SPATIAL MODELLING OF PRETERM BIRTH NEAR THE SYDNEY TAR PONDS, NOVA SCOTIA, CANADA

SPATIAL MODELLING OF PRETERM BIRTH NEAR THE SYDNEY TAR PONDS, NOVA SCOTIA, CANADA

By

ISMAILA AFISI S., B.Sc. (Hons)

A Project Report
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2004)
(Statistics)

MCMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Spatial Modelling of Preterm Birth near the Sydney Tar Ponds, Nova Scotia, Canada

AUTHOR: A. S. Ismaila, B.Sc. (University of Lagos)

SUPERVISOR: Dr. A. Canty

NUMBER OF PAGES: x, 79

Abstract

The major objective of the research is to assess the risk of preterm birth associated with maternal proximity to hazardous waste and pollution from the Sydney Tar Pond sites in Nova Scotia, Canada. The design is spatial modelling of risks of preterm birth in population living in the Cape Breton regional municipality in 1996. The subjects are: 1604 observed cases of preterm birth out of total population of 17559 at risk in 1996. The analysis was done using both the frequentist and the Bayesian approaches. In the frequentist approach, the Poisson model for aggregated data was fitted using the quasi-likelihood approach to accommodate over-dispersion. Weighted regression was also used. In order to accommodate both the random effect and the anticipated spatial effects, Bayesian hierarchical modelling was also used to fit the Poisson model. The result of the Bayesian modelling shows that there is no significant spatial association of risk in the area studied. All the models also show that there is no decrease in risk of preterm birth as we move from the Tar Pond site to other region. None of the other covariates in the model show any significant association with increase risk of preterm birth either. There was no obvious clustering of risk in any region or part.

Acknowledgements

I express my profound gratitude to my supervisor, Dr. Angelo Canty for his guidance and direction and for providing the financial support towards the completion of this project. Many thanks to Prof. Pavlos Kanaroglou for giving us the data used in this project and allowing the use of the Center for Spatial Analysis at McMaster University.

My thanks to Prof. Peter MacDonald for introducing me to most of the statistical software used for this project; and most importantly, for his fatherly advice and support. To my lecturers: Prof. Charlie Goldsmith and Dr. Aaron Childs, I say thank you for allowing me the opportunity to learn from you. My appreciation also goes to Patrick Deluca, a member of the Center for Spatial Analysis, for taking time out of his busy schedule to introduce me to the use of Arc-view and Spacestat. I also thank all the other members of the center: Ron Buliung, Laura Labate, Dimitri Potoglou, Jamie Spinney, and Hanna Maoh for being very good hosts.

I am also using this opportunity to thank the Department of Mathematics and Statistics, and McMaster University for providing the scholarship for my studies. Thanks to Ayaz Hyder, Charles Adeyanju and Rabe Makwarela for their helpful advice. Thanks to all my departmental secretaries: Cheryl, Connie, Pasty, Diana and others. To the members of my class: Ahmed, Anisha, Amanda, Cathlin, Chen, Christina, David, Katherine, Keon, Yvonne and others, I say thank you (you guys are the greatest). I must not fail to thank Nofisah Balogun for her support. Finally, my utmost thanks go to my parents Mr. and Mrs. S. Ismaila for their support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Research Background and Context

1.1 Introduction

Reproductive health is defined "as a condition in which the reproductive process is accomplished in a state of complete physical, mental and social well being: it is not merely the absence of disease or disorders of the reproductive process" (Michal *et al.*, 1993: page 1). This definition is rooted in the World Health Organization definition of health (cited in Rootman and Raeburn, 1994: page 58). Public awareness about potential environmental hazards has continued to grow in recent years. This concern has led to an increased demand for public health authorities to investigate potential clustering of diseases around putative sources of hazards. The attention given to this topic by the mass media has necessitated research on the possible effects of hazardous waste on the people living near the waste sites (Dolk *et al.*, 1997; Dolk *et al.*, 1998; Elliot *et al.*, 1996).

An assessment of the effect of human exposure to particular substances may be a very difficult task for two reasons: firstly because multiple chemicals are usually involved so it is very difficult to discern the specific agent responsible for a particular health concern; and secondly extraneous factors, like cultural and socioeconomic, may confound the effect of direct exposure to a waste site. Michal *et al.* (1993), provide a summary of some of the environmental factors that may affect reproductive health. They say:

> Chemical Pollutants are considered to be the greatest threat to reproductive health in developed countries. However, as a global problem, the major factors in

> descending order of importance are infection, malnutrition, chemicals, radiation, and stress. In less developed countries, the added effects of socioeconomic and cultural influences become more evident (Michal *et al.*, 1993: page 2).

Deriving from the works of two theorists of society, Beck (1992) and Giddens (1991), 'risk' is not only perceived in late/high modern societies, it is 'real'. For Beck (1992), modernity is constituted by 'risks', most especially risks emanating from "pollution, nuclear and chemical productive forces" (Beck, 1992: page 22). Risk in late modernity, from the perspective of Beck (1992), does not only manifest at the 'material'/'physical' level, but also at the psychic realm. According to him:

> Risks such as those produced in the late modernity differ essentially from wealth. By risks I mean above all radioactivity, which completely evades human perceptive abilities, but also toxins and pollutants in the air, the water and foodstuffs, together with the accompanying short-and long term effects on plants, animals and people. They induce systematic and often irreversible harm, generally remain invisible, are based on causal interpretations, and thus initially only exist in terms of the (scientific or anti-scientific) knowledge about them. They can thus be changed, magnified, dramatized or minimized within knowledge, and to that extent they are particularly open to social definition and construction. Hence the mass media and the scientific and legal professions in charge of defining risks become key social and political positions (Beck, 1992: page 22–23).

What is central to the 'risk society thesis' of Giddens (1991) and Beck (1992) is the source of 'anxiety' around environmental issues in modern industrial societies. Thus, exposure to chemical and byproducts of industrialization in Sydney, Nova Scotia may have constituted a danger to reproductive activities in the area. As Beck (1992) indicates, the mass media as a major source of information in modern society play an important role in making people become aware of risks and danger. Not only this, the media can also construct a problem around medical issues to incite panics and anxieties in the public (Seal, 2002).

1.2 The Research Problem

The history of the Tar Pond site in Sydney, Nova Scotia, and the health consequences are well documented (Tara, 2002; Nova Scotia Department of Health and the Cape Breton District Health Authority, 2001). The Tar Pond is a tidal estuary of 33 hectares in the Cape Breton regional municipality of Nova Scotia. This site, considered to be the most toxic site in Canada, is a result of 100 years of steel manufacturing and other allied industries in the area. The byproducts from these industries include BTEX (benzene, toluene, ethylbenzene, and xylene), PAH (polycyclic aromatic hydrocarbons), PCB (polychlorinated biphenyl) and particulate laden with toxic metals, such as arsenic, lead, and other heavy metals. This led to the contamination of soil and other sources of natural water in the surrounding areas. Studies have shown that exposure to these kinds of contaminants (in particular PCB) may have constitute a danger to reproductive outcomes in the area (Baibergenova *et al.*, 2003; Rylander *et al.*, 2000).

This study examines how proximity to the Tar Pond site affects one of the reproductive outcomes: preterm birth. This current project was undertaken at the Department of Mathematics and Statistics of McMaster University as one part of a large multi-phased research project to investigate the association between preterm birth and other adverse reproductive outcomes in Sydney and proximity to Tar Pond site. Other groups involved are: The Center for Spatial Analysis at McMaster University, McMaster Institute of Environment and Health, and St. Joseph's Health Care Center in Hamilton.

1.3 Research Questions

1. Is maternal proximity to hazardous waste and pollution from the Sydney Tar Pond sites associated with increased risk of preterm birth?

2. How much of the variation in preterm birth can be explained by socioeconomic inequalities across the study region?

1.4 Research Objectives

The research objectives are:

1. To explore the spatial distribution of preterm birth among women of childbearing ages in Cape Breton Regional Municipality of Nova Scotia. The hypothesis of interest is

    - H_o: maternal proximity to the Tar Pond sites does not influence the risk of preterm birth in Cape Breton municipality of Nova Scotia, Canada. This will be tested against

    - H_1: maternal proximity to the Tar Pond sites does influence the risk of preterm birth in Cape Breton municipality of Nova Scotia, Canada.

2. To investigate the presence of clusters of health outcomes that may be of significance in testing the above hypothesis; and

3. To compare different methods for the analysis of aggregated spatial data.

1.5 Data Description

Cape Breton Regional Municipality is made up of 158 enumeration districts but aggregated counts of preterm birth are only available for 144 enumeration districts in the Municipality based on the 1996 census data. This data are not available for various reasons and throughout our analysis we shall be working with information from 144 enumeration districts. There are 1604 observed cases of preterm birth out of a total population of 17559 at risk of preterm birth. Other covariates include: population in 1996; the proportion of persons who have no high school diploma; the rate of unemployment to population; average income; the proportion of persons who are separated, divorced or widowed; the proportion of single-parent families and the proportion of people living alone. All these variables were extracted from the 1996 census data. The data is summarized in Table 1.1:

Table 1.1: *Table of variables*

variables	meaning
d	The distance from the Tar Pond
x_1	The rate of unemployment to population
x_2	The proportion of persons who are separated, divorced or widowed
x_3	The proportion of persons who have no high school diploma
x_4	The proportion of people living alone
x_5	The proportion of single-parent families

1.6 Methods of Analysis

- From the given data, the centroid of each enumeration districts will be calculated. The distance of each centroid from the Tar Pond centroid will be measured. This variable will be labelled "distance (d_i)".

- The analysis will start with an exploratory data analysis to examine the first order variations in attribute values. Choropleth maps of the aggregated data will be drawn using all the important variables like population in 1996, counts of preterm birth, standardized incidence ratios and so on to see if there is large scale variation within.

- The next stage is to examine second order properties, which involves spatial dependency i.e., test for spatial autocorrelation. Two ways of doing that are to use the Moran's I spatial correlogram or Geary's C correlogram.

- The third stage is to model the data. At this stage, both the frequentist and Bayesian approaches will be employed. Spatial weighted regression models or generalized least squares models will be fitted to the data to examine whether there are covariates that can explain the spatial variations in preterm birth. Various transformations will be made accordingly.

1.7 Computer Packages

The analysis will be done using the following packages: Arcview, Spacestat, S-plus, WINBUGS and R. Arcview has features that allow among other things conversion of data into maps (Choloropleth maps) for easy visualization of patterns and exploratory data analysis. It also has extensions for easy integration of other packages like Spacestat and S-plus. The Spacestat has features designed to speed up exploratory data analysis; detect spatial autocorrelation; and fit spatial regression models. S-plus and R allow flexible coding, which make it possible for other programs/routines to be developed or written.

1.8 Chapter Outlines

- Chapter 2 will contain a detailed review of some relevant literature to this research. This will be divided into two parts
  1. A review of relevant literature on the geographical approaches used in the analysis
  2. A review of spatial statistics literatures and methods from statistical point of view,
- Chapter 3 will examine detailed exploratory data analysis of all covariates used in the model,
- Chapter 4 will contain the result of the Bayesian analysis,
- Chapter 5 will examine detailed analysis of the frequentist approaches used in the project,
- Chapter 6 will discuss summary of our findings and possible directions for further work.

1.9 Definition of Key Terms

Gestational age: the interval between the first day of the mother's last normal menstrual period and the date of delivery

Preterm birth: a gestational age less than 37 completed weeks (less than 259 days).

Preterm birth rate: the number of preterm births per 100 live births in any given year.

Low birth weight: a birth weight less than 2500g.

Congenital Anomalies: these are structural abnormalities inborn errors of metabolism, physiological disturbances, mental retardation, and cellular and molecular abnormalities that are present at birth.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we will review some of the work done in relation to maternal proximity to waste landfills and risk of adverse reproductive outcomes. We will also review some of the methodological and theoretical background of these studies. Relevant English-language papers published between 1980 and 2003 were found using computerized literature searches on the Medline database. In addition, articles were traced using references cited in previous reviews (Morris and Wakefield, 2000; Tara, 2002; Upon, 1989; Vrijheid, 2000), and some unpublished and ongoing research works were also examined. All the studies relating to adverse reproductive outcomes were critically appraised with respect to the study design, exposure measure, source of health data, control for confounders, and reported findings.

These searches identified a number of studies in relation to maternal proximity to hazardous waste sites and risk of adverse reproductive outcomes. While some of the studies reviewed have reported a statistically significant association between maternal proximity to hazardous waste sites and risk of having low birth-weight births (see Berry and Bove, 1997; Elliot *et al.*, 2001; Goldberg *et al.*, 1995; Goldman *et al.*, 1985; Vianna and Polan, 1984), some other studies have reported otherwise (Baker *et al.*, 1988; Fielder *et al.*, 2000; Kharazi *et al.*, 1997; Shaw *et al.*, 1992).

A lot of studies have also reported a significant association between congenital anomalies and maternal proximity to waste sites (see Dolk *et al.*, 1998; Elliot *et al.*, 2001; Fielder *et al.*, 2000; Geschwind *et al.*, 1992; Gilbertson and Brophy, 2001; Goldman *et al.*, 1985). Nevertheless, these studies have been criticized by a lot of authors on the basis that they have failed to consider the chemical composition of the waste site and failure to identify which chemicals are responsible for the observed health effects (Baibergenova *et al.*, 2003; Vrijheid, 2000; Rylander *et al.*, 2000).

In response to the shortcomings identified in previous studies, further studies have been done to assess the association between adverse reproductive outcomes and maternal proximity to sites contaminated by polychlorinated biphenyls (PCB) or other volatile organic compounds (Baibergenova *et al.*, 2003; Rylander *et al.*, 2000). In particular, recent studies by Baibergenova *et al.* (2003) and Rylander *et al.* (2000) have shown that women exposed to PCB are at increased risk of giving birth to an infant with low birth weight.

All in all, general weaknesses in the literature studied can be stated as follows: First, a lack of direct exposure measurements can increase bias. Second, in some of the literature reviewed, residents near waste sites have reported cases of adverse reproductive outcomes or symptoms associated with it. However, it is difficult to conclude whether these cases or symptoms are effects of direct exposure to waste sites, stress and fear, or reporting bias. Third, the use of surrogate or indirect measures of exposure measurements in most of the studies can lead to misclassification of exposure, which may decrease the sensitivity of the study for finding a true effect (Vrijheid, 2000). This situation is a major source of bias in some of the case-controlled studies reviewed, especially the ones done by Dolk *et al.* (1998); Geschwind *et al.* (1992); and Shaw *et al.* (1992).

Fourth, the strength of a cross sectional design is enhanced if the survey is administered in both the population of interest and a control community. The difficulty of finding an appropriate control community limits the strength of most of the cross-sectional studies reviewed in this literature. Finally, socio-economic factors may be a major confounder in the study of reproductive health (Michal *et al.*, 1993; Sullivan, 1993), but have not been properly accounted for in some of the studies reviewed in the literature (Berry and Bove, 1997; Shaw *et al.*, 1992). In order to correct some of these shortcomings, a lot of work has also been done on the improvement of methodologies and the theoretical aspects of studies involving proximity to waste landfills and risk of adverse health. We will now review some of them in more details.

2.2 Theoretical Background and Context

2.2.1 Poisson Model for Aggregated Data

Let Y_i denote the number of observed cases of the disease, and N_i the population at risk in area A_i, $i = 1, \ldots, n$. Let E_i denote the expected number of cases in area A_i obtained by multiplying the population at risk, N_i, by the national rate (r). The national rate is a measure of the probability that a healthy person will develop a disease during a specific period of time. This is usually calculated by dividing the number of new cases of a disease over a period of time by the population at risk at that time. In most cases, these rates are age-standardized to adjust for differences in age composition of various populations. This is because age has a marked effect on mortality and morbidity. One of the most common methods for adjusting these rates is to stratify them by age-group. So that E_i is calculated using the age-specific rates.

Following Clayton and Kaldor (1987), we assume that for observed count, Y_i, in the area A_i.

$$Y_i|\lambda_i \sim \text{Poisson}(E_i\lambda_i) \quad i = 1, \ldots, n, \tag{2.1}$$

where λ_i denotes the relative risk of the disease for the study region A_i compared to the whole country (or a chosen reference region). Based on this assumption, the distribution of Y_i can be written as

$$f_{Y_i}(y;\lambda_i) = \frac{\exp(-E_i\lambda_i)(E_i\lambda_i)^y}{y!}; \quad y = 0, 1, \ldots; \quad 0 < \lambda_i < \infty$$

Hence, the maximum likelihood estimator of the relative risk (λ_i) in area A_i is given by

$$\hat{\lambda}_i = \frac{y_i}{E_i}$$

with $\text{Var}(\hat{\lambda}_i) = \lambda_i/E_i$ which can be estimated by y_i/E_i^2. This quantity is generally referred to as the standardized morbidity ratio (SMR) or standardized incidence ratio (SIR). It is an unbiased estimator of λ_i and one of the most widely used in measures of incidence of diseases in spatial epidemiology.

Recently, a number of authors (Datta *et al.*, 2000; Morris and Wakefield, 2000; Lawson *et al.*, 2000; Best *et al.*, 1999) have argued against the use of crude SIR without making adjustments. One of the disadvantages of the use of crude SIR is that it tends to be unstable and may not reveal the underlying structure in the data when the population at risk is small. This is because the standard error of $\hat{\lambda}_i$ is proportional to E_i^{-1} and so for very rare events and/or small areas (and hence a small E_i) the SIR may be very unstable. Alternative approaches have also been proposed for adjusting the crude SIR to improve its stability. These include smoothing models, linear Bayes methods, Bayesian models and empirical Bayes models.

2.2.2 Hypothesis Testing

Morris and Wakefield (2000) represent the null hypothesis that proximity to source does not influence risk by

$$H_0 : \lambda_i = \eta \quad \text{for} \quad i = 1, \ldots, n.$$

This definition is based on the assumption that all other sources of variability in risk have been accounted for. Let (x_0, y_0) denote the centroid of the putative source, (x_i, y_i) the centroid of area A_i and d_i the distance from the source to the centroid of area A_i. In the absence of an exposure measure that may be attached to each A_i, Morris and Wakefield (2000) define a natural additive distance/risk model by

$$\lambda_i = \eta \; \{1 + f(d_i; \theta)\}$$

where η is the background relative risk and $f(d_i; \theta)$ is a function of distance, such that $f(d_i; \theta) \rightarrow 0$ as $d_i \rightarrow \infty$. We will use a reparameterization of the form

$$\lambda_i = \eta \; g(d_i; \theta) \tag{2.2}$$

so that this model will be consistent with Bithel (1995) which will be discussed later. With this reparameterization, $g(d_i; \theta) \rightarrow 1$ as $d_i \rightarrow \infty$.

Now suppose a $q \times 1$ vector of area-level risk factors which may be denoted by z_i is available. This may be incorporated through a regression model of the form

$$\lambda_i = \eta \; g(d_i; \theta) \exp(z_i^T \phi).$$

Morris and Wakefield (2000) note that a regression approach will correctly adjust standard errors of estimated relative risks but may be inefficient due to sparsity of data.

One of the problems is that of over-dispersion (actual variance exceeding the nominal variance under the assumed probability model) which is common when using Poisson regression to account for both distance/risk and known covariates. One of the methods for accommodating this extra poisson variability is the use of the quasi-likelihood approach proposed by McCullagh and Nelder (1989) which specifies $E(Y_i|\lambda_i) = E_i\lambda_i$ and $\text{Var}(Y_i|\lambda_i) = \kappa E(Y_i|\lambda_i)$, the overdispersion parameter κ is then estimated. The most common method, however, is to follow a hierarchical modelling approach and to model the spatial dependence between the λ_i. This method will be explained in more detail later.

2.3 Conventional Epidemiological Methods

In this section, we review some of the conventional methods proposed for cluster detection and clustering analysis. We will also review some of the work done in relation to the assessment of disease risk for putative sources of hazard. While acknowledging the difficulties of defining "clustering", Wakefield *et al.* (2000: page 129) refers to it as "the pattern of the location of disease cases, relative to the pattern of the non-cases". They further describe spatial clustering as "residual spatial variation in risk". Wakefield *et al.* (2000) note that cluster detection has to start with a simple exploratory data analysis to see whether the data exhibits overdispersion. They define overdispersion in terms of

1. heterogeneity: independent Y_i with $\text{Var}(Y_i|\lambda_i) > E(Y_i|\lambda_i)$ for $i = 1, \ldots, n$ and

2. spatial dependence or clustering: dependence between Y_i and Y_j that is related to the geographical position of areas i and j for $i, j = 1, \ldots, n$ and $i \neq j$

and described methods for detecting it. We will now describe some of them in more detail.

2.3.1 Tests of Heterogeneity

Before carrying out a formal test to assess whether there is increased risk in any region, it is important to test whether the rates of disease differ from one Enumeration District to another. To test this, Wakefield *et al.* (2000) define the null hypothesis (H_o) as

$$H_o : \lambda_1 = \lambda_2 = \ldots = \lambda_n = \eta$$

and $H_1 : \lambda_i \neq \lambda_j$ for $i \neq j$ and describe two methods for testing this hypothesis.

