Use of spatial models and the MCMC method for investigating the relationship between road traffic pollution and asthma amongst children

Yong Zhang

BSc
Peking University

MSc
Peking University

A thesis submitted in partial fulfilment of the requirement of the University of Greenwich for the degree of Doctor of Philosophy

Statistics and Operational Research Group
School of Computing and Mathematical Sciences
The University of Greenwich

October 2000

Abstract

This thesis uses two datasets: NCDS (National Child Development Study) and Bartholomew's Digital road map to investigate the relationship between road traffic pollution and asthma amongst children. A pollution exposure model is developed to provide an indicator of road traffic pollution. Also, a spatially driven logistic regression model of the risk of asthma occurrence is developed. The relationship between asthma and pollution is tested using this model. The power of the test has been studied.

Because of the uncertainty of exact spatial location of subjects, given a post-code, we have considered error-in-variable model, otherwise known as measurement error model. A general foundation is presented. Inference is attempted in three approaches. Compared with models without measurement error, no improvement on log-likelihood is made. We suggest the error can be omitted.

We also take a Bayesian approach to analyse the relationship. A discretized MCMC (Markov Chain Monte Carlo) is developed so that it can be used to estimate parameters and to do inference on a very complex posterior density function. It extends the simulated tempering method to 'multi-dimension temperature' situation. We use this method to implement MCMC on our models. The improvement in speed is remarkable.

A significant effect of road traffic pollution on asthma is not found. But the methodology (spatially driven logistic regression and discretized MCMC) can be applied on other data.

Acknowledgements

I would like to thank my supervisors, Professor Keith Rennolls and Dr Anyue Chen for their guidance and support throughout my studies.

I am also grateful for the help from Dr. Ramesh Nadarajah and Mrs. Patel Swatee.

For financial support, I thank for the University of Greenwich for awarding me the bursary for this research.

Finally, I would like to thank my parents and my girlfriend for their support and belief in me.

Declaration

I certify that this work has not been accepted in substance for any degree, and is not concurrently submitted for any degree other than that of Doctor of Philosophy (PhD) of the University of Greenwich. I also declare that this work is the result of my own investigations except where otherwise stated.

Table of contents

List of figures

List of tables

Chapter 1 Introduction and Overview

In this chapter we will give a general introduction to the problems at which this thesis is directed, the methods of statistical analysis which have been adopted and developed to tackle these problems, and the data upon which these methods have been applied to obtain substantive results.

1.1 Statement of the problem

Epidemiology is defined (in Elizabeth Martin 1998) as "*the study of the occurrence, distribution, and control of infectious and non-infectious diseases in populations, which is a basic part of public health medicine. Originally restricted to the study of epidemic infectious diseases, such as smallpox, it now covers all forms of disease that relate to the environment and ways of life. It includes the study of the links between smoking and cancer, diet and coronary disease, as well as communicable diseases*". Statistical methods are often used in epidemiology studies. This thesis discusses the statistical methodologies in spatial statistics and uses them to study the relationship between asthma and road traffic pollution.

Asthma is one of the main chronic, treatable conditions in Western Europe, which is increasing in frequency. Many sources, from research studies to anecdotal newspaper reports, have suggested that there is a strong causal link between road traffic pollution and occurrence of asthma, particularly for children. The main question addressed in this thesis is: Does the available evidence prove scientifically/statistically that such a causal relationship exits? The problem is to access and make use of sufficient relevant available data in an appropriate statistical analysis.

If the answer to the question posed above is clearly affirmative, then there should be major implications for Government policies both in Health and in Transport. There would conceivably be major national issues of finance and the national economy involved. Clearly with such major implications possible it will be necessary that any conclusion drawn must have very high levels of confidence. In statistical terms this amounts to choosing a type I error probability of tests to be very low.

