largest (red) and smallest (yellow) estimates of the random effect associated with the non-resistant component are indicated in c and d.
Chapter 1

Introduction

1.1 Overview

Spatial modelling is useful in a variety of settings including disease mapping and species abundance studies. Hierarchical spatial models provide a flexible framework for modelling spatially correlated data in the form of, for example, counts or proportions. In this thesis we use hierarchical spatial models and extensions of such models for the analysis of health and environmental data.

In a variety of applications, identification of areas of extreme risk are of particular interest. These areas may arise in proximity to one another in a smooth spatial surface, or they may arise as isolated ‘hot spots’ or ‘low spots’ which are quite distinct from those of neighbouring sites. The latter may provide important information regarding areas of potential concern. Hierarchical spatial models often utilize a so-called conditional autoregressive (CAR) prior to accommodate spatial correlation, however this distribution may oversmooth such spatial extremes. Here, the CAR model is extended in order to accommodate and identify ‘spatial outliers’.

Further, many health and environmental applications consider datasets containing a large number of zero values. For instance, species abundance studies often observe many zero counts in a given area or at a particular location. Such zeros are interesting in that
they may provide important clues to physical characteristics associated with, for example, habitat suitability or individual immunity. There has recently been a great deal of interest in developing zero-inflated spatial models and this is one of the main areas considered in this thesis.

1.1.1 Hierarchical Spatial Models

The use of hierarchical spatial models has become increasingly popular due to computational advancements such as MCMC methods and the freely available WinBUGS software with its associated spatial package, GeoBUGS. Wikle and Anderson (2003) describe the hierarchical model in three stages:

i) the data given the spatial process and the process parameters

ii) the spatial process given the process parameters

iii) the prior distributions of the process parameters

Thus, we have a series of conditional distributions for: the data conditioned on the spatial process and parameters defining the spatial dependencies between locations, the spatial process conditioned on the parameters, and the parameters themselves. Estimation is based on sampling from the posterior distribution which is the joint distribution of the process and the parameters given the data:

\[
\{\text{process, parameters} | \text{data}\} \\
\sim \{\text{data} | \text{process, parameters}\} \{\text{process} | \text{parameters}\} \{\text{parameters}\}
\]

where \( \cdot \) denotes a probability distribution and \( [y|z] \) refers to the distribution of \( y \) given \( z \). This joint distribution describes the behaviour of the process simultaneously at all spatial locations.

Markov random fields (MRF) are a special class of spatial models that are suitable for data on discrete (countable) spatial domains in which a joint distribution is determined by a set of locally specified conditional distributions for each spatial unit conditioned on its
neighbours. MRFs include a wide class of spatial models such as auto-Gaussian models for
spatial Gaussian processes, auto-logistic models for binary spatial processes, auto-Gamma
models for non-negative continuous processes and auto-Poisson models for spatial count
processes. One very popular auto-Gaussian model is the conditional autoregressive (CAR)
model (Besag, York, Mollié 1991) which is analogous to an autoregressive time series model.

Within the hierarchical framework, spatial dependence may be introduced at the second
level of the hierarchy through spatially correlated random effects, \( b = (b_1, b_2, \ldots, b_N) \) where
each \( b_i, i \in 1, \ldots, N \), is associated with a particular location in \( \mathbb{R}^2 \). Such random effects
may account for heterogeneity which represents missing spatially structured covariates. A
convenient distributional form for the random effects is the multivariate normal distribution
with mean 0 and spatially structured covariance matrix \( \Sigma \).

1.1.2 The CAR model

The CAR model specifies \( \Sigma \) indirectly through a set of conditional distributions. The \( N \)
conditional distributions are assumed to be univariate normal:

\[
\begin{align*}
    b_i | b_j &\neq i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2), & i, j = 1, \ldots, N
\end{align*}
\] (1.1)

where \( \sigma_i^2 > 0 \) is the conditional variance, \( \mu_i = \sum_{j \in \delta_i} w_{ij} b_j \), and \( \delta_i \) is a set of neighbours of
area \( i \). The weights, \( w_{ij} \geq 0, w_{ii} = 0, i, j = 0, 1, \ldots, N \) can be based on adjacency indicators
for a regular or irregular lattice, or on the distance between points \( i \) and \( j \).

Results of Besag (1974) can be used to show that the joint distribution of \( b \) can be
written as:

\[
b \sim N(0, (I - W)^{-1}M)
\] (1.2)

where \( W = (w_{ij}) \) and \( M = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2) \). In order to ensure symmetry of \( \Sigma \), we
require \( w_{ij} \sigma_j^2 = w_{ji} \sigma_i^2 \). Further, \( (I - W) \) must be invertible and \( (I - W)^{-1}M \) must be
positive-definite. A popular weighting choice is that used with intrinsic autoregression:
\( w_{ij} = c_{ij}/c_i \) and \( \sigma_i^2 = \sigma^2/c_i \) where the \( c_{ij} \) are user defined weights and \( c_i = \sum_j c_{ij} \). With
$c_{ij} = 1$, if two sites or regions are defined as neighbours and 0 otherwise, we can write the variance of the joint distribution of the random effects as $(I - W)^{-1} \sigma^2$.

The conditional specification of the CAR model facilitates Markov Chain Monte Carlo estimation. In particular, Gibbs sampling requires one to sample from the full conditional distributions which are obtained from the joint posterior distribution. The terms in each full conditional can conveniently be obtained from a Directed Acyclic Graph (DAG) (Mollié 1996). For instance, consider the convolution prior (Besag, York, Mollié 1991) where both CAR spatial random effects and independent random effects are included in the hierarchical model. One can determine the conditional distributions via:

$$P(v|V - v) \propto P(v|\text{parents of } v) \ast \prod P(w|\text{parents of } w)$$

where $v$ is a node on the graph (Figure 1.1) and $w$ denotes the children of $v$.

The full conditional distributions may not take a standard form. However, log-concave distributions, as above, can be sampled via adaptive rejection sampling (Gilks and Wild, 1992).

Carlin and Banerjee (2002) and Gelfand and Vounatsou (2003) present multivariate extensions of the CAR model. We use a multivariate CAR model in Chapter 4 when we consider zero-inflated spatial models and such multivariate models will be discussed there.

1.1.3 Zero-inflation

Zero-inflated data is commonly modelled by a mixture model or a conditional model. Conditional models consist of separately modelling the zero mass and a truncated form of a standard discrete distribution such as the binomial, Poisson or negative binomial. Orthogonality of this parameterization simplifies computation and allows for simpler interpretation of covariate effects. In the ecology setting, one may interpret the indicator of a zero count as representing habitat suitability, and the conditional mean as representing mean abundance given suitable habitat. On the other hand, the mixture model formulation (Lambert 1989) of the zero-inflated model allows us to distinguish between structural zeros which arise
Figure 1.1: Direct Acyclic Graph of the hierarchical Bayesian spatial model with a convolution Gaussian prior on the log relative risks
due to individual immunity or unsuitable habitat, and random zeros which arise simply by chance.

Consider $\theta$, the probability of a true zero, and $\mu$, the mean parameter for the probability mass function $f$ associated with the random variable $Y$. The mixture model is formulated as,

$$Y \sim \theta I_\theta + (1 - \theta) f(Y|\mu)$$  \hspace{1cm} (1.3)

where $I_\theta$ is the degenerate distribution taking the value zero with probability one. This formulation encompasses distributions such as the binomial, Poisson, and generalized Poisson. Additional flexibility may be added to such models by incorporating random effects into $\theta$ and the distribution of $f$.

### 1.2 Overview of Thesis

This thesis considers PQL and MCMC estimation techniques, modelling and identifying spatial outliers, as well as modelling zero-inflated spatial data. It embodies three research contributions as well as a discussion of extensions of these. The first contribution addresses the performance of PQL estimation of the relative risk for a spatial hierarchical model as compared with MCMC estimation in the context of disease mapping. The second contribution develops a CAR model with a discrete clustering component which accommodates extreme risks on a smooth surface. The third reviews the literature on zero-inflation, develops several zero-inflated spatial models and compares their use and interpretation in an ecological application.

#### 1.2.1 Approximate Inference for Disease Mapping

Disease mapping is an important area of statistical research. Contributions to the area over the last twenty years have been instrumental in helping to pinpoint potential causes of mortality and to provide a strategy for effective allocation of health funding. Because of the complexity of spatial analyses, new developments in methodology have not generally found
application at Vital Statistics agencies. Inference for spatio-temporal analyses remains computationally prohibitive, for routine preparation of mortality atlases. Chapter 2 considers whether approximate methods of inference are reliable for mapping studies, especially in terms of providing accurate estimates of relative risks, ranks of regions, and standard errors of risks. These approximate methods lie in the broader realm of approximate inference for generalized linear mixed models. Penalized quasi-likelihood is specifically considered here. The main focus is on assessing how close the penalized quasi-likelihood estimates are to target values, by comparison with the more rigorous and widespread Bayesian Markov Chain Monte Carlo methods. No previous studies have compared these two methods. The quantities of prime interest are small-area relative risks and the estimated ranks of the risks which are often used for ordering the regions. It will be shown that penalized quasi-likelihood is a reasonably accurate method of inference and can be recommended as a simple, yet quite precise method for initial exploratory studies.

1.2.2 Detection of Outliers from Smooth Maps

In mapping studies, extreme risk areas may arise in proximity to one another in a smooth spatial surface. They may also arise as isolated ‘hot spots’ or ‘low spots’, which are quite distinct from those of neighbouring sites. Chapter 3, develops spatial methods which encompass both types of extreme risks. The former is modelled by a spatially smooth surface using a conditional autoregressive model; the latter is addressed with the addition of a discrete clustering component, which offers the flexibility of accommodating extreme isolated risks and is not limited by spatial smoothness. The autoregressive component incorporates the spatially correlated risk as a baseline surface, acknowledging that environmental activity, often spatially correlated, influences risk responses. The discrete component identifies hot spots/low spots of activity beyond the spatially correlated baseline risk surface. Both types of extreme risk are important, but isolated extremes may provide insight into areas with potential of being a center for future spatially correlated extreme risks. Hence these may be particularly important in terms of surveillance. A Bayesian approach to inference is
employed and graphical techniques for isolating extremes are illustrated. Model assessment is conducted via cross-validation posterior predictive checks. Three examples demonstrate the utility of the methods and case studies show the procedures to be useful for pinpointing extreme risks. In addition, sensitivity to priors is investigated.

1.2.3 Zero-inflated Spatial Models

Many environmental applications, such as species abundance studies, rainfall monitoring or tornado count reports, yield data with a preponderance of zero counts. This leads to what is called ‘zero-inflation’. Such zeros are interesting as they provide important clues to physical characteristics associated with, for example, habitat suitability or individual immunity. Chapter 4 reviews the zero-inflation literature, particularly models for correlated data. The main focus is the development of several zero-inflated spatial models used to highlight specific data features. The spatial process is modelled with normal conditional autoregressive random effects, discrete random effects or autocovariates. Models are formulated in the exponential family framework to encompass a variety of distributions for the data. The analysis of white pine weevil infestation data for spruce trees illustrates the unique features distinguished by each model. Of particular interest are the features identified by the probability of belonging to the zero component, ‘resistance’. For instance, one model focuses on individually resistant trees located among infested trees, while another focuses on clusters of resistant trees which are likely located in protective habitats. We discuss such unique features identified by the zero-inflated spatial models and make recommendations regarding application.

1.2.4 Discussion

The thesis ends with a discussion of the methods presented and a list of future research projects which have been inspired by this work.
Chapter 2

Approximate Inference for Disease Mapping

2.1 Introduction

Mapping disease or mortality risks is a useful way of displaying geographic variation in risks and identifying regions with high and low risks for further follow-up in a surveillance context. Raw risks, however, are not reliable quantities to map as they tend to be highly variable and imprecise for small areas. With the emergence of high-speed computing and sophisticated packages which blend mapping and analysis seamlessly, there have been many advances in the development of methodology for spatial analyses. See for example, Lawson, Bohning, Lesare, Biggeri, Viel & Bertollini, (1999) or Bohning (1999). However, much of this relies on complicated statistical theory and analysis. The routine adoption of such methods for exploratory work has been slow.

The problem of mapping and estimating risks typically lies in a generalized linear mixed model (GLMM) framework. The use of GLMM involves estimating random effects which represent relative risks. In the last ten years, the use of approximate methods for generalized linear mixed models, such as penalized quasi-likelihood, generalized estimating equations, and quasi-likelihood, have become commonplace because of their ease of implementation,
robustness and precision. Generally, independent random effects have been considered; here we focus on spatially correlated random effects.

Typically, maximum likelihood estimation for GLMMs with proportions or counts as outcomes requires numerical integration for the calculation of the log-likelihood, score equations and the information matrix (Breslow & Clayton, 1993). By using a Laplace approximation to the quasi-likelihood, penalized quasi-likelihood (PQL) avoids numerical integration (Breslow & Clayton, 1993). Instead, a series of weighted least squares regressions can be solved using standard software such as Splus.

PQL can be expected to perform well for nearly normal responses such as moderately large counts or binomial proportions with large denominators (Breslow & Clayton, 1993; Leroux, Lei & Breslow, 1999). PQL is computationally simple and estimates can be computed quickly, with few convergence problems. Thus, it seems to proffer a viable approximate method of inference for the analysis of count data. In the context of mapping risks, it would be particularly helpful to have available a simple inferential tool such as PQL for exploratory studies.

Markov chain Monte Carlo (MCMC) methods have been used in spatial statistics for over a decade. MCMC is a simulation technique which allows us to make approximate draws from high dimensional probability distributions which may arise from realistic statistical modelling. MCMC uses Markov chains to draw samples from the required distribution and the sample averages are used to approximate expectations. MCMC draws these Monte Carlo samples by running a cleverly constructed Markov chain for a long time. The 1996 volume edited by Gilks, Richardson and Spiegelhalter presents a variety of MCMC applications including an example of Bayesian disease mapping by Mollié (1996).

There have been no validation studies which compare PQL with standard Bayesian MCMC methods. This chapter aims to reduce that gap. Our particular focus is how well small-area risks are estimated and how well the highest and lowest risks are identified. The yardstick of good performance is that of the MCMC estimators and true relative risks. Although PQL has been considered extensively in the literature, inference has been restricted
to global parameter estimates, such as covariate effects in the general mean term or variance components. Since we specifically focus on relative risk estimation here, interval estimates for the penalized quasi-likelihood relative risk estimators are developed.

The chapter is organized as follows. Section 2.2 presents the spatial model under consideration. Section 2.3 outlines PQL methodology and develops standard errors for PQL relative risk estimators while Section 2.4 outlines MCMC methodology. Section 2.5 discusses an analysis of infant mortality data for British Columbia as well as a simulation study based on the scenario of the infant mortality analysis which evaluates the estimators. Section 2.6 concludes with a discussion and recommendations.

2.2 The Spatial Model

We consider the basic spatial model (Besag, York and Mollié, 1991) which uses adjacencies to define neighbourhoods in a conditional specification of the model. Suppose the map under study is divided into $I$ contiguous regions labelled $i = 1, \ldots, I$. For example, the province of British Columbia is partitioned into 79 local health areas. Let $y_i$ be the number of deaths (for a particular time period, age group and disease) in the $i$th region. Let $e_i$ be the ‘expected’ number of deaths, and $r_i$ be the unknown relative risks of mortality in area $i$. It is assessment of these $r_i$ which is the focus of this chapter. The expected number of deaths may be calculated based on external information such as published rates, or based on the mean over all regions, calculated from the data at hand. Conditional on the random region effects, the number of deaths in each area, $Y_i$, is assumed to be Poisson distributed with mean $\mu_i = e_i r_i$: $y_i | r_i \sim \text{Poisson}(e_i r_i)$. Here we use $e_i = n_i m$, where $m$ is a fixed effect, the overall mean rate, and $n_i$ is the population count in the $i$th region.

The conditional log-linear model specifies $\log \mu_i = \log n_i + \log m + b_i$, $b_i = \log r_i$. We allow that $b_i$ accommodate local spatially structured variation, $u_i$, and unstructured variation, $v_i$, in the relative risks. That is, we decompose $b_i$ such that $b_i = u_i + v_i$. The spatially structured variation of $u_i$ is captured through an intrinsic Gaussian autoregression model (Besag, York and Mollié, 1991; Besag and Kooperberg, 1995) while the unstructured variation of $v_i$ is
captured through a simple Gaussian distribution. Thus, we have a Poisson mixture model incorporating two normal random effects: one exhibiting full spatial autocorrelation and one with independent errors modelling unstructured heterogeneity.

The model can be written in the usual terminology of a generalized linear mixed model for linkages with descriptions of PQL estimation already present in the literature. The elements of the linear predictor, \( \eta = X\alpha + Zb + \text{offset} \), are related to the conditional means, \( \mu_i = e_i r_i \), by the link function \( g(\mu_i) = \log(\mu_i) = \eta_i \). Conditional on the random effects vector, \( b \), the observed incidence counts, \( Y \), follow a log-linear generalized linear model with conditional mean, \( \mu_i \), given by

\[
\log(\mu) = X\alpha + Zb + \text{offset}
\]

where the offset is \( \log(n) \), \( X \) and \( Z \) are design matrices for the fixed and random effects respectively, and \( \alpha \) is a vector of covariates. For the case herein, the design matrix \( X \) is a unit vector of length \( I \). Thus, \( X\alpha \) is an intercept term corresponding to \( \log(m) \). The design matrix \( Z \) is an identity matrix of dimension \( I \). The non-spatial component is distributed as an independent normal variate: \( v_i \sim N(0, \sigma_v^2) \). The spatial random component \( u_i \) can be interpreted conditionally given \( u_{-i} \), the set of spatially structured random region effects excluding the \( i \)th: \( u_i|u_{-i} \sim N(\bar{u}_\delta, \sigma_u^2/\delta_i) \), where \( \bar{u}_\delta \) is the mean of the random effects corresponding to the regions in the ‘neighbourhood’ of the \( i \)th region and \( \delta_i \) is the number of regions forming this neighbourhood. In the examples considered, neighbourhoods are defined by regions bordering or sharing a common boundary with a given region. However, many alternate definitions may be envisioned (Besag, York and Mollie, 1991; MacNab and Dean, 2000) depending on the context of the analysis.