Pearson's Chi-squared Statistic

The test statistic can be calculated from

$$T = \sum_{i=1}^{n} \frac{(Y_i - E_i^*)^2}{E_i^*} \quad \text{where} \quad E_i^* = E_i \times \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} E_i}$$

so that under H_o, the distribution of T is asymptotically a chi-square with *n-1* degrees of freedom. Hence, large values of T will result if there is heterogeneity. Wakefield *et al.* (2000) also explain that the significance of the test statistic can also be assessed by computing the empirical p-values based on a Monte Carlo test. This method starts with the random simulation of observations Y_i under the null hypothesis. Next, the test statistic is calculated under each simulation and this procedure is repeated a large number of times. Finally, the calculated test statistic is compared with the observed test statistic from original data. The drawbacks of this method were also identified as:

1. The test does not give any information about the location of the cluster but large $Y_i - E_i^*$ may be examined for clues.

2. The power of the test against any realistic alternatives is not very clear.

Potthoff and Whittinghill's method

The second method described by Wakefield *et al.* (2000) for detecting heterogeneity was based on the work by Potthoff and Whittinghill (1996). The test statistic was defined as

$$T = \left(\sum_{i=1}^{n} E_i\right) \sum_{i=1}^{n} \frac{Y_i(Y_i - 1)}{E_i}$$

where $Y_i(Y_i - 1)$ is the number of unordered pairs of observed cases in each area. Hence, an area will contribute to T if two or more cases occur. Under H_o, T has a mean $\sum_{i=1}^{n} Y_i(\sum_{i=1}^{n} Y_i - 1)$ and variance $2(n-1)(\sum_{i=1}^{n} Y_i(\sum_{i=1}^{n} Y_i - 1))$ with large values of T indicating heterogeneity. The distribution of T may be taken to be asymptotically normal but the empirical p-values based on a Monte Carlo test discussed in the previous section is more straight forward and preferable (Wakefield *et al.*, 2000).

2.3.2 Distance/adjacency Method

After testing for heterogeneity, the next step is to test whether Y_i exhibits some spatial structure. Wakefield *et al.* (2000) describe some of the methods for testing spatial dependency:

Autocorrelation Statistics

These statistics are based on a chosen measure of closeness, W_{ij}, between areas i and j. In the simplest form, a binary coefficient is used, such that $W_{ij} = 1$ if areas i and j share a common boundary, and $W_{ij} = 0$ otherwise. In general, W_{ij} may be selected based on the kind of spatial dependency that is anticipated (Wakefield *et al.*, 2000). Let $Z_i = Y_i/E_i$ denote the standardized incidence ratio of area i. Wakefield *et al.* (2000) describe three statistics for assessing spatial autocorrelation. These are Moran's I, Geary's C and D. Walter Test (Walter, 1993).

Moran's I

One of the most popular measure of spatial autocorrelation is Moran's I statistic defined as

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij})\sum_{k}^{n}(Z_j - \bar{Z})^2}.$$

This statistic is closely related to the conventional correlation coefficient and it is an approximate measure of the spatial dependence. When Z_i does not exhibit any spatial pattern, I will be close to zero and values of I close to 1 indicate clustering.

Geary's C

Another measure of spatial autocorrelation is the Geary's C statistic which is based on the weighted sum of square difference between observations and defined as:

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(Z_i - Z_j)^2}{2(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij})\sum_{k=1}^{n}(Z_k - \bar{Z})^2}.$$

When there is spatial dependence, the term in the numerator will be small and the value of the statistic will be close to zero. The absence of spatial dependence will result in large value for the numerator and hence, C will be close to 1.

D. Walter Test (Walter 1993)

This is a non-parametric rank-based method, denoted by D. The statistic is obtained by ranking Z_i and denoting the ranks by $Z_i^\star$. The non-parametric measure of spatial dependence can be calculated from

$$D = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} |Z_i^\star - Z_j^\star|}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}}$$

with small values of D indicating positive dependence. The major drawback of all the three statistics is that they do not allow for unequal variance of the SIR and hence may be misleading.

2.3.3 'Near Versus Reference' Comparisons

This is one of the simplest approaches of assessing the risk in relation to a point source. It involves the direct comparison of risk in the exposed population (i.e. lying within a certain distance of the point source) to that in the reference population (e.g the national rate). This approach suffers from the same problems as the use of crude SIR. Morris and Wakefield (2000) advise that the approach must be used with other confirmatory methods since it is rather exploratory. Two drawbacks were identified:

1. The problem of identifying the exposed population is not a clear one and may be very crucial in the analysis.

2. A significant increase in risk cannot be attributed to the exposure alone, since there is a very high likelihood that the two populations also differ in respects other than exposure.

In order to solve these problems a near versus far comparison was proposed.

2.3.4 'Near Versus Far' Comparisons

This involves dividing the study region into 'near' and 'far' regions which correspond to the exposed and unexposed population respectively. This method may be unreliable if the population at risk is small. In such cases, some of the methods used for adjusting the crude SIR may be applicable. In the study of disease rate, it is common to adjust for age/social economic status because of their effects on mortality and incidence of disease. One of the most common methods of doing this is to stratify the population by age-group or social economic status. Now, suppose we stratify the population at risk (N_i) by age-group into J strata. Then, N_{ij} is the population at risk in age-group j in area i and Y_{ij} is the observed number of cases of disease in age-group j in area i.

Based on this stratification, Morris and Wakefield (2000) consider a rare disease in which the number of cases observed and population at risk in stratum j in the near region are denoted by Y_{1j} and N_{1j}, respectively. The corresponding numbers in the far region are denoted by Y_{2j} and N_{2j}, respectively. Furthermore, they use Z_j and M_j to denote the number of observed cases and population at risk in stratum j in a standard population. The direct standardized rate was defined by

$$\sum_{j=1}^{J} \frac{M_j \hat{p}_{ij}}{M}, \qquad i = 1, 2$$

and the indirect standardized rate by

$$\frac{Y_i}{\sum_{j=1}^{J} N_{ij} \hat{q}_j} \times \frac{Z}{M}, \qquad i = 1, 2$$

where $\hat{p}_{ij} = Y_{ij}/N_{ij}$, $\hat{q}_j = Z_j/M_j$, $Y_i = \sum_j Y_{ij}$, $Z = \sum_j Z_j$ and $M = \sum_j M_j$. The directly standardized rate corresponds to a 'counter-factual' argument in which the estimated rates within the area of interest are applied to the standard population. The indirectly standardized rate applies the estimated relative risk to the rate in the standard population (Morris and Wakefield, 2000). Simple summaries of the rate ratios in the near or far regions are compared to the standard population by taking the ratios of the direct and indirect standardized rates to Z/M to give the comparative mortality figure (CMF) or the standardized incidence ratio (SIR), respectively:

$$\text{CMF}_i = \frac{\sum_{j=1}^{J} M_j \hat{p}_{ij}/M}{Z/M}, \quad \text{and} \quad \text{SIR}_i = \frac{Y_i}{\sum_{j=1}^{J} \hat{q}_j} \quad \text{for} \quad i = 1, 2.$$

A comparison of the near and far regions is provided by the ratios $\text{CMF}_1/\text{CMF}_2$ or $\text{SIR}_1/\text{SIR}_2$.

2.4 Semi-parametric Tests

2.4.1 Besag and Newell's method

Morris and Wakefield (2000) describe a version of the method proposed by Besag and Newell (1991) that is appropriate for cluster detection in relation to a pre-specified point source. The null hypothesis is that the cases are distributed at random over the population at risk in region A. They assume that the region A is divided into n disjoint areas as in the basic Poisson model. The area containing the centroid of the source (x_0, y_0) was labelled A_1 and all other areas $A_2 \ldots, A_n$ by increasing distance from A_1 based on the area centroid. With the assumption that there is no clustering, they defined the aggregated number of observed cases (D_i) and population within the nearest i areas (u_i) as

$$D_i = \sum_{l=1}^{i} y_l \quad \text{and} \quad u_i = \sum_{l=1}^{i} N_l \quad i = 1, \ldots, n.$$

The test statistic is $M = \min\{i : D_i \geq k\}$, the number of areas required to accrue at least k cases. A small observed value of M indicates that there is clustering around (x_0, y_0). Suppose m is the observed value of M, then the significance level of the test is $Pr(M \leq m)$ under the null hypothesis. Besag and Newell (1991) note that M will only be greater than m if and only if fewer than k individuals among u_m have the disease. Hence, under H_o, the hypergeometric probability that exactly s individuals among u_m have the disease can be closely approximated by the Poisson term if the disease is rare (Besag and Newell, 1991). It follows that the significant level for each potential cluster can be calculated from

$$Pr(M \leq m) = 1 - \sum_{s=0}^{k-1} \frac{\exp(-E_i)E_i^s}{s!}, \tag{2.3}$$

where $E_i = N_i q$ and q is an estimate of risk obtained through internal or external standardization. Each term in the sum represents the probability of observing s cases from a Poisson distribution with mean E_i. Morris and Wakefield (2000) note that a small p-value may result if the risk in the whole study region is high relative to the reference region from which q is derived and propose internal standardization or replacement of E_i by $E_i^{\star} = E_i \times Y/N$, where Y and N represent the total number of cases and the population at risk in the study region respectively.

In order to adjust for known risk factors, Morris and Wakefield (2000) consider the cumulative number at risk in stratum j within the nearest i areas defined by $u_{ij} = \sum_{l=1}^{i} N_{lj}$. For this adjustment to work, they propose replacing $E_i = N_i q$ in (2.3) by $E_i = \sum_j N_{ij} q_j$ where q_j are a set of stratum-specific reference probabilities. Alternatively, E_i may also be replaced by $E_i^\star$ where

$$E_i^\star = E_i \times Y/E \tag{2.4}$$

where $Y = \sum_i Y_i$ and $E = \sum_i E_i$ so that the overall difference between the risks in the study and reference areas have been removed. The method of Besag and Newell is simple to apply but its major drawback, as pointed out by Morris and Wakefield (2000), is that it may produce many false positives and does not provide an estimate of the risk around the putative source. This problem is a direct consequence of the fact that the method was originally designed for detection of clusters by detecting discrepancies between numerators and denominators due to differences in risk or data inaccuracies.

2.4.2 Stone's Test

Stone's test (Stone, 1988) is based on the following assumptions

1. $Y_i \sim \text{Poisson}(E_i \lambda_i)$ for $i = 1, \ldots, n$
2. The risk is a non-increasing function of distance.
3. Areas are ordered by increasing distance from the putative sources so that $i = 1$ corresponds to the closest area

The null hypothesis for the unconditional test is

$$H_o : \lambda_1 = \ldots = \lambda_n = 1$$

An alternative is to estimate λ_i subject to the order restriction:

$$H_1 : \lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$$

with at least one strict inequality holding. The estimation of parameters under this restriction is achieved analytically using the theory of isotonic regression (regression with order restriction) and implementation is carried out using either the min-max formulae (Stone, 1988) or "the pooled adjacent violators" (PAV) algorithm described by Stone. The hypothesis is tested using a generalized likelihood ratio test statistic based on the Poisson likelihood under the null and alternative hypotheses. The observed significance level of the test is calculated via Monte Carlo simulation.

One of the limitations of the unconditional test is that H_0 may be rejected simply because the study region as a whole has elevated or lowered risk compared to the reference (or national) rate used to compute E_i (Morris and Wakefield, 2000). To solve this problem, Bithel and Stone (1989) suggest a replacement of E_i by $E_i^\star$ as in (2.4) which allow for adjustment in known risk. Alternatively, Shaddick and Elliot (1996) suggest a conditional test with the null hypothesis defined as

$$H_o : \lambda_1 = \ldots = \lambda_n = \eta$$

so that the method of significance level estimation in Stone's test can be modified to allow for the unknown constant η.

As a way of avoiding the Monte Carlo test, the Poisson maximum test originally designed by Stone is always used. The Stone test has been applied widely in epidemiological literature, in particular those studies conducted by Small Area Health Statistics Unit (SAHSU) (Elliot *et al.*, 1992). Its major advantage is that it avoids the need to assume a fully specified distance/risk relationship.

2.4.3 Score Tests

With the assumption that w_i denote an 'exposure' associated with area A_i and $E_i^\star$ is as defined in (2.4), Morris and Wakefield (2000) define a test statistic which may be used to compare the null hypothesis of constant risk in all areas versus the general monotonic alternative by

$$\frac{\{\sum_{i=1}^n w_i(Y_i - E_i^\star)\}^2}{\sum_{i=1}^n w_i^2 E_i^\star - \frac{(\sum_{i=1}^n w_i E_i^\star)^2}{Y}} \tag{2.5}$$

The distribution of this statistic under H_0 is approximately chi-squared with a single degree of freedom. In the absence of specific quantitative exposure they assume that $w_i = i$ (with the areas ranked according to distance from (x_o, y_o)).

Lawson (1993) and Waller *et al.* (1992) suggest the use of a class of locally most powerful tests based on the likelihood score. These tests are based on the additive excess risk model:

$$\lambda_{1i} = 1 + \varepsilon g_i$$

where g_i represents a surrogate for exposure and may be prespecified constants or may be modelled using a parametric function of distance (Bithel, 1995).

2.5 Regression Methods

Regression is one of the most widely used methods in spatial epidemiology. In this section, we will discuss various models and methods of estimation.

2.5.1 Poisson Regression Models

In general for rare diseases and aggregated data where Y_i denotes the observed counts of diseases. The most widely used model is to assume Equation (2.1), where λ_i are the area-specific rate ratios and E_i are expected counts of events. It should be noted that if λ_i are not equal, then the data Y_i will display extra-poisson variation. In modelling disease rates in relation to a point source, we may assume the generalized linear model that incorporates both area-specific covariates and a measure of the spread of the risk from source. This model may be written as

$$\log \lambda_i = \log \eta + \log\ g(d_i;\theta) + z_i^T \phi \tag{2.6}$$

where z_i is a $q \times 1$ vector of area-specific covariates, η is a measure of the overall inflation of risk in the region under study and $g(d_i)$ is a decreasing function of distance. The parameters of the model may be estimated using the likelihood or the Bayesian approach.

2.5.2 Choice of g(d_i)

As explained earlier $g(d_i)$ must be defined such that as $d_i \to \infty$, $g(d_i) \to 1$. Bithell (1995) further clarify most of the controversies surrounding the choice of $g(d_i)$. Bithell proposes the following distance functions as suitable forms for $g(d_i)$.

$$\begin{aligned} g_1(d_i) &= \exp(\alpha/d_i) && (2.7)\\ g_2(d_i) &= (1+\xi\exp(-d_i/\beta)) && (2.8)\\ g_3(d_i) &= (1+\xi\exp(-(d_i/\gamma)^2)) && (2.9)\\ g_4(d_i) &= (1+\xi/(1+d_i/\delta)) && (2.10) \end{aligned}$$

where d_i may be the distance of the centroid of the subregion from the origin, or may denote any surrogate measure of inverse risk e.g. rank of distance. He further defines α, β, γ, and δ to represent decay rate. For $g_2(d_i)$ to $g_4(d_i)$, $1+\xi$ is a measure of the ratio of relative risk at source to that at infinity. Setting

$$\alpha_o = \log\ \eta, \qquad (2.11)$$

equation (2.6) becomes

$$\log \lambda_i = \alpha_o + \log\ g(d_i;\theta) + z_i^T\phi$$

Some functions $g(d_i;\theta)$ are very worthy of mention. Diggle (1990) defines $g(d_i;\theta)$ as

$$g(d_i;\theta) = 1 + \xi\ \exp(-\beta_o\ d_i^2)$$

where $\theta = (\xi, \beta_o)$. Diggle *et al.* (1997) propose an extension to the model by including a disc around the source of unknown radius δ, within which the risk remains constant. They also reparameterize the model by using $\beta = \beta_o^{-1/2}$ so that β is measured in the same units as distance. This leads to

$$g(d_i;\theta) = \begin{cases} 1+\xi & d_i \le \delta \\ 1+\xi\ \exp[-(d_i-\delta)/\beta^2] & d_i > \delta \end{cases}$$

where $\theta = (\xi, \beta, \delta)$. Here, $1+\xi$ is used as a measure of the proportion of elevated risk at source, δ is the radius of the plateau of maximal risk and β represents the distance from the rim of plateau at which the risk has decreased by a factor of $\exp(-1) \approx 0.36$.

2.5.3 Area-specific Covariates

One of the most effective methods for measuring the socio-economic status of a community is the use of deprivation index. The importance of some socio-economic factors in the prediction of disease incidence and mortality has been emphasized by a lot of researchers (Jolley *et al.*, 1992; Pampalon and Raymond, 2000; Townsend, 1987). Townsend defined deprivation as "a state of observable and demonstrable disadvantage relative to the local community or the wider society or nation to which the individual, family or group belongs" (Townsend, 1987: page 125). In modelling disease risk in relation to a point source, socio-economic variables have to be taken into account because of confounding effects (Jolley *et al.* 1992). A lot of methods based on some socio-economic variables have been proposed and used as a measure of deprivation in the community. We describe some of these in more detail

The Townsend index (Townsend, 1987) involves four variables: unemployment; absence of a car; housing tenure; and overcrowding. These variables are standardized and log transformation is done to ensure normality, unit weights are then attached to the standardized variables to obtain the combined index. Carstairs and Morris (1991) developed another index for measuring deprivation based on the following variables: persons in households with more than one person per room; persons in households where the head is economically active and from social class IV or V (semi-skilled and unskilled laborers); economically active male seeking work, and persons in private households without access to a car. These variables are standardized and unit weights are then attached to the standardized variables to obtain the combined index.

Some of the variables used in Townsend's index and Carstairs' index are not readily available in the Canadian census data, so Pampalon and Raymond (2000) propose one for health and welfare planning in Quebec. The index is based on the following socio-economic variables: the proportion of persons who have no high school diploma; the rate of unemployment to population; average income; the proportion of persons who are separated, divorced or widowed; the proportion of single-parent families; and the proportion of people living alone. Using principal component analysis (with varimax rotation), they are able to derive two independent scores (material and economic), which shows a very significant association with life expectancy at birth among men and women in Quebec.

2.6 Parameter Estimation

The parameters of the model (2.6) can be estimated using the likelihood approach or Bayesian Hierarchical modelling. The likelihood method will be explained in detail in Chapter 5. We will now introduce the Bayesian approach and leave the full discussion for Chapter 4:

2.6.1 Bayesian Hierarchical Modelling

Following Wakefield and Morris (2001), Wakefield *et al.* (2000), Datta *et al.* (2000), Best *et al.* (1999) and Besag *et al.* (1991) we define the stages as

First-stage Model

$$\log \lambda_i = \alpha_o + \log\ g(d_i;\theta) + z_i^T\phi + V_i + U_i \tag{2.12}$$

where V_i and U_i denote the non-spatial and spatial random effects respectively which are generally assumed to be independent.

The function $g(d_i;\theta)$ is a function of distance d_i from the center of the point source such as those defined earlier. Different form of $g(d_i;\theta)$ have been used in modelling of diseases risk in relation to point source. For example Datta *et al.* (2000) define $g(d_i;\theta)$ as

$$g(d_i;\theta) = \exp(\alpha/d_i)$$

where $\theta = \alpha$. Wakefield and Morris(2001) define $g(d_i;\theta)$ in terms of $g_2(d_i)$ to give

$$g(d_i;\theta) = 1 + \xi\ \exp\left[-\left(\frac{d_i}{\beta}\right)^2\right]$$

where $\theta = (\xi,\beta)$. It should be noted that $\xi = 0$ corresponds to no relationship between distance and risk.

Second-stage Model

At the stage we try to address some of the problems of instability of MLE $\hat{\lambda} = Y_i/E_i$ when the data is sparse. The usual approach is to allow the estimates of each of λ_i to 'borrow strength' from the remaining estimates of λ_j, $j \neq i$ by specifying a joint model for $\lambda = (\lambda_1,\ldots,\lambda_n)$. This is achieved by specifying a multivariate probability distribution for λ (Wakefield *et al.*, 2000).

Third Stage: Prior Distributions

At this stage we specify prior distributions for all the parameters in the first and second stages. In general for all parameters in the model normal priors with large variance are usually specified to represent vague beliefs. Another possibility is to specify improper uniform priors (Datta *et al.*, 2000; Wakefield *et al.*, 2000).

Chapter 3

Exploratory Data Analysis

3.1 Introduction

In this chapter, exploratory data analysis of all the covariates will be carried out. Maps of each of these variables will be plotted to see if there are any obvious clusters. A confirmation of these patterns will be done using some of the methods discussed in chapter 2. Areas that are not shaded on the maps show the 14 missing values explained earlier in section 1.5.

3.2 Standardized Incidence Ratio

Preterm births only occur in females within the child bearing age and the condition is not infectious. Hence, it is reasonable to assume that each case occurred independently, so that the distribution of observed counts, Y_i, $i = 1, \ldots, 144$ is as defined in Equation (2.1). The expected counts (E_i) for each enumeration district was calculated from the Canada preterm birth rate of 7.1 per 100 live births in 1996 (source: Population and Public Health Branch, Health Canada).