On the other hand, if a causal relationship cannot be clearly (significantly) demonstrated, then research efforts into the aetiology of asthma must look elsewhere than at road traffic pollution. However it might conceivably be the case that an important causal relationship between road pollution and asthma really does exist, but that either the relationship is weak, the environmental variability is high, or the size of the available data is insufficient to statistically demonstrate the existence of the relationship. In such circumstances we are in danger of coming to dangerous false negative conclusions. This is of course the issue of 'power', which is considered in a later chapter.

1.2 The data used

In Chapter 3 we present in detail the data used in this project. There are two main datasets. The first dataset is the National Child Development Study (NCDS) which is a continuing longitudinal study which follows the lives of all those living in Great Britain who were born between 3 and 9 March 1958, inclusive (for details see Chapter 3). The spatial positions of subjects' houses are derived from their home post-codes available from NCDS when the subjects were aged 16 years. The second dataset is Bartholomew's digital map of Great Britain which provides the grid references of ends of the linear road segments which constitute the road map. The map also distinguishes between 22 different road types. However, we collapse these into four types of road (motorway, trunk, principle and minor road) for which average traffic intensities are available from the Department of Transport.

This thesis reports on work which has attempted to use the above data to investigate the relationship between occurrence of asthma by the age of 16 years, and the influence of road pollution as determined at the home locations of the NCDS subjects at age 16 years.

1.3 Modelling approaches available

In epidemiological studies the spatial distribution of a disease is often studied in relation to the spatial distribution of covariates. The study of spatial variables might involve study of their joint distribution in order to clarify the relationship between the variables.

However, it is often of interest to investigate the relationship between a spatially distributed disease-related outcome variable and spatially distributed 'causal' (or regressor) variables. In the context of the problem and data available in this study there is an important primary question about how we treat the spatial location data. This spatial location data is the best available, but we know it is NOT completely precise. We have to decide if we wish our model to make the assumption of locational accuracy, or whether we wish to model the inaccuracy of our spatial location data. Hence modelling approaches can be classified initially as being from either of the following frameworks:

- Accurate Location Models, or
- Inaccurate Location Models, (or 'measurement error' or 'error in location' models).

1.3.1 Accurate Location Models

In this framework, there are three possible modelling approaches to the analysis of the spatial epidemiology data, with the choice between them depending on the nature

of the data available and the extent to which the analyst wishes to develop and elaborate models. The three approaches are:

- the conditional spatial location approach,
- the point process modelling approach,
- the count data modelling approach.

We introduce these approaches briefly in this chapter and review them more carefully in Chapter 2. The nature of our data means that the conditional modelling approach is feasible and this is the line of the main development in this thesis. However, the alternatives are still reviewed for completeness, and to possibly aid in the comparison with other studies.

1.3.1.1 The conditional spatial location approach

This approach treats the occurrence of disease (asthma) of subjects as independent Bernoulli variables, conditionally on their locations. The probability of a subject having disease may be modelled with a logistic regression model, in which spatial variables are included as covariates. This model is flexible, can be used to study putative point, linear, and network (and even more complicated) pollution sources.

1.3.1.2 The point process modelling approach

In contrast to the conditional approach is the 'point-process' approach. In this approach the locations of the cases are treated as a realization of a spatial point process.

The intensity function $\lambda_c(\mathbf{x})$ of the spatial point process for ***cases*** is defined as

$$\lambda_c(\mathbf{x}) = \lim_{|d\mathbf{x}|\to 0}\left\{\frac{E[N(d\mathbf{x})]}{|d\mathbf{x}|}\right\},$$

in which, d$\mathbf{x}$ is an infinitesimal region which contains the point $\mathbf{x}$, $|d\mathbf{x}|$ is the area of the infinitesimal region, $N(d\mathbf{x})$ is the number of cases in region dx. $\lambda_c(\mathbf{x})$ is then modelled as a function of the causal covariates in which we are interested (Diggle 1990). For example, $\lambda_c(\mathbf{x}) = f(\lambda_s(\mathbf{x}), y(\mathbf{x}))$ where $\lambda_s(\mathbf{x})$ is the