We can decompose the covariance of the random effects in terms of the relative importance of spatial variation to non-spatial variation. Let the parameter \( \lambda \) represent the proportion of variability attributable to spatial correlation \( \lambda = \sigma_u^2/\sigma^2 \), \( \sigma^2 = \sigma_u^2 + \sigma_v^2 \), and consider \( D \), the covariance matrix of \( b \): \( D = \sigma^2(\lambda Q^{-1} + (1-\lambda)I_d) \), \( I_d \) is an identity matrix, \( Q \) has \( i \)th diagonal element equal to the number of neighbours of the \( i \)th region while for
\[
Q_{ij} = \begin{cases} 
-1 & \text{if } i \neq j \text{ and } i \text{ and } j \text{ are neighbours}, \\
0 & \text{otherwise.}
\end{cases}
\]

### 2.3 Penalized Quasi-Likelihood Inference for Relative Risk Estimates

PQL is a straightforward technique to implement for GLMMs such as the one considered here. The notation used in the previous section allows for a seamless transition to the implementation of PQL as described by Breslow and Clayton (1993). Appendix B provides details regarding PQL.

Inference for relative risks is important here. The estimate of the relative risk is the posterior mean, \( D Z' V^{-1} (Y_{(w)} - X\alpha) \), evaluated at the PQL estimates: \( \hat{\alpha}, \hat{\lambda} \) and \( \hat{\sigma}^2 \). Here, \( Y_{(w)} \) has \( i^{th} \) element \( Y_{(w)i} = \eta_i + (Y_i - \mu_i)g'(\mu_i) - \log(n_i) \), and is termed a ‘working response vector’ by Breslow and Clayton (1993); \( V = W^{-1} + ZDZ' \), and \( W = \text{diag}(\mu_i) \).

The working response vector represents a Taylor expansion of \( \log(Y) \) in the log-linear model, for example, and is treated as normally distributed in the development of PQL. The posterior variance of the relative risks, \( D - DZ' V^{-1} ZD \), may not provide a good estimate of the variability in the relative risks as this estimate ignores variability due to estimation of \( \alpha \) and the variance components \( \theta = (\lambda, \sigma^2) \).

Below we develop methods for generating accurate standard errors for PQL regional relative risk estimates by accounting for the variability introduced through estimation of \( \alpha \) and \( \theta \). From examples considered, it seems that accounting for the variability due to estimation of \( \theta \) is more important than accounting for that of \( \alpha \). However, it is simple to account for both as presented below.

Letting \( \zeta = (\lambda, \sigma^2, \alpha) \) we use as a variance estimator (c.f. Dean and MacNab, 2001)

\[
E(\text{Var}(b|\hat{\zeta})) + \text{Var}(E(b|\hat{\zeta}))
= E(\hat{D} - DZ' V^{-1} ZD) + \text{Var}(\hat{b}).
\]
The first term is approximated by $\hat{D} - \hat{D}Z\hat{V}^{-1}Z\hat{D}$ while the second is calculated as:

$$\left(\frac{\delta\hat{b}}{\delta\zeta}\right)\text{Var}(\zeta)\left(\frac{\delta\hat{b}}{\delta\zeta}\right)^t |_{\zeta=\hat{\zeta}}$$

where $\frac{\delta\hat{b}}{\delta\zeta} = (\frac{\delta\hat{b}}{\delta\lambda}, \frac{\delta\hat{b}}{\delta\sigma^2}, \frac{\delta\hat{b}}{\delta\alpha})'$, with components

$$\frac{\delta\hat{b}}{\delta\lambda} = \sigma^2(Q^{-1} - I_d)ZV^{-1}(Y(w) - X\alpha) - DZ^tV^{-1}ZZ^t\sigma^2(Q^{-1} - I_d)V^{-1}(Y(w) - X\alpha),$$

$$\frac{\delta\hat{b}}{\delta\sigma^2} = \{\lambda(Q^{-1} - I_d) + I_d\}ZV^{-1}(Y(w) - X\alpha) - DZ^tV^{-1}ZZ^t\{\lambda(Q^{-1} - I_d) + I_d\}V^{-1}(Y(w) - X\alpha),$$

and

$$\frac{\delta\hat{b}}{\delta\alpha} = -DZ^tV^{-1}X.$$

Splus 2000 was used to program PQL and compute variance estimates. Starting values for the fitting algorithm were $\sigma = 0.5$, $\lambda = 0.5$, and 0 for the regression coefficients. Iteration was terminated when changes in parameter estimates were less than some specified tolerance level (for example, we used 0.000001). It was helpful to take a single Newton step toward the estimated variance component before returning to update $\alpha$ and $b$. The Marquardy technique of inflating the diagonal terms of the information matrix may be used to control the step size in the early iterations.

### 2.4 MCMC Estimation

Bayesian approaches to the analysis of disease incidence or mortality risks view the distribution of the frailties or random effects as prior information on the variability of disease risks in the overall map. Inference about the relative risks is based on the posterior distribution $[r|y] \propto [y|r][r]$. The prior distribution $[r]$ is parameterized by the hyperparameter $\gamma$. A typical prior for the relative risks is a Gamma distribution, while a typical prior for $\log(r)$ is a normal distribution. The marginal posterior distribution of $r$, given the data, is $[r|y] = \int [y|r][r|\gamma][\gamma]d\gamma$. Direct evaluation is not always possible. However, Markov Chain Monte Carlo methods permit samples to be drawn from the joint posterior distribution $[r, \gamma|y]$ and then, from the marginal posteriors $[r|y]$ and $[\gamma|y]$. Gibbs sampling (Gelfand and
Smith 1990; Gelfand 2000) is useful when the joint posteriors are complicated but the full conditional distributions have simple forms, as is the case here.

For the disease mapping problem, the parameters of the prior distribution on the relative risks are $\sigma^2_u$ and $\sigma^2_v$. We used conjugate Gamma$(a, b)$ prior distributions on these hyperparameters; here Gamma$(a, b)$ denotes the gamma probability density function, with mean $a/b$ and variance $a/b^2$. This is a general class of hyperpriors for the inverse variance. Setting $a$ and $b$ to 0 is equivalent to the uninformative uniform hyperprior, $U(-\infty, \infty)$; we used $a = 0.5$ and $b = 0.0005$. With a gamma hyperprior, Gamma$(a_1, b_1)$ on $\sigma^{-2}_v$, the full conditional on $\sigma^{-2}_v$ becomes a Gamma$(a_1 + \frac{n}{2}, b_1 + \frac{1}{2}v^2v)$ distribution (Mollié, 1996). Likewise, choosing a gamma hyperprior, Gamma$(a_2, b_2)$, on $\sigma^{-2}_u$, the full conditional on $\sigma^{-2}_u$ becomes Gamma$(a_2 + \frac{n}{2}, b_2 + \frac{1}{2}\sum_{i=1}^n \sum_{j<i} w_{ij}(u_i - u_j)^2)$ distribution.

WINBUGS 1.3 software (freely available at http:www.mrc-bsu.cam.ac.uk/bugs) was used for such MCMC analyses. Initial values were $1/\sigma^2_u = 0.2, 1/\sigma^2_v = 0.5, \alpha = 0, u_i = 0,$ and $v_i = 0$ for all $i$. For each analysis, empirical posterior distributions were generated from the last 8,000 of the 10,000 samples. The posterior distributions provided point (mean and median) and interval estimates for $\lambda, \sigma^2, r_i$, and the rank of the relative risks.

### 2.5 Small-Sample Properties of the Estimators

Infant mortality data from British Columbia, Canada, for the years 1985-1994 are used to set the scenario for the investigation. The data contain the number of infant deaths and the infant population sizes for the 79 local health areas of the province. Infants are defined as children less than one year of age and local health areas are administrative health units in British Columbia. This data set is discussed in detail by MacNab & Dean (2002). The population sizes considered here are quite small and have much smaller denominators for constructing relative risks than for mortality from most other causes. Compounding this is the fact that infant mortality, though important to monitor, is a relatively rare event.
CHAPTER 2. APPROXIMATE INFERENCE FOR DISEASE MAPPING

2.5.1 Infant Mortality Analysis

Both the PQL and Bayesian analyses of the infant mortality data suggest a substantial amount of spatial variation. Estimates of $m$, $\lambda$ and $\sigma^2$ from both analyses are provided in Table 2.1. The raw risks or standardized mortality ratios (SMRs) range from 0.82 to 2.06.

Table 2.1: PQL and MCMC estimates for the infant mortality data

<table>
<thead>
<tr>
<th></th>
<th>PQL ($\hat{m} = 0.007$)</th>
<th>MCMC ($\hat{m} = 0.007$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Mean</td>
<td>Posterior Median</td>
</tr>
<tr>
<td>$\lambda$ (S.E.)</td>
<td>0.764 (0.253)</td>
<td>0.619 (0.395)</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(0.267, 1.000)</td>
<td>(0.004, 0.998)</td>
</tr>
<tr>
<td>$\sigma^2$ (S.E.)</td>
<td>0.077 (0.035)</td>
<td>0.079 (0.039)</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(0.009, 0.146)</td>
<td>(0.031, 0.177)</td>
</tr>
</tbody>
</table>

Relative risks range from 0.73 to 1.48 for PQL and from 0.75 to 1.48 for MCMC (Figure 2.1). The last graph in Figure 2.1 displays the strong correlation between PQL and MCMC relative risks. A map of the relative risk estimates from the PQL analysis is provided in Figure 2.2. The largest risks occur on the West Coast of Vancouver Island, along the coast, in the northern parts of the province, and in downtown Vancouver. The lowest risks occur in the southern parts of the province. This includes the Greater Vancouver Area where lower risks may be associated with larger population densities and consequently better access to hospitals and specialized health care.

Convergence of the 79 MCMC relative risk estimates was determined by review of trace and Gelman-Rubin plots (Brooks & Gelman, 1998, see Figure 2.3 for samples). The relative risk estimates are not correlated with population size under either PQL or MCMC estimation. However, the widths of the 95% confidence intervals of the relative risks appear to decrease as population sizes increase, particularly for MCMC estimation.

PQL and MCMC rank the same ten regions as having the largest relative risks with the first seven ranks being identical. The top ten relative risk estimates are very similar (range: 1.27 to 1.48, maximum difference=0.02) for PQL and MCMC. Nine of the ten regions with
Figure 2.1: Infant mortality relative risks (RR)

- **PQL RR Estimates**
- **PQL RR by Population Size**
- **PQL Width of 95% C.I.**

- **MCMC RR Estimates**
- **MCMC RR by Population Size**
- **MCMC Width of 95% C.I.**

- **Ordered PQL RR**
- **Ordered MCMC RR**
- **MCMC vs PQL RR**
Figure 2.2: Relative risk of infant mortality for local health areas in British Columbia
Figure 2.3: Gelman-Rubin plots for assessing convergence of selected parameter estimates in the infant mortality analysis.
the smallest relative risks are the same for PQL and MCMC. The smallest PQL and MCMC relative risk estimates are also very similar (range: 0.73 to 0.85, maximum difference=0.04).

Posterior distributions (not shown here) of the relative risk ranks provide interval estimates for the ranks. For Prince Rupert, the region with the largest estimated relative risk, the distribution of the ranks has median 75 while its 25th percentile is 71. For Delta, which has the smallest estimated relative risk, the corresponding distribution has a median of six with 75% of the MCMC estimates having a rank less than 12.

2.5.2 Simulation Study

Two hundred and fifty datasets were generated under the spatial model in Section 2.2. Population sizes and neighbourhood structures were identical to those of the infant mortality data (Figure 2.4a). The infant mortality analysis was used as a basis for choosing the parameter values: \( m = 0.01, \sigma^2 = 0.10 \) and \( \lambda = 0.75 \). Figure 2.4b displays the resulting 19,750 true relative risks (250 simulated values for each of 79 regions) used for our simulations. About six percent of these are quite large, greater than 1.5.

Due to the large number of simulations, it was not feasible to carry out a detailed analysis of convergence for each dataset. Instead, datasets with the largest and smallest sum of squared differences between PQL and MCMC ranks as well as three randomly selected datasets were selected for a more thorough investigation of convergence. Investigation of Gelman-Rubin plots for the individual cases examined suggest a sufficient run length for the MCMC process.

Ranks and Relative Risks

The mean estimated rank of the region with the true highest rate is 74.81 from PQL. MCMC provides two estimates for this quantity: the posterior mean, 74.70, and the posterior median, 74.54. Corresponding quantities related to the second through tenth highest and the ten lowest ranked regions are given in Table 2.2. Based on these overall measures, there is very little difference in performance of the three estimators in terms of identifying the
Figure 2.4: Population sizes and true relative risks for the simulation study
highest and lowest risks.

Table 2.2: Mean estimated ranks of the highest and lowest relative risks

<table>
<thead>
<tr>
<th>True Rank</th>
<th>PQL Mean</th>
<th>PQL Median</th>
<th>MCMC Mean</th>
<th>MCMC Median</th>
<th>b) LOWEST TEN RELATIVE RISKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>74.81</td>
<td>74.70</td>
<td>74.54</td>
<td>1</td>
<td>7.76</td>
</tr>
<tr>
<td>78</td>
<td>72.72</td>
<td>72.60</td>
<td>72.42</td>
<td>2</td>
<td>8.92</td>
</tr>
<tr>
<td>77</td>
<td>70.55</td>
<td>70.65</td>
<td>70.40</td>
<td>3</td>
<td>11.48</td>
</tr>
<tr>
<td>76</td>
<td>70.85</td>
<td>70.91</td>
<td>70.65</td>
<td>4</td>
<td>11.06</td>
</tr>
<tr>
<td>75</td>
<td>68.45</td>
<td>68.33</td>
<td>68.12</td>
<td>5</td>
<td>13.67</td>
</tr>
<tr>
<td>Mean = 77</td>
<td>71.48</td>
<td>71.44</td>
<td>71.23</td>
<td>Mean = 3</td>
<td>10.58</td>
</tr>
<tr>
<td>74</td>
<td>67.48</td>
<td>67.49</td>
<td>67.32</td>
<td>6</td>
<td>15.60</td>
</tr>
<tr>
<td>73</td>
<td>65.54</td>
<td>65.60</td>
<td>65.43</td>
<td>7</td>
<td>14.19</td>
</tr>
<tr>
<td>72</td>
<td>65.54</td>
<td>65.27</td>
<td>65.14</td>
<td>8</td>
<td>15.68</td>
</tr>
<tr>
<td>71</td>
<td>64.52</td>
<td>64.42</td>
<td>64.37</td>
<td>9</td>
<td>16.18</td>
</tr>
<tr>
<td>70</td>
<td>62.71</td>
<td>62.86</td>
<td>62.79</td>
<td>10</td>
<td>18.54</td>
</tr>
<tr>
<td>Mean = 74.5</td>
<td>68.32</td>
<td>68.28</td>
<td>68.12</td>
<td>Mean = 5.5</td>
<td>13.31</td>
</tr>
</tbody>
</table>

The largest true relative risk is ranked in the top ten 215/250 times by PQL and 214/250 times by MCMC but, it is in the lowest 50% twice for both PQL and MCMC. The difference between the PQL and MCMC ranks is large only for true relative risks of approximately one. When the true relative risk is quite high or low, the agreement in the PQL and MCMC ranks is excellent. Thus, we can expect both PQL and MCMC to be good at identifying the extreme regions. More detailed investigation of agreement for the top and bottom four true relative risks (not shown here) reveal good agreement between the PQL and MCMC rankings. The variability in agreement is largest for true relative risks close to one.

The mean square errors of the relative risks are presented in Table 2.3 by risk percentile. PQL and MCMC mean square errors are quite similar although, the PQL estimates are somewhat smaller for large relative risks.
Table 2.3: Mean square errors (MSE) of relative risks (RR) by percentile of true relative risk (PQL MSE = 0.025, MCMC MSE = 0.033)

<table>
<thead>
<tr>
<th></th>
<th>Highest</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RR</td>
<td>1.73</td>
<td>1.59</td>
<td>1.45</td>
<td>1.30</td>
<td>1.24</td>
<td></td>
</tr>
<tr>
<td>PQL</td>
<td>0.102</td>
<td>0.078</td>
<td>0.056</td>
<td>0.050</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>MCMC</td>
<td>0.134</td>
<td>0.102</td>
<td>0.074</td>
<td>0.066</td>
<td>0.045</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Lowest</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RR</td>
<td>0.58</td>
<td>0.64</td>
<td>0.70</td>
<td>0.73</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>PQL</td>
<td>0.033</td>
<td>0.030</td>
<td>0.024</td>
<td>0.022</td>
<td>0.017</td>
<td></td>
</tr>
<tr>
<td>MCMC</td>
<td>0.033</td>
<td>0.031</td>
<td>0.026</td>
<td>0.024</td>
<td>0.020</td>
<td></td>
</tr>
</tbody>
</table>

**Coverage Probabilities**

The coverage probabilities corresponding to 90% confidence intervals for the relative risks are 89.2% and 81.3% for PQL and MCMC respectively. Similarly, the 95% coverage probabilities for PQL and MCMC are 94.5% and 87.9%. PQL relative risk coverage probabilities are closer to target than those from the MCMC analysis.