This rate is assumed fixed for 1996 and might have been calculated by including data from the Cape Breton Regional Municipality, but we will assume that the effect of this can be ignored. The expected counts for each enumeration district were calculated by multiplying the population at risk in each enumeration district by the national rate of 7.1% and this is denoted by E_i. Hence, E_i is the expected number of preterm birth from all other sources

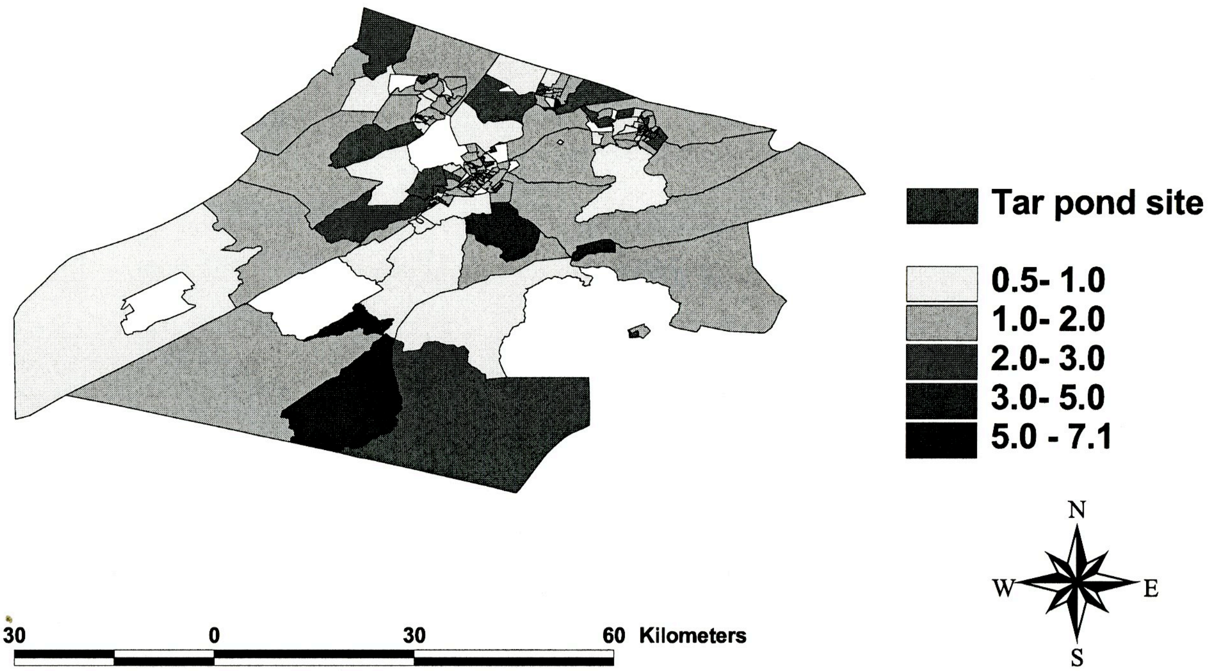

Figure 3.1: *Maximum likelihood estimates of the relative risks (SIR) for preterm birth*

of risk other than pollution from the Sydney Tar Pond. Figure 3.1 shows the map of the maximum likelihood estimators of the relative risk (SIR), $\hat{\lambda}_i = Y_i/E_i$. From the map, areas with $\lambda_i \leq 1$ indicate no risk or absolute risk reduction while $\lambda_i > 1$ indicate high risk of preterm birth compared to the rest of Canada. But as explained earlier, care has to be taken when interpreting the crude map of SIRS. To illustrate this, we will plot the SIR against the population at risk (see Figure 3.2). This graph clearly shows that areas with low population at risk tend to show high variability in SIR. This can be adequately accounted for using the Poisson model for aggregated data. This will be explained in the next two chapters.

3.3 Area-specific Risk

Following Pampalon and Raymond (2000), the following area-specific variables were considered for the analysis: The proportion of persons who have no high school diploma, the rate of unemployment to population, average income, the proportion of persons who are separated,

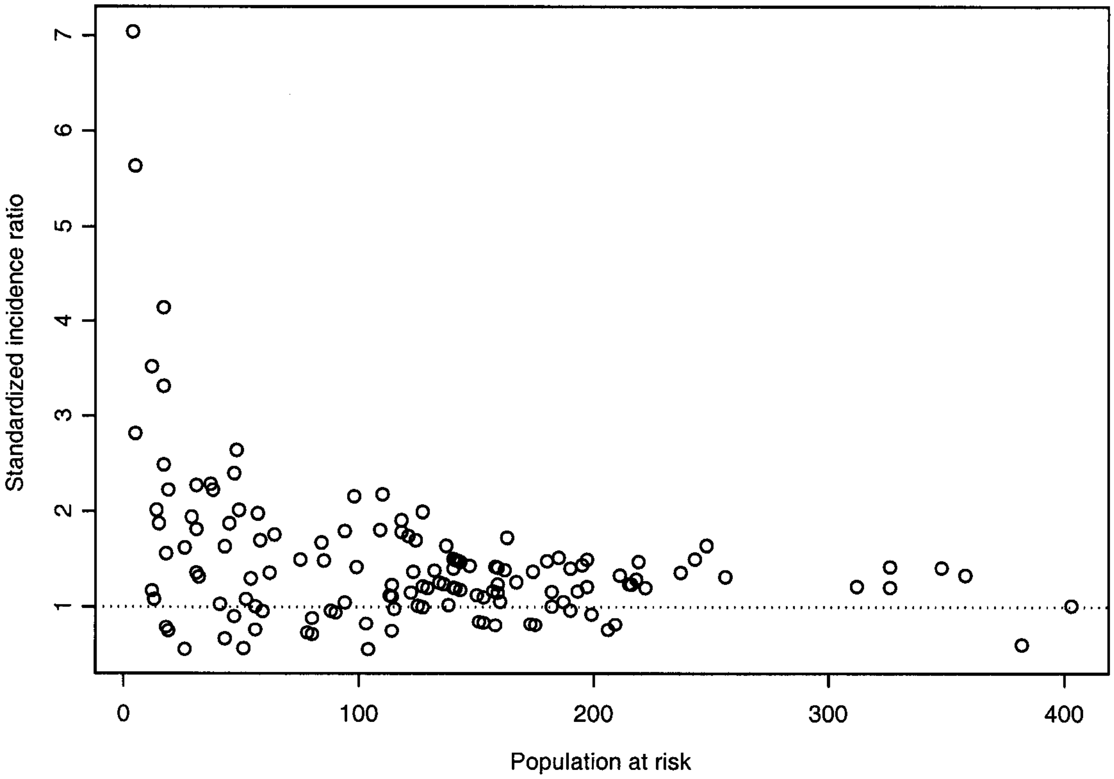

Figure 3.2: *Plot of SIR against population at risk*

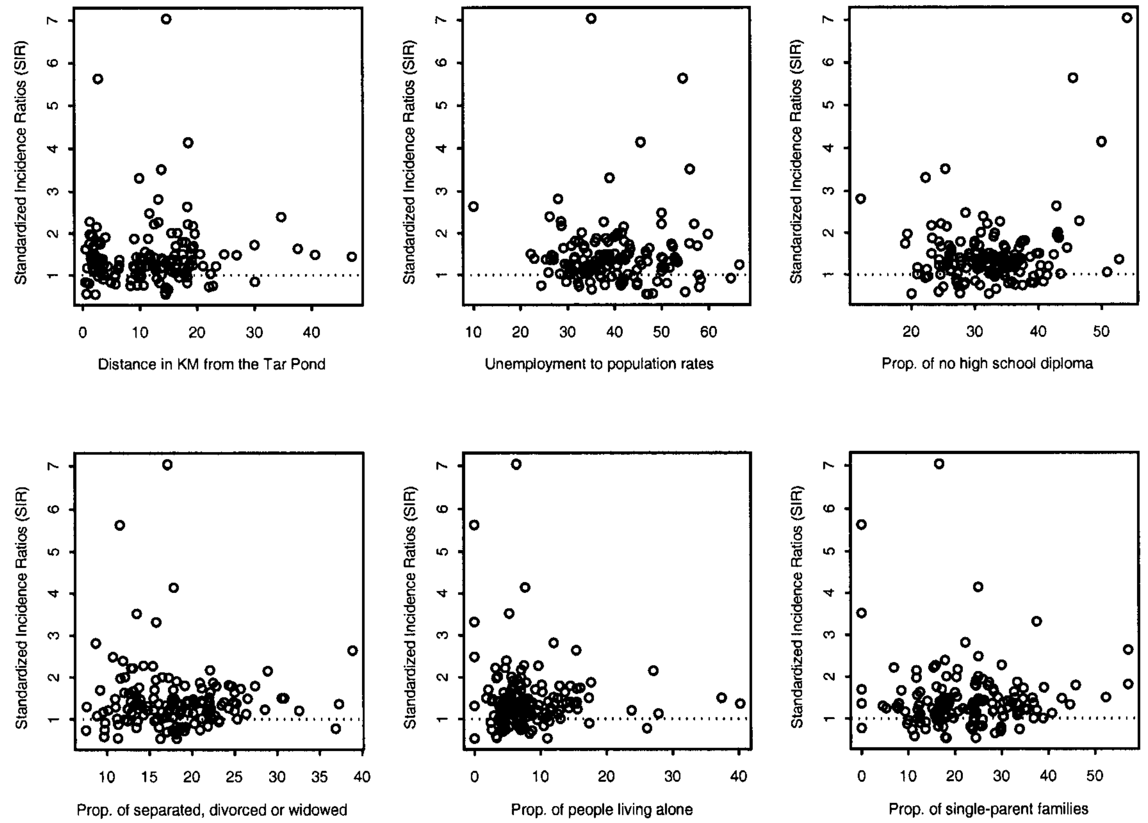

Figure 3.3: *Plot of SIR versus distance in km from the centroid of the Tar Pond and other area-specific covariates.*

divorced or widowed, the proportion of single-parent families and the proportion of people living alone.

Only five of the variables are available at all the 144 EDs with average income available only in 130 EDs. So we could not compute an adequate measure of deprivation based on the method proposed by Pampalon and Raymond, we decided to assess the effect of each of the variables separately leaving out average income. Distance from the Tar Pond site and all the area-specific variables were plotted against SIR to assess the effect of each. The plots are given in Figure 3.3.

As explained earlier, points below the doted line indicates no risk or absolute risk reduc-

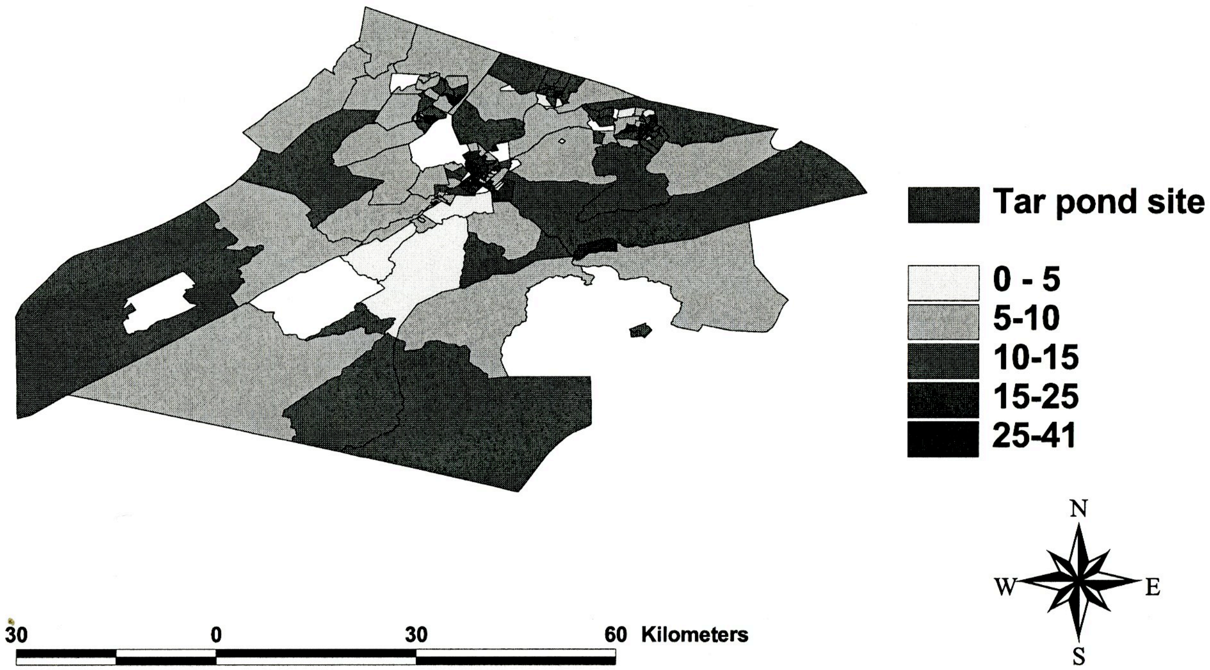

Figure 3.4: *The percentage of people living alone*

tion and vice versa. All the high values of SIR occurred within the 20 km distance from the Tar Pond. There is a slight evidence of decrease in risk from source as we move further away but this will be tested statistically in the next chapter.

The plot of SIR and the rate of unemployment to population show an upward trend with high unemployment rates associated with high SIR. A similar pattern is displayed by the plot of SIR and proportion of persons with no high school diploma. In the plot of the SIR and proportion of separated, divorced and widowed; areas with low proportion of separated, divorced and widowed tend to have low SIR. A similar pattern is seen in the plot of SIR and proportion of people living alone. There is no obvious pattern in the plot of SIR and proportion of single parent families.

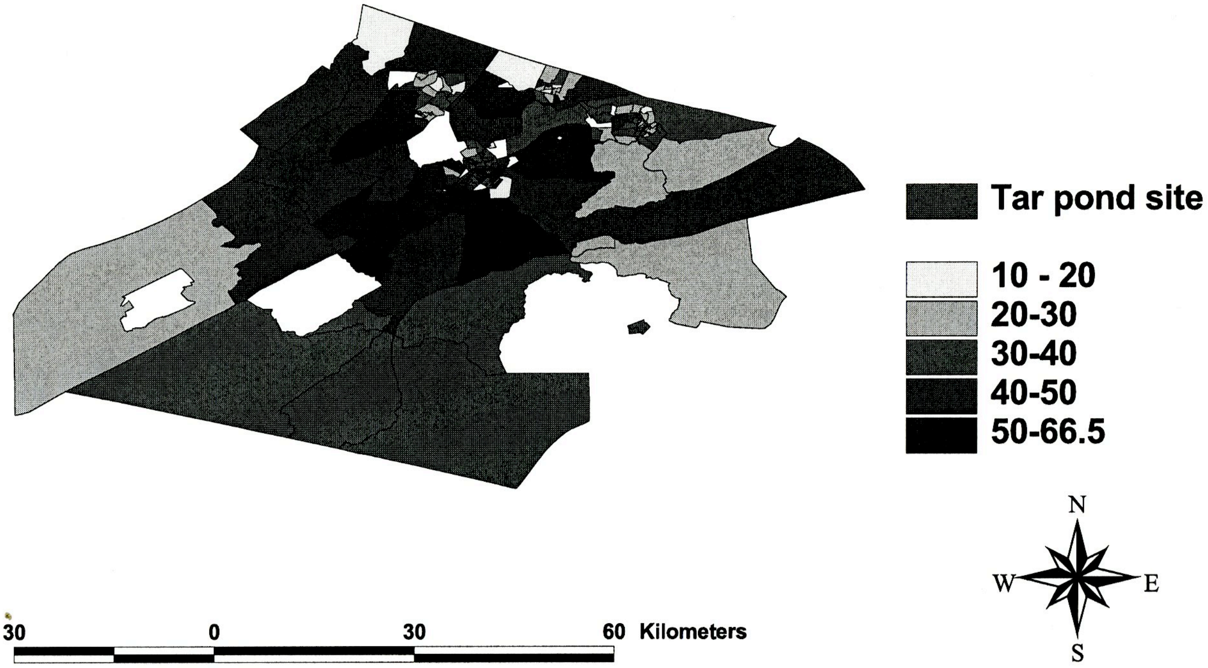

Figure 3.5: *The rate of unemployment to population*

3.4 Test for Spatial Dependency

One of the objectives of this study is to check for any obvious clustering of events around the Tar Pond that may be significant in explaining the variation in preterm birth rates. This can be done by plotting the maps of all the variables and visually assessing whether there is any cluster or carrying out a formal test using some of the methods discussed in the last chapter like Moran I or Geary C statistics.

From the map of SIR in Figure 3.1, we would expect a cluster of high SIR around the Tar Pond or a decrease in the SIR as we move further away from the Tar Pond but neither of the two is obvious from Figure 3.1. The maps of all the area-covariates were plotted to see whether there is any spatial pattern . The plots are displayed in Figures 3.4 to 3.8. Figure 3.4 shows a pattern with the highest proportion of people living alone occurring within the 20km radius of the Tar Pond site.

Figure 3.5 also shows that the unemployment to population ratio decreases as we move further away from the Tar Pond. From Figure 3.6 we can see a small cluster of proportion

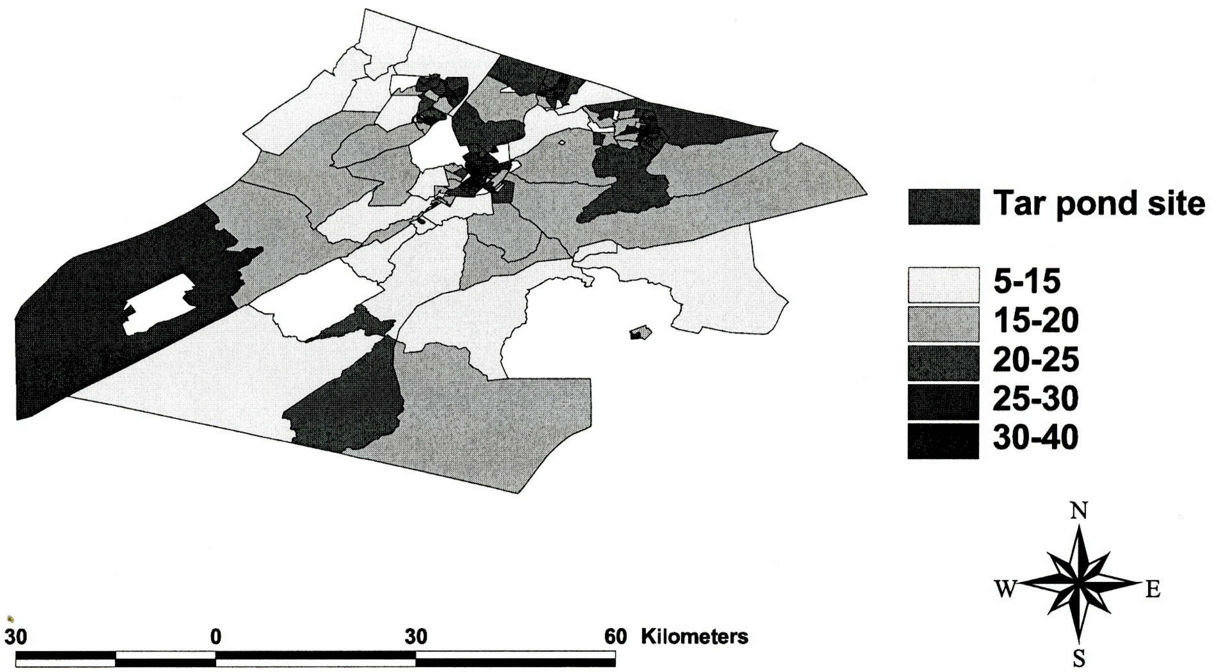

Figure 3.6: *The percentage of persons who are separated, divorced or widowed*

of separated, Divorced of widowed. Figure 3.7 show that the proportion of single parent family is relatively spread except for three noticeable clusters of which two are close to the Tar Pond site. Finally, we can see that the proportion of persons who have no high school diploma shown in Figure 3.7 displays some spatial pattern with some of the area close to the Tar Pond having high proportion.

Existence of spatial autocorrelation was also tested formally using the Moran I test. This test was carried out using the S-Plus extension in Arcview which allow coordinates of these maps to be exported to S-Plus. Results of the spatial autocorrelation analysis are given in Table 3.1 with variables defined as in Table 1.1. These results show the correlation, variance, normal statistic and p-value. The only variable that is not significant based on the associated p-value is SIR. This confirms the result of the visual examination of maps.