intensity function of the subjects in the (NCDS) study, (defined as above), and $y(\mathbf{x})$ is a vector covariate at $\mathbf{x}$. The simplest functional form, with no dependence of the disease on the environment, is $\lambda_c(\mathbf{x}) = p\lambda_s(\mathbf{x})$, where p is the proportion of the subjects in the study who have the disease (asthma). In this way the two datasets (the spatial distribution of cases and subjects, and of covariates) are linked. Once $\lambda_c(\mathbf{x})$ *and* $\lambda_s(\mathbf{x})$ are estimated (such as, by kernel estimation), the relationship is estimated.

1.3.1.3 The count data modelling approach.

In this approach the disease data is aggregated by regions, (or quadrates) to produce counts of cases, subjects, etc in the regions. The counts of cases would usually be assumed to follow Poisson distributions with parameters depending on local pollution-related covariates. The expected count of cases, N_i, in subregion i, is modelled as

$$E(N_i) = e_i\, m(f_i\, \alpha), \qquad i=1,\ldots,n$$

Where the e_i as the background rate for the ith subregion. The function $m(.)$ is a link to spatial and other covariates in f_i, which is a 1 x q vector. f_i includes r_i, the radial distance of subregion i from the source. Likelihood ratio test may be used to do inference. Several alternative tests (Stone 1988, Lawson 1993) have been developed.

Following the detailed review of approaches, in Chapter 2, the conditional approach is adopted and developed in Chapter 4. The initial models and results are presented in Chapter 4.

In Chapter 4 we use a logistic model as the asthma incidence model in which some of the covariates are pollution exposures from different types of roads. Suggested partly by physical models of pollution dispersal, such as that of Pasquill (1983), a pollution exposure dispersal model is proposed. Using it, a model of the pollution exposure of subjects from road traffic is obtained. From the logistic regression, parameter estimation is achieved using maximum likelihood estimation (MLE), and inference is based on changes in the -2log-likelihood, (equivalent to likelihood ratio tests). In

Chapter 5 the power of the inferences are studied with the conclusion that more subjects are needed for a conclusive study.

1.3.2 Inaccurate Location Models
(or 'measurement error' or 'error in location' models)

The above mentioned models (considered in Chapters 4 and 5) assume home locations are exact. This is not true in our data. Hence in Chapter 6 we explore this issue and develop a 'measurement error model' for asthma/road-pollution, where the measurement error is in spatial location determination. Other measurements are assumed to be exact. (See Kendal and Stuart (1979), p399-443, for an introduction, and Hoschel (1989), for a more recent review of 'functional relation' models).

The measurement error model is fitted to the data using three methods. The first method uses a general maximum likelihood approach to parametric estimation. The second method attempts to obtain estimates more quickly by approximating the pollution exposure for an individual at a particular home location in a given post-code by the average exposure over all points in that post-code. The third simulate the true house location, then do inference and estimation based on simulations. The results are compared with results from models in Chapter 4. Little changes in the log-likelihood measure of goodness-of-fit result.

1.4 The Bayesian Approach, and Markov Chain Monte Carlo

In Chapter 7 the Bayesian approach is adopted and Markov Chain Monte Carlo (MCMC) is used to estimate the model parameters from their posterior distribution.

MCMC (Robert and Casella 1999) achieves a sampling from the posterior joint distribution of the model parameters. There are two basic MCMC methods: the Gibbs sampler (Geman and Geman 1984) and the Metropolis-Hastings Algorithm (Metropolis *et al.* 1953, Hastings 1970). The basic idea is building a Markov chain

whose equilibrium distribution is the target (posterior joint) distribution. Then after a long period, the chain's distribution converges to the target distribution.