It is instructive to consider further details regarding these coverage probabilities. They are segregated here by quartiles of the regional population sizes (0 to 983, 984 to 2668, 2669 to 5693, and 5694 to 52,856), and by the size of the true relative risk being estimated (5th, 10th, 15th, 20th, 40th, 60th, 80th, 85th, 90th, and 95th percentiles). PQL coverage probabilities are stable across the four population size quartiles (Figure 2.5a), while MCMC coverage probabilities are smaller for estimates corresponding to the third and fourth quartiles. The MCMC estimates are on target for relative risks close to one, while the PQL coverage probabilities are somewhat higher than targeted for these risks. The accuracy of the coverage probabilities show a distinct relation to relative risk size for both PQL and MCMC estimates (Figure 2.5b). Confidence intervals show poorer coverage when the true risks are both quite small and quite large.
Figure 2.5: Coverage probabilities by population size quartiles and by relative risk size
2.6 Discussion

The results of our study indicate that PQL is a computationally simple, fast technique which seems to yield reasonably accurate estimates of the relative risks and fair coverage probabilities for their confidence intervals. PQL has approximately the same ability to identify high and low risk regions as does MCMC.

Convergence is relatively quick for PQL, generally requiring fewer than 20 iterations. Even for the small population sizes considered here, PQL performs as well as the Bayesian MCMC method. The overall performance of the PQL and MCMC interval estimates derived here is fair, but the coverage probabilities may be untrustworthy if the risks are small or large.

Note, however, that PQL may not perform as well with smaller means. In this case, the Poisson distribution is approximately binomial. Bias corrections have been recommended when using PQL with binary data (Lin & Breslow, 1996).

MCMC analyses were run without extensive consideration of convergence for each dataset. Typically, a more detailed analysis of convergence criteria and autocorrelation of the estimates is undertaken for individual data analyses. In this way sampling can be adjusted to ensure convergence. Because such detailed investigations were not conducted for each run of the simulation, our estimators may be more biased and may tend to have larger standard errors than would have occurred if each run in the simulation had been examined. It may be that MCMC would perform better than PQL in individual cases when convergence is stringently diagnosed. However, detailed investigations of relative risks for several selected simulation runs indicated that convergence was achieved. This is in agreement with previous experience fitting similar spatial models (see for example, Eberly & Carlin, 2000).

The relative risks and ranks of the relative risks, which are the focus of this chapter, are well identified under the full spatial model. However, the spatial, \( u_i \), and non-spatial, \( v_i \), components in the model are unidentified if vague priors are used for these parameters (Eberly & Carlin, 2000). In this case, the proportion of variation due to spatial correlation, \( \lambda = \sigma^2_u / (\sigma^2_v + \sigma^2_u) \) is also unidentified. Besag, Green, Higdon & Mengersen (1995) argue
that it is legitimate to sample from unidentified spaces so long as the samples are only used to summarize components of the proper posterior. Indeed, in spite of the identifiability problem, Eberly and Carlin (2000) found that this parameter was well estimated in many of the situations they considered.

A strong advantage of MCMC is that estimation of alternate measures to describe the model is simple. The posterior distributions of many quantities, such as relative risk ranks, are easy to construct. The development of WINBUGS software has made implementation of MCMC readily available for the specific models considered here.

There are several alternate approaches to disease mapping which were not considered. Militino, Ugarte & Dean (2001) analyzed British Columbia infant mortality data for the years 1981-1985 using a non-parametric mixture model which assumes a discrete mixture of Poisson distributions. A time component is added to the non-parametric mixture approach by Bohning, Dietz & Schlattmann (2000) who accommodate spatio-temporal correlations through dynamic mixtures (DMDM).

Accounting for spatial and spatio-temporal correlations in disease risk can be quite helpful in overcoming the problems inherent in the use of crude estimators for disease mapping and surveillance. Both PQL and MCMC methods are effective for inference in spatial models, within the limitations described. For exploratory studies, such as routine atlas productions, PQL is recommended over the crude estimators traditionally employed.
Chapter 3

Detection of Local and Global Outliers in Mapping Studies

3.1 Introduction

Mapping rates is now a common activity for ecological studies. Often a major focus of mapping is the identification of extreme rates. Indeed, in the study of environmental data, such extremes offer powerful evidence of trends. Detecting outliers in space is generally an important concern for monitoring the health profile of a set of individuals (persons or trees) or for species abundance studies. Rapid identification of outliers permits a quick response to increases in rates of adverse events and to decreases in species abundance. Traditional tools, for mapping and for monitoring rates or proportions broadly, utilize models such as the conditional autoregressive model. This model has gained widespread acceptance in the spatial statistics literature. However, it is not unusual to find such models routinely used for analysis without the checks and precautions suggested by the authors of such tools. The BYM model (Besag, York, and Mollié, 1991) has become commonplace in ecological and health studies at various agencies. This model is popular because it is readily and freely available in WinBUGS software (Spiegelhalter, Thomas, Best, and Lunn, 2003). The BYM model has been shown to be flexible and robust (Lawson, Biggeri, Boehning, Lesaffre, Viel,
Clark, Schlattmann and Divino, 2000). It is most useful for obtaining an overall picture of smooth spatial trends, but it is often applied to analyses of health and environmental data when the focus is identification of extremes.

Increased flexibility can be obtained by replacing the independent normal prior of the BYM model with a more variable distribution, such as the t-distribution used by Pascutto, Wakefield, Best, Richardson, Benardinelli, Staines and Elliott (2000). On the other hand, sharp discontinuities in spatial risk surfaces have been accommodated through the use of discrete mixtures of Poisson distributions or mixture models in general. Schlattmann and Boehning (1993) use an empirical bayes approach to such a clustering analysis. Their methodology does not incorporate the spatial arrangement of the sites so that members of a cluster may be scattered throughout the map. Knorr-Held and Rasser (2000) develop a Bayesian non-parametric clustering procedure using a variable number of cluster centers and variable cluster risks, but with spatially contiguous clusters. Thus, they primarily focus on discontinuities in the risk surface. Fernandez and Green (2002) also incorporate spatial structure in a cluster model by using spatially correlated priors for the mixture component weights. Finally, Lawson and Clark (2002) address the problem of discontinuities in the risk surface through a mixture of spatial models; one component incorporates a smooth spatial surface, the second reflects a surface with spatial jumps in the risks. In one limiting case, the Lawson and Clark (LC) model reduces to the BYM model. At the other extreme, the risk surface jumps from site to site.

In the traditional framework of model fitting, diagnostics may help to identify sites whose risks differ from elsewhere on the surface. Deviance measures are often used to assess overall fit. For a discussion of these in the Bayesian context see Spiegelhalter, Best, Carlin and van der Linde (2002). Posterior predictive \( p \)-values (Gelman, Meng, and Stern, 1996) are useful for focused investigations of fit. A ‘discrepancy statistic’ is chosen to suit the type of model departures of interest. Stern and Cressie (2000) discuss posterior predictive model checks for disease mapping with small areas. These authors consider the difficulty of distinguishing between extrema which occur by chance and those which arise from truly
CHAPTER 3. DETECTION OF OUTLIERS IN MAPPING STUDIES

3.2 A Spatial Model with Discrete Components for Accommodating Local Outliers

Our model assumes an underlying smooth surface where rates may be either homogeneous or flow along a smooth gradient reflecting spatially contiguous environmental effects. Flexibility in modelling extremes is accommodated through an independent discrete random effects. We use the term ‘global outliers’ to refer to extreme rates arising from the spatially correlated random effects which are modelled here with a conditional autoregressive distribution; ‘local outliers’ refer to extreme risks arising from the discrete random effects and represent extreme risks which are discrepant from neighbouring risk estimates.
Poisson models are typically used in mapping studies because rare events are often of interest. However, binomial models are required when events are not rare. With environmental data, it is not unusual that a binomial model is more appropriate. Thus, we consider both types of models in the applications.

Here we develop a generalized formulation which encompasses both the Poisson and binomial frameworks. Conditional on the random effects, \( \mathbf{b} \), the observations, \( y_i, i = 1, 2, \ldots N \) are independent with expectation \( E(y_i|\mathbf{b}) = \mu_i^b \) and variance \( \text{var}(y_i|\mathbf{b}) = \nu_i(\mu_i^b) \) where \( \nu_i \) is a specified variance function. The link function \( g(\mu_i^b) = \eta_i^b \) relates the conditional mean to the linear predictor \( \eta_i^b = \alpha + \mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{b} \) where \( \mathbf{x}_i \) and \( \mathbf{z}_i \) are covariates associated with fixed and random effects, \( \beta \) and \( \mathbf{b} \) respectively. In our applications, \( \alpha = g(m) \) where \( m \) is the overall mean rate. In cases where an offset is required, such as when population sizes differ across sites, \( \alpha = g(m) + \text{offset} \). Although we have not considered covariates here, in general, they may be included. Conditional on the random effects, the observations are assumed to have been drawn from a linear exponential family with mean \( \mu_i^b = h(\eta_i^b) \) where \( h = g^{-1} \).

We contrast several models in Table 3.1. They have a variety of structures for additive random effects and encompass both the Poisson and binomial distributions. For a Poisson loglinear model, \( g(\mu_i^b) = \log(\mu_i^b) \) and the random effects represent the logarithm of the relative risk for each site. For a binomial logistic model, \( g(\mu_i^b) = \log(\mu_i^b/(1-\mu_i^b)) \) and the regional random effects correspond to the logarithm of the odds ratio for each site. Under the BYM model, \( u_i \) and \( v_i \) are continuous; \( v_i \sim N(0, \sigma_v^2) \), accommodates unstructured heterogeneity, while the spatial random effect, \( u_i \), accommodates local spatially structured variation. The distribution of \( u_i \) can be described and interpreted conditionally given \( u_{-i} \), the set of spatially structured random site effects excluding the \( i \)th: \( u_i|u_{-i} \sim N(\bar{u}_{\delta_i}, \sigma_u^2/\delta_i) \), where \( \bar{u}_{\delta_i} \) is the mean of the spatial random effects corresponding to sites in the ‘neighbourhood’ of the \( i \)th site and \( \delta_i \) is the number of sites in this neighbourhood.

The discrete random effects, \( d_i \), in model M1, take values \( \log(R^1), \log(R^2), \ldots, \log(R^k) \) with probability \( \theta^1, \theta^2, \ldots, \theta^k \) respectively, \( \sum_{j=1}^k \theta^j = 1, \theta = (\theta^1, \theta^2, \ldots, \theta^k) \). In this parameterization, we write the discrete random effects as taking distinct values on the log
Table 3.1: Model specification: $\mu_i^b$ is the mean, conditional on the linear combination of random effects, $b$

<table>
<thead>
<tr>
<th>Model</th>
<th>Conditional Distribution</th>
<th>Linear Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYM</td>
<td>$y_i</td>
<td>b \sim f(\mu_i^b)$</td>
</tr>
<tr>
<td>M1</td>
<td>$y_i</td>
<td>b \sim f(\mu_i^b)$</td>
</tr>
<tr>
<td>M2</td>
<td>$y_i</td>
<td>b \sim 0$ with prob $\theta^*$</td>
</tr>
</tbody>
</table>

$b$ encompasses the linear combinations of the random effects, $u, v, and d$

- $u_i \sim \text{CAR}$, $v_i \sim N(0, \sigma_v^2)$
- $d_i = \log(R_j)$ with probability $\theta^j$, $j = 1, 2, \ldots, k$
- $u_i, v_i$ independent, $u_i, d_i$ independent

scale so that these values represent relative risks or odds ratios. The discrete random effects allow sites to have excessively elevated or deflated risks relative to the underlying smooth map, as they tend to be more flexible than those constrained to be normally distributed.

Model M2 also accommodates a large number of zero counts, as is typical in environmental studies of animal sightings or species abundance. In this case, we have $\theta = (\theta^*, \theta^1, \ldots, \theta^k)$, with $\theta^*$ being the probability of the component which yields $y_i = 0$ with probability 1, and $\theta^* + \sum_{j=1}^k \theta^j = 1$. In the analyses, setting $k = 4$ or $5$ provides sufficient flexibility.

### 3.2.1 Priors

The inverse of the variance components $\sigma_u^2$ and $\sigma_v^2$ were assigned conjugate Gamma($a, b$) priors. Here Gamma($a, b$) denotes the gamma probability density function with mean $a/b$ and variance $a/b^2$. This is a general class of hyperpriors for the inverse variance. Setting $a$ and $b$ to 0 is equivalent to an uninformative, uniform hyperprior, $U(-\infty, \infty)$. We used values of $a$ and $b$ equal to 0.5, 0.001, or 0.0005.

For model M1, $\theta$ was assigned a Dirichlet(1, 1, ..., 1) prior distribution. In addition, the prior for $R^1$ was a Normal distribution with mean one and variance four, which was suitably standardized for truncation at zero. For computational stability, we used $R^{j+1} = R^j + \delta^j$.
with $\delta^j, j = 1, 2, \ldots, k - 1$, also having truncated Normal prior distributions, but with mean zero and variance four. Note that Viallefont, Richardson, and Green (2002) provide a useful comparison of various priors in the context of discrete Poisson mixtures. They recommend weakly informative priors as chosen here.

All analyses were run in WinBUGS 1.4 (Spiegelhalter et al., 2003) using two parallel Markov chains with disperse starting values. Due to the large lag in the autocorrelation of the parameter estimates, we thinned the samples after an initial burn-in of at least 30,000 iterations. Convergence was assessed via trace plots and Gelman and Rubin plots (Gelman and Rubin, 1992).

3.3 Application to Codling Moth Data

The codling moth (CM), Cydia pomonella L, is the main insect that pesters apples and pears around the world. The Okanagan-Kootenay Sterile Insect Release program (SIR) investigated autocidal control of these pests (Vernon, Thistlewood, Smith and Kabaluk, 2006). The number of wild moths caught in 506 pheromone traps at sites in the Okanagan Valley, British Columbia, was monitored over the summers of 1999 to 2001. We focus on an analysis of the total moth count in the first 14 weeks of summer, 1999 at these 506 sites. For a detailed discussion of the data see Vernon et al. (2006), Nathoo, Ainsworth, Gill and Dean (2006), Daigle, Duchesne, Reny-Nolin, and Rivest (2006), and Esterby, Thistlewood, Vernon and Smith (2006).

Interestingly, there are both a large proportion of zeros and several large outlying values in the moth data. Counts range from 0 to 58, with an average of 2.1 and a median of one. Model M2, which handles excess zero counts, is particularly useful here as almost half (251/506 or 0.496) of the traps remained empty during the 14 week period.

Neighbourhoods are defined as traps within 400m in the longitude-latitude plane. Eight special cases have neighbours slightly farther away. This definition evolved through inspection of the map, preliminary analysis, discussion with scientists involved, and a study of the distribution of the number of neighbours (the intention was to keep neighbourhoods
relatively small). Our definition results in between one and 30 neighbours for each trap (median = 12). An alternative definition based on the distance between traps is also useful in these moth count analyses.

Table 3.2 presents the estimates and credible intervals from fits of the BYM model and M2. Under model M2, five percent of the traps are estimated to have approximately a three-fold relative risk beyond that accounted for by the spatial variation in this model. We expected the BYM model to be lacking in this situation because of the severity of the extreme counts. Surprisingly, the BYM model seems to ‘stretch itself’ to these extremes reasonably well. In fact, posterior distributions of the log relative risks from the BYM model fit are not normally distributed and thus display the flexibility of this model in accommodating large values.

In stark contrast to the very high proportion of zero counts in the data (50%), model M2 estimates the probability of belonging to the zero component at 4%. In fact, there are large spatially contiguous clusters of traps with very low means which are modelled by the spatially correlated component of the model. The probability of belonging to the zero
component is estimated to be zero for all active traps (those which were not empty over the period) but ranges from 0.05 to 0.39 (mean = 0.09) for the empty traps.

Figure 3.1 displays the locations of the active and empty traps as well as traps with the largest and smallest posterior probabilities of membership in the zero component. The empty traps with the smallest posterior probability of membership in the zero component tend to be located within the spatially correlated clusters of empty traps, whereas those with the largest probabilities tend to have very active neighbours. This is also illustrated in Figure 3.2a, which displays the relationship between the mean of counts in neighbouring sites and the posterior probability of membership in the zero component. The locally outlying empty traps are useful targets for investigations aimed at isolating local rather than global features, such as landscape and climate patterns, which tend to be unattractive to these moths.

Figure 3.2b displays the posterior probabilities of component membership. The ten most extreme local ‘lowspots’ are tracked in red. Although their posterior probabilities of membership in component one are larger than that for the zero component, the zero component values are more widely separated from the values of the remaining traps.
CHAPTER 3. DETECTION OF OUTLIERS IN MAPPING STUDIES

Figure 3.1: Map of moth trap locations identifying active traps (counts \( \leq 10 \), blue circles; counts \( > 10 \), solid blue circles) empty traps with the largest (red) and smallest (yellow) posterior probabilities of membership in the zero component, and remaining empty traps (green)
Figure 3.2: Mean of moth counts for neighbouring traps and posterior probabilities of component membership with the 10 largest posterior probabilities of membership in the zero component indicated in red.
3.4 Application to Weevil Infestation Data

A white pine weevil infestation study was conducted over a seven-year period (1996 to 2002), in a 21,960m$^2$ spruce tree plantation in British Columbia. Susceptible trees were annually inspected for the presence of weevil attack ($N = 2662$). We consider the number of attacks per tree over the seven-year period in a mixed binomial analysis using model M2.

There are many possible neighbourhood definitions. He and Alfaro (1997) and Nathoo (2005) discuss neighbourhood definitions in the context of weevil infestations. He and Alfaro (1997) found spatial correlations attenuated at different distances depending on the stage of a weevil infestation cycle. We define a neighbourhood as all trees within six meters. Assessment of the sensitivity of our outlier detection techniques to this definition indicates robustness (results not presented here).