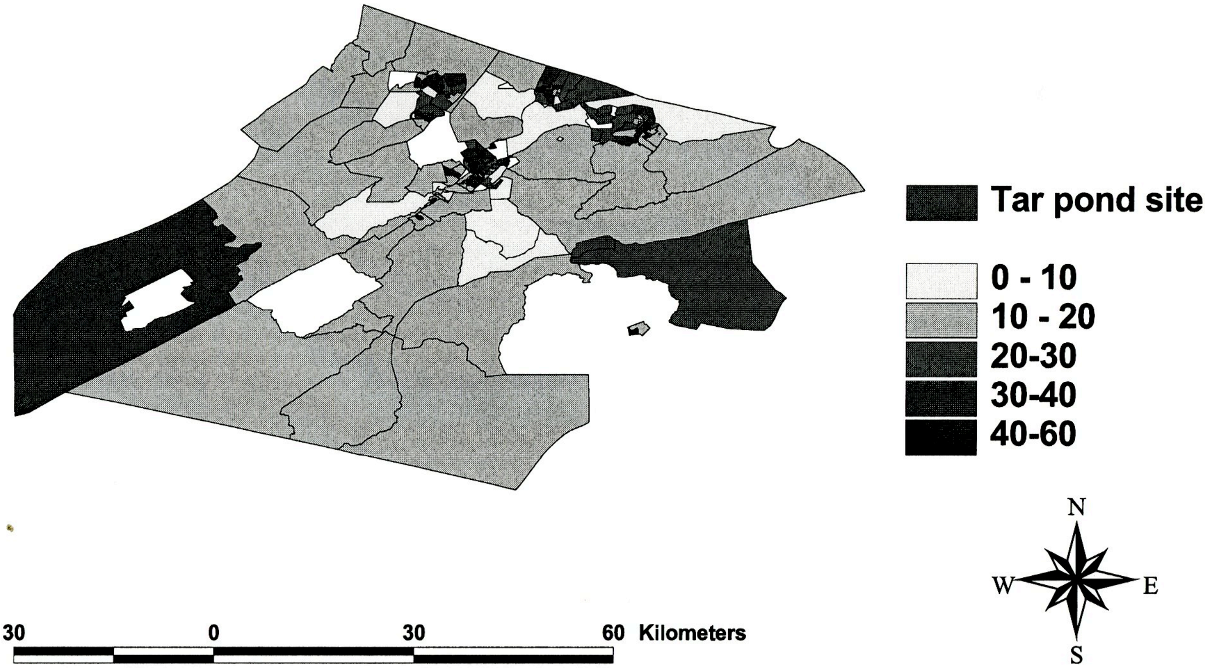

Figure 3.7: *The percentage of single-parent families*

Table 3.1: *Results of Spatial Autocorrelation Analysis using Moran I statistics*

Variables	Correlation	Variance	Std. Error	Normal statistic	Normal p-value
SIR	-0.03798	0.002541	0.05041	-0.6148	0.5387
x_1	0.348	0.002541	0.05041	7.043	1.888e-12
x_2	0.4582	0.002541	0.05041	9.229	2.732e-20
x_3	0.1924	0.002541	0.05041	3.955	7.659e-5
x_4	0.4051	0.002541	0.05041	8.174	2.984e-16
x_5	0.2932	0.002541	0.05041	5.955	2.607e-9

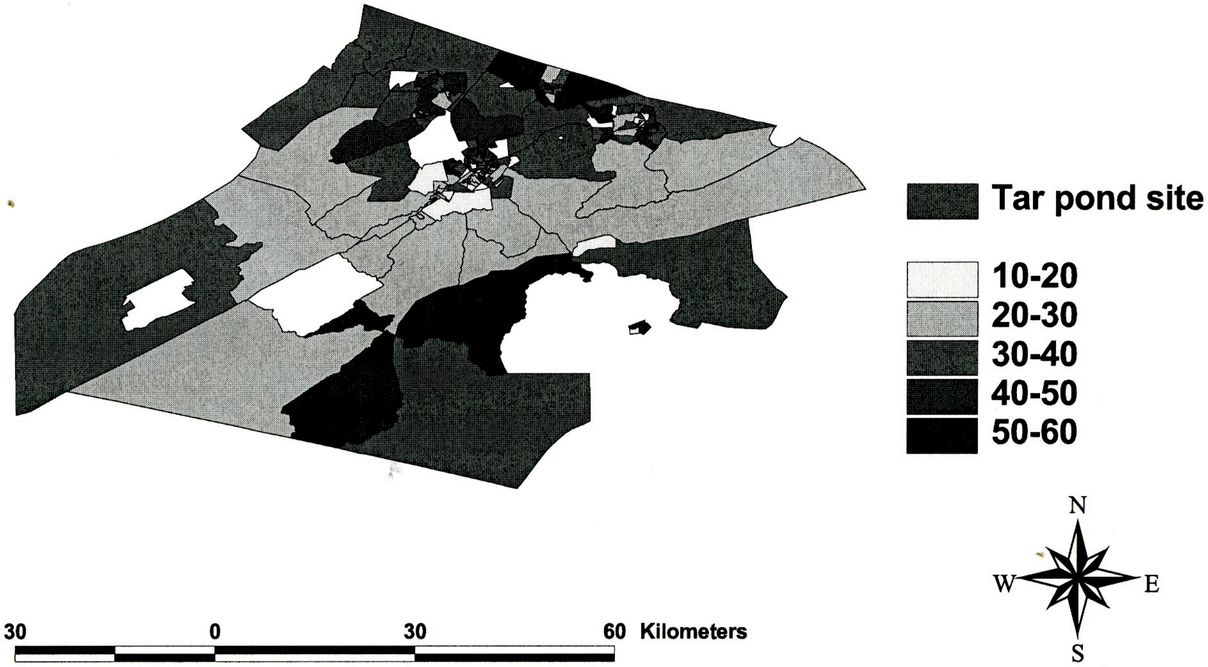

Figure 3.8: *The percentage of persons who have no high school diploma*

Chapter 4

Bayesian Hierarchical Modelling

4.1 Introduction

In order to model the data while accommodating the expected over dispersion and also including the spatial components (location or relative position of data values) of the data, Bayesian hierarchical modelling was used. The implementation of this modelling was done with WINBUGS and GeoBugs software for modelling aggregated data with plots and convergence diagnostic test done with **coda** package in R (Plummer *et al.*, 2004). One of the major advantages of this method of modelling disease risk is that it combines information from the data (likelihood) with the prior distribution of the disease risks. The mean or the median of the posterior distribution is used as a point estimate of disease risk for each area. The two basic assumptions underlying the use of this method for aggregated data are: First, disease in each enumeration district is assumed rare and non-infectious. Hence, occurrences are independent. Second, the risk is assumed to be constant in each enumeration district. The modelling is explained in the following three stages:

4.2 Description of the Model

First-stage Model

Following Datta *et al.* (2000) and Bithel (1995), we defined

$$g(d_i;\theta) = \exp(\alpha/d_i)$$

so that equation (2.12) becomes

$$\log \lambda_i = \alpha_o + \alpha/d_i + z_i^T\phi + V_i + U_i \tag{4.1}$$

where d_i is the distance of the ith enumeration district (ED) from the centroid of the Tar Pond, $\eta = \exp(\alpha_o)$ is a measure of the overall inflation of risk in the region under study, α represents the decay rate and ϕ is a vector of parameters of the area-specific covariates. V_i are unstructured random effects included in the model to capture the effects of unknown or unmeasured area level covariates. Hence, $\exp(V_i)$ will be equal to the residual or unexplained relative risk in area A_i after adjusting for known area-specific covariates. We have included U_i in the model to capture our belief that the unstructured random effects (V_i) may exhibit some spatial structure. The second stage of our analysis is to model the expected overdispersion in the model by defining appropriate structured and unstructured random effects for the model.

Second stage: overdispersion model

As discussed earlier in Chapter 2, there are two types of over-dispersion: heterogeneity and spatial dependence. In this section, we assume that the unstructured random effects which is a measure of heterogeneity is of the form

$$V_i \overset{iid}{\sim} N(0,\sigma_v^2) \quad i = 1,\ldots,n$$

where σ_v^2 is a measure of the between-area variability of the V_i. Next, we specify the spatial random effect to model the anticipated spatial dependence of the log of relative risk. For a detailed review on the modelling of the spatial variability we turn to Wakefield *et al.* (2000). The problem of accounting for spatial dependence is a bit more complex than that of heterogeneity. This is because we are interested in modelling an n-dimensional random vectors $\mathbf{U} = (U_1,\ldots,U_n)^T$ while making allowance for dependence between U_i and U_j for $i \neq j$. One way of doing this modelling is by specifying the joint distribution of $\mathbf{U}$. The second, which will be discussed, is the use of univariate conditional distributions of $U_i|U_j = u_j$, $j \neq i, i = 1, \ldots, n$. Wakefield *et al.* (2000) define

$$\mathbf{U} \sim N_n(\mathbf{0}_n, \sigma_u^2 \Sigma) \tag{4.2}$$

where Σ is an $n \times n$ positive definite correlation matrix. The parameter σ_u^2 is a measure of the overall variance of the U_i. They also define a matrix $\mathbf{Q} = \Sigma^{-1}$ and denote element (i, j) of this matrix by Q_{ij} for $i, j = i = 1, \ldots, n$.

Following Besag and Kooperberg (1995) and the standard properties of the multivariate normal, Wakefield *et al.* (2000: page 124–125) give a detailed derivation of the conditional distribution of $U_i|U_j$ from equation (4.2). The general form is given by

$$U_i|U_j = u_j, j \neq i \sim N\left(\sum_{j=1}^{n} W_{ij} u_j, \sigma_u^2 D_{ii}\right) \tag{4.3}$$

where $W_{ii} = 0$, $W_{ij} = -Q_{ij}/Q_{ii}$, and $D_{ii} = Q_{ii}^{-1}$. This equation defines a Markov random field (MRF) model because spatial dependency is modelled through the conditional distribution of $U_i|U_j$ (Wakefield *et al.*, 2000; Wakefield and Morris, 2001). The use of equation (4.3) always starts with the specification of a spatial weight (W_{ij}) which defines the set of neighbours that contributes positive weights to the conditional expectation of U_i.

One of the most common methods of specifying the MRF model is the use of intrinsic conditional autoregressive (CAR) proposed by Besag *et al.* (1991) and defined by

$$U_i|U_j = u_j, j \neq i \sim N(\bar{u}_i, \frac{\sigma_u^2}{m_i})$$

where $\bar{u}_i = \sum_{j \neq i} u_j/m_i$ and m_i is the number of neighbours. Comparing this with equation (4.3) shows that $D_{ii} = 1/m_i$ and $W_{ij} = 1/m_i$ for neighbours and zero otherwise. The most challenging aspect of this modelling is in how to define neighbours and choose W_{ij}. In our case, we have defined areas i and j as neighbours if they share a common boundary (see Wakefield *et al.*, 2000; Clayton and Kaldor, 1987; Besag *et al.*, 1991). We have also defined the spatial weights $\{W_{ij} : i = 1, \ldots, n\}$ as a 0-1 contiguity matrix in which $W_{ij} = 1$ for neighbours and $W_{ij} = 0$ otherwise. Furthermore, $W_{ii} = 0$ and the constraint $\sum_{i=1}^{n} U_i = 0$ is imposed for identifiability.

Third Stage: Prior Distributions

Table 4.1: *Priors*

Authors	$\sigma_v^{-2} priors$	σ_u^{-2} priors
Wakefield *et al.*, 2000	Gamma $(0.5, 0.0005)$	Gamma $(0.5, 0.0005)$
Wakefield and Morris 2001	Gamma $(0.5, 0.0005)$ Gamma $(0.5, 0.0005)$	Gamma $(0.5, 0.0005)$ Gamma $(0.1, 0.1)$
Best *et al.*, 1999	Gamma $(0.001, 0.001)$	Gamma $(0.1, 0.1)$

At this stage all the parameters (α_o, α, ϕ, σ_v^{-2} and σ_u^{-2}) of the model are assigned a prior distribution. α_o was assigned a flat prior which corresponds to a uniform distribution over the whole real line. α, and ϕ_i were assigned a normal $(0, 10^{-5})$. In WINBUGS, a normal distribution is always specified in terms of its mean and precision. Hence a normal $(0, 10^{-5})$ is another way of making a uniform distribution out of normal by specifying a large variance.

The choice of prior for σ_v^{-2} and σ_u^{-2} is a very challenging one and it has to be done carefully. Also, sensitivity test has to be done with these priors. Many authors have favoured the use of gamma(a,b) for both σ_v^{-2} and σ_u^{-2} because it is a conjugate prior to the normal but the choice of a and b is what they have not agreed on (Wakefield *et al.*, 2000; Wakefield and Morris, 2001; Clayton and Kaldor, 1987; Besag *et al.*, 1991; Best *et al.*, 1999; Datta *et al.*, 2000). In the Table 4.1, we give some of these priors. In our case, we have assigned Gamma $(0.1, 0.1)$ to both σ_v^{-2} and σ_u^{-2} and carry out sensitivity analysis with all the priors given in Table 4.1.

4.2.1 Relative Risk Estimates

What is the true relative risk (λ_i) of preterm birth in each enumeration districts for 1996 compared with the reference population? In order to answer this question, we define the general form of the relative risk function (λ_i) from equation (4.1) as

$$\lambda_i = \eta\ g(d_i; \theta) \exp\{z_i^T \phi + V_i + U_i\}. \tag{4.4}$$

Wakefield and Morris (2000) advise that this general form should be used to estimate area-level relative risk when doing model checking and also to factor in our belief that the random effects V_i and U_i are contributing significantly to λ_i. An alternative form is

$$\lambda_i = \eta \; g(d_i; \theta) \exp\{z_i^T \phi\}.$$

which is preferred when we believe that V_i and U_i are mainly accounting for data anomalies. In our case, we have estimated λ_i from equation (4.4).

4.3 Implementation

Let us define a parameter vector $\theta = (\alpha_o, \alpha, \phi, \sigma_u^2, \sigma_v^2)^T$ and denote the prior distribution of θ by $\pi(\theta)$ and the likelihood for data D given θ by $L(D|\theta)$. Then, the posterior distribution of θ and D which is the "object of all Bayesian inference" (Gilks *et al.*, 1996) is defined as

$$\pi(\theta|D) \;\propto\; \pi(\theta) \; L(D|\theta) \tag{4.5}$$

The next step is to generate a sample from the posterior distribution $\pi(\theta|D)$. This may be a bit complex because the normalization constant of equation (4.5) defined by $\int \pi(\theta) \; L(D|\theta) d\theta$ is high dimensional and may not be easy to evaluate analytically. One way of doing effective sampling from this posterior distribution is to use the Markov chain Monte Carlo (MCMC) simulation. Roberts (1996: page 41) defines a Markov chain X as

> a discrete time stochastic process $\{X_0, X_1, \ldots\}$ with the property that the distribution of X_t given all previous values of the process, $X_0, X_1, \ldots X_{t-1}$ only depends upon X_{t-1}. Mathematically, we write
>
> $$P[X_t \in A | X_0, X_1, \ldots X_{t-1}] = P[X_t \in A | X_{t-1}]$$
>
> for any set A, where $P[.|.]$ denotes a conditional probability.

Roberts (1996) further gives three important properties that must be satisfied by the distribution of X_t before it can converge to a stationary distribution. These are:

1. It must be "irreducible". This means that the ability of the Markov chain to reach any non-empty set with positive probability, in some number of iterations should not be influenced by the starting position.

2. The chain must be "aperiodic". This is a condition that will make it impossible for the Markov chain to oscillate between different sets of states in a regular periodic movement.

3. The last and most important condition is that the chain must be "positive recurrent". This means that if the initial value X_0 is sampled from $\pi(.)$, where $\pi(.)$ is a stationary distribution, then all subsequent iterates will also be distributed according to $\pi(.)$.

Applying this concept to our case, we need to construct a Markov chain with state space θ_c, where $\theta \in \theta_c \subset \Re^k$. This process is then used to generate random samples from the joint posterior distribution of $\pi(\theta|D)$. Suppose we denote these samples by $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(R)}$.

The next step is to use Monte Carlo integration to approximate the expectation of a function $f(\theta)$. This is defined as

$$E[f(\theta)] \approx \frac{1}{R}\sum_{i=1}^{R} f(\theta^{(i)}).$$

The main idea here is that the population mean of $f(\theta)$ is been estimated by a sample mean. One of the methods of constructing the needed Markov chain is the use of the Metropolis-Hastings algorithm. Another method is the use of the Gibbs sampling algorithm. This second approach is a special case of the single-component Metropolis-Hastings algorithm and it can be implemented in the WINBUGS statistical software. Output from the MCMC is usually summarized in terms of ergodic averages, which provide an estimate of the posterior means of θ_k $\{k = 1, \ldots, l\}$, where l is the number of parameters. Hence, the posterior mean can be estimated by the sample posterior mean $\bar{\theta}_k = \sum_{i=1}^{R} \theta_k^{(i)}/R$. The posterior variance is estimated by the sample posterior Monte Carlo variance $\sigma_\theta^2 = \sum_{i=1}^{R}(\theta_k^{(i)} - \bar{\theta}_k)^2/R$.

4.3.1 Methods of Sampling from Posterior Distribution

There are basically two approaches agreed on by a lot of authors on how to sample from the posterior distribution. One approach involves running one long chain for a long period and assessing the convergence of the chain to the required expectation (posterior distribution). This method is considered more efficient but assessment of convergence may be difficult. The second approach involves running more than one chain and starting from different points in the parameter space. This method is very good for convergency assessment. Gelman and Rubin (1992) suggest running 3 to 5 chains starting from "overdispersed" positions in the posterior distribution and drawing inference from all the chains. This is to avoid a situation in which one is stuck around 1 local posterior mode.

4.3.2 Convergence diagnostics

In this section we will discuss some of the methods of assessing convergence of chain(s) in MCMC analysis to the target distribution. Theoretically posterior means can only be obtained at infinity. In practise however, a reasonable approximation is good enough. The main question is that, at what point can we say that a chain or chains have converged? We have answered this question by running five independent chains starting at different initial values. Assessment of convergence in WINBUGS can be done informally by checking the time series plots. This can then be confirmed formally with the Gelman and Rubin's method (Gelman and Rubin, 1992). Once convergence is reached, we expect the samples to look like a random scatter plot about a stable mean value. This can easily be seen in the time series plots. Gelman and Rubin's method monitors convergence by estimating the factor by which the scale parameter might shrink if sampling were continued indefinitely. This is defined as

$$\sqrt{\hat{R}} = \sqrt{(\frac{n-1}{n} - \frac{m+1}{mn}\frac{B}{W})\frac{df}{df-2}}$$

where B is the variance between the means from m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of the approximating t distribution.