MCMC is computationally demanding. It calculates the density function many times. If the density function takes too much time to evaluate (such as in our study; each posterior /likelihood-function takes seven minutes to evaluate), it is practically impossible to implement MCMC. In Chapter 7 a discretization-method which involves concepts of 'multi-dimensional temperature' to implement MCMC on certain kind of complex posterior density is developed. The improvement in computational speed is remarkable. Conditions in which similar methods may be used in other modelling situations are indicated in Chapter 7. The results of Bayesian analysis is no different from the likelihood analysis i.e. no significant relationship between asthma and traffic pollution is found.

The concluding Chapter 8 discusses the evidence from our data of a link between asthma and road traffic pollution. It discusses the limitations of our models, data, and suggests further studies. It also suggests our models and methods can be used in other situations.

Chapter 2 Review of the literature

The prevalence of asthma has been increasing in recent years particularly amongst children. There has been much speculation on the cause of this increase. Some research suggests traffic pollution is part of the reason. In this chapter we review previous studies on asthma, and in particular its relationship with air pollution and traffic. We find there is a link missing in previous research. These earlier research papers fall into two groups:

(i) they consider the distribution of air pollution, using for example NO_2, SO_2, etc as indictors of pollution (Baxter *et. al.* 1983, Spinaci *et. al* 1955 and Shy *et. al* 1971). However such pollution are only partly generated from traffic vehicles; there are other sources, such as industrial sites. Hence any asthma effects cannot be assigned to traffic pollution causes alone.

(ii) they consider the effect of 'roads' or 'traffic' represented only as a simple categorical variable, typically having values 'high' and 'low' (Speizer and Ferris 1973, Waldron *et al.* 1995, Edwards *et al.* 1994, Wjst *et al.* 1993, Weiland *et al.* 1994 etc.). It would seem that the putative driving variable, traffic pollution, needs to be more refined.

Research is required to examine the relationships between asthma and pollution exposure specifically arising from traffic on roads. A suitable model for pollution dispersal from road traffic also needs to be developed.

2.1 Asthma

2.1.1 General information

Asthma affects people of all ages; it can be severe and is sometimes fatal. It is primarily caused by inflammation of the airways which causes victims to be hyper-

irritated and respond with mucus production and decreased airflow (Burney 1992). This irritability may be associated with coughing, wheezing, shortness of breath and mucus production. The tendency to have extra-irritable airways may have an inherited component, or may be acquired. This tendency is then acted on by any of a number of stimuli (singly or in combination) including: air temperature, respiratory viruses, pollution, odors, allergens, stress, chemicals, dust and cigarette smoke (active or passive) (Busse *et. al.* 1993, Department of Health 1995). The two most important triggers for most people are respectively viruses and cigarette smoke (Strachan 1995).

There has been an increase of about 50% in the prevalence of childhood asthma over the last 30 years (Committee on the medical effect of air pollution 1995). Burr *et al.* (1989) carried out a survey on 12-years-old children in one area in Wales in 1973, repeated in 1988. They found the proportion of asthma sufferers had increased from 4.9% to 9.1%. Two studies (Ninan and Russell 1992) in Aberdeen, Scotland, in 1964 and 1989 suggested that the prevalence of wheezing between this period had doubled. Though there are no precise figures for current asthma prevalence in the UK, there has been at least a ten-fold increase in hospital admissions for asthma amongst children (Committee on the medical effect of air pollution 1995). About 2,000 people die of asthma every year. The total annual cost of asthma in UK is nearly £1billion. There are many suggested causes for this increasing trend: the increased population of bed mites due to increased central heating usage; increasing levels of poverty, or simply because changes in medical practice have increased the awareness of asthmatic disease.

2.1.2 Asthma and air pollution

Air carries not only life-giving oxygen, but also noxious pollutants that reach unhealthy levels, such as ozone, carbon monoxide, fine particles, sulphur dioxide, nitrogen dioxide and lead (Department of Health 1992). In June 1980 (Baxter *et. al.* 1983), the pollution level in Washington DC rose to an unusually high level, and the

number of patients with asthma was five times higher than normal during that period. Though there is laboratory evidence that air pollution could have a role in the initiation of asthma, there is no firm proof that road traffic pollution is the real cause of increasing asthma. The alleged relationship between asthma and traffic pollution is largely based on anecdotal newspaper reports of higher morbidity and mortality on the days of high pollution. For example, in an article in The Times (Vaughan Freeman 1996):

"The polluted air is so thick you can cut it with a knife. Our two girls have asthma, and we have had our share of traumatic times when they have suffered attacks. I am convinced that traffic pollution has a lot to do with asthma."