Parameter estimates ($\hat{\theta} = (0.02, 0.24, 0.33, 0.23, 0.18)$ and $\hat{R} = (0.45, 0.87, 1.28, 1.98)$) suggest a small proportion of lowspots. Although 27% of the trees never experienced an attack in the seven years, an estimated 2% belong to the zero component. Figure 3.3 displays the spatial location of the trees. Trees with the largest posterior probability of belonging to the zero component are indicated in red (top 2% of the data). They tend to be located in the northern half of the plantation, distant from other trees which were never infested (indicated in green), and close to those which were infested many times (circle size represents the proportion of years infested). These trees are considered by the scientists to be resistant and of importance for further study.
CHAPTER 3. DETECTION OF OUTLIERS IN MAPPING STUDIES

Figure 3.3: Map of spruce tree locations identifying uninfested trees (in green) highly resistant trees (in red) and infested trees (circles size represent the proportion of years infested)
3.5 Application to Infant Mortality Data

The number of infant deaths and the infant population sizes for each of the 79 local health areas in the province of British Columbia (BC), Canada are examined routinely for spatial disturbances. Here we consider data for the period 1985-1994. Infants are defined as children less than one year of age and local health areas (LHAs) are administrative health units. This data set is discussed in detail by MacNab and Dean (2000).

Many of the LHAs in BC have small population sizes (ranging from 123 to 52,860; 25th percentile = 983; median = 2668). Thus, they have much smaller denominators for constructing relative risks than mortality from most other causes. Compounding this, infant mortality, though important to monitor, is a relatively rare event (LHA estimates range from six to nine deaths per thousand infants).

The LHA standardized mortality ratio (SMR = \( \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} n_i} \)) for infant death ranges from 0 to 2.06 with an interquartile range of (0.82, 1.25). Neighbours are defined here as LHAs sharing a boundary and the number of neighbours of a local health area ranges from 2 to 12 (median = 4).

The BYM model fit accounts for a large proportion of variation through the spatial random effect (\( \hat{\sigma}_u^2 = 0.074 \) and \( \hat{\sigma}_v^2 = 0.014 \)). Table 3.3 presents estimates and credible intervals from the fit of the BYM model and M1. The 95% credible intervals for the estimated variance of the spatial random effects differ somewhat between the two models; the upper limit of the interval from the BYM model fit is quite a bit larger than that from the fit of model M1. Credible intervals for the risks associated with discrete mixture components one and five, \( \hat{R}_1 \) and \( \hat{R}_5 \) respectively, do not include unity, indicating the presence of some low and high outlying risks.

Figures 3.4a and 3.4b compare estimates from the fits of the BYM model and M1 (circle radius in the scatterplots equals the square root of population size). The relative risk estimates (\( \hat{r} \)) for the LHAs with the largest population sizes are robust to model choice. On the other hand, model M1 allows for more spatial disparity from the mean of the neighbouring risk estimates in LHAs with small to moderate population sizes. Large differences between
Table 3.3: Analysis of infant mortality: posterior mean estimates and 95% credible intervals for model parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>parameter</th>
<th>estimate</th>
<th>95% credible interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYM</td>
<td>( \sigma_u^2 )</td>
<td>0.074</td>
<td>(0.002, 0.179)</td>
</tr>
<tr>
<td>BYM</td>
<td>( \sigma_y^2 )</td>
<td>0.014</td>
<td>(0.0003, 0.052)</td>
</tr>
<tr>
<td>M1</td>
<td>( \sigma_u^2 )</td>
<td>0.03</td>
<td>(0.002, 0.11)</td>
</tr>
<tr>
<td>M1</td>
<td>( \theta^1 )</td>
<td>0.27</td>
<td>(0.002, 0.78)</td>
</tr>
<tr>
<td>M1</td>
<td>( \theta^2 )</td>
<td>0.31</td>
<td>(0.010, 0.77)</td>
</tr>
<tr>
<td>M1</td>
<td>( \theta^3 )</td>
<td>0.21</td>
<td>(0.006, 0.68)</td>
</tr>
<tr>
<td>M1</td>
<td>( \theta^4 )</td>
<td>0.13</td>
<td>(0.002, 0.44)</td>
</tr>
<tr>
<td>M1</td>
<td>( \theta^5 )</td>
<td>0.08</td>
<td>(0.001, 0.32)</td>
</tr>
<tr>
<td>M1</td>
<td>( R^1 )</td>
<td>0.76</td>
<td>(0.11, 0.97)</td>
</tr>
<tr>
<td>M1</td>
<td>( R^2 )</td>
<td>0.96</td>
<td>(0.75, 1.34)</td>
</tr>
<tr>
<td>M1</td>
<td>( R^3 )</td>
<td>1.17</td>
<td>(0.88, 1.67)</td>
</tr>
<tr>
<td>M1</td>
<td>( R^4 )</td>
<td>1.56</td>
<td>(0.97, 4.02)</td>
</tr>
<tr>
<td>M1</td>
<td>( R^5 )</td>
<td>2.36</td>
<td>(1.26, 6.34)</td>
</tr>
</tbody>
</table>

the estimated risks from the two models tend to occur for LHAs whose SMR estimates differ most from the mean of the SMR estimates of neighbouring LHAs. Finally, the LHAs whose estimated SMR differs greatly from the mean of the estimates for its neighbours tend to have the small population sizes. Smoothing models such as BYM borrow strength from neighbours for precisely these LHAs. Thus, model M1 provides more flexibility for these LHAs to emerge as outliers.

### 3.5.1 Hotspot identification

Posterior probabilities of component membership are displayed in Figure 3.4c. Figure 3.4d displays the posterior probability of membership in the high risk component (five) versus the posterior probability of non-membership in low risk component (one) for each LHA. LHAs close to the top and left edges have the clearest assignment to the extreme components. Our discussion of outliers in the context of the infant data focuses on the LHAs with the largest posterior probability of belonging to component five. These LHAs are flagged in red in Figure 3.4. They are: Alberni, Vancouver, Ladysmith, Prince Rupert, Vancouver Island
Figure 3.4: Analysis of infant mortality: \( \hat{r} \) is the relative risk estimate (subscripts indicate the model), a superscript ‘nbr’ indicates the mean of the estimates of neighbouring LHAs, circle radius equals the square root of the population size, and flagged LHAs are indicated in red.
North, Smithers, Hope, New Westminster, Central Coast and Princeton.

Table 3.4 displays SMR estimates, the mean of SMR estimates for neighbouring LHAs, population sizes, relative risk estimates under the BYM model and M1, and the number of neighbours for each of the flagged LHAs. Model M1 provides greater flexibility for these outliers, especially for LHAs with small to moderate population sizes. As seen in Figure 3.4b, there is less smoothing toward the estimates of neighbouring LHAs under model M1. The outliers are accommodated by the discrete random effect, which is more extreme than the independent normal random effect in the BYM model (not shown). Draws from the posterior predictive distributions are used to assess the fit for the flagged LHAs. Model M1 provides a slightly better fit in each case.

Table 3.4: Estimates for selected LHAs from the analysis of infant mortality: \( \hat{\tau} \) is the relative risk estimate (subscripts indicate the model), a superscript ‘nbr’ indicates the mean of estimates for neighbouring sites

<table>
<thead>
<tr>
<th>LHA</th>
<th>( SMR )</th>
<th>( SMR^{nbr} )</th>
<th>Population</th>
<th>( \hat{\tau}_{M1} )</th>
<th>( \hat{\tau}_{BYM} )</th>
<th># neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberni</td>
<td>1.78</td>
<td>1.00</td>
<td>4508</td>
<td>1.57</td>
<td>1.45</td>
<td>6</td>
</tr>
<tr>
<td>Vancouver</td>
<td>1.33</td>
<td>0.77</td>
<td>52856</td>
<td>1.31</td>
<td>1.30</td>
<td>3</td>
</tr>
<tr>
<td>Ladysmith</td>
<td>1.98</td>
<td>1.08</td>
<td>1634</td>
<td>1.55</td>
<td>1.42</td>
<td>4</td>
</tr>
<tr>
<td>Prince Rupert</td>
<td>1.76</td>
<td>1.04</td>
<td>3523</td>
<td>1.56</td>
<td>1.49</td>
<td>4</td>
</tr>
<tr>
<td>North Vanc. Island</td>
<td>1.74</td>
<td>1.43</td>
<td>2668</td>
<td>1.50</td>
<td>1.43</td>
<td>4</td>
</tr>
<tr>
<td>Smithers</td>
<td>1.71</td>
<td>1.19</td>
<td>2800</td>
<td>1.50</td>
<td>1.44</td>
<td>4</td>
</tr>
<tr>
<td>Hope</td>
<td>1.81</td>
<td>1.21</td>
<td>1011</td>
<td>1.27</td>
<td>1.13</td>
<td>6</td>
</tr>
<tr>
<td>New Westminster</td>
<td>1.24</td>
<td>0.77</td>
<td>5326</td>
<td>1.17</td>
<td>1.04</td>
<td>5</td>
</tr>
<tr>
<td>Central Coast</td>
<td>2.05</td>
<td>1.14</td>
<td>757</td>
<td>1.41</td>
<td>1.37</td>
<td>4</td>
</tr>
<tr>
<td>Princeton</td>
<td>2.06</td>
<td>0.97</td>
<td>547</td>
<td>1.23</td>
<td>1.10</td>
<td>5</td>
</tr>
</tbody>
</table>

3.6 Case Study

This study evaluates the discussed methods for identifying outliers in the case of artificially generated hotspots. Count data, \( y_i \), for site \( i \) was generated from a Poisson distribution with conditional mean \( n_i m_e^{u_i + d_i} \) with \( u_i \) a spatially correlated random variable, as in Section
3.2. For all but one site, \( d_i = 0 \), so that the underlying risk surface is a spatially correlated smooth map and the emphasis is detecting a local outlier.

The spatial arrangement of sites and the population sizes \( n_i \) were the same as for the infant mortality data in Section 3.5. To provide spatial risk estimates which accurately reflect this data, we analyzed the infant mortality data under model M1, but with no discrete component (\( d_i = 0 \) for all \( i \)). We used the estimated log relative risks, \( \hat{u}_i \), as the true log risks for the case study (\( e^{u_i} \) ranged from 0.71 to 1.45). In separate case analyses, specific risks corresponding to sites with small, moderate or large population sizes were inflated to produce a local outlier. Specifically, these sites correspond to Arrow Lakes (population = 619, \( SMR = 0.91 \)), Saanich (population = 4,981, \( SMR = 0.91 \)), and Burnaby (population = 19,034, \( SMR = 0.72 \)). For each of these sites the conditional Poisson mean was assigned \( n_i e^{u_i + d_i} \) with \( d_i = \log(R) \) taking each of the values \( R = 2, R = 3, \) and \( R = 5 \) for a total of nine case studies. Poisson counts were then generated and the data analyzed using the BYM model, model M1, and the LC model (Lawson and Clark, 2002). Model fit was assessed using measures based on posterior predictive distributions.

As expected, the BYM model tends to smooth the relative risk estimate of the inflated site more than model M1. As seen in Table 3.5, model M1 is particularly helpful for the site with a small population size. Figure 3.5 compares estimates of the posterior probabilities of belonging to the outlying components from the fit of model M1. The outlier in Arrow Lakes (small population size) is detectable with \( R = 3 \) and obvious when \( R = 5 \). An outlier in a site with moderate or large population size is clearly located for a small risk inflation, \( R = 2 \).

We contrast our posterior probabilities of component membership with the estimated weights from the fit of the LC model, \( \hat{p}_{LC} \). These represent the weight assigned to the spatially smooth component as opposed to the spatial jump component from the LC model. Comparisons for two of the case studies are presented in Figures 3.6 and 3.7. Figure 3.6 compares posterior probabilities of membership in component five under model M1, \( \hat{p}_{LC} \), and the mean of the \( \hat{p}_{LC} \) for neighbouring sites. The inflated site has the largest value of
Table 3.5: Estimates for the case studies: \( r_{true} \) is the true relative risk, \( \hat{r} \) is the relative risk estimate (subscripts indicate the model), a superscript ‘\( nbr \)’ indicates the mean of estimates for neighbouring sites

<table>
<thead>
<tr>
<th>Population</th>
<th>R</th>
<th>( r_{true} )</th>
<th>( \hat{r}_{BY,M} )</th>
<th>( \hat{r}_{M1} )</th>
<th>( \hat{r}<em>{BY,M} - \hat{r}</em>{nbr,BY,M} )</th>
<th>( \hat{r}<em>{M1} - \hat{r}</em>{nbr,M1} )</th>
<th>( SMR - SMR_{nbr} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>5</td>
<td>4.56</td>
<td>1.18</td>
<td>3.53</td>
<td>0.22</td>
<td>2.60</td>
<td>2.89</td>
</tr>
<tr>
<td>Moderate</td>
<td>2</td>
<td>1.95</td>
<td>1.64</td>
<td>1.86</td>
<td>0.63</td>
<td>0.89</td>
<td>1.12</td>
</tr>
<tr>
<td>Moderate</td>
<td>3</td>
<td>2.92</td>
<td>2.40</td>
<td>2.76</td>
<td>1.16</td>
<td>1.59</td>
<td>1.65</td>
</tr>
<tr>
<td>Large</td>
<td>2</td>
<td>1.53</td>
<td>1.46</td>
<td>1.55</td>
<td>0.52</td>
<td>0.63</td>
<td>0.68</td>
</tr>
</tbody>
</table>

\( \hat{p}_{LC} \) in both cases. Interestingly, the mean of the estimates of neighbours stands out from the remaining sites as much, or more than \( \hat{p}_{LC} \). It may be that the LC model pinpoints the neighbouring values as disjoint. Figure 3.7 compares posterior probabilities of membership in component five with the difference between various estimates and the mean of the estimates in neighbouring sites (\( SMR, \hat{p}_{LC} \) and relative risk estimates under the LC model and M1). The difference between a site’s estimated relative risk and the mean of the relative risk estimates for its neighbours is a useful quantitative measure for identifying the outliers under both the LC model and M1.
Figure 3.5: Case studies: posterior probabilities of membership in the extreme components (inflated site in red); each sub-plot displays the posterior probability of non-membership in component 1 versus the posterior probability of membership in component 5
Figure 3.6: Case studies for a site with moderate population size (Saanich): comparison of the LC model component weight ($\hat{p}_{LC}$), the mean of the estimated weights of neighbours ($\bar{p}_{LC}^{nbr}$) and model M1 posterior probabilities of membership in component 5, inflated site in red.
Figure 3.7: Case studies for a site with moderate population size (Saanich): identification of outliers using estimates from the LC model and M1; \( \hat{r} \) is the relative risk estimate (subscript indicates model), \( \hat{p}_{LC} \) is the estimated spatial component weight under the LC model, a superscript \( 'nbr' \) indicates the mean of the estimates of neighbours, and inflated values are indicated in red.
The overall fit of the BYM model and M1, as measured by the chi-square discrepancy measure, $T(Y|b) = \sum_i (Y_i - E(Y_i|b))^2 / Var(Y_i|b)$, does not indicate obvious inadequacies of either model. However, an assessment of Pearson residuals and the posterior predictive distributions from the BYM model and M1 fits indicates that model M1 provides a much better fit to the inflated values. Table 3.6 provides the estimated mean from the posterior predictive distribution for selected case studies. Though the BYM model tends to be quite robust, M1 seems to provide a better fit even for outliers in a site with a large population. Finally, the ‘leave-one-out’ posterior predictive probabilities clearly identify the inflated sites as outliers for the fits from both the BYM model and model M1. In all cases, at least 99 percent of the time, the replicate obtained from the ‘leave-one-out’ posterior predictive distribution was smaller than the observed value.

Table 3.6: Case studies: estimated mean of posterior predictive distributions

<table>
<thead>
<tr>
<th>Case Study</th>
<th>BYM model mean</th>
<th>model M1 mean</th>
<th>Case Study Observed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Population $R = 5$</td>
<td>5.2</td>
<td>15.4</td>
<td>17</td>
</tr>
<tr>
<td>Moderate Population $R = 2$</td>
<td>58.2</td>
<td>67.0</td>
<td>72</td>
</tr>
<tr>
<td>Moderate Population $R = 3$</td>
<td>84.7</td>
<td>98.0</td>
<td>100</td>
</tr>
<tr>
<td>Large Population $R = 2$</td>
<td>197.9</td>
<td>210.0</td>
<td>217</td>
</tr>
</tbody>
</table>

3.7 Sensitivity to Prior Specification

Practical implementation of Bayesian methods requires particular care, especially when the model involves many weakly identified parameters. A sensitivity analysis was undertaken to investigate the impact of choice of prior distributions. Here we are primarily interested in the posterior probabilities of membership in components one and five. The robustness of these estimates, as well as relative risk estimates, to choice of prior is illustrated using the moth and infant mortality datasets.

We chose Gamma or truncated Normal priors for the discrete random effect. Normal and
Gamma prior distributions considered here are: truncated Normal(1, 4) for $R_1$ and truncated Normal(0, 4) for $\delta_j$, $j = 1, 2, \ldots, k - 1$; Gamma(1, 2) for $R_j$ and $\delta_j$; and Gamma(2, 1) for $R_j$ and $\delta_j$. Under the Gamma(1, 2) distribution, $R_j$ shifts slowly from component one to five. The Normal prior and the Gamma(2, 1) are similar, with these priors shifting quickly across components. The main difference between the latter two priors is the distribution of $R_1$; the Gamma(2, 1) distribution concentrates approximately twice as much mass below one.

The Moth data has many small counts and a few very large values. Thus, we consider the Gamma(2, 1) prior as an alternative to the truncated Normal priors used in the analysis presented in Section 3.3. The two priors lead to similar interpretations. The correlations between estimates of posterior probabilities of membership in components one and five are very large ($r = 0.998$ and $r = 0.994$ respectively) and the mean squared difference between relative risk estimates is small, 0.0003.