4.3.3 Autocorrelation function

This is a measure of how the values within the chains are related. High autocorrelation may occur if the parameters in our model are highly correlated. This is a very serious problem because it may slow down the Gibbs sampling process and increase the time needed to fully explore the entire posterior distribution. A very simple diagnosis of the autocorrelation is the use of the autocorrelation plots which is available in WINBUGS. The presence of high

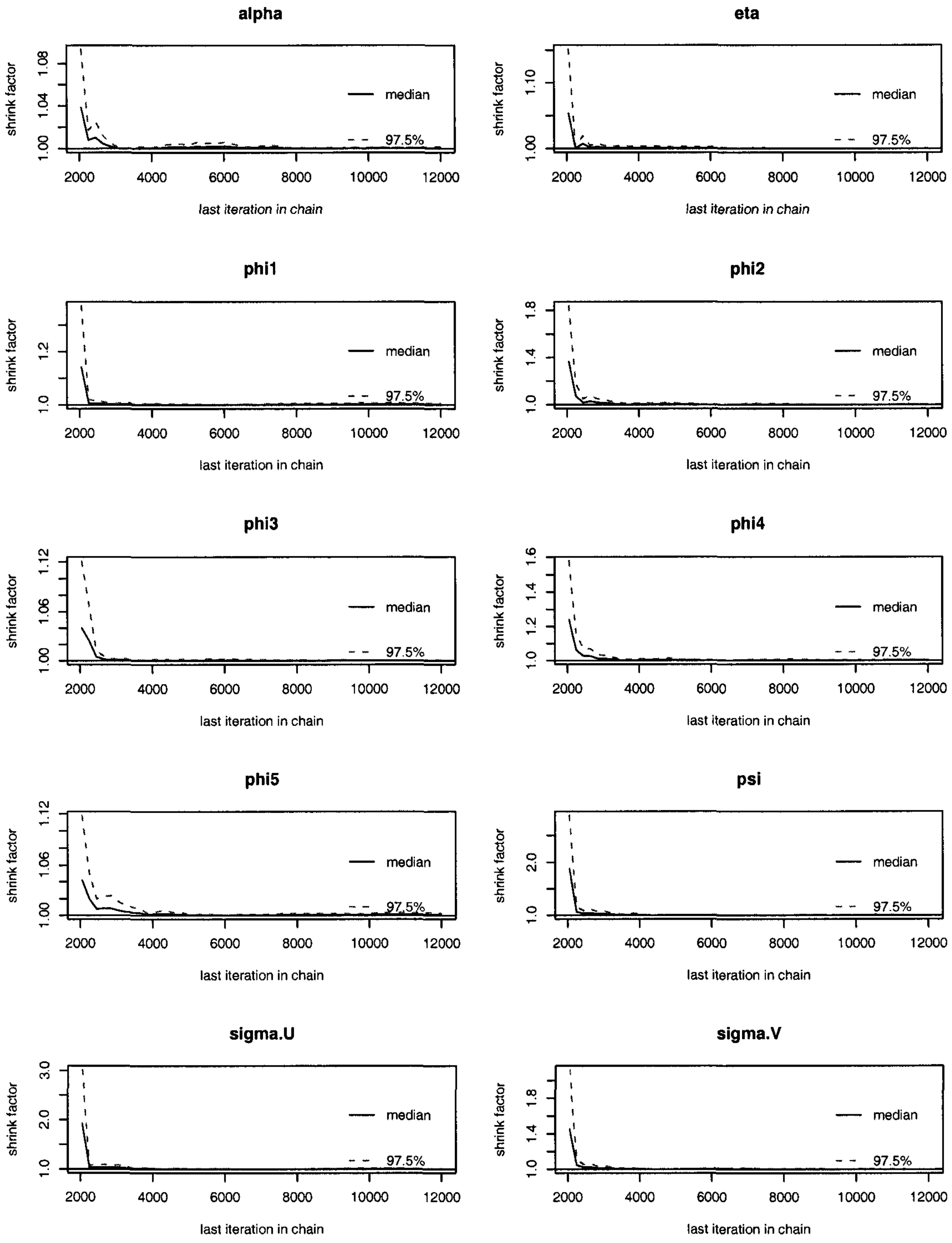

Figure 4.1: *Gelman Rubin Plots from five parallel chains. Convergence is suggested when the medians and the 97.5 percentiles approach 1*

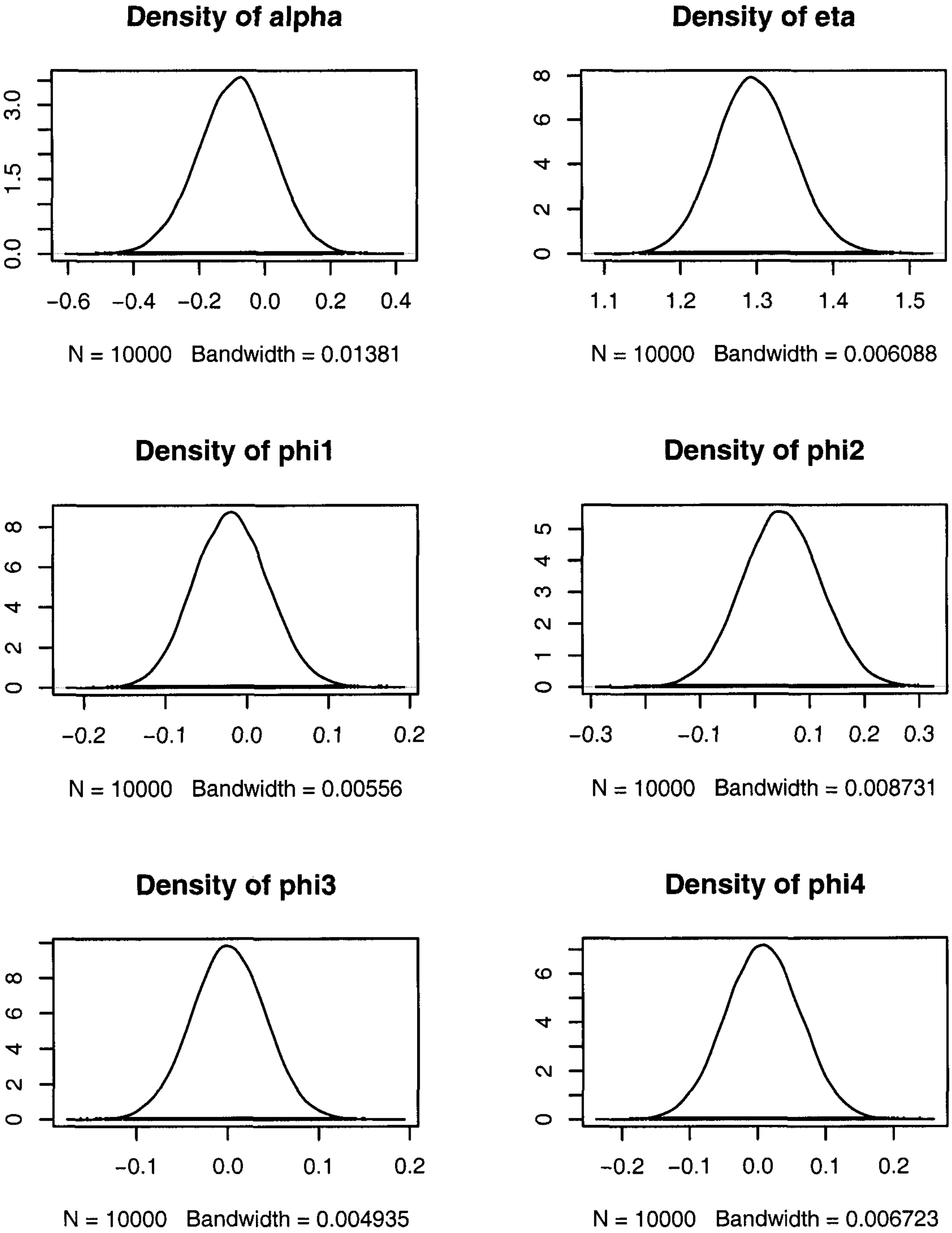

Figure 4.2: *Kernel density plots of sampled values for parameters of model 4 based on five pooled chains.*

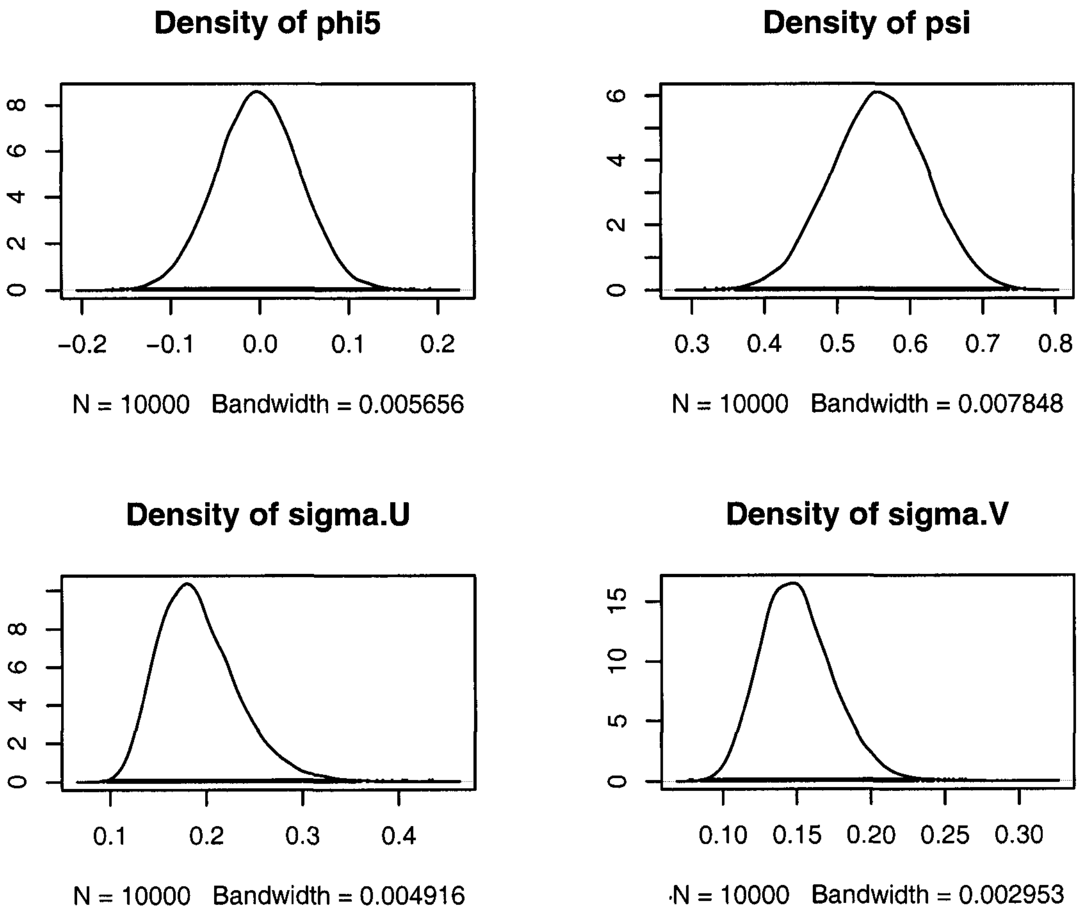

Figure 4.3: *Kernel density plots of sampled values for parameters of model 4 based on five pooled chains.*