The results of the scientific studies are mixed. Speizer and Ferris (1973) compared 128 Boston policemen heavily exposed to traffic and 140 policeman posted at an "outskirts station" where 87% had never worked outdoors in traffic. No statistically significant differences in respiratory symptoms were found. Waldron *et al.* (1995) found prevalence of wheezing in one-year period was significantly lower among 448 teenage children from an area close to the M23 and the M25 than in 952 teenagers from other parts of East Surrey. Edwards *et al.* (1994) did a case-control study of pre-school children admitted to hospital. They found children admitted with asthma were more likely to live in an area with high traffic flow than children admitted for non-pulmonary disease. Other researches conducted in Germany, (Wjst *et al.* 1993, Weiland *et al.* 1994) and Japan (Ishizaki *et al.* 1987) found no solid conclusion.

After a review of previous research studies on the relationship between asthma and traffic pollution, the Committee on the medical effect of air pollution (1995) concluded,

"*In most of the few studies so far published, there is a consistent, though modest, association between exposure to traffic and asthma prevalence in children. Assuming that this association is real (and not due to reporting bias, for example) it is unclear whether this is due to an increase in the initiation or provocation of asthma*".

In most of the research reviewed, only "high traffic" or "low traffic" was used an indicator of traffic-related pollution. Traffic pollution exposure was not treated as a continuous covariate.

In our research, a model is developed first to represent the spatially distributed level of traffic pollution exposure, and second to model and test the relationship between asthma incidence and traffic pollution exposure.

2.2 Previous statistical models in environmental epidemiology

Modelling the effect of covariates on an outcome variable is a central component of statistical methodology. In the recent years, with the degradation of environment locally, regionally, globally, and increasing pollution, the elucidating the effect of the degrading environment on human health has become a substantial challenge to statistical modelling methodology.

In the past decade, there has been considerable progress in developing the techniques for studying the relationship between diseases and pollution sources. But the most sophisticated techniques have yet to be used widely in practical studies. Barnett (1997), reviewing the use of statistical methods in pollution research in the past, stated that 65% of papers in the area used simple descriptive methods, or significance tests, and a further 25% used standard regression, multivariate, or simplistic spatial techniques. That is, a total of 90% of papers used only simple or simplistic statistical techniques. (Researches reviewed in Section 2.1 belong to these 90%). The use of more sophisticated and novel statistical techniques has tended to have "*rather limited (even cursory) application to real problems*", (Elliot *et al.* 1995). Among the 10% usage of sophisticated models, Diggle (1990) and Diggle and Rowlingson (1994) described methodology for fitting models for raised incidence of rare disease around point sources of environmental pollution for a point process realization of subject-case locations. In these studies, the primary data are is the spatial distribution of a

disease. The aim of these studies is to see whether this spatial distribution is random or depends on the pollution source.

Lawson, Biggeri and Williams (1999), Lawson and Waller (1996) classified the available spatial data as being of two kinds: locations of disease or counts of disease. These two kinds of data need different models. Lawson (1993), Lawson and Williams (1994), and Lawson and Waller (1996) considered the use of spatial models of pollution exposure around point sources, and Stone (1988) proposed a test of excess risk for count data.

McCready *et al.* (1997) considered the road network as the pollution source in relation to incidence of asthma. A geographic information system (GIS) was used with a simplistic 'Zone of Influence' model of the effect of roads on the incidence of asthma. Though these analyses did produce some significant results, concerning the effect of road pollution as a cause of asthma, the results were rather ambiguous. For example, the proximity to road traffic did have a significant effect upon reported ever-asthma at age 33, but there was no significant effect detected on current severe asthma at age 33. It was therefore felt that further analysis was necessary, which makes use of more appropriate spatial models of the effect of road pollution on asthma.