The analysis of the infant mortality data also reveals important correspondence between estimates obtained using Gamma(1, 2) priors and truncated Normal priors. There is a consistent difference between the estimated posterior probabilities of membership in components one and five. However, they are highly correlated ($r = 0.997$ and $r = 0.998$ respectively) and equality is not essential here, as we are primarily interested in identifying those sites with the largest values. The ranks of the posterior probabilities of membership in components one and five are nearly identical under the two sets of prior distributions. Finally, the mean squared difference between relative risk estimates is small, 0.00005.

### 3.8 Discussion

The proposed additive models combine a spatial random effect for modelling an underlying smooth map and a discrete random effect in order to accommodate and identify local and global outliers. These hotspots/lowspots may first appear in isolation but subsequently create changes in the spatial structure. Thus, it is important to consider such potential values for early review.
Posterior probabilities of component membership are a key quantitative measure for identifying extremes. A graphical comparison of the posterior probabilities of membership in components one and five are useful for flagging outlying locations. These plots are attractive as they are similar to receiver operator curves (ROC) which are commonly used to assess sensitivity and specificity of diagnostic tests. As well, posterior probability of component membership and estimated values of the discrete component provide simple, interpretable measures which can be used to set cut points for flagging locations of interest in a surveillance context.

Posterior predictive probabilities are useful for investigating fit in particular locations while ‘leave-one-out’ posterior predictive distributions are useful for assessing the significance of the outliers. Unfortunately, in practice, it is not practical to assess every site. Stern and Cressie (2000) propose the use of importance weighting and importance sampling in order to approximate the cross-validation posterior predictive distributions and reduce the computation time involved in assessing many sites. The methods discussed here provide a complementary flagging system which identifies a relatively small number of locations for assessment.

The proposed methods are particularly useful when some sites have small denominators (population sizes or length of follow-up). Moderate to large outliers in these regions seem to be detected under model M1 whereas they are smoothed out by the BYM model.

The methods proposed here are easily implemented in freely available software. They are flexible and useful for exploratory analysis and surveillance purposes. Although not evaluated here, the methods also perform well for identification of several outlying values. Natural extensions include joint isolation of extremes in spatial and temporal dimensions.
Chapter 4

Zero-inflated Spatial Models

4.1 Introduction and Overview

A common concern when modelling count data with a parametric distribution, is overdispersion, where the observed and theoretical variance differ. One particular cause of overdispersion is zero-inflation, where a large number of zero counts are observed. This may arise, for example, in ecological studies of species abundance where unsuitable habitat may be included within the study area. In our application, which monitors the presence of weevil attacks on spruce trees in a plantation in British Columbia, a large number of trees were observed to be infestation-free over the seven-year study period. In addition, these zero counts arise in two forms: from unsuitable habitat, which leads to zero counts which are spatially correlated; and secondly, from resistant properties of specific trees, which may lead to isolated zeros among heavily infested regions. This motivates a consideration of zero-inflated models, as well as models which incorporate a variety of structures on the zero counts in order to accommodate these features.

The remainder of the introduction to this chapter provides an overview of the literature on zero-inflation, with a brief review of overdispersion and a particular focus on methods for zero-inflated correlated data. This also includes a discussion of the variety of application contexts in which the various models arose. Those interested in zero-inflation models in
a broader context, for example, multivariate zero-inflated models, or computational packages available, are referred to Appendix C. In section 4.2, we develop zero-inflated spatially correlated models that are formulated in the exponential family framework to encompass a variety of distributions for the data. Section 4.3 illustrates these approaches through an analysis of white pine weevil infestations in spruce trees. Section 4.4 provides a discussion of the unique features highlighted by the zero-inflated spatial models developed here, recommendations regarding applications, and ideas for future research.

4.1.1 Overdispersion

Overdispersion occurs when the observed variance of a random variable is larger than the variance of the theoretical model. The problem arises as basic models for count data are constrained such that the variance is functionally related to the mean and, hence, completely determined once the mean is defined. Much of the overdispersion literature focuses on handling extra-Poisson and extra-binomial variation. However, the problem arises for any distribution where the mean and variance are functionally related.

Failing to account for overdispersion, when it occurs, by using Poisson or binomial maximum likelihood or moment estimating equations for estimation of the parameters in the mean, leads to consistent estimates; however, standard errors are underestimated. Several approaches provide adjustments for heterogeneity when data are overdispersed. A simple approach utilizes quasi-likelihood, where a scale parameter is included in the model and adjusts the standard errors (McCullagh and Nelder, 1989). This robust technique does not require strong distributional assumptions and is readily available in standard software packages such as SAS®, Splus (2000) and R (2005).

Parametric methods using mixture models have been considered by Wang, Puterman, Cockburn and Le (1996) and Lawless (1987), for example. Lawless (1987) discusses the negative binomial distribution, which is constructed by allowing the Poisson mean to follow a gamma distribution. Similarly, overdispersion with respect to the binomial distribution can be handled through the use of the beta-binomial distribution (Williams, 1982), which
permits the binomial mean to follow a beta distribution. Another alternative to a Poisson model is the generalized Poisson distribution discussed by Consul and Jain (1973). More broadly, generalized linear mixed models, with a variety of random effects to represent clustering at various levels in the model, have proved quite useful for handling overdispersion. Breslow and Clayton (1993) discuss such models in the context of developing simple methods for approximate inference. These methods generally assume a continuous distribution for the random effects and are only suitable for unimodal data.

4.1.2 Zero-inflated Models

The presence of excess zeros is a special case of overdispersion. When overdispersion is induced by an extra mass at zero, transformations of the data or traditional methods for handling overdispersion, as described above, may not be satisfactory since these excess zeros can create a bimodal distribution. In addition, the analysis of the counts of four species of birds by Martin et al. (2005) demonstrates that failure to account for zero-inflation can reduce the ability to detect relationships and lead to incorrect inference.

Zero-inflated models are important in a wide range of applications, such as environmetrics, ecology, health, manufacturing, economics and psychology. In a manufacturing context, Lambert (1992) proposes zero-inflated Poisson (ZIP) regression to study covariate effects on soldering defect data for wiring boards. In the fields of psychology and health, zero-inflated models have been applied to risks associated with marijuana use (Simons, Neal and Gaher, 2006), to cognitive functioning in children (Cheung, 2002, 2006) and to the effect of supplemental formula feeding of nursing infants (Lee, Wang, Scott, Yau and McLachlan, 2006). In ecology, zero-inflated models are a natural way to model species abundance data where, at many sampling sites, the species may not be observed. Examples of such analyses are presented by Welsh, Cunningham, Donnelly and Lindenmayer (1996) and Martin et al. (2005).

In practice, the parameterization of a model has implications for interpretation. Ridout, Demetrio and Hinde (1998), Martin, et al. (2005) and Kuhnert, Martin, Mengersen and
Possingham (2005) make the distinction between different types of zeros in the ecological setting. True zeros may arise due to immunity of individuals or due to ecological effects that create unsuitable species habitat. On the other hand, sampling zeros occur simply by chance. In addition, false zeros arise when an observer fails to detect an occurrence. Specification of a zero-inflated model as a mixture of a distribution for count data and a zero mass allows one to assess the likelihood that a zero arose from the degenerate zero mass. In contrast, the conditional or two-part model (Heilbron, 1994; Mullahy, 1986) handles the zeros separately from the non-zeros and specifies a zero-truncated distribution for the non-zero data. The former model is useful for distinguishing sampling zeros from true zeros; the latter treats all zeros in the same manner.

4.1.3 Zero-inflated Models for Correlated Data

There is currently a great deal of interest in extending zero-inflated models in order to account for correlation structures arising in, for example, longitudinal, clustered or spatial data. Two approaches to handling such correlated data are the use of generalized estimating equations (GEE) (Liang and Zeger, 1986) and the introduction of random effects. Dobbie and Welsh (2001) extend the zero-inflated Poisson model to allow for longitudinal data. These authors use a two-part model and allow responses to be correlated through the construction of GEEs for each component. They apply their methodology to the analysis of noisy friarbird counts, which were recorded as part of the Canberra garden bird survey. If the serial dependence in such repeated measures is ignored, the standard errors are underestimated and, thus produce misleadingly small p-values.

Whereas Dobbie and Welsh (2001) use a GEE for the observed response, Hall and Zhang (2004) incorporate a GEE into an EM algorithm for analysis of a mixture model where the serial correlation pertains to the response prior to zero-inflation. They replace the weighted generalized linear model (GLM) in the maximization step with weighted GEEs. This is a generalization of the EM algorithm referred to as the ES algorithm by Rosen, Jiang and Tanner (2000), who studied a more general form of mixtures of marginal GLMs for correlated
Hall (2000) develops a zero-inflated binomial (ZIB) model and includes random effects to account for both within-subject correlation and between-subject heterogeneity. The EM algorithm with Gaussian quadrature is used to handle integrations. Hall (2000) calculates the conditional expectation of the latent indicator variable for the zero component, \( Z \), given the counts and covariate effects. This leads to the use of an unweighted logistic regression as opposed to the weighted logistic regression required by Lambert (1992). Zero-inflated binomial models have also been considered in the analysis of cognitive growth in children (Cheung, 2006) and biological assay control (Vieira, Hinde and Demetrio, 2000).

Yau and Lee (2001) introduce independent random effects into both components of a two-part zero-inflated Poisson regression model for longitudinal data. Similarly, Wang, Yau and Lee (2002) incorporate independent normal random effects in both components of a ZIP model in order to account for clustering in length of stay data collected at multiple hospitals. Lee et al. (2006) extend this to a multi-level ZIP regression model with random effects. Levels include observations, individuals and clusters. These authors use an EM algorithm in conjunction with the penalized likelihood and REML estimating equations for variance components.

Kuhnert et al. (2005) perform an analysis of the impact of commercial cattle grazing on bird counts, incorporating random effects for handling variability between species, grazing regimes and species within a specific regime. Both Bayesian ZIP and zero-inflated negative binomial (ZINB) models are used and expert opinion is incorporated in the prior distributions.

4.1.4 Zero-inflated Models for Spatially Correlated Data

In many applications, zero-inflated count data are spatially oriented. In an ecological context, this may lead to: spatial correlation in the counts, representing abundance at each site; in the probability of a zero count, representing non-suitable habitat; or in both. Rathbun and Fei (2006) use a zero-inflated Bayesian Poisson model with excess zeros generated by a
spatial probit model. On the other hand, Agarwal, Gelfand and Citron-Pousty (2002) discuss Bayesian methods for zero-inflated Poisson regression models where spatial correlation is included in the Poisson distribution. Agarwal et al. (2002) provide a discussion of the issues of posterior propriety, informative prior specification, well behaved simulation-based model fitting and handling data with a large proportion of zeros. Posterior propriety under improper priors on the regression coefficients in the Poisson case led these authors to devise techniques for proper prior specification of regression parameters.

Wikle and Anderson (2003) and Velarde, Migon and Pereira (2004) discuss zero-inflated spatio-temporal models. Wikle and Anderson (2003) use a spatio-temporal zero-inflated Poisson model within a hierarchical Bayesian framework. They use a spectral representation of the spatial process on the Poisson mean. The spatial process related to zero-inflation is considered a ‘nuisance’, but is modelled via two indicator variables for their so-called ‘data rich’ and ‘data poor’ grid boxes. In this way, they account for boundary cases where no data is observed. Velarde et al. (2004) consider spatio-temporal models applied to point referenced rainfall data. These authors use a zero-inflated continuous exponential family distribution. Random effects with conditional autoregressive (CAR) priors are considered for modelling both the probability of rain and the amount of rain. Further, the chance of rain at time $t$ is allowed to depend on the occurrence of rain at one or more previous times.

### 4.2 Zero-inflated Spatial Models

Much of the literature on zero-inflation focuses on the Poisson distribution. In environmental studies, counts may not be rare enough to justify the Poisson approximation to the binomial distribution or the counts may have an upper bound, as is the case in our application. This leads to our focus on binomial models. However, we formulate the zero-inflation models in terms of a generalized linear mixed model in order to facilitate generalization to other distributions, such as the Poisson, negative binomial and beta binomial distributions.

Our zero-inflation models consist of two components: the true zero component, termed the ‘resistant’ component, and the ‘non-resistant’ component. This links with the context
of the example considered later. Spatial correlation in the two components is modelled via random effects or an autocovariate. For all models, spatially correlated random effects in the non-resistant component are modelled by a conditional autoregressive distribution. The random effects associated with the resistant component are modelled using either a conditional autoregressive distribution, a discrete random effect or an autocovariate. In addition, we explore the use of a multivariate conditional autoregressive distribution for modelling the correlation between the two sets of spatially correlated random effects. As noted in section 4.1, these models focus on different aspects of the resistant component. This is discussed in more detail in section 4.3 within the context of the application.

Formulation of the zero-inflated models is facilitated by the introduction of a latent variable, $z_i$. Let $z_i = 1$ if the observation arises from the zero mass, a distribution which generates zero with probability 1, and let $z_i = 0$ if the observation arises from an alternate distribution $f()$. Conditional on random effects, $b_i = (b_{\theta i}, b_{\mu i})'$, which may represent spatially correlated or clustering effects, the observations $y_i, i = 1, 2, \ldots N$ are distributed

\[ y_i | b_i \sim \begin{cases} f(\mu_i^*) \text{ with probability } 1 - \theta_i^* \\ 0 \text{ with probability } \theta_i^* \end{cases} \]

where $f()$ denotes a distribution from the exponential family, such as the binomial or Poisson distribution, $\mu_i^*$ is its conditional mean given $b_i$ and $z_i = 0$, and $P(z_i = 1 | b_{\theta i}) = \theta_i^*$. Conditional on $b_i$ and $z_i = 0$, the observations, $y_i$, are independent with expectation $E(y_i | b_i, z_i = 0) = \mu_i^*$ and variance $Var(y_i | b_i, z_i = 0) = \nu_i(\mu_i^*)$ where $\nu_i$ is a specified variance function. The link function $g(\mu_i^*) = \eta_i^*$ relates the conditional mean to the linear predictor $\eta_i^* = \alpha + x_i' \beta + b_{\mu i}$ where $x_i$ is a vector of covariates associated with fixed effects, $\beta$ is its coefficient, and $b_{\mu i}$ is the spatial random effect. We also model the probability of resistance as a function of explanatory terms, spatial or discrete random effects, or autocovariates. We use the logit link function, $\text{logit}(\theta_i^*) = \alpha_0 + x_{0i}' \beta_0 + b_{\theta i}$ where $x_{0i}$ is a vector of covariates associated with the fixed effects, $\beta_0$ is its coefficient, and $b_{\theta i}$ is the random effect or the autocovariate in this binary component. Let $b_{\mu}$ and $b_{\theta}$ be the full suite of random effects for the two components, where $b_{\mu i}$ and $b_{\theta i}$ are the $i^{th}$ elements of $b_{\mu}$ and $b_{\theta}$
respectively.

Table 4.1 provides a general formulation for each model considered here, without covariates. These models have a variety of structures for the additive random effects and encompass distributions from the exponential family. For instance, for a binomial logistic model, \( f() \) is the probability density function for the binomial distribution, \( g(\mu^*_i) = \log(\mu^*_i/(1-\mu^*_i)) \) is the link function, and the spatial random effects correspond to the logarithm of the odds ratio at each site. The two basic models would be a standard binomial model (1A) and a beta-binomial model (1B) which accommodate overdispersion while the remaining models incorporate zero-inflation. Thus, in general, the random effects accommodate a variety of local spatially structured or clustering effects as described below.

| Non-inflated Models: \( y_i \sim f(\mu_i) \) |
| Model 1A, GLM | \( g(\mu_i) = \alpha \) |
| Model 1B, overdispersed | \( \mu_i \sim Beta(\alpha^*, \beta^*) \), in the application |

| Zero-inflated Models: |
| Conditioned Distribution |
| \( y_i | b_i \) \{ \begin{align*} & \sim f(\mu^*_i) \quad \text{with probability} \quad 1 - \theta^*_i \\ & \sim 0 \quad \text{with probability} \quad \theta^*_i \end{align*} \} |
| Linear Predictors |
| \( g(\mu^*_i) = \alpha + b_{\mu i} \) |
| \( \text{logit}(\theta^*_i) = \omega_0 + b_{\theta i} \) |
| Random Effects |
| \( b_{\mu} \) and \( b_{\theta} \) |
| Model 2A, \( ZI \) | \( b_{\mu i} = 0 \) with prob 1 \( b_{\theta i} = 0 \) with prob 1 |
| Model 2B, \( ZI_{MCAR} \) | \( b_{\mu i} = u_i \) \( b_{\theta i} = u_{0i} \) |
| Model 2C, \( ZI_{MCAR} \) | \( b_{\mu i} = u_{im} \) \( b_{\theta i} = u_{0im} \) |
| Model 3, \( ZI_{ds}, \alpha_0 = 0 \) with prob 1 | \( b_{\mu i} = u_i \) \( b_{\theta i} = d_i, \alpha_0 = 0 \) |
| Model 4, \( ZI_{autocov} \) | \( b_{\mu i} = u_i \) \( b_{\theta i} = \beta \sum_{j=1}^{\delta_i} I(y_j = 0) / \delta_i \) |

Let \( g() \) be a link function and \( f() \) be the pdf of a distribution from the exponential family \( u_i, u_{0i} \sim \text{CAR}, u, u_0 \) independent \( u_{im}, u_{0im} \sim \text{Multivariate CAR} \) \( d_i = \log(R^2) \) with probability \( \gamma^j, j = 1, 2, \ldots, k \) \( \delta_i \) is the number of neighbours
The intrinsic conditionally autoregressive (CAR) model (Besag, York and Mollié, 1991) is commonly used for handling spatial correlation. Let \( \mathbf{u} \) represent a vector of conditionally autoregressive random effects which may be used to describe spatial correlation in either \( \mathbf{b}_\mu \) or \( \mathbf{b}_\theta \). The distribution of \( u_i \) can be described and interpreted conditionally, given \( u_{-i} \), the set of spatially structured random site effects excluding the \( i^{th} \) site, \( u_i | u_{-i} \sim N(\bar{u}_{\delta_i}, \sigma_u^2 / \delta_i) \), where \( \bar{u}_{\delta_i} \) is the mean of the spatial random effects corresponding to sites in the ‘neighbourhood’ of the \( i^{th} \) site, and \( \delta_i \) is the number of sites in this neighbourhood. Neighbourhoods may be defined in a variety of ways based on distance from a site. In model 2B (see Table 4.1), we use two independent sets of conditionally autoregressive random effects, labelled \( \mathbf{u} \) and \( \mathbf{u}_0 \), whose \( i^{th} \) elements are \( u_i \) and \( u_{0i} \), with \( b_{\mu i} = u_i \) and \( b_{\theta i} = u_{0i} \), and corresponding variance parameters, \( \sigma_u^2 \) and \( \sigma_{u_0}^2 \). In addition, conditionally autoregressive random effects are used for \( b_{\mu i} \) in models 2B, 3, and 4.