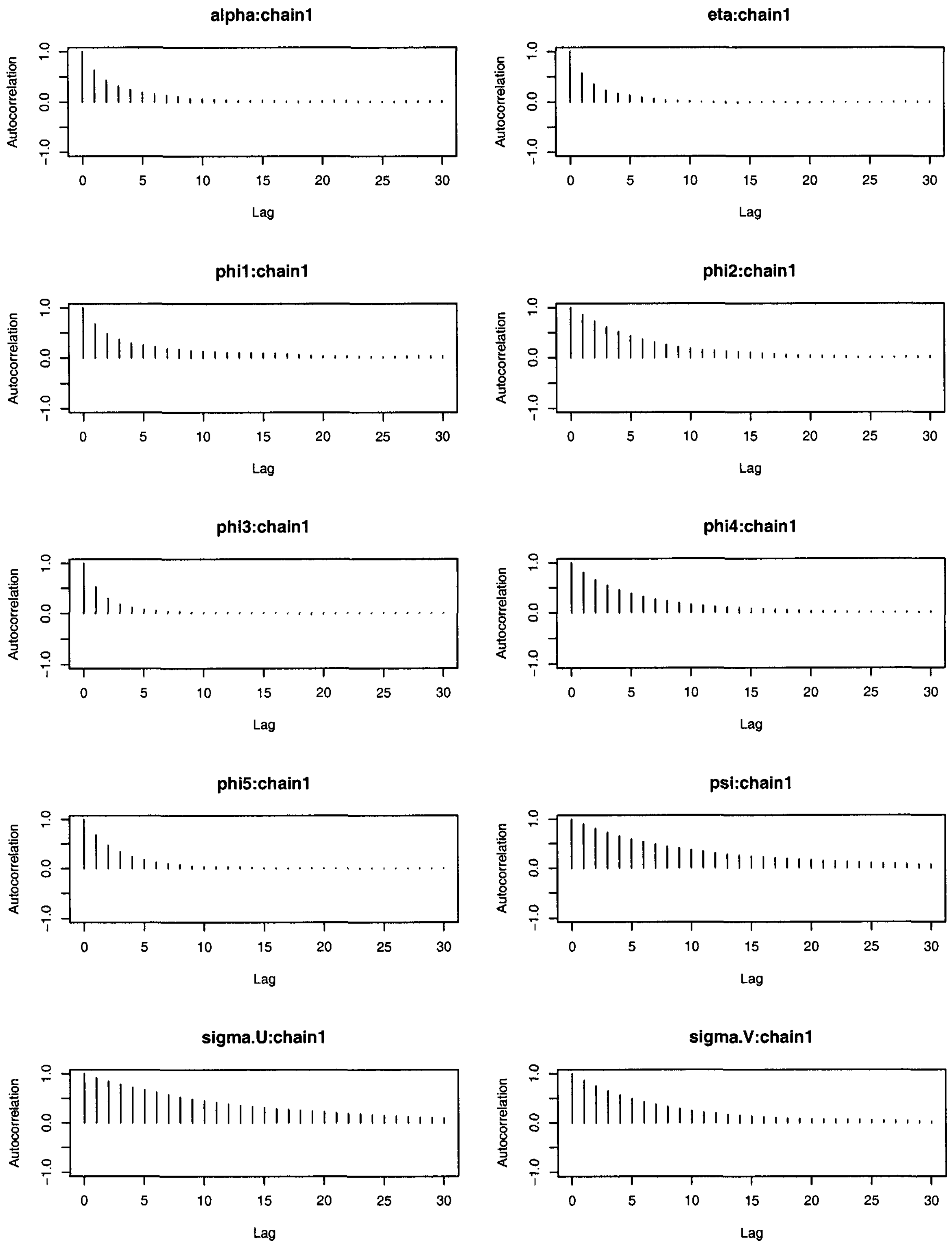

Figure 4.4: *Autocorrelation for chain 1*

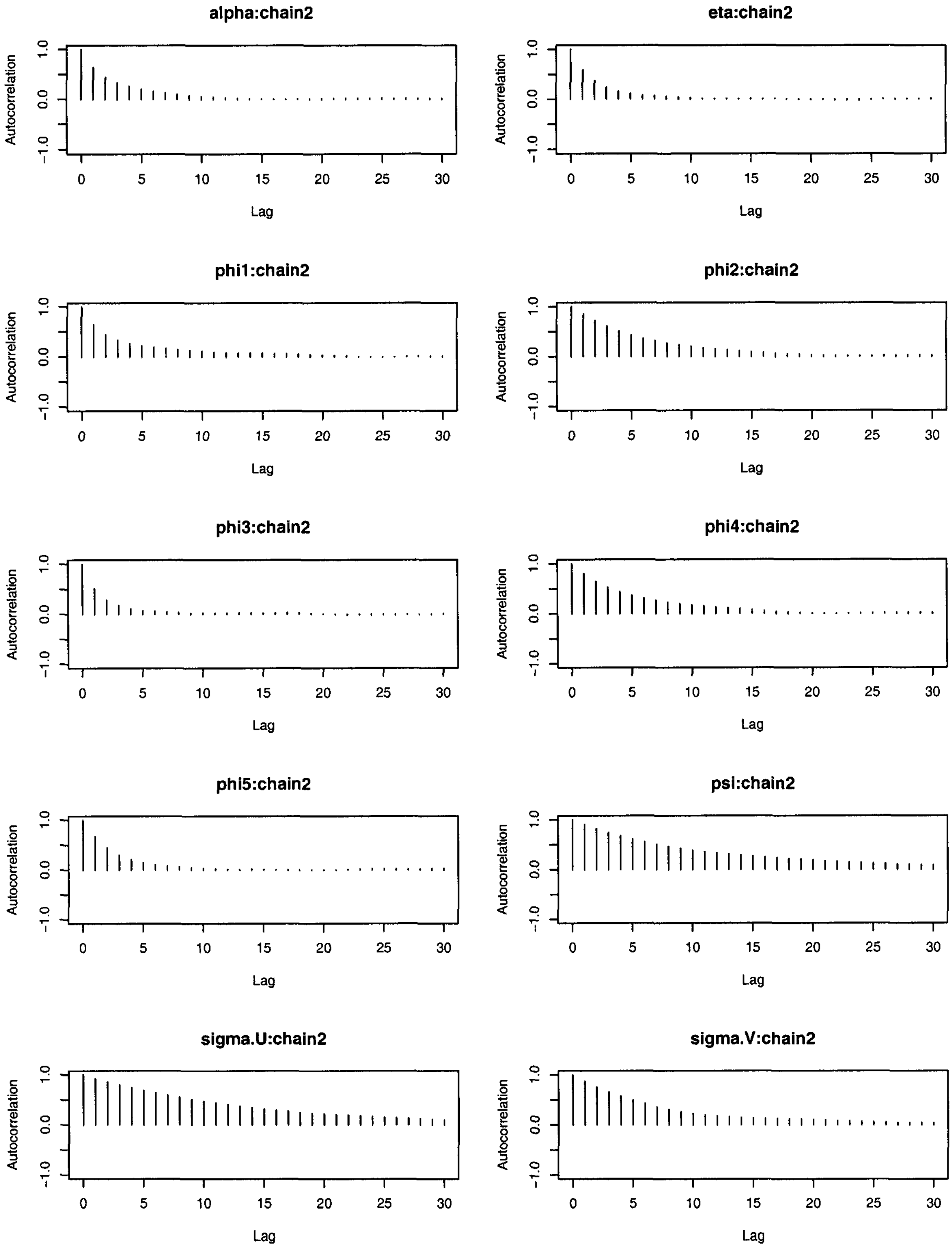

Figure 4.5: *Autocorrelation for chain 2*

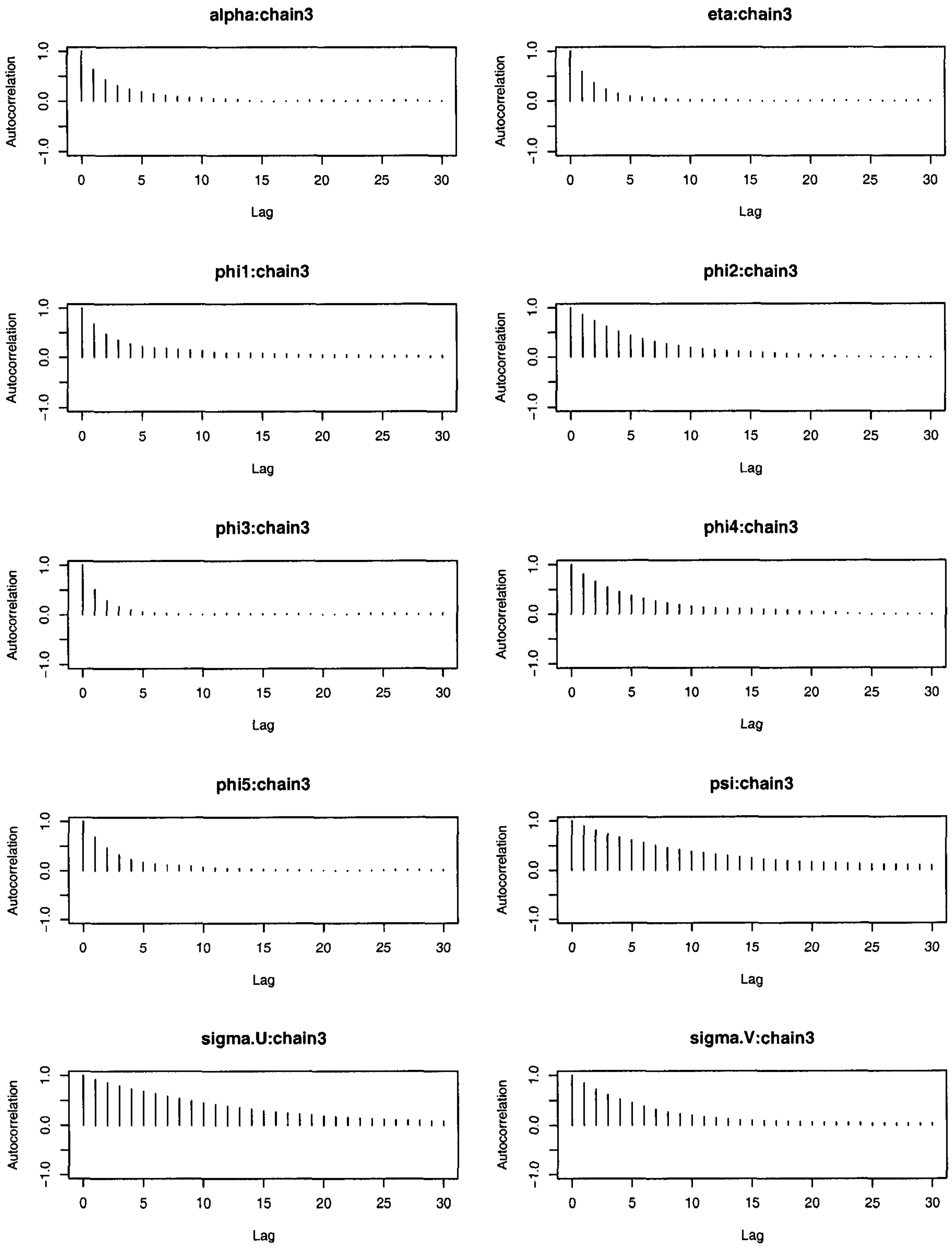

Figure 4.6: *Autocorrelation for chain 3*

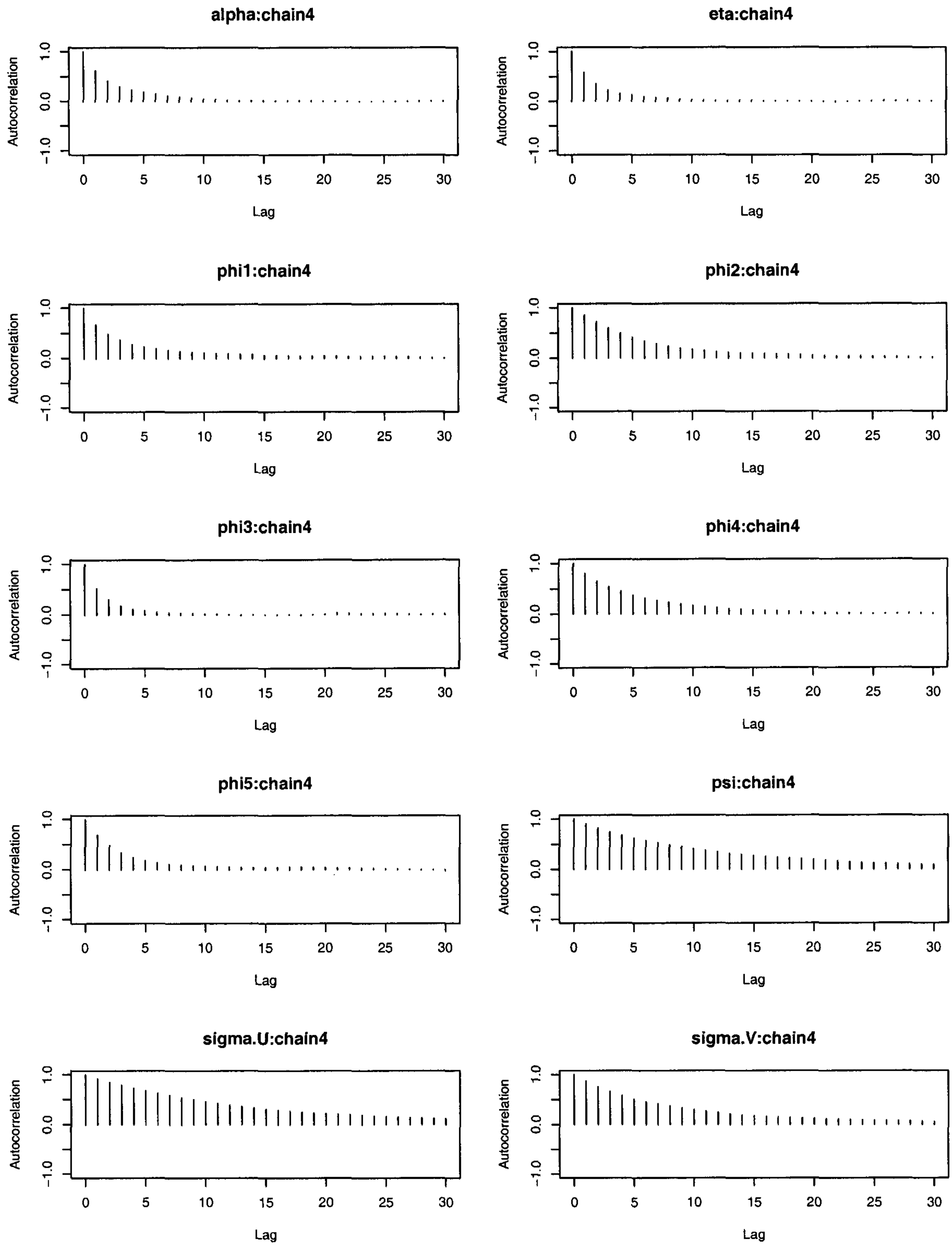

Figure 4.7: *Autocorrelation for chain 4*

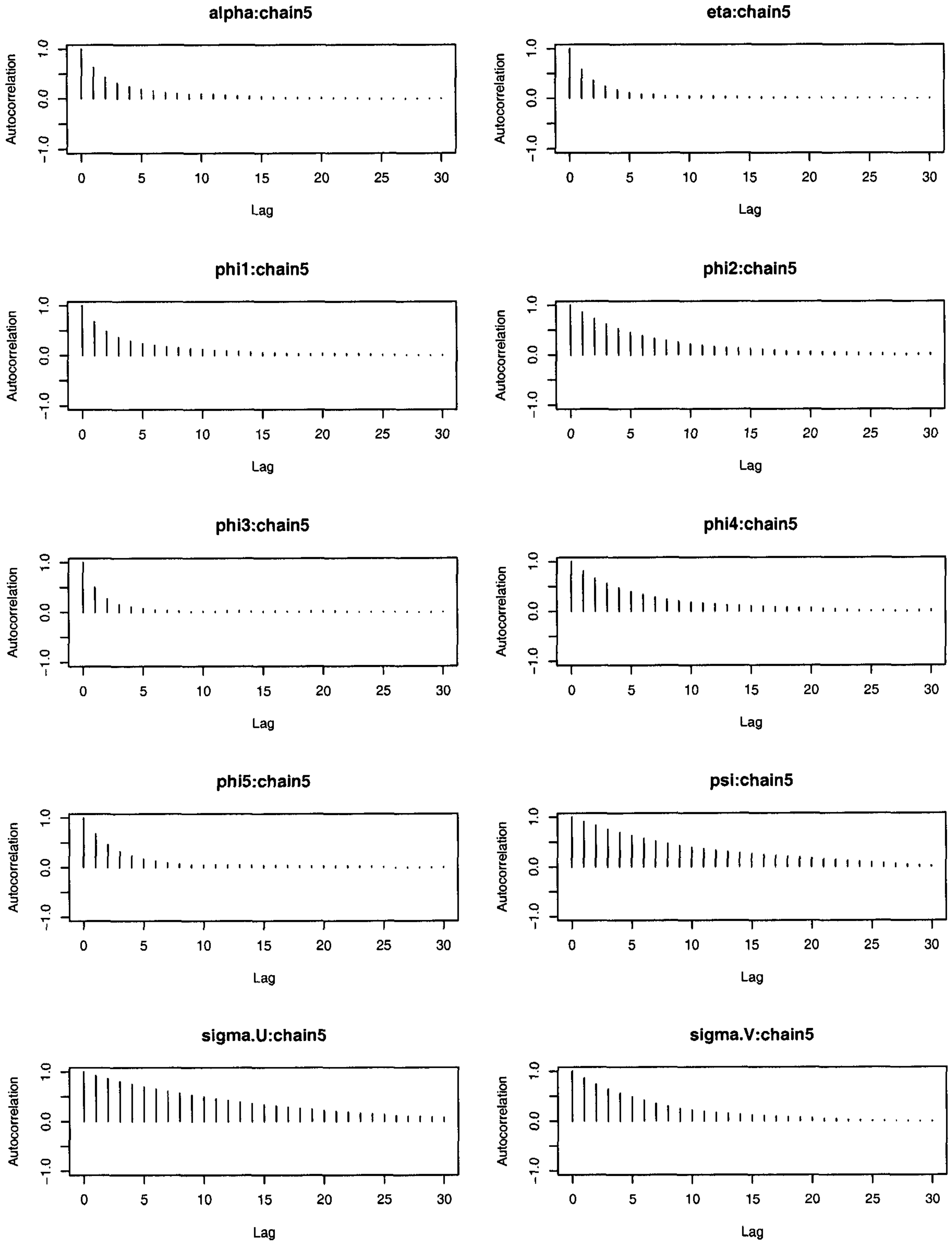

Figure 4.8: *Autocorrelation for chain 5*

Table 4.2: *Posterior median (95% credible interval) for parameters of each model and summaries of model fit (DIC) and complexity (p_D)*

Nodes	Model 1	Model 2	Model 3	Model 4
α	-	-0.097 (-0.326,0.120)	-	-0.087 (-0.317,0.130)
α_o	0.246 (0.188,0.305)	0.268 (0.193,0.343)	0.241 (0.182,0.300)	0.260 (0.1834,0.336)
ϕ_1	-	-	-0.019 (-0.108,0.070)	-0.019 (-0.107,0.072)
ϕ_2	-	-	-0.001 (-0.080,0.077)	0.001 (-0.079,0.080)
ϕ_3	-	-	0.051 (-0.091,0.195)	0.049 (-0.092,0.189)
ϕ_4	-	-	0.008 (-0.102,0.118)	0.008 (-0.101,0.116)
ϕ_5	-	-	-0.002 (-0.093,0.090)	-0.002 (-0.092,0.090)
ψ	0.557 (0.428,0.676)	0.559 (0.434,0.679)	0.555 (0.426,0.677)	0.558 (0.430,0.682)
η	1.279 (1.207,1.356)	1.307 (1.212,1.409)	1.272 (1.200,1.349)	1.297 (1.201,1.400)
σ_u	0.187 (0.125,0.281)	0.189 (0.127,0.283)	0.1847 (0.124,0.282)	0.187 (0.126,0.287)
σ_v	0.149 (0.109,0.204)	0.149 (0.110,0.204)	0.1486 (0.108,0.204)	0.149 (0.108,0.203)
DIC	727.934	728.672	732.164	734.653
p_D	38.419	39.208	42.915	41.094

autocorrelation indicates that the sample needs to be larger in order to fully explore the posterior distribution. It should be noted that low autocorrelation or absence of autocorrelation does not indicate convergence of the chains.

4.4 Bayesian analysis result

Using the prior distributions of the previous section, the analysis of incidence of preterm birth in the proximity of the Tar Ponds was carried out. The following models were fitted using the five area covariates available at all the 144 EDs and a measure of proximity (d_i):

1. model with no covariates which corresponds to the null model,

2. model with only distance measure alone,

3. model with only deprivation covariates alone, and

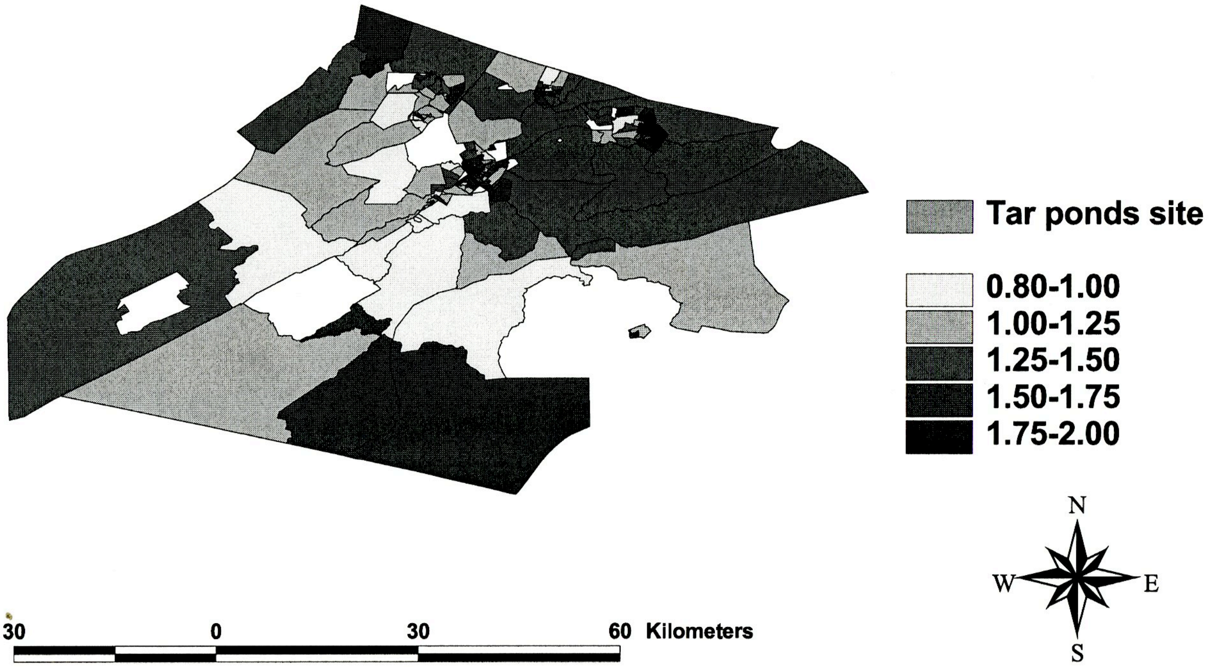

Figure 4.9: *Posterior mean of the relative risk of preterm births for model 4*

4. finally, model with distance and deprivation covariates.

The models were fitted using MCMC simulation method discussed earlier. Five separate chains starting from different initial values were run for each model. Convergence was assessed by visual examination of time series plots for each parameter and by carrying out the Gelman and Rubin diagnostic based on the ratio of between to within chain variances for each model. The time series plots with all the five chains superimposed were examined to see whether the chains were mixing well. Figure 4.1 shows the Gelman Rubin Plots with the "shrinking factor". This clearly shows that "shrinking factor" for each parameter approaches 1. Hence, all chains have escaped the influence of their starting points. Figure 4.2 and 4.3 show the posterior density of each parameter after convergence. The autocorrelation plot shown in Figures 4.4-4.8 show that autocorrelation decreases very fast from lag 1. All the plots were produced with the **coda** package for R (Plummer *et al.*, 2004). On this basis, the first 2000 samples of each chain were discarded as 'burn-in'; each chain was run for a further 10000 iterations, and posterior estimates were based on pooling the 5 × 10000 samples for

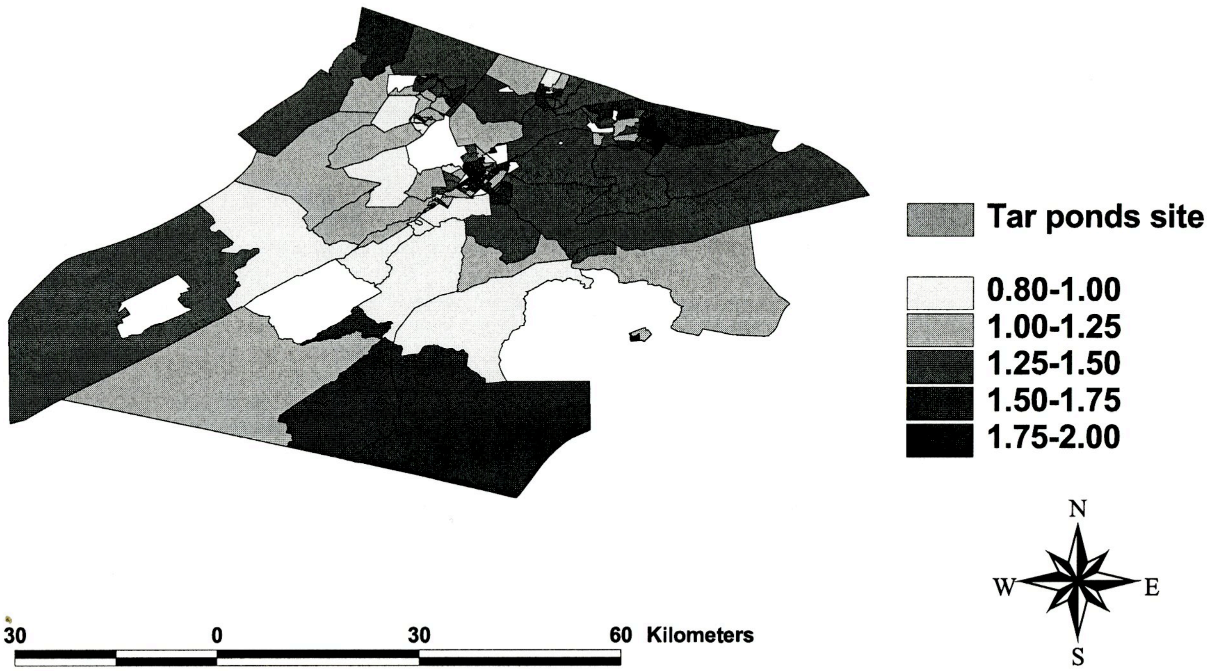

Figure 4.10: *Posterior median of the relative risk of preterm births for model 4*

each model. This gave Monte Carlo standard errors that are less than 1% of the posterior standard deviation for each parameter in the models.

Table 4.2 gives the summaries of the posterior distribution under each model. From Table 4.2, we can see that estimates of α in both models 2 and 4 is negative, and the 95% credible interval contain zero which is evidence that there is no increase in risk from source. The 95% credible interval for ϕ_i $(i = 1, \ldots, 5)$ in models 3 and 4 also contain zero which shows that the risk cannot be explained by any of the socio-economic covariates. For each of the models η which is a measure of the overall risk was found to be greater than 1 which is evidence that there is an increased risk of preterm birth in each of the enumeration districts compared to the rest of Canada.

The parameters, σ_u and σ_v only change slightly over the 4 models. Following Best *et al.* (1999), we defined a quantity $\psi = \sigma_u/(\sigma_u + \sigma_v)$ as a measure of the relative contribution of U_i and V_i to the total overdispersion. So that as $\psi \to 1$, spatial variation dominates, while as $\psi \to 0$, spatial variation becomes negligible. From Table 4.2, the 95% credible intervals

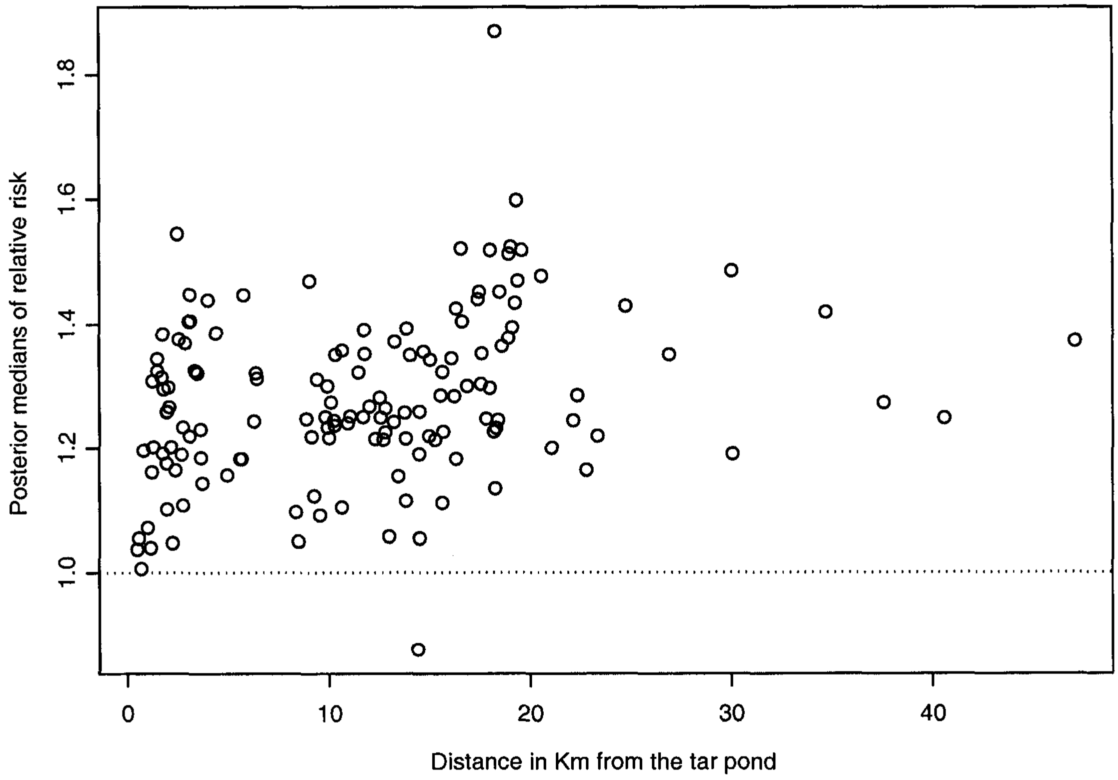

Figure 4.11: *Plot of the posterior medians of relative risk against distance from Tar Pond in km*

for ψ for each model contain 0.5. Hence, there is no clear evidence that the spatial structure dominates the random effect in any of the model.

Goodness of Fits

Spiegelhalter *et al.* (1998) proposed the use of Deviance Information Criterion (DIC) which consists of two terms, one is a measure of goodness of fit and the other is a penalty for increasing model complexity so that smaller values of DIC indicate a better-fitting model. From the result of Table 4.2, the DIC increases as more variables are added into the model. Hence, Model 1 is better than all the three other models.

Predicted Relative Risk

Finally, the posterior median and mean of the relative risk of preterm birth were plotted. These plots are shown in Figures 4.9 and 4.10. We can now compare this figures with the crude SIR plot in Figure 3.1. The plot shows that high relative risk of preterm birth in almost all the enumeration districts. However, the risk is not as high as shown in Figure 3.1. Also the posterior median was plotted against distance from the Tar Pond in Figure 4.11. There is no clear distance risk relationship.

Chapter 5

Frequentist Methods

5.1 Introduction

In this chapter, two frequentist methods will be used to fit the models in Chapter 4. First, Poisson models will be fitted using the quasi-likelihood approach. This method will be used to accommodate the expected over dispersion while excluding the spatial component of the data. Second, weighted linear regression will be fitted and the residuals of the fit will be tested for spatial autocorrelation. If these residuals exhibit spatial properties then a spatial linear regression will be fitted. We will expect the result to be very close because quasi-likelihood is a special case of weighted least squares (McCullagh and Nelder, 1989).

5.2 Poisson Regression

For $Y_i \sim \text{Poisson}(\mu_i)$, where $\mu_i = \lambda_i \text{E}_i$ $(i = 1, \ldots, n)$, we assume the generalized linear model (McCullagh and Nelder, 1989). Four models were fitted for the log relative risk ($\log \lambda_i = \log \mu_i - \log E_i$) in terms of a constant, area-level covariates (see Table 1.1) and the reciprocal of distance. The fitted models are:

$$\log \lambda_i = \alpha_o \tag{5.1}$$

$$\log \lambda_i = \alpha_o + \alpha/d_i \tag{5.2}$$

$$\log \lambda_i = \alpha_o + \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \phi_4 x_4 + \phi_5 x_5 \tag{5.3}$$

$$\log \lambda_i = \alpha_o + \alpha/d_i + \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \phi_4 x_4 + \phi_5 x_5 \qquad (5.4)$$

These are the same model as used in Chapter 4 except no random effects or spatial effects. This is also called a log-linear model, because the log of the mean is assumed to be a linear function of covariates. In each of the fitted models, $\log \mathrm{E}_i$ is used as an offset to account for variations in λ_i over the study region. An offset is a covariate in linear predictor whose coefficient is not estimated, but assumed to be equal to 1. The results of the fit obtained by using the quasi-likelihood approach in SAS package are summarized in Table 5.1.