2.2.1 Point process models for point data

It is assumed that the disease is recorded as the locations (points) of disease/patients.

To start with we present some of the basic terminology of spatial point process in two-dimensional space.

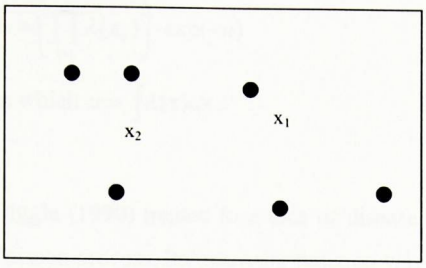


Figure 2.1 An example of a realization of a spatial point process within a window

A spatial point process is a stochastic mechanism, which generates a countable set of points in the plane. Figure 2.1 shows that part of a realization of a spatial point process that is within a given viewing window. Denote a point by **x**, the vicinity of **x** by d**x**, with area |d**x**|, and the number of points in an area A by N(A).

The first-order intensity function $\lambda(\mathbf{x})$ is defined by

$$\lambda(\mathbf{x}) = \lim_{|d\mathbf{x}|\to 0}\left\{\frac{E[N(d\mathbf{x})]}{|\,d\mathbf{x}\,|}\right\} \qquad (2.1)$$

$\lambda(\mathbf{x})$ represents the mean number of points per unit area in the immediate vicinity of **x**.

Inhomogeneous Poisson process is often used to model the locations of cases. A ***heterogeneous (non-stationary) Poisson process*** **(HEPP)** is a point process with properties:

1. N(A) has a Poisson distribution with mean $\int_A \lambda(\mathbf{x})d\mathbf{x}$.
2. Given N(A) = n, the n points in A form independent random samples from the distribution on A with density function proportional to $\lambda(\mathbf{x})$, $\mathbf{x}\in$A.

The likelihood function

Suppose $\{\mathbf{x}_i \in \text{A}, i=1, \ldots, n\}$ is part of a realization of a heterogeneous Poisson process with intensity function $\lambda(\mathbf{x})$.

Given data $\{\mathbf{x}_i \in \text{A}, i=1, \ldots, n\}$, the density function of the heterogeneous Poisson process is (for the proof see Appendix C):

$$L = \left(\prod_{i=1}^{n} \lambda(\mathbf{x}_i) \right) \cdot \exp(-u) \qquad (2.2)$$

in which $u = \int_A \lambda(\mathbf{x}) d\mathbf{x}$.

Diggle (1990) treated locations of disease cases as a realization of a heterogeneous Poisson process. Its' intensity function was factored to two components: the intensity of population and the distance from the pollution source.

The intensity $\lambda(\mathbf{x})$ may be parameterized as:

$$\lambda(\mathbf{x}) = f(\mathbf{x}) \lambda_0(\mathbf{x}) \qquad (2.3)$$

where $\lambda_0(\mathbf{x})$ is the background intensity of the population at risk at point $\mathbf{x}$, and $f(\mathbf{x})$ is a parameterized function of risk relative to the location of the pollution source. Put (2.3) in (2.2), the log-likelihood function is

$$L = \sum_{i=1}^{n} \log f(\mathbf{x}_i) \lambda_0(\mathbf{x}_i) - n \log \int_A f(\mathbf{x}) \lambda_0(\mathbf{x}) d\mathbf{x} \qquad (2.4)$$

$f(\mathbf{x})$ can be in different forms. For example, Diggle (1990) used the isotropic form

$$f(\mathbf{x}, \alpha, \beta) = 1 + \alpha \exp\{-\beta g((\mathbf{x} - \mathbf{x}_0)'(\mathbf{x} - \mathbf{x}_0))\} \qquad (2.5)$$

where $\alpha \geq 0, \beta \geq 0$ and g(.) is monotone non-decreasing with g(0) = 0. x_0 is the location of pollution source.