A multivariate CAR distribution (Jin, Carlin, and Banerjee, 2005) may be used to model the correlation between two sets of random effects, \( \mathbf{u}_m \) and \( \mathbf{u}_{0m} \). This may be helpful in permitting correlation between the spatial processes used for the two components of the model, \( \mathbf{b}_\mu \) and \( \mathbf{b}_\theta \). Let \( \mathbf{U}_i = (u_{m i}, u_{0m i}) \) be a 2-dimensional vector of spatially correlated Gaussian random effects at site \( i \). Here, the CAR prior gives \( \mathbf{U}_i | \mathbf{u}_{m(-i)}, \mathbf{u}_{0m(-i)} \sim N_2(\bar{U}_{\delta_i}, \Sigma / \delta_i) \), where \( N_2 \) denotes the bivariate normal distribution, \( \bar{U}_{\delta_i} \) is the mean of the spatial random effects corresponding to sites in the neighbourhood of the \( i^{th} \) site, \( \delta_i \) is the number of sites in this neighbourhood, and the 2x2 positive definite, symmetric matrix, \( \Sigma \), represents the conditional, within region covariance of the random effects. The diagonal elements of \( \Sigma \), \( \sigma_u^2 \) and \( \sigma_{u_0}^2 \), represent the conditional variance parameters corresponding to \( u_m \) and \( u_{0m} \) respectively, while the unrestricted off-diagonal element represents the conditional correlation. Model 2C uses such a multivariate CAR distribution to link the random effects in the two components.

Alternatively, discrete random effects may be used to accommodate the heterogeniety in the probability of resistance (see model 3 in Table 4.1). These are particularly useful as they allow sites to have excessively elevated or deflated posterior probabilities of resistance and tend to be more flexible than those constrained to be normally distributed. Let the discrete
random effects, \( d_i \), take one of \( k \) values, \( \log(R^1) \), \( \log(R^2) \), \ldots, \( \log(R^k) \), with probability \( \gamma^1, \gamma^2, \ldots, \gamma^k \), respectively, and \( \sum_{j=1}^{k} \gamma^j = 1 \), \( \gamma = (\gamma^1, \gamma^2, \ldots, \gamma^k) \).

We also examine the use of autocovariates to model the binary spatial process associated with resistance. Such an approach is inspired by the autologistic model which has been proposed for handling spatially correlated binary data (Besag, 1974) and applied in some settings (Hoeting, Lancaster, and Bowden 2000; Wintle and Bardos, 2006). Model 4 (Table 4.1) uses the proportion of neighbouring sites with \( y_i = 0 \) as an autocovariate in modelling the probability of resistance. In a Bayesian analysis, we use the pseudo-likelihood for the binary component of the autologistic model as suggested by Wintle and Bardos (2006).

### 4.3 Application to Pine Weevil Infestation Data

A white pine weevil infestation survey conducted over a seven-year period (1996 to 2002), in a 21,960 m\(^2\) spruce tree plantation in British Columbia, annually inspected the presence of weevil attack in \( N = 2662 \) susceptible trees. Here we analyze the number of attacks per tree over the seven-year period. For this data a large number of trees was observed to have no infestations (715/2662 = 0.269) over the study period. These zeros could simply be due to random fluctuation. On the other hand, they may be indicative of resistance of the tree itself or of unsuitable habitat. Zero-inflated binomial models developed in the previous section are applied here. The zeros are modelled as arising either from a binomial distribution (random fluctuation), or as arising from a zero mass (resistant trees). In addition, we illustrate how the various models isolate different features of resistance and discuss this more fully in the context of this example.

Seven models (Table 4.1) are fit to this data. We use two basic models: 1A, a standard binomial model; 1B, a beta-binomial model, to accommodate overdispersion. The remaining models incorporate zero-inflation. Model 2A is a simple ZIB model. The final four models are ZIB spatial models. As discussed in the previous section, the first two spatial models use a CAR prior on the random effects associated with the counts as well as the random effects associated with the probability of resistance. Model 2B uses independent CAR priors on
each of the two sets of random effects, while model 2C explicitly models correlation between
the two sets of random effects via a multivariate CAR prior. The final two spatial models
utilize a CAR prior for the random effects associated with the counts; however, these models
accommodate heterogeneity in the probability of resistance by: a discrete random effect,
model 3, or an autocovariate, model 4.

A variety of neighbourhood definitions can be used to model spatial effects. He and
Alfaro (1997) discuss neighbourhood definitions in the context of weevil infestations. These
authors found that spatial correlations attenuated at different distances depending on the
stage of a weevil infection cycle. Exploratory analysis, including variograms for the total
number of infestations, the number of infestations given at least one, and whether or not
the tree was ever infested, as well as biological expertise, led us to define a neighbourhood
as all trees within a radius of 6m. For two trees, the nearest neighbour was located further
than 6m away. That is, the closest neighbour was located at a distance of 6.1 and 11.3
meters respectively. These nearest neighbours were defined as the single neighbour for each
of these trees. The number of neighbours ranges from 1 to 37 with a median of 15.

Specification of the models is completed by assigning prior distributions. Vague priors
may be useful in a variety of settings unless specific informative priors are available. The
intercept terms, \( \alpha \) and \( \alpha_0 \) were given flat priors, Uniform(\(-\infty, \infty\)). For the beta-binomial
model, the parameters of the beta distribution were given Gamma (1,0.001) priors. Here,
Gamma \((a, b)\) denotes the gamma probability density function with mean \(a/b\) and variance
\(a/b^2\). For the spatial models, the precision parameter \(1/\sigma_u^2\) was assigned a conjugate Gamma
\((0.001, 0.001)\) while \(1/\sigma_{u0}^2\) was assigned a prior with a smaller variance, Gamma \((0.05, 0.05)\).
The precision matrix for the multivariate CAR distribution was given a Wishart prior with
matrix elements \((0.001, 0, 0, 0.05)\).

For model 3, \( \gamma \) was assigned a Dirichlet\((1,1,...1)\) prior distribution. In addition, the
prior for \( R^1 \) was assigned to be \( N(1,10) \), suitably standardized for truncation at zero. For
computational stability, we used \( R^{j+1} = R^j + \delta^j \) with \( \delta^j, j = 1, 2, \ldots, k - 1 \), having zero
truncated \( N(0,10) \) prior distributions. Here, \( k = 3 \) provided sufficient flexibility, and \( R^1, R^2, \ldots \).
and \( R^3 \) reflect low, medium and high probabilities of resistance.

All analyses were run in WinBUGS 1.4 (Spiegelhalter et al., 2003) using two parallel Markov chains with dispersed starting values. Due to the large lag in the autocorrelation of the parameter estimates in these sorts of studies, we recommend thinning the samples to no less than every 10 samples after an initial burn-in. Here a burn-in of at least 10,000 iterations was used. Convergence was assessed via trace plots and Gelman and Rubin plots (Gelman and Rubin, 1992).

Table 4.2 presents transformed estimates of the intercept terms associated with the linear predictors. These yield mean estimates of the proportion of years a tree is infested or of the probability of resistance. The standard binomial model and the four models which account for spatial correlation provide similar estimates of the proportion of years infested. Standard errors are smaller under model 1A, as expected. The two models which account for spatial correlation in the zero component through a CAR prior, models 2B and 2C, provide similar estimates and standard errors of the proportion of zeros arising from the resistant component. Under the ZIB model, 2A, both the estimated proportion of infestations and the estimated proportion of resistant trees are larger. As we will see later, neither model 1A nor model 2A have sufficient flexibility to fit this data well. Under model 3, the probability of resistance takes the three values \( (R^j/(1 + R^j), j = 1, 2, 3) \), 0.025, 0.45 and 0.72, with probabilities \( (\gamma^j, j = 1, 2, 3) \), 0.69, 0.24 and 0.07, respectively. This model allows a few trees (7%) to have a very large probability (0.72) of resistance. Note that under model 4, the interpretation of \( \alpha_0 \) is quite different from that of the other models. Of greater interest for this model is the estimate of the coefficient of the autocovariate which is 3.58 with 95% credible interval, (2.22, 4.98) exhibiting a high degree of spatial association in the zeros.

The last two columns of Table 4.2 provide estimates of the variability in the CAR random effects. The values of \( \hat{\sigma}^2_u \) from models 2B and 2C are quite similar. In model 3, isolated resistant trees are accommodated by the discrete random effects, \( b_\theta \), so that the estimated variability in the CAR random effects associated with the non-resistant component is much smaller than models 2B and 2C.
Table 4.2: Posterior mean estimates of proportion of years infested, proportion of resistant trees, and variability of random effects. CI is the credibility interval

<table>
<thead>
<tr>
<th>Model</th>
<th>$\frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}}$ (95% CI)</th>
<th>$\frac{e^{\hat{\alpha}<em>{0}}}{1+e^{\hat{\alpha}</em>{0}}}$ (95% CI)</th>
<th>$\hat{\sigma}^2_u$ (95% CI)</th>
<th>$\hat{\sigma}^2_{u_0}$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0.22 (0.209,0.221)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2A</td>
<td>0.26 (0.250,0.265)</td>
<td>0.16 (0.141,0.182)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2B</td>
<td>0.22 (0.203,0.231)</td>
<td>0.05 (0.016,0.091)</td>
<td>2.01 (1.29, 2.96)</td>
<td>3.46 (0.13, 13.21)</td>
</tr>
<tr>
<td>2C</td>
<td>0.22 (0.203,0.233)</td>
<td>0.06 (0.019,0.112)</td>
<td>1.95 (1.03, 2.93)</td>
<td>0.43 (0.05, 0.93)</td>
</tr>
<tr>
<td>3</td>
<td>0.22 (0.205,0.227)</td>
<td>-</td>
<td>1.43 (1.20, 1.69)</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0.23 (0.214,0.239)</td>
<td>0.45 (0.017,0.059)</td>
<td>1.47 (0.86, 2.20)</td>
<td>-</td>
</tr>
</tbody>
</table>

The estimates of $\sigma^2_{u_0}$ are very different in models 2B and 2C. The difference between these two models lies in the estimates of the random effects associated with the resistant component. The estimates of $b_{\mu i}$ for the spatial models, 2B and 2C are highly correlated with a Pearson correlation of $r = 0.99$, whereas those for $b_{\theta i}$ are only moderately correlated ($r=0.60$). Note that the estimates of the two sets of random effects, $b_{\mu i}$ and $b_{\theta i}$, have only a moderate, negative correlation ($r=-0.51$) under model 2B. On the other hand, model 2C, the multivariate CAR model, imposes strong correlation structure on these two sets of random effects; estimated random effects are almost perfectly, negatively correlated ($r=-0.995$).

Recall that model 3 uses discrete classes of probabilities for the zero component; for this model we define the posterior probability of resistance as $\gamma^3_i$ which is the probability corresponding to the largest odds of belonging to the resistant component. Figure 4.1 shows the relationship between the estimated posterior probabilities of resistance under the spatial models and highlights the difference in the estimates from model 3. Further insight into the unique characteristics of model 3 is obtained from Figure 4.2 which shows the relationship between the posterior probability of resistance for infestation-free trees and two measures on neighbouring trees: mean proportion of years of infestations observed, and the proportion of trees which were never infested. Weak associations are observed under model 2B, whereas the posterior probability of resistance has a strong relationship to both measures under models 2C, 3 and 4. Large posterior probabilities of resistance are associated with a
Figure 4.1: Correspondence of the posterior probability of resistance among the spatial models.
Figure 4.2: The relationship between the posterior estimate of resistance for trees with no infestations and measurements on neighbouring trees: mean proportion of infestations observed for neighbouring trees and the proportion of neighbouring trees which were never infested.
small infestation rate in neighbouring trees and having many non-infested neighbours under models 2C and 4. Similar, but weaker, relationships between these measures are observed for model 2B. However, under model 3, these relationships are reversed; model 3 assigns large posterior probabilities of resistance to those trees which are surrounded by many highly infested trees. For this model, spatial correlation in resistance is captured through the random effects in the non-resistant component, $b_\mu$, by permitting spatial clusters of very low means while the resistant component identifies isolated resistant trees. Note that under model 4, the proportion of neighbours with no infestations is functionally related to $\hat{\theta}$ for those trees with no infestations.

Figures 4.3a and 4.3b highlight the 100 trees with the largest estimated probabilities of resistance in yellow; the remaining non-infested trees are indicated in blue; and infested trees are indicated in black with larger circles indicating a larger proportion of years infested. Model 2B locates spatial clusters of resistant trees in the south-west corner of the plantation. On the other hand, model 3 identifies resistant trees surrounded by trees with many infestations, especially in the north. This is more clearly seen by comparison with Figure 4.3c (and Figure 4.3d) which highlights in red (and yellow), the 100 trees with the largest (smallest) posterior estimates of the random effect associated with the non-resistant component – indicating sites with the largest (smallest) estimated probability of infestation. The trees with the largest estimated random effects tend to be located in the northern corner of the plantation. Note that Figures 4.3c and 4.3d plot estimates from model 3; those from model 2B are almost perfectly correlated with estimates from model 3.

Goodness-of-fit is assessed in a variety of ways. Standardized residuals (not shown here) for each model, defined as the posterior mean of $r_i = (y_i - E(y))/\sqrt{V(y)}$, indicate that models 1A and 2A tend to underestimate the largest infestation rates. Histograms of the posterior predictive p-values, $P(y_i < y^{rep})$ (Gelman, Meng and Stern, 1996), for each model (not shown here) reveal that model 1B has very few extreme p-values. However, the p-values associated with models 2B, 2C, 3, and 4 show a more uniform distribution.

Table 4.3 displays measures of goodness-of-fit for the various models. The first column
Figure 4.3: Largest and smallest posterior estimates of the probability of resistance (a and b) and the random effect associated with the non-resistant component (c and d). Infested trees are indicated in black with larger circles indicating a larger proportion of years infested. In a and b, trees with the largest probability of resistance are indicated in yellow; remaining non-infested trees are indicated in blue. The 100 trees with the largest (red) and smallest (yellow) estimates of the random effect associated with the non-resistant component are indicated in c and d.
displays the number of trees observed to have 0, 1, 2, . . . , 7, infestations, \( O_j \), \( j = 0, 1, \ldots, 7 \). We generated data using the current estimates at each iteration of the MCMC sampler and we then calculated the median number of trees \( E_j \) with \( j \) infestations, \( j = 0, 1, \ldots, 7 \) over the runs of the sampler. This is displayed for each model in columns 3-9. Based on this generated data we calculated \( GOF = (O_j - E_j)^2/E_j \). Table 4.3 also reports the p-value corresponding to a \( \chi^2 \) discrepancy measure, \( \chi^2_{\text{disc}} \), proposed by Gelman, Meng, and Stern (1996). This measure compares a statistic, \( T \), calculated using the observed data, \( T_{\text{obs}} \), to the same statistic calculated using data generated from the posterior predictive distribution, \( T_{\text{rep}} \). Here we use \( T = (y_i - E(y_i|b))^2/\text{Var}(y_i|b) \). The proportion of times \( T_{\text{obs}} \) is less or equal to \( T_{\text{rep}} \) is then recorded across iterations of the MCMC sampler. Finally, Johnson (2004) develops a goodness-of-fit statistic, \( R^B \), which has a \( \chi^2 \) distribution if the model is accurate. At each iteration of the MCMC sampler, \( R^B \) is calculated and compared to the the 95\(^{th} \) percentile of the \( \chi^2 \) distribution. The posterior mean of \( R^B \), as well as proportion of times \( R^B \) exceeds this critical value, are reported in the last two rows of Table 4.3.