The quasi-likelihood approach is used to account for the overdispersion that might occur in the data set. This is implemented by specifying $E(Y_i|\lambda_i) = \mu_i$ and $\mathrm{Var}(Y_i|\lambda_i) = \kappa\mu_i$ and estimating κ, using a hierarchical modelling approach with the assumption that λ_i are random variables from a probability distribution. Where κ is the dispersion parameter with value greater than 1 for overdispersion. The conventional estimate of κ is the mean Pearson χ^2 statistic. We have explained the use of the Pearson χ^2 for goodness-of-fit in detecting heterogeneity of relative risk in section 2.3.1. This statistic compares the fit of the current model to that of a saturated model and it is defined by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

Hence, $\hat{\kappa} = \chi^2/(N - p)$ where N and p respectively denotes the number of observation (length of Y_i) and parameters in the model. For each of the fitted model κ was estimated to be approximately equal to 1 (see Table 5.1), a condition that shows that there is no evidence of overdispersion.

5.2.1 Analysis

The results of all the four models are displayed in Table 5.1. The Wald confidence intervals shown in Table 5.1 are based on the asymptotic normality of the parameter estimators. They are sometimes called the normal confidence intervals. The 95% Wald interval for any unknown parameter (θ) is given by $\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}$. Where $\hat{\theta}$ is the maximum likelihood estimate of θ and $\hat{\sigma}_{\hat{\theta}}$ is the standard error estimate of $\hat{\theta}$. From the table, we can see that the estimated α in both model 2 and 4 is negative, and the 95% Wald confidence intervals contain zero which is evidence that there is no increase in risk from source.

Table 5.1: *Parameter estimates (95% Wald C.I.), residual deviance, and over-dispersion parameter*

parameter	Model 1	Model 2	Model 3	Model4
α	-	-0.0878(-0.2519,0.0763)	-	-0.075 (-0.239,0.089)
α_o	0.2520	0.2707 (0.2111,0.3303)	0.2163 (-0.3427,0.7753)	0.226 (-0.334,0.785)
ϕ_1	-	-	-0.0034 (-0.0103,0.0035)	-0.003 (-0.010,0.004)
ϕ_2	-	-	-0.0008 (-0.0099,0.0083)	-0.0005 (-0.0096,0.0086)
ϕ_3	-	-	0.0115 (-0.0074,0.0305)	0.011 (-0.008,0.030)
ϕ_4	-	-	0.0007 (-0.0128,0.0142)	0.0006 (-0.0129,0.0141)
ϕ_5	-	-	-0.0011 (-0.0079,0.0057)	-0.0010 (-0.0078,0.0058)
Deviance	132	130.56	122.9983	122.18
Df	143	142	138	137
κ	0.99	0.9942	0.9887	0.9906

The 95% confidence intervals for ϕ_i $(i = 1, \ldots, 5)$ in models 3 and 4 also contain zero which shows that the covariates are not significant factors in risk of preterm birth. This result is confirmed by the Wald Chi-square test, the square ratio of each, parameter estimate divided by its standard error is a measure of the individual effects in the fitted models. The results of the test are given in Table 5.2. This result shows that none of the variables has significant contributions to the explanation of the variation in risk. Now combining equations (5.1) and (2.11), we have

$$\log \lambda_i = \log \eta$$

therefore $\eta = \lambda_i = \mu_i/E_i$. This is referred to as the overall mean of the relative risk. For each of the models, Table 5.3 gives the estimates of the overall risk together with its 95% confidence intervals. The overall mean of the relative risk is greater than 1 for each model which indicates that there is elevated risk of preterm birth across the whole of Cape Breton Municipality.

Table 5.2: *Type III (Wald) Tests*

Models	Effect	DF	Wald χ^2	Pr $>$ χ^2
Model 2	d_i	1	1.0993	0.2944
Model 3	x_1	1	0.9435	0.3314
	x_2	1	0.0310	0.8603
	x_3	1	1.4209	0.2333
	x_4	1	0.0093	0.9232
	x_5	1	0.0971	0.7554
Model 4	x_1	1	0.9306	0.3347
	x_2	1	0.0110	0.9164
	x_3	1	1.3270	0.2493
	x_4	1	0.0083	0.9274
	x_5	1	0.0802	0.7771
	d_i	1	0.8058	0.3694

Table 5.3: *Overall mean of the relative risk (η) and its 95 % Confidence intervals*

Parameter	Model 1	Model 2	Model 3	Model 4
η	1.287 (1.225, 1.351)	1.311 (1.235, 1.391)	1.241 (0.710, 2.171)	1.254 (0.716, 2.192)

Table 5.4: *Summary of* Y_{132}. x_i *are in percents*

Y_i	N_i	$\hat{\lambda}$	d_i	x_1	x_2	x_3	x_4	x_5
16	382	0.59	14.37	55	24.26	9.66	3.27	11.29

5.2.2 Description of Plots

For each model, the following plots were produced in S-plus package: Deviance residuals versus fitted values; observed counts versus fitted values; predicted values versus the square roots of the absolute values of the deviance residuals; and Pearson Residuals of the fitted model versus Quantiles of Standard normal. The plots are shown in figures (5.1)–(5.4).

Deviance Residuals Plots

Deviance residuals is a measure of fit in a generalized linear model. They are defined as

$$r_D = \text{sgn}(y_i - \mu_i)\sqrt{\delta_i},$$

where δ_i is the contribution of the ith observation to the deviance. Hence, r_D increase (or decreases) with $y_i - \mu_i$. For the Poisson distribution,

$$\delta_i = 2(y_i \ln(y_i/\mu_i) - y_i + \mu_i).$$

This residuals are useful for detecting observation(s) that are having undue effects on the fitted models. A look at the plots of Deviance Residuals versus fitted for each model shows no systematic trend except for one observation, y_{132}, that is far away from the rest.

Other Plots

With the exception of y_{132}, the plot of observed counts versus fitted values for each model did not show any great departure from the model. Pearson Residuals of the fitted model versus Quantiles of Standard normal for each model does not show any instability. It should be noted that observation, y_{132}, needs to be investigated. The observed value of y_{132} is 16 and the minimum fitted values of 30.63 is almost twice the observed. The summary of y_{132} is given in Table 5.4 This shows that Y_{132} is 14.37 km away from the Tar Ponds and has a low relative risk of preterm birth (SIR = 0.59). It has a high rate of unemployment to

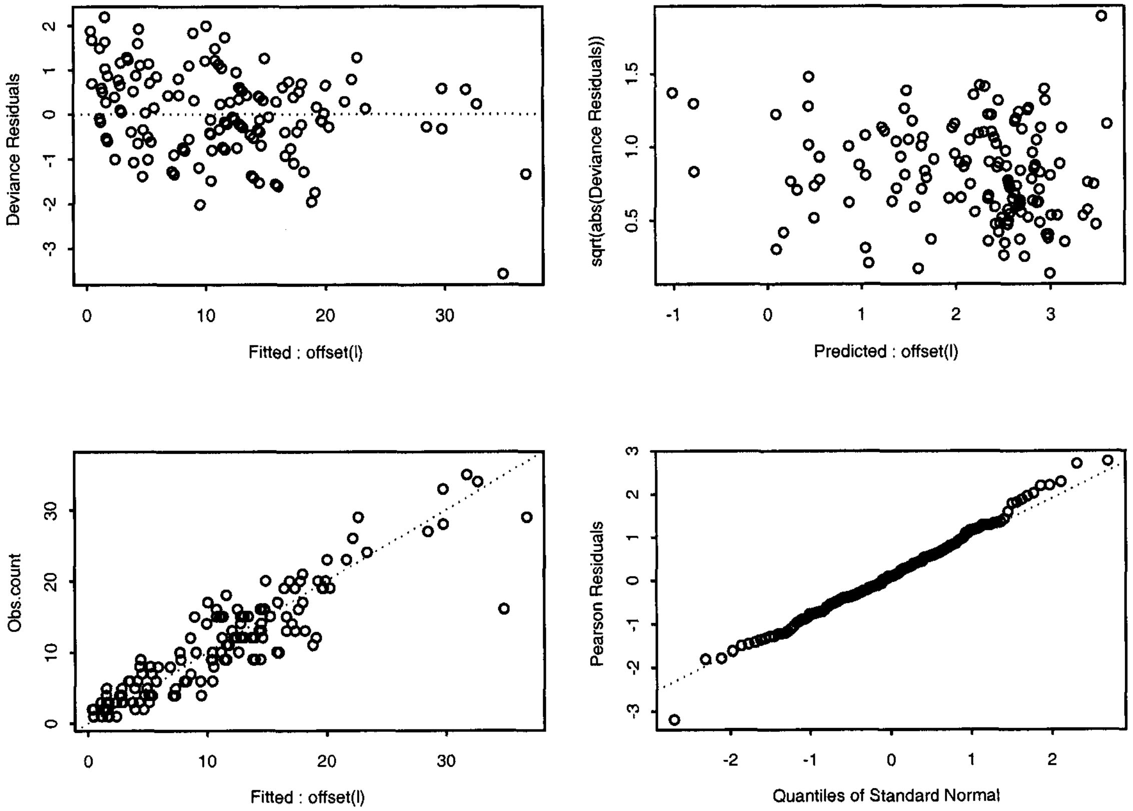

Figure 5.1: *Diagnostic plots for model 1*

population and a high proportion of persons who are separated, divorced or widowed. This ED also has one of the lowest proportion of persons with no high school and proportion of people living alone.

5.3 Weighted Linear Regression

Here we have fitted a modified version of Model 4, equation (5.4) using the weighted regression approach. This was done to account for the dispersion that might result from the violation of the constant variance assumption in the least squares approach. The weight (w_i) was set equal to $E_i / \sum_{i=1}^{n} E_i$ and λ_i was replaced by the SIR ($\hat{\lambda}_i = Y_i/E_i$) so that the error

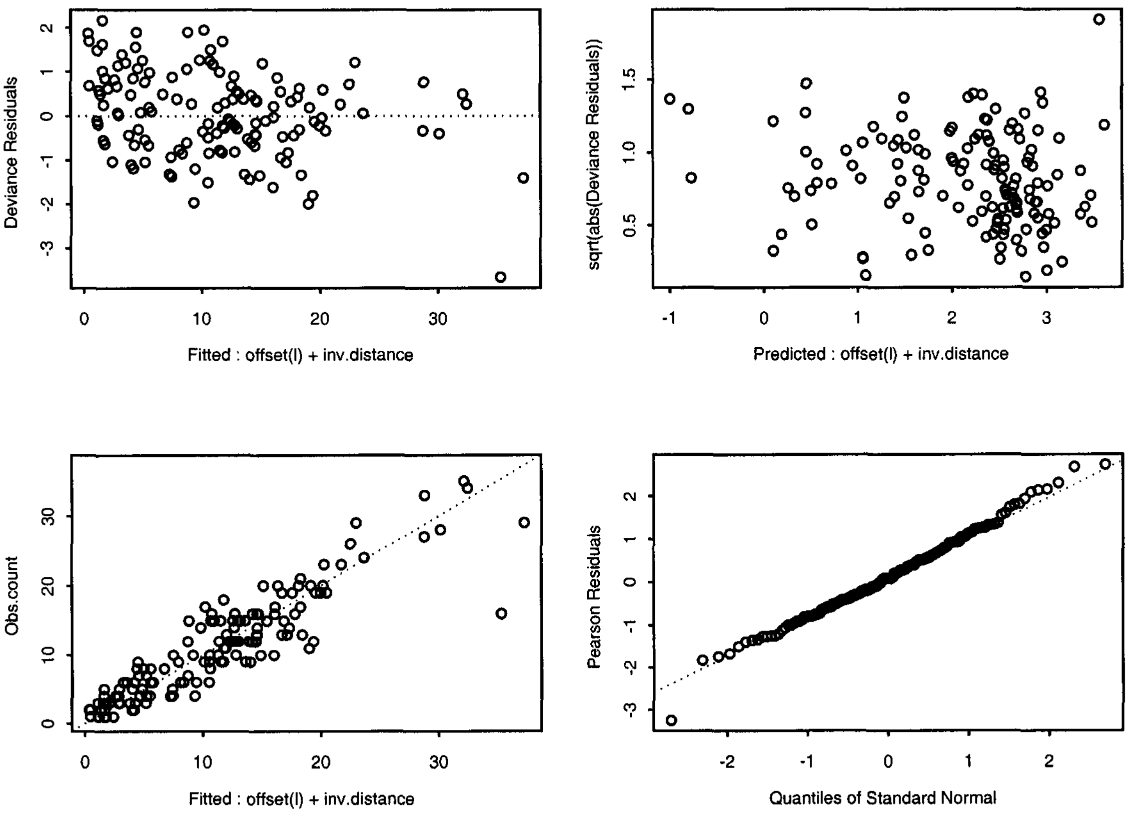

Figure 5.2: *Diagnostic plots for model 2*

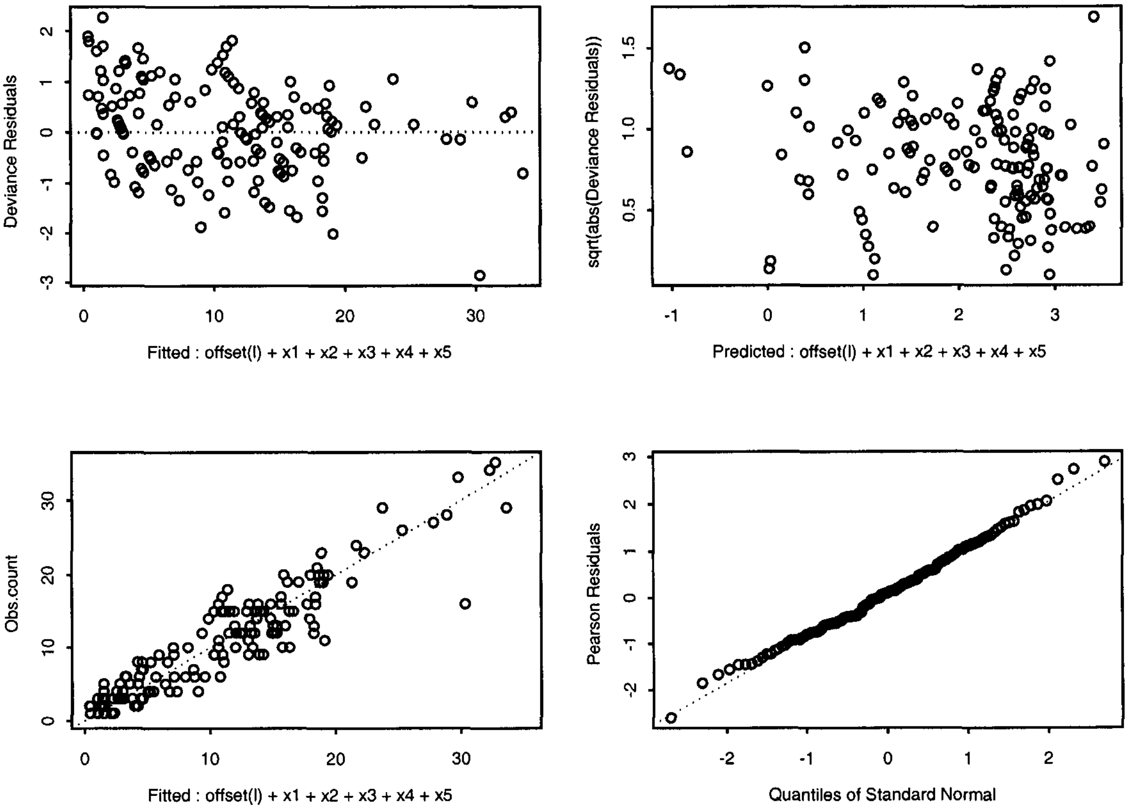

Figure 5.3: *Diagnostic plots for model 3*

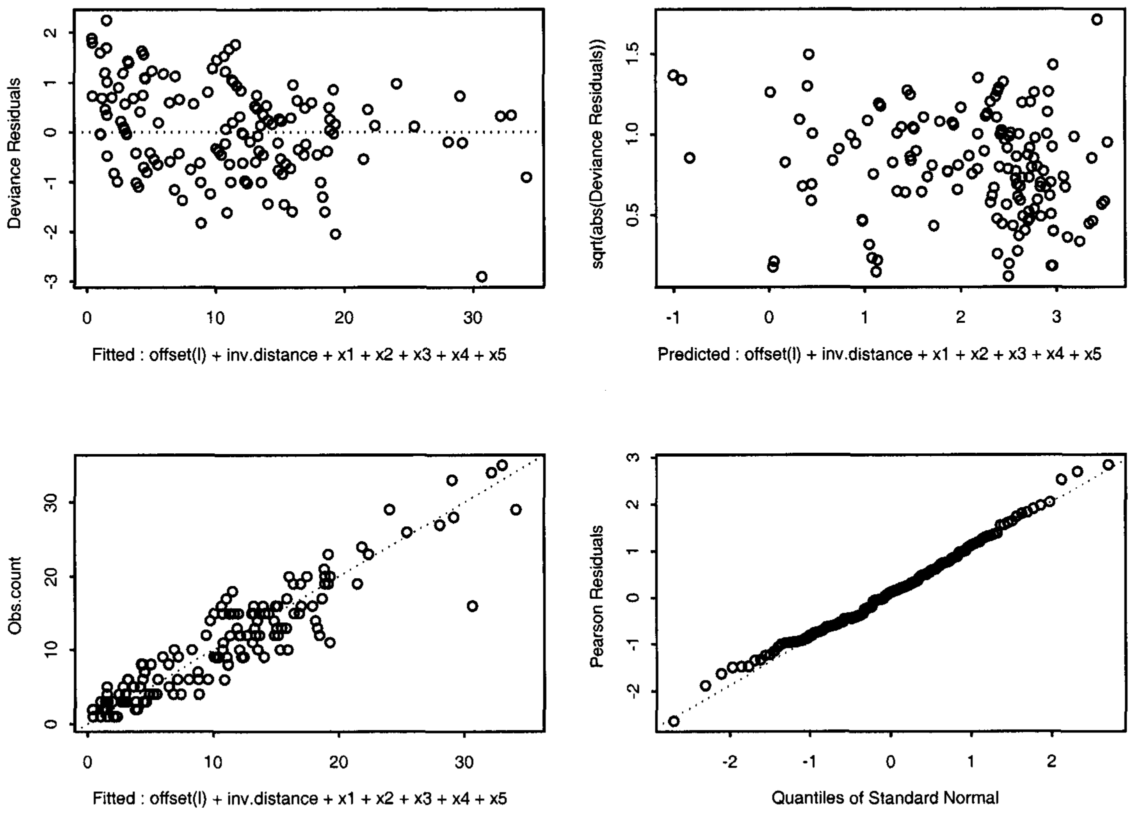

Figure 5.4: *Diagnostic plots for model 4*

Table 5.5: *Weighted regression result*

Parameters	Value	Std. Error	t value	Pr(> $\|t\|$)
α_o	0.2180	0.2708	0.8052	0.4221
α	-0.0878	0.0758	-1.1582	0.2488
ϕ_1	-0.0045	0.0033	-1.3817	0.1693
ϕ_2	-0.0001	0.0044	-0.0252	0.9800
ϕ_3	0.0106	0.0094	1.1330	0.2592
ϕ_4	0.0025	0.0067	0.3751	0.7081
ϕ_5	-0.0011	0.0034	-0.3194	0.7499

sum of squares (Q) of the weighted linear regression can be written as

$$Q = \sum_{1=1}^{n} w_i\{\log \hat{\lambda}_i - (\alpha_o + \alpha/d_i + \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \phi_4 x_4 + \phi_5 x_5)\}^2.$$

Here, we have not included the spatial component of the model because, we have seen in Chapter 3 that the SIR does not exhibit spatial dependency.

The result of the fit is given in Table 5.5. From the t-value and the associated p-value, it appears that none of the variables is significant in the explanation of increase risk of preterm birth. The residual standard error of the model was estimated to be 0.02347 on 137 degrees of freedom. Multiple R-Square is 0.09795 which shows that the variables in the model are only able to explain less than 10% of the total variation in the risk. The F-statistic for the regression relationship was estimated to be 2.479 on 6 over 137 degrees of freedom and the associated p-value is 0.0262. This shows that at least one of the parameters (α, and ϕ_i) does not equal zero. Hence, there is an existence of a regression relationship between the dependent variable (Y_i) and the Independent variables (X_i).