Using maximum likelihood estimation to maximize log-likelihood function L, parameters α and β can be estimated. Testing H_0: $\alpha = 0$ vs. H_1: $\alpha > 0$ can examine the dependence of disease on pollution source. In (2.4) λ_0 can be estimated from the location of a 'control' disease, possibly using kernel estimation.

Diggle (1990) used this method to analyse the spatial distribution of cancer in the Chorle-Ribble area of Lancashire, England.

2.2.2 Models for count data

This approach has been summarised in Lawson, Biggeri and Williams (1999), and Lawson *et al.* (1996). In this approach the disease data is aggregated by regions, (or

quadrates) to produce counts of cases, subjects, etc in the regions (Figure 2.2). The counts of cases would usually be assumed to follow Poisson distributions with parameters depending on local pollution-related covariates.

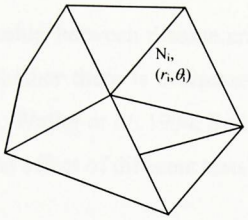


Figure 2.2 An example of count data

Assume there are n small regions. N_i $(1 \le i \le n)$ is the count of cases in the ith region (Figure 2.2). The region counts assumes to be independent Poisson random variables with parameters λ_i, i=1,…,n. The model follows from an assumption that the location of individual event follows an HEPP, any non-overlapping regionalization of an HEPP leads to independently Poisson distribution regional event counts with means

$$\lambda_i = \int_{A_i} \lambda(x)dx, \qquad i=1,\ldots,n \tag{2.6}$$

where $\lambda(x)$ is the first order intensity of the HEPP and A_i is the ith subregion.

The expected count of disease, N_i, in the ith region is modelled as

$$E(N_i)= \lambda_i = e_i\, m(f_i\, \alpha), \quad i=1,\ldots,n \tag{2.7}$$

Where the e_i acts as the background rate for the ith subregion. The function $m(.)$ is a link to spatial and other covariates in f_i .which is a 1 x q vector. The parameter vector α has dimension q x 1. Define the polar coordinates of the subregion centre as (r_i, θ_i), relative to the pollution source. Often, only r_i, the radial distance from the source, is included in f_i. When this is used alone, an additive link such as $m(.)=1 + \exp(.)$, is often used. However directional variables (such as, $\cos\theta_i$, $\sin\theta_i$, $r_i\cos\theta_i$, etc.) representing preferred direction and angular linear correction can also be useful in detecting directional preference resulting from preferred directions of pollution outfall. Lawson and Williams (1994) discussed the use of direction variables and the appropriate form of function $m(.)$ and α.

A likelihood ratio test may be used to do inference. Stone (1988) proposed a test based on the assumption that the substance levels are non-increasing with geographical distance from the point source. Lawson (1993) developed a score test against a non-monotone peaked alternative.

The relationship between disease and pollution point source can also be studied by detecting whether there is a disease clustering around the point source (Besag and James 1991, Waller *et al.* 1994, Smith and Neutra 1993). Waller and Lawson (1995) compared the effect of different tests for detecting disease clustering.

2.2.3 A conditional spatial location approach

This approach treats the occurrence of disease of subjects as independent Bernoulli variables conditionally on their locations. The probability of a subject having disease may be modelled with a logistic regression model, in which the spatial variables are included as covariates.

Diggle and Rowlingson (1994) proposed a method by regarding the control and case (given their locations) as realizations of a Bernoulli distribution in which the probability is

$$p(\mathbf{x}) = \rho\, f(\mathbf{x}\text{-}\mathbf{x}_0;\alpha,\beta)/\{1+\rho\, f(\mathbf{x}\text{-}\mathbf{x}_0;\alpha,\beta)\} \qquad (2.7)$$

This method is easy to use when a point map of a control disease is available. It is flexible, can be used to study linear and more complicated network sources. It also can include covariates about individual subject in the model.