Table 4.3: Classification of observed data and posterior median number of values generated for each category under each model

<table>
<thead>
<tr>
<th>Infestations</th>
<th>Observed</th>
<th>1A</th>
<th>1B</th>
<th>2A</th>
<th>2B</th>
<th>2C</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>715</td>
<td>496</td>
<td>695</td>
<td>715</td>
<td>703</td>
<td>697</td>
<td>700</td>
<td>713</td>
</tr>
<tr>
<td>1</td>
<td>767</td>
<td>941</td>
<td>810</td>
<td>681</td>
<td>779</td>
<td>785</td>
<td>785</td>
<td>752</td>
</tr>
<tr>
<td>2</td>
<td>616</td>
<td>765</td>
<td>600</td>
<td>698</td>
<td>628</td>
<td>629</td>
<td>626</td>
<td>636</td>
</tr>
<tr>
<td>3</td>
<td>347</td>
<td>348</td>
<td>340</td>
<td>398</td>
<td>347</td>
<td>346</td>
<td>345</td>
<td>358</td>
</tr>
<tr>
<td>4</td>
<td>157</td>
<td>95</td>
<td>151</td>
<td>137</td>
<td>145</td>
<td>144</td>
<td>145</td>
<td>147</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>15</td>
<td>52</td>
<td>28</td>
<td>46</td>
<td>46</td>
<td>47</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
GOF = \frac{(O_j - E_j)^2}{E_j}
\]

\[
p-value for \chi^2_{\text{disc}}
\]

\[
R^B
\]

\[
R^B \text{ exceeding } \chi^2_{6.0.95}
\]

All goodness-of-fit measures indicate that the simplest models, 1A and 2A, provide poor
fits to this data. Although model 2A regenerates about the right number of zeros, it does not provide a good fit to the non-zero counts. Accounting for overdispersion beyond that arising from zero-inflation appears to be necessary here. Surprisingly, model 1B, the beta-binomial model, which accounts for overdispersion in general, but does not account for zero-inflation specifically, provides a reasonably good fit to this data by these measures. This is likely due to the fact that this data is unimodal. If the mean of the non-zero counts were slightly larger or the proportion of zeros were slightly greater, the resulting bimodal distribution would not be accommodated by the beta-binomial. All of the spatial models are able to accommodate both the excess zeros and the overdispersion in the counts and provide a reasonable fit to the data, particularly models 2B and 3.

4.4 Discussion

This chapter considers a rich and flexible class of zero-inflated mixture models which incorporate random effects for modelling spatial correlation and clusters of resistance. These complementary models provide a variety of structures for modelling the probability of the zero component, resistance, and thus allow an exploration of individual characteristics, as well as habitat characteristics, that may provide protection from infestation. The simple ZIB model, although it accommodates the zero counts well, does not provide enough flexibility for modelling the overdispersed counts. Although the beta-binomial model does not provide insight regarding resistant trees, it surprisingly provides a great deal of flexibility in modelling zero-inflation relative to the binomial model; in situations where the number of zeros is not excessive, such as presented here, it may provide a reasonable overall fit to the data.

Extensions of the zero-inflated models presented here are being considered for the temporal context, where splines are used for modelling trends in both the probability of resistance and the mean count. For binomial spatio-temporal analyses where the upper bound on counts is not too large, a dynamic state-space structure with certain sites being in an absorbing state may provide a useful modelling approach.
Chapter 5

Future Work

Future work will consider extensions and modifications of the models presented in Chapters 3 and 4 as well as alternate estimation techniques. One topic of interest is the use of spline smoothers in an extension of the discrete mixture models (Chapter 3) and zero-inflated models (Chapter 4) developed here, to accommodate temporal trends. Second, alternate distance definitions for hierarchical spatial models will be developed in the context of oceanographic and fisheries applications. Such distance measures will be based on water flow and velocity. Third, the use of generalized estimating equations (GEE) will be considered for estimation of the spatial models considered in Chapters 4.

5.1 Spatio-temporal Models

Temporal extensions to hierarchical spatial models have been considered by such authors as Waller, Carlin, Xia, and Gelfand (1997), MacNab and Dean (2001) and Silva, Dean, Niyonsenga and Vanasse (2006). Waller, Carlin, Xia, and Gelfand (1997) consider extensions to the spatial models of Besag et al. (1991) in order to accommodate general temporal effects and space-time interactions. These authors ‘nest’ heterogeneity and spatial effects within time. Thus, heterogeneity and spatial clustering may vary over time. MacNab and Dean (2001) propose a generalized additive mixed model (GAMM), an additive extension of
CHAPTER 5. FUTURE WORK

GLMMs, for the analysis of infant mortality rates. These authors include spatial random effects as well as fixed and random temporal components in their spatio-temporal model. Their development of B-spline smoothing over the temporal dimension provides a flexible means of accommodating overall time effects as well as region specific time effects. Fixed effects spline smoothing is used for the overall temporal trends and random effect splines are used to isolate small-area trends. Advantages of this method are that estimation is computationally straightforward, simply requiring an extension of the design matrices used for a GLMM analysis, and it allows the temporal trends to have arbitrary shapes. Further, it provides a simple mechanism for identifying spatial patterns as well as identifying those regions with temporal effects which differ from the overall mean time trend. Silva, Dean, Niyonsenga, and Vanasse (2006) consider a spatio-temporal odds model for the revascularization data considered in Chapter 3. These authors include heterogeneity random effects, spatial random effects, an overall time trend and a region specific time trend. As developed by MacNab and Dean (2001), Silva et al. (2006) model time trends using a cubic B-spline.

Splines represent complicated curves by combining relatively simple polynomial segments, where each polynomial is of order \( D \). Constraints are imposed in order to ensure the smoothness of the composite curve at the points at which the segments join, the inner knots. Thus, the curve has continuous derivatives of all degrees \( \leq D - 1 \). Cubic B-splines are used by MacNab and Dean (2001) and Silva et al. (2006) to model both overall time trends, and region specific time trends. Such smoothing splines are useful for modelling nonlinear temporal effects.

Future research will utilize cubic B-splines to extend the models used to identify spatial outliers in Chapter 3. In particular, the use of an overall time trend as well as a region specific time trend will extend the models presented to handle the temporal dimension. Such extensions to the discrete mixture models of Chapter 3 will allow identification of both spatial and temporal outliers. That is, the discrete component of the model is expected to identify values which stand out from the fluid changes to the smooth map. This is expected to be particularly useful in health and environmental surveillance contexts where outliers
in space or time may indicate potentially important changes. Similarly, the zero-inflated models of Chapter 4 will be extended to zero-inflated spatio-temporal models. Cubic B-splines will be used to model time effects in both the resistant and non-resistant components of the model. Further developments of the above models might consider adaptations to allow for space-time interactions.

5.2 Distance Measures

In many instances, spatial random effects can act as a surrogate for unknown or unmeasurable environmental variables. In the context of fisheries and oceans data, water characteristics, such as temperature, salinity, mixing depth, and upwelling, may be important environmental influences on measurements such as salmon abundance. For instance, the salinity and temperature of the water has an impact on plankton productivity which leads to a major food sources for juvenile salmon. Recently, spatial hierarchical models have been applied to fisheries and oceans data by Su, Peterman, and Haeseker, (2004). These authors consider salmon abundance, using the great circle distance to model the spatial similarity between sampling points. Their results indicate a strong relationship between SST and salmon survival rates. Such characteristics of ocean water are strongly related to weather systems and may provide the most important environmental measures for outcomes such as salmon abundance.

Applications presented here consider the spatial relationships between health regions or trees. Environmental variables associated with such points tend to be stable. For example in the health context, regions are rural or urban and the distance to a major health facility will remain relatively stable over time. In the forestry context, trees remain on the same slope with the same aspect. On the other hand, the environmental variables associated with a specific sampling point in the ocean are dynamic. In particular, pink, chum, sockeye and coho salmon usually occupy the upper 10 m of the water (Jaenicke and Celewycz, 1994). In this context, the environment of interest is the water, which is constantly moving. Thus, it may be that distance measures based on hydrologic relationships, such as flow direction
and velocity, are more informative than distances based on the physical distance between two points. For instance, the Alaska Coastal Current is a freshwater-driven current which begins along the B.C. coast and flows north then west within 20 km of shore into the Bering Sea. The strength of the current is affected by local precipitation, wind, air temperature and other meteorological conditions. It has been hypothesized that this current represents a critical early feeding habitat for juvenile salmon (Jaenicke and Celewycz, 1994). Thus, future research on hierarchical spatial models applied to fisheries data could consider the use of distance measures which are based on dominant features of water circulation, such as the Alaska Coastal Current.

Temporal extensions to spatial models which use hydrologic distance measures may provide unique challenges in that the distance between any two physical locations will change as the currents changes with time. This may be an interesting area of development. In addition, zero-inflated models may be of particular interest in the fisheries context as catch data over short time periods may contain a large number of zero counts (Jaenicke and Celewycz, 1994).

### 5.3 Generalized Estimating Equations (GEE)

A general class of regression models, marginal models (Liang and Zeger, 1986), allow responses to be correlated, but do not require the joint probability distribution to be fully specified. The focus is on the relationship between the explanatory variables and the outcome variable while correlations, such as spatial or temporal correlations, are included simply to obtain better estimates of the expected response. In order to adapt GEE methods to the analysis of spatial data, an appropriate correlation structure must be chosen. Semivariogram estimates provide a means of investigating appropriate forms for the variance structure. A popular choice is the parametric exponential model where correlations decreases exponentially as a function of the distance between two locations.

Albert and McShane (1995) use a GEE approach with spatially correlated binary responses obtained from neuroimaging data. Each subject’s brain image is partitioned into
grids and the binary outcomes at each grid location are correlated within subjects. Gumpertz, Wu and Pye (2000) investigate the underlying environmental factors which influence the proportion of years a county experiences southern pine beetle outbreaks over a 31 year period. These authors use a GEE approach for parameter estimation of their marginal logistic regression model. This model allows for spatial autocorrelation among counties and adjusts variance terms to account for temporal autocorrelation in presence/absence data at each time point. Finally, Dobbie and Welsh (2001) use a GEE approach to model longitudinal data which exhibit zero-inflation. These authors use a conditional formulation for the zero-inflated model. Future work will investigate extending the above models to accommodate spatio-temporal zero-inflated data.
Bibliography


Public Health Data’. Statistics in Medicine, 19, 2333-2344.


study of growth and development’. Statistics in Medicine, 21, 1461-1469.

inflated proportion models’. Statistics in Medicine, 25(17), 3011-3022.


with regression effects on the mean, dispersion and zero-inflation level applied to patent
outsourcing rates’. Statistical Modelling, in press.

de la topographie et des caractéristiques des vergers sur l’efficacité du programme
d’épandage d’insectes stériles pour le carpocapse de la pomme (Laspeyresia pomonelle)’.


## Appendix A

## Infant Data

<table>
<thead>
<tr>
<th>Region</th>
<th>Population Size</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fernie</td>
<td>2379</td>
<td>10</td>
</tr>
<tr>
<td>Cranbrook</td>
<td>2940</td>
<td>19</td>
</tr>
<tr>
<td>Kimberly</td>
<td>862</td>
<td>6</td>
</tr>
<tr>
<td>Windermere</td>
<td>937</td>
<td>4</td>
</tr>
<tr>
<td>Creston</td>
<td>1442</td>
<td>6</td>
</tr>
<tr>
<td>Golden</td>
<td>1200</td>
<td>5</td>
</tr>
<tr>
<td>Kootenay Lake</td>
<td>436</td>
<td>3</td>
</tr>
<tr>
<td>Nelson</td>
<td>2587</td>
<td>19</td>
</tr>
<tr>
<td>Castlegar</td>
<td>1515</td>
<td>11</td>
</tr>
<tr>
<td>Arrow Lakes</td>
<td>619</td>
<td>4</td>
</tr>
<tr>
<td>Trail</td>
<td>2565</td>
<td>14</td>
</tr>
<tr>
<td>Grand Forks</td>
<td>793</td>
<td>5</td>
</tr>
<tr>
<td>Kettle Valley</td>
<td>408</td>
<td>2</td>
</tr>
<tr>
<td>Revelstoke</td>
<td>1283</td>
<td>8</td>
</tr>
<tr>
<td>Salmon Arm</td>
<td>3118</td>
<td>13</td>
</tr>
<tr>
<td>Armstrong Spallumcheen</td>
<td>925</td>
<td>8</td>
</tr>
<tr>
<td>Vernon</td>
<td>6002</td>
<td>45</td>
</tr>
<tr>
<td>Enderby</td>
<td>713</td>
<td>5</td>
</tr>
<tr>
<td>Southern Okanagan</td>
<td>1533</td>
<td>15</td>
</tr>
<tr>
<td>Penticton</td>
<td>3787</td>
<td>20</td>
</tr>
<tr>
<td>Keremeos</td>
<td>391</td>
<td>2</td>
</tr>
<tr>
<td>Princeton</td>
<td>547</td>
<td>8</td>
</tr>
<tr>
<td>Central Okanagan</td>
<td>12791</td>
<td>92</td>
</tr>
<tr>
<td>Summerland</td>
<td>954</td>
<td>2</td>
</tr>
<tr>
<td>Kamloops</td>
<td>10972</td>
<td>72</td>
</tr>
<tr>
<td>North Thompson</td>
<td>732</td>
<td>6</td>
</tr>
</tbody>
</table>
## APPENDIX A. INFANT DATA

<table>
<thead>
<tr>
<th>Region</th>
<th>Population Size</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lillooet</td>
<td>744</td>
<td>7</td>
</tr>
<tr>
<td>South Cariboo</td>
<td>1136</td>
<td>11</td>
</tr>
<tr>
<td>Merritt</td>
<td>1700</td>
<td>12</td>
</tr>
<tr>
<td>Hope</td>
<td>1011</td>
<td>13</td>
</tr>
<tr>
<td>Chilliwack</td>
<td>8162</td>
<td>43</td>
</tr>
<tr>
<td>Abbotsford</td>
<td>13639</td>
<td>68</td>
</tr>
<tr>
<td>Agassiz Harrison</td>
<td>790</td>
<td>7</td>
</tr>
<tr>
<td>Langley</td>
<td>13012</td>
<td>70</td>
</tr>
<tr>
<td>Maple Ridge</td>
<td>9305</td>
<td>63</td>
</tr>
<tr>
<td>Mission</td>
<td>4761</td>
<td>39</td>
</tr>
<tr>
<td>Surrey</td>
<td>41091</td>
<td>247</td>
</tr>
<tr>
<td>Delta</td>
<td>11874</td>
<td>58</td>
</tr>
<tr>
<td>New Westminster</td>
<td>5326</td>
<td>47</td>
</tr>
<tr>
<td>Coquitlam</td>
<td>22340</td>
<td>130</td>
</tr>
<tr>
<td>Sechelt</td>
<td>2648</td>
<td>20</td>
</tr>
<tr>
<td>Powell River</td>
<td>2318</td>
<td>16</td>
</tr>
<tr>
<td>Howe Sound</td>
<td>3254</td>
<td>20</td>
</tr>
<tr>
<td>Cowichan</td>
<td>5750</td>
<td>55</td>
</tr>
<tr>
<td>Lake Cowichan</td>
<td>649</td>
<td>4</td>
</tr>
<tr>
<td>Ladysmith</td>
<td>1634</td>
<td>23</td>
</tr>
<tr>
<td>Nanaimo</td>
<td>9320</td>
<td>75</td>
</tr>
<tr>
<td>Qualicum</td>
<td>2754</td>
<td>14</td>
</tr>
<tr>
<td>Alberni</td>
<td>4508</td>
<td>57</td>
</tr>
<tr>
<td>Courtenay</td>
<td>5636</td>
<td>33</td>
</tr>
<tr>
<td>Campbell River</td>
<td>4946</td>
<td>39</td>
</tr>
<tr>
<td>Vancouver Island West</td>
<td>720</td>
<td>7</td>
</tr>
<tr>
<td>Vancouver Island North</td>
<td>2668</td>
<td>33</td>
</tr>
<tr>
<td>Cariboo Chilcotin</td>
<td>5856</td>
<td>49</td>
</tr>
<tr>
<td>Quesnel</td>
<td>3568</td>
<td>37</td>
</tr>
<tr>
<td>Central Coast</td>
<td>757</td>
<td>11</td>
</tr>
<tr>
<td>Queen Charlotte</td>
<td>1155</td>
<td>6</td>
</tr>
<tr>
<td>Prince Rupert</td>
<td>3523</td>
<td>44</td>
</tr>
<tr>
<td>Smithers</td>
<td>2800</td>
<td>34</td>
</tr>
<tr>
<td>Kitimat</td>
<td>2041</td>
<td>13</td>
</tr>
<tr>
<td>Stikine</td>
<td>335</td>
<td>4</td>
</tr>
<tr>
<td>Terrace</td>
<td>5209</td>
<td>37</td>
</tr>
<tr>
<td>Nishga</td>
<td>365</td>
<td>4</td>
</tr>
<tr>
<td>Telegraph Creek</td>
<td>123</td>
<td>0</td>
</tr>
<tr>
<td>Peace River South</td>
<td>4967</td>
<td>34</td>
</tr>
<tr>
<td>Peace River North</td>
<td>4813</td>
<td>31</td>
</tr>
</tbody>
</table>
## APPENDIX A. INFANT DATA

<table>
<thead>
<tr>
<th>Region</th>
<th>Population Size</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fort Nelson</td>
<td>1221</td>
<td>9</td>
</tr>
<tr>
<td>Burns Lake</td>
<td>1407</td>
<td>14</td>
</tr>
<tr>
<td>Nechako</td>
<td>2976</td>
<td>31</td>
</tr>
<tr>
<td>Prince George</td>
<td>15362</td>
<td>137</td>
</tr>
<tr>
<td>Vancouver</td>
<td>52856</td>
<td>498</td>
</tr>
<tr>
<td>Burnaby</td>
<td>19034</td>
<td>97</td>
</tr>
<tr>
<td>North Vancouver</td>
<td>15070</td>
<td>87</td>
</tr>
<tr>
<td>West Vancouver Bowen Island</td>
<td>3570</td>
<td>22</td>
</tr>
<tr>
<td>Richmond</td>
<td>16706</td>
<td>93</td>
</tr>
<tr>
<td>Greater Victoria</td>
<td>20620</td>
<td>138</td>
</tr>
<tr>
<td>Sooke</td>
<td>6903</td>
<td>47</td>
</tr>
<tr>
<td>Saanich</td>
<td>4981</td>
<td>32</td>
</tr>
<tr>
<td>Gulf Islands</td>
<td>1030</td>
<td>7</td>
</tr>
</tbody>
</table>
Appendix B