5.3.1 Weighted Regression Diagnostic Plots

The diagnostic plot are shown in Figure 5.5. The residual plots (first row, first plot) does not show any obvious trend. Three observations are identified as outliers. These are Y_{141}, Y_{64} and Y_{62}. The plot of residuals versus quantiles of standard normal (second row, first plot) shows a slight deviation from normality but not sufficient to reject the assumption of

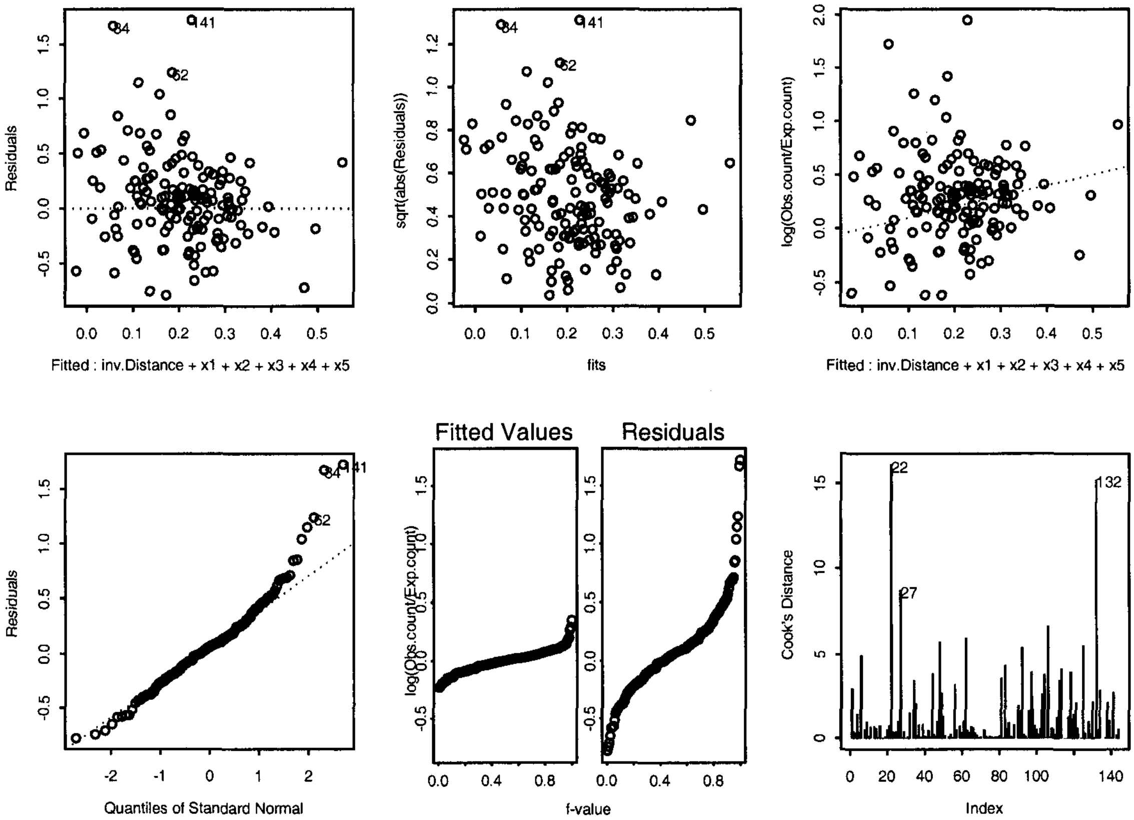

Figure 5.5: *Weighted Regression Diagnostic plots*

normality.

Residual and fit spread plot (Second row, second plot) shows some weakness in the model because the spread of the residual is greater than the spread of the fitted values. We actually expect the opposite to happen, if the model is fitting perfectly well. The Cook's distance plot (second row, last plot) shows that the three observations (Y_{22}, Y_{27} and Y_{132}) are having great influence on the regression coefficient. It should be noted here that observation Y_{132} was also identified as outlier in the quasi-likelihood Poisson model fitted earlier.

5.3.2 Test for Autocorrelation

Next, Moran's I test was also carried out to examine whether there is spatial autocorrelation in the residuals. The result gave a correlation of -0.01628, variance of 0.002541 and standard error of 0.05041. In addition, the normal test statistic was -0.1843 with associated 2-sided p-value equal to 0.8538. These results are sufficient to conclude that there is no spatial autocorrelation in the residuals. Hence, there was no need to use spatial regression modelling.

5.3.3 Where there is Autocorrelation

In practice, a typical spatial regression modelling will start with the examination of the dependent variable for spatial dependency. This can be done with Moran's I statistics or Geary C statistics. If there is no spatial pattern, then ordinary least square or weighted least square is sufficient to model the data.

On the other hand if the dependent variable shows spatial patterns. Then, the first order spatial pattern can be incorporated at the beginning of the modelling using adjacency matrix described in Chapter 4. The major question is: what will happen if a spatial modelling was carried out when in fact there was no justification? We actually carried it out. It produced a different result but what is actually interesting is that when test of autocorrelation was carried out on the residual after fitting the model, it produced the following results: correlation of 0.02341, variance of 0.002541 and standard error of 0.05041. In addition, the normal test statistic was 0.6031 with associated 2-sided p-value equal to 0.5465. Comparing this with our last result shows that the correlation only shifted and the p-value reduced.

This shows that great care has be taken when using spatial modelling.

Chapter 6

Conclusion

This research is part of a big project done to assess the effect of maternal proximity to the hazardous waste from the Sydney Tar Pond, Nova Scotia. Two question have been addressed in this project:

1. Is maternal proximity to hazardous waste and pollution from the Sydney Tar Pond sites associated with increased risk of preterm birth?
2. How much of the variation in risk of preterm birth can be explained by socioeconomic inequalities across the study region?

In addressing these questions frequentist and Bayesian methods were employed. In the frequentist approach, Poisson regression for aggregated data and weighted least squares were fitted using distance from the Tar Pond and the following area specific-covariates: the proportion of persons who have no high school diploma; the rate of unemployment to population; the proportion of persons who are separated; divorced or widowed; the proportion of single-parent families; and the proportion of people living alone. The same models were fitted using a Bayesian Hierarchical modelling incorporating both structured and unstructured random effects to account for model overdispersion.

Our intention was to combine all of the area covariates to form the deprivation index, but income data was not available in 14 of the 144 enumeration districts included in the study. So the effect of each variable was assessed independently. The overall estimate of relative risk of preterm birth was found to be greater that 1 for almost all the enumeration districts.

Also, none of the area covariates in the model is significant in the explanation of the risk of preterm births.

There was no evidence of any decrease in risk as we move away from the Tar Pond. The result of both the weighted least square and the quasi-likelihood Poisson regression agrees with the result from the Bayesian Hierarchical modelling which incorporates the spatial effects. The result of the Bayesian modelling shows that there is no significant spatial association of risk in the area studied. There was no obvious cluster of outcome around the Tar Pond significant enough to explain an association between maternal proximity to the Sydney Tar Ponds and risk of preterm birth.

6.1 Threats to Internal Validity

The following are some of the limitations of this research

- Data are not available for 14 of the Enumeration districts. Hence, they are omitted from our analysis but the effects of this on spatial dependency or our conclusion are not known;
- We have based our analysis on the 1996 data but we do not have any evidence of whether the exposure from the Tar Pond has decreased before 1996;
- The problem of imprecise geographical matching and data aggregation may have created a source of bias during data collection;
- Ecological bias which can occur due to the differences between individual and group-level estimates of disease risk. This is a major limitation of all studies based on aggregated data;
- Under-ascertainment/duplication of cases may have occurred; and
- Migration of women between exposure and pregnancy outcome may be a source of bias which may lead to underestimation of the risk.

6.2 External Validity

The methodology can be generalized but the result may not be easy to generalize. This is because landfill sites differ enormously in the conditions that render them hazardous; and conditions that determine the exposure to and resulting health risks posed by any waste site are likely to be unique to that particular site. Hence, the results of this study are not intended for direct use in decision-making with respects to other landfill sites. Rather they are to serve as a guide.

6.3 Ethical Considerations

Unlike observational and experimental studies where human beings are involved, as subjects of study, this study only makes use of aggregated data. Aggregated data by their nature do not reveal the identities of individuals involved. Therefore, confidentiality of the cases involved is automatically guaranteed.

6.4 Future Research Plans

The future plans are:

- To aggregate the data for up to ten years and model using other forms of $g(d;\theta)$
- To work more on the statistical properties of most of the estimators used in the cluster analysis of outcomes
- Further research in this area is needed to improve our understanding of the impact of social factors, fear and risk perceptions on both actual and perceived ill health by people living in the vicinity of waste sites. The use of mixed model, incorporating both qualitative and quantitative methods may be very good approach for future studies.
- There is an elevated risk of preterm births, which appears to be uniform across the whole of Cape Breton regional municipality as shown by all the methods used. This shows that the pollution may be occurring at a wider scale and overtime may have

affected the ability to differentiate the EDs in terms of amount of exposure. A direct comparison of Cape Breton regional municipality with other close municipalities may help answer some of the remaining questions.

BIBLIOGRAPHY

Baibergenova, A., Kudyakov, R., Zdeb, M. and Carpenter, D.O., (2003). Low birth weight and residential proximity to PCB-contaminated waste sites, *Environmental Health Perspectives*, **111**, 1352–1357.

Baker, D. B., Greenland, S., Mendlein, J. and Harmon, P. (1988). A health study of two communities near the Stringfellow waste disposal site. *Archives of Environmental Health* **43**: 325–334.

Beck, U. (1992). *Risk Society: Towards a New Modernity*, London: Sage.

Berry, M., and Bove, F., (1997). Birth weight reduction associated with residence near a hazardous waste landfill. *Environmental Health Perspectives* **105**, 856–861.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems . *Journal of Royal Statistical Society*, Series B, **36**, 192–236.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733–746

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases . *Journal of the Royal Statistical Society, Series A*, **154**, 143–55

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the institute of Statistics and Mathematics*, **43**, 1-59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian Models for Spatially Correlated Disease and Exposure Data. In *Bayesian Statistics*

6 (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds.), New York: Oxford University Press, pp 131– 156.

Bithell, J. F.(1995). The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, **14**, 2309-22

Bithell, J. F., and Stone, R. A. (1989). On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *Journal of Epidemiology and Community Health*, **43**, 79-85

Carstairs, V. and Morris, R. (1991). *Deprivation and Health in Scotland.* Aberdeen University Press, UK.

Clayton, D.G. and Kaldor, J.(1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671–82

Cressie, N. A. C. (1993). *Statistics for spatial data.* Wiley, New York.

Datta, G., Ghosh, M. and Waller, L. A. (2000). Hierarchical and Empirical Bayes Methods for Environmental Risk Assessment. In *Handbook of Statistics* (P.K. Sen and C.R. Rao, eds.), **18**, 223–245, Elsevier Science B.V.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomena in the vicinity of a prespecified point. *Journal of the Royal Statistical Society*, series A, **153**, 340–362.

Diggle, Morris, S., Elliot, P. and Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society.* Series A, **160**, 491–505.

Dodds L. and Sevoiur R. (2001). Congenital anomalies and other birth outcomes among infants born to women living near a hazardous waste site in Sydney, Nova Scotia. *Canadian Journal of Public health* **92**, 331-334

Dolk, H., Shaddick, G., Walls, P., and Thakrar, B. (1997). Cancer incidence near radio and television transmitters in great Britain : All high power transmitters. *American Journal of Epidemiology*, **145**, 10–17

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianchi, F., Garne, E., Nelen, V., Robert, E., Scott, J. E .S., Stone, D., and Tenconi, R., (1998). Risk of Congenital anomalies near hazardous waste landfill sites in Europe: the EUROHAZCON study. *The Lancet* **352**, 423– 427.

Elliot, P., Briggs, D., Morris, S., de Hoogh, C., Hurt, C., Jensen, T.K., Maitland, I., Richardson, S., Wakefield, J. and Jarup, L. (2001). Risk of adverse birth outcomes in populations living near landfill. *British Medical Journal* **323**, 363–68.

Elliot, P., Cuzick, J., English, D. and Stern, R. (1992). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press.

Elliot, P., Shaddick, G., Kleinschmidit, I., Jolley, D., Walls, P., Beresford, J. and Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, **73**, 702–770.

Fielder, H. M. P., Poon-King, C. M., Palmer, S, R., Moss, N. and Coleman, G. (2000). Assessments of impact on health of residents living near the Nant-y-Gwyddon landfill site: retrospective analysis. *British Medical Journal* **320**,19–23.

Gelman, A. and Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Geschwind, S. A., Stolwijk, J. A. J., Bracken, M., Fitzgerald, E., Stark, A., Olsen, C. and Melius, J., 1992. Risk of congenital malformations associated with proximity to hazardous waste sites. *American Journal of Epidemiology* **135**, 1197–1207.

Giddens, A. (1991). *Modernity and Self-Identity: Self and Society in the Late Modernity Age*, Cambridge: Polity.

Gilbertson, M. and Brophy, J., 2001. Community health profile of Windsor, Ontario, Canada: Anatomy of a Great Lakes Area of Concern. *Environmental Health Perspectives* **109**(suppl 6), 827–43.

Gilks, W. R., Richardson, S., and Spiegelhater, D. J., (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S.Richardson and D.J. Spiegelhalter), London: Chapman & Hall, pp. 1–17

Goldberg, M. S., Goulet, L., Riberdy, H. and Bonvalot , Y., (1995). Low birth weight and preterm births among infants born to women living near a municipal solid waste landfill site in Montreal, Quebec. *Environmental Research* **69**, 37–50.

Goldman, L. R., Paigen, B., Magnant, M. M. and Highland, J. H. (1985). Low birth weight, prematurity and birth defects in children living near the hazardous waste site, Love Canal. *Hazardous Waste and Hazardous Materials* **2**, 209–223.

Jolley, D., Jarman, B., and Elliot, P.(1992). Socio-economic Confounding. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies* (P. Elliot, J. Cuzick, D. English and R. Stern,eds.), New York: Oxford press, pp 115-124.

Kharrazi, M., von Behren, J., Smith, M., Lomas, T., Armstrong, M., Broadwin, R., Blake, E., Mclaughin, B., Worstell, G. and Goldman, L. (1997). A community-based study of adverse pregnancy outcomes near a large hazardous waste landfill in California. *Toxicology and Industrial Health* **12**, 211-224.

Lawson, A. B.(1993). On the analysis of Mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society, Series A*, **56**, 363-77

Lawson, A. B., Biggeri, A. B., Boehning, D, Lesaffre, E., Viel, J-F., Clark, A., Schlattmann, P., Divino, F.(2000). Disease mapping models: an empirical evaluation, *Statistics in Medicine*, **19**, 2217-2241.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edn.). Chapman and Hall, London.

Michal, F., Grigor, K. M., Negro–Vilar, A. and Skakkebaek N. E.(1993). Impact of the Environment on Reproductive Health: Executive Summary. *Environmental Health Perspectives* **101**(Suppl. 2), 159–167.

Morris, S. E. and Wakefield J. C. (2000). Assessment of disease risk in relation to a prespecified source. In *Spatial Epidemiology: Methods and Application* (P. Elliot, J.C. Wakefield, N.G. Best and D.J. Briggs, eds.) New York: Oxford University press, pp. 153-184.

Nova Scotia Department of Health and the Cape Breton District Health Authority, 2001. Lead and Arsenic Biological Testing Program in Residential Areas Near the Coke Ovens Site[online]. Available :
http://www.muggah.org/site/projects/reports/archives/10.pdf

Pampalon R. and Raymond G. (2000). A Deprivation Index for Health and Welfare Planning in Quebec, *Chronic Diseases in Canada* , **21**(3), 104–113.

Plummer, M., Best, N. G., Cowles, M. K. and Vines, S. K. (2004). *Output analysis and diagnostics for Markov chain Monte Carlo: version 0.7-1* (available at http: // www.fis.iarc.fr/coda)

Potthoff, R. F. and Whittinghill, M. (1996). Testing for homogeneity in the Poisson distribution. *Biometrika*, **53**, 183–190.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S.Richardson and D.J. Spiegelhalter), London: Chapman & Hall, pp. 45–58

Rootman, I., and Raeburn, J.(1994). The Concept of Health. In A. Pederson, M. O'Neill and I. Rootman(Eds.). *Health Promotion in Canada: Provincial, National and International Perspectives.* Toronto: W. B. Saunders, pp. 56–71.

Rylander, L., Stromberg, U., Hagmar, L., (2000). Lowered birth weight among infants born to women with a high intake of fish contaminated with persistent organochlorine compounds. *Chemosphere* **40**, 1255–1262.

Seal, C. (2002). *Media and Health*, London: Sage Publications.

Shaddick, G. and Elliot, P.(1996). Use of stone's method in studies of disease risk around point sources of environmental pollution. *Statistics in Medicine*, **15**, 1927-34.

Shaw, G, M., Schulman, J., Frisch, J.D., Cummins, S. K., Harris, J. A., (1992). Congenital malformations and birth weight in areas with potential environmental contamination. *Archives of Environmental Health* **47**, 147-154.

Spiegelhalter, D. J., Best, N. G., and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models (available at http://www.med.ic.ac.uk/divisions/60/biostat/dic.ps).

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1998), "WINBUGS User Manual" version 1.2, Cambridge, UK (available at http://www.mrc-bsu.cam.ac.uk/bugs)

Stone, R. A.(1988). Investigation of excess environmental risks around putative sources: Statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–60.

Sullivan, F.M. (1993). Impact of the environment on reproduction from conception to parturition. *Environmental Health Perspectives* **101**(suppl 2), 13-18

Tara A. R. Burra (2002). Reproductive and Psychological Health of Women living in the vicinity of the Tar Ponds, Sydney, Nova Scotia. *M.sc. project, Dept. of Geography and Geology, McMaster University.*

Townsend P. (1987) Deprivation. *Journal of Social Policy*, **16**(2), 125-146

Upton, A. C., (1989). Public health aspects of toxic chemical disposal sites. *Annual review of Public health* 10, 1-25.

Viana, N. J., and Polan, A. K. (1984). Incidence of low birth weight among Love Canal Residents. *Science* **226**, 1217-1219.

Vrijheid, M., (2000). Health effects of residence near hazardous waste landfill sites: a review of epidemiologic literature. *Environmental Health Perspectives*, **108** (suppl.1), 101–112.

Wakefield, J. C., Best, N. G. and Waller, L.(2000). Bayesian approaches to disease mapping. In *Spatial Epidemiology: Methods and Application* (eds. P. Elliot, J.C. Wakefield, N.G. Best and D.J. Briggs) New York: Oxford University press, pp. 105–126.

Wakefield, J. C., Kelsall, J. E. and Morris, S. E.(2000). Clustering, cluster detection, and spatial variation in risk. In *Spatial Epidemiology: Methods and Application* (eds. P. Elliot, J.C. Wakefield, N.G. Best and D.J. Briggs) New York: Oxford University press, pp. 129–152.

Wakefield, J. C. and Morris, S. E. (2001). The Bayesian modelling of disease risk in relation to a point source. *Journal of the American Statistical Association*, **96**, 77–91

Waller, L. A., Turnbull, B. W., Clark, I.C., and Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: application to leukaemia incidence and TCE-contaminated dump sites in upstate New York. *Environmetrics*, **3**, 281-300.

Walter, S. D.(1993). Assessing spatial patterns in disease rates. *Statistics in Medicine*, **12**, 1885–1894

Appendix A

Model Specification in WINBUGS

```
model; {

#poisson regression model
    for (i in 1:N) {
        obs_count[i] ~ dpois(mu[i])
#Model 1: model with no covariates
       log(mu[i])<- log(E[i]) + alpha0 + V[i] + U[i]
#Model 2: model with only distance covariate
      log(mu[i])<- log(E[i])+ alpha0 + alpha*(1/d[i]) + V[i] + U[i]
#Model 3: model with only area-level covariate
     log(mu[i])<- log(E[i])+ alpha0 + phi1*x1[i] + phi2*x2[i]
                 + phi3*x3[i] + phi4*x4[i] + phi5*x5[i] + V[i] + U[i]
#Model 4: full model
   log(mu[i])<- log(E[i])+ alpha0 + alpha*(1/d[i]) + phi1*x1[i]
   + phi2*x2[i]+ phi3*x3[i] + phi4*x4[i] + phi5*x5[i] + V[i] + U[i]
# Predicted area-specific relative risk
   lambda[i] <-  mu[i]/ E[i]
# Unstructured random effects
  V[i]~dnorm(0,tau.V)
    }
#CAR prior distribution for spatial random effects:
    U[1:N] ~ car.normal(adj[],weights[],num[],tau.U)
    for(k in 1:sumNumNeigh){
        weights[k]<- 1
        }
# other priors:
    alpha~dnorm(0,1.0E-5)
    alpha0~dflat()
    phi1~dnorm(0,1.0E-5)
    phi2~dnorm(0,1.0E-5)
    phi3~dnorm(0,1.0E-5)
    phi4~dnorm(0,1.0E-5)
    phi5~dnorm(0,1.0E-5)
    tau.V~dgamma(0.1,0.1)
    tau.U~dgamma(0.1,0.1)
#  variance and standard deviation of unstructured random effect
    var.V<-1/tau.V
    sigma.V<- sqrt(1 / tau.V)
#  variance and standard deviation of spatial random effect
    var.U<-1/tau.U
    sigma.U <- sqrt(1 / tau.U)
# other estimates
    eta<-exp(alpha0)  # scale parameter
    psi<-sigma.U/(sigma.U+sigma.V)
}
```

Each of the models was run separately