Penalized Quasi-likelihood

Recall that $b_i = u_i + v_i$ is a random region effect where $u_i$ represents spatial variation, $u_i|u_{-i} \sim N(u_i, \sigma^2_u/\delta_i)$, and $v_i$ represents spatially unstructured variation, $v_i \sim N(0, \sigma^2_v)$. Let $\sigma^2 = \sigma^2_u + \sigma^2_v$ and $\lambda = \sigma^2_u / (\sigma^2_u + \sigma^2_v)$. We assume that $b$ has a multivariate normal distribution with mean $\theta$ and covariance matrix $D = D(\theta)$ which depends on an unknown vector, $\theta = (\lambda, \sigma^2)$ of variance components. Conditional on $b$, the $y_i$ are assumed independent with

$$E(Y|b) = h(X\alpha + Zb + offset) = \mu^b$$

where $h$ is the inverse of the link function, $g()$, defined in Section 2.2 above. Here we use the logarithmic link; $g(\cdot) = \log(\cdot)$

To estimate $(\alpha, \theta)$, the following integrated quasi-likelihood function can be used

$$e^{ql} \propto |D|^{1/2} \int \exp \left[ - \frac{1}{2} \sum_{i=1}^{N} d_i(y_i, \mu^b_i) - \frac{1}{2} b^t D^{-1} b \right] \, db$$

(B.1)

where $\mu^b_i = Var(y_i|b), d_i(y, \mu) = -2 \int_{y}^\mu \frac{y - u}{u} \, du$, Equation B.1 can be rewritten as:

$$e^{ql}(\alpha, \theta) = c |D|^{-1/2} \int e^{-k(b)} \, db$$

where $c$ is a multiplicative constant and $k(b) = \frac{1}{2} \left[ \sum d_i(y_i, \mu^b_i) + b^t Db \right]$. Note that conditioning on $b$, the observations $Y$ are drawn from a Poisson distribution with variance function $Var(\mu^b_i) = \mu^b_i$. Breslow and Clayton (1993) proposed a Laplace approximation to
APPENDIX B. PENALIZED QUASI-LIKELIHOOD

this integrated quasi-likelihood. A quadratic expansion of \(-k(b)\) about its maximizing value is made before integration. That is,

\[
e^q_l \approx c|D|^{-1/2} \int e^{-[k(b)+\frac{1}{2}(b-\hat{b})^t k''(\hat{b})(b-\hat{b})]}db
\]

where \(\hat{b} = \hat{b}(\alpha, \theta)\) denotes the solution to \(k'(b) = -\sum_{i=1}^N (y_i - \mu_i^b)z_i + D^{-1}b\) which maximizes \(-k(b)\). The matrix of second-order partial derivatives of \(k\) with respect to \(b\) is given by

\[
k''(b) = \sum_{i=1}^N z_i z_i^t \mu_i^b + D^{-1} = Z'WZ + D^{-1}
\]

where \(W = \text{diag}(\mu_i)\), the GLM iterated weights. Ignoring the multiplicative constant \(c\), the log quasi-likelihood can be approximated by

\[
ql(\alpha, \theta) \approx -\frac{1}{2} \log(I + Z'WZD) - k(\hat{b}) \tag{B.2}
\]

where \(W\) denotes \(W\) evaluated at \(b = \hat{b}\). If the GLM iterative weights vary slowly as a function of the mean, as is the case here, we can ignore the first term in equation B.2 and choose \(\alpha\) to maximize \(-k(b)\). Thus, we need \((\hat{\alpha}, \hat{b})\) to jointly maximize Green’s (1987, in Breslow & Clayton, 1993) PQL equation:

\[
-\frac{1}{2} \sum_{i=1}^n d_i(y_i, \mu_i^b) - \frac{1}{2} b'D^{-1}b. \tag{B.3}
\]

For a fixed \(\theta\), differentiation with respect to \(\alpha\) and \(b\) leads to the following score equations for the mean parameters:

\[
U_\alpha(\alpha, b) = X'(Y - \mu) = 0 \tag{B.4}
\]

\[
U_\beta(\alpha, b) = Z'(Y - \mu) - D^{-1}b = 0 \tag{B.5}
\]

Breslow and Clayton (1993) define a working vector, \(Y_{(w)} = \eta_i + (Y_i - \mu_i)g'(\mu_i) - \log(n_i).\) A Fisher scoring algorithm is then used to solve equations B.4 and B.5. This is done by iteratively solving the linear system:
(X'WX X'WZD) (α) = (X'WY(w))

where b = Dν.

Updated estimates of the variance parameters are obtained by a Newton-Raphson or a Fisher scoring step as follows:

(\hat{\lambda}, \hat{\sigma}^2)'_{\text{new}} = (\hat{\lambda}, \hat{\sigma}^2)'_{\text{new}} + I^{-1}U.

U is the score vector (B.8) and I is either the observed or expected information based on the REML likelihood for Y(w):

\[ L = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2}(Y - X\hat{\alpha})'V^{-1}(Y - X\hat{\alpha}) \]

Thus, the REML equations for estimating the variance components (Breslow and Clayton, 1993) are

\[ U = \frac{1}{2}[(Y(w) - X\hat{\alpha})'V^{-1}(\frac{dV}{d\theta_r})V^{-1}(Y(w) - X\hat{\alpha}) - tr(P(\frac{dV}{d\theta_r}))] = 0 \]

where r = 1, 2 and P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}

Equivalently, one may apply the normal theory model to Y(w) = Xα + Zb + ϵ, where α is a fixed effect, ϵ ∼ N(0, W^{-1}), b ∼ N(0, D), and ϵ and b are independent. Hence, Y(w) ∼ N(Xα, V), with V = W^{-1} + ZDZ', and the log-likelihood is:

\[ l = -\frac{1}{2}N\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(Y - X\alpha)'V^{-1}(Y - X\alpha) \]

Then one simply needs to solve a series of weighted least squares regressions. The estimating equation for α is:

\[ \frac{dl}{d\alpha} = X'V^{-1}(Y - X\alpha) = 0. \]

Thus,

\[ \hat{\alpha} = (X'V^{-1}X)^{-1}X'V^{-1}Y(w) \]
APPENDIX B. PENALIZED QUASI-LIKELIHOOD

The random effect, \( b \), is estimated as the posterior mean given the data. This is easily obtained from the joint distribution of \( (Y_w, b) \sim N(\mu, \Sigma) \) by using the well-known properties of the multivariate normal distribution (Mardia, Kent and Bibby, 1979).

Let

\[
\mu = \begin{pmatrix} X\alpha \\ 0 \end{pmatrix}.
\]

and

\[
\Sigma = \begin{pmatrix} V & ZD \\ DZ' & D \end{pmatrix}.
\]

Then the conditional distribution of \( b|Y_w \) is

\[
b|Y_w \sim N(DZ'V^{-1}(Y_w - X\alpha), D - DZ'V^{-1}ZD)
\]

and \( b \) is estimated as

\[
\hat{b} = DZ'V^{-1}(Y_w - X\hat{\alpha}). \tag{B.10}
\]

Equations B.9, B.10, and B.8 are updated iteratively.
Appendix C

Review of Literature on Analysis of Zero-inflated Data

Here we review additional literature related to zero-altered models. This literature emphasizes the more common zero-inflation case, while zero-deflation tends only to be considered in general formulations which are meant to encompass a broad class of models.

The use of mixture models and conditional models are two common approaches to handling zero-inflation within the context of ecological and health data. The well known ZIP model (Lambert 1992) is a mixture of a degenerate zero mass and a Poisson distribution. On the other hand, Welsh, Cunningham, Donnelly and Lindenmayer (1996) formulate a two-component conditional model where the presence/absence data is modelled with a binomial distribution and the abundance at active sites is modelled using a truncated Poisson or truncated negative binomial distribution.

Conditional models, also known as two-part models, hurdle models, and compatible models are discussed by Mullahy (1986) and Heilbron (1994), for example. Conditional models consist of a zero mass, the so-called ‘hurdle’, and a truncated form of a standard discrete distribution such as the binomial, Poisson or negative binomial. Most applications of hurdle models assume independence between the linear predictor of the probability of overcoming the hurdle and the conditional mean of the counts, given that the hurdle is overcome. This orthogonality simplifies computation and provides a straightforward interpretation of the covariate effects. In the ecology setting, the hurdle may be habitat suitability, and the conditional mean may represent mean abundance given suitable habitat. Thus, the covariate...
effects on habitat suitability can be interpreted independently of the effects on abundance within suitable habitat. Under the conditional model, the likelihood is the sum of two independent likelihoods with no terms in common. Therefore, it is fully efficient to fit the two components separately.

Lambert’s (1992) mixture model approach to zero-inflated Poisson regression also allows the design matrices associated with the Poisson mean, \( \mu \), and the probability of arising from the zero mass, \( \theta \), to contain different sets of experimental factors and covariates. Through the use of a latent variable representing membership in the zero component, a complete likelihood is formed and an EM algorithm (Dempster, Laird and Rubin 1977) used to obtain maximum likelihood estimates (MLE). The complete log-likelihood conveniently splits into the sum of two exponential family log likelihoods so that weighted logistic and Poisson regressions can be used to obtain parameter estimates. Interval estimates rely on likelihood asymptotics and are based on normal approximations which require the log-likelihood surface to be approximately quadratic near the MLE.

Thus, we have two specifications of the zero-inflated model: 1) a mixture of a degenerate distribution with mass at zero and a non-degenerate distribution such as the binomial or Poisson distribution, and 2) a conditional specification where the the zero mass and the truncated distribution of the non-zero counts are modelled independently.

Consider \( \theta \), the probability of a true zero, and \( \mu \), the mean parameter for the probability mass function \( f \) associated with the random variable \( Y \). Then the mixture model specification of the zero-altered model is:

\[
Y \sim \theta I_\theta + (1 - \theta)f(Y | \mu)
\]

where \( I_\theta \) is the degenerate distribution taking the value zero with probability one. This formulation encompasses distributions such as the binomial, Poisson, and generalized Poisson with \( E(Y|\theta, \mu) = (1 - \theta)E_f(Y | \mu) \) and \( Var(Y|\theta, \mu) = \theta(1 - \theta)E_f(Y | \mu)^2 + (1 - \theta)Var_f(Y | \mu) \) where \( E_f(Y | \mu) \) and \( Var_f(Y | \mu) \) denote the expectation and variance of a random variable with probability mass function \( f \). Under this formulation, it is possible for \( \theta \) to be negative, in which case we have a model which provides for zero-deflation, though the interpretation of the distribution as arising from a mixture would be lost. Additional flexibility may be added to such models by incorporating random effects into the distribution of \( f \), or \( \theta \), or both.
Under the conditional model we specify \( Z = 1 \) if \( Y = 0 \) and \( Z = 0 \) if \( Y > 0 \). Then,

\[
Z \sim \theta_c^Z (1 - \theta_c)^{1-Z}
\]

\[
Y | Z = 0 \sim f_{\text{trunc}}(Y | \mu)
\]

where \( f_{\text{trunc}} \) is a truncated distribution such as a truncated binomial or truncated Poisson distribution, and \( \theta_c \) is the probability of a zero count.

Ridout, Demetrio and Hinde (1998), Martin et al. (2005) and Kuhnert, Martin, Mengersen and Possingham (2005) make the distinction between different types of zeros in the ecological setting. True zeros may be structural. This would be the case for immune individuals or when an ecological effect creates unsuitable species habitat. On the other hand, sampling zeros occur simply by chance and false zeros may arise due to a failure to detect an occurrence because of observing too small an area or because of limited ability of a species to disperse to all parts of the region. Martin et al. (2005) explain that the type of zero represented by a particular observation depends on the study objective. If the goal is to quantify the instantaneous location of a species, and the species is temporarily absent from the study site, then the recording would not be considered a false zero. On the other hand, if the goal is to determine which sites are inhabited by a species, then an absence would constitute a false zero.

Thus, the probability of a zero count under the conditional specification, \( \theta_c \), corresponds to the probability of either a true or false zero, whereas the probability \( \theta \) under the mixture specification corresponds to the probability of a true zero. That is, structural zeros and random zeros are not distinguished under the conditional specification, whereas under the mixture model, we can examine the different sources of error (Kuhnert et al., 2005). Martin et al. (2005) suggest the use of Bayesian methods in order that information on the contribution of false zeros may be incorporated through an informative prior on the detection probability.


A multivariate version of the zero-inflated Poisson distribution has been developed by Li et al. (1999). These authors assume that most of the data arises from the perfect state. However, outside of the manufacturing context in which they work, this is not always a reasonable assumption. Without simplifying assumptions, such extensions to the multivariate
situation require a large number of parameters.

Another approach to modelling zero-inflated data is the use of a mixture of Poisson distributions leading to the Neyman Type A distribution (Dobbie and Welsh 2001). This distribution is particularly helpful for multi-modal data. These authors consider three parameterizations of the Neyman type A distribution and extend each to incorporate covariates. Unfortunately, choice of parameterization is important and model fitting is complicated by infinite sums. Further, this distribution is not a member of the exponential family so that the associated advantages are not available. Finally, good initial parameter estimates are needed but these are often difficult to obtain.

Birth processes and threshold models are also used to model zero-inflated data. See Ridout et al. (1998) for a review of these models. Other approaches to zero-inflation include the spatial probit model for zero-inflated data (Rathbun and Fei 2006) and the zero-inflated modified power series distribution used by Gupta, Gupta, and Tripathi (1996). References related to the development of zero-inflated models for continuous data can be found in Martin et al. (2005).

C.1 Overdispersion and zero-inflated models

In practice, it has been found that even after modelling excess zeros, some overdispersion related to the counts may remain. In the standard Poisson regression setting, ignoring overdispersion relative to the Poisson model generally leads to consistent parameter estimates. However, with a truncated distribution, as used in hurdle models, ignoring overdispersion leads to inconsistent parameter estimates (Grogger and Carson 1991). In order to avoid such problems, Gurmu (1997) developed a semi-parametric hurdle model.

Ridout, Hinde and Demetrio (2001) also discuss the bias in parameter estimates when non-zero counts are overdispersed in relation to the Poisson distribution. In the context of mixture models, one solution is the use of a negative binomial, generalized Poisson or beta binomial distribution for $f$.

Zero-inflated negative binomial models have found a wide range of applications in the literature. Martin et al. (2005) compare ZIP and ZINB models of bird counts; Welsh et al. (1996) compare the use of a truncated Poisson and a truncated negative binomial distribution for modelling abundance of Leadbeater’s Possum in south-eastern Australia; and Nobtvedt et al. (2002) use the ZINB model in a longitudinal study of gastrointestinal
parasite burdens in Canadian dairy cows; while Simons, Neal and Gaher (2006) use the ZINB model to study problems associated with drug use among college students.

The generalized Poisson distribution can accommodate underdispersion as well as overdispersion in count data through the addition of an extra parameter. For $\lambda_1 > 0, |\lambda_2| < 1$, the generalized Poisson distribution (Consul and Jain, 1973) is defined by

$$p_x(\lambda_1, \lambda_2) = \lambda_1 (\lambda_1 + x\lambda_2)^{x-1} e^{-(\lambda_1 + x\lambda_2)}/x!, \quad x = 0, 1, 2, \ldots$$  \hspace{1cm} (C.1)

such that

$$p_x(\lambda_1, \lambda_2) = 0 \quad \text{for} \quad x \geq m \quad \text{if} \quad \lambda_1 + m\lambda_2 \leq 0$$  \hspace{1cm} (C.2)

zero-inflated versions of the generalized Poisson distribution (ZIGP) have been utilized in the frequentist setting by, for example, Gupta, Gupta and Tripathi (1996) and Famoye and Singh (2006), while Angers and Biswas (2003) approach the ZIGP from a Bayesian perspective. Czado, Erhardt and Min (2006) discuss an extension of zero-inflated generalized Poisson (ZIGP) regression models to allow for regression on the overdispersion and zero-inflation parameters.

Famoye and Singh (2006) note that, in practice, inadequacies of the ZIP model and computational concerns related to the zero-inflated negative binomial distribution are common. As noted by other authors, iterative techniques used to estimate the parameters of the ZINB model often fail to converge. Thus, ZIGP models provide a nice alternative to the ZINB model in these cases.

C.2 Computing

Bohning (1998) discusses using C.A.MAN software as a diagnostic device for ZIP models. Considering the ZIP model as a special case of a wide class of mixture models with an unknown number of components, one can use C.A.MAN to obtain maximum likelihood estimates for a non-parametric mixing distribution. On the other hand, Dietz and Boehning (2000) present an EM algorithm for maximum likelihood estimation of the zero-truncated Poisson model using standard software. This algorithm can also be used to estimate the parameters of a zero-deflated Poisson model and to calculate a likelihood ratio test of zero modification. Finally, the ‘zicounts’ package for R software can be used to fit regression
models for zero-inflated count data. The package can fit a Poisson, negative binomial, zero-inflated Poisson or zero-inflated negative binomial model. Covariates can be used with both the zero and non-zero components.

Spatial models can also be formulated in a Bayesian framework. Freely available WinBUGS software (Speigelhalter, Best and Lunn 2003) with the associated GeoBUGS package, can be used to fit such models. Any Bayesian analysis must be carefully scrutinized. Careful consideration must be paid to parameter convergence. As well, WinBUGS routinely reports the MC errors which indicate technical errors in computation. It is suggested that the MC error divided by the standard deviation of the posterior distribution ought to be less than 5%.

C.3 Score Tests

Finally, there is a large body of literature on score tests. Dean (1992) develops a score test for overdispersion in Poisson and binomial models. Many authors have considered score tests in the ZIP context: van den Broek (1995), Deng and Paul (2005), Xiang, Lee, Yau, McLachlan (2006). Ridout et al. (2001) discuss score tests for comparing ZIP models and ZINB models, while Hall and Berenhaut (2002) discuss score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. Gupta, Gupta, Tripathi (2004) and Famoye and Singh (2006) discuss score tests for the zero-inflated generalized Poisson distribution and Lee, Xiang, Fung (2004) study the influence of outliers on the score test for comparing a ZIP model against a ZINB model.