In our research, roads are spatial networks, locations of cases of our survey are known, additional covariates about subjects (gender, etc) are included. Hence the conditional approach is adopted. The work of McCready *et al.* (1997) is extended, by taking the spatial locations of the homes of NCDS subjects as given, (the conditional approach of Diggle and Rowlingson 1994). Spatial models of the exposure to air

pollution caused by road traffic, and the effect of this exposure on asthma incidence are developed. In contrast to Diggle and Rowlingson (1994), who considered a number of point sources of pollution, the source of pollution considered in this study is a large network of roads with varying traffic volumes. A parametric two-dimensional spatial model for the distribution of exposure levels around a point on a road is specified. This is integrated over road segments, and weighted by traffic intensities, to give pollution exposures at the locations of NCDS subjects. Covariate information about the subjects is also considered in the modelling. The resulting models for the occurrence of asthma in the study regions are non-linear spatial risk models, which have been investigated using maximum-likelihood methods and MCMC.

Chapter 3 Data and general analyses

Three datasets are used in this research study: (i) NCDS (The National Child Development Study), (ii) Bartholomew's digital map of Great Britain and (iii) the post-code file.

NCDS and Bartholomew's digital map of Great Britain are stored in MIMAS (Manchester Information and Associated Services). MIMAS is a JISC (Joint Information System Committee)-supported national data centre run by Manchester Computing Service at the University of Manchester, specialising in flexible on-line access to strategic research and teaching datasets, key bibliographic information, software packages, specialist support and training, and large-scale computing resources for the UK academic community. MIMAS is a free service for higher education throughout the UK.

In this research, information about subjects, their asthma outcome and some other information i.e. social status are obtained from NCDS. The locations of the homes of these subjects are stored in post-code form. Using the post-code file we can obtain the locations in National Grid co-ordinate. Bartholomew's digital map contains the locations of major and minor roads in Britain (It does not however include very small roads). They are stored in the same National Grid co-ordinate as post-code file. Combining these data, we have the home locations (where subjects live) and the spatial distributed road network. In Chapter 4 we use the data to model pollution exposure from road traffic for each subject and hence study the relationship between asthma occurrence and pollution exposure.

3.1 The National Child Development Study

The National Child Development Study (NCDS) is a continuing longitudinal study which is seeking to follow the lives of all those living in Great Britain who were born

between 3 and 9 March 1958, inclusive (about 18,000 individuals). The aim of the cohort study is to improve understanding of the factors affecting human development over the whole lifespan.

NCDS has its origins in the Perinatal Mortality Survey (PMS) sponsored by the National Birthday Trust Fund. This was designed to examine the social and obstetric factors associated with stillbirth and death in early infancy among the 18,000 children born in Great Britain in that one week.

To date there have been five attempts to trace all members of the original study in order to monitor their physical, educational and social development. These were carried out by the National Children's Bureau in 1965 (NCDS1: age 7), in 1969 (NCDS2: age 11), in 1974 (NCDS3: age 16), in 1981 (NCDS4: age 23), and by the Social Statistics Research Unit, City University in 1991.

For the birth survey, information was obtained from the mother and from medical records of the midwife. For the purposes of the first three NCDS surveys, information was obtained from parents (who were interviewed by health visitors), head teachers and class teachers (who completed questionnaires), the school health services (who carried out medical examinations) and the subjects themselves (who completed tests of ability and, latterly, questionnaires).

The 1981 survey differs in that information was obtained from the subject (who was interviewed by a professional survey research interviewer). Similarly, the 1991 survey relied on survey research interviewers to collect information from all cohort members; from husbands, wives or cohabiters; and also from the natural and adopted children of cohort members and their mothers in a "one in three" random sample. Extensive use was also made of self-completion questionnaires.

The NCDS is used for a wide range of research, including medical/health research. The data covers a long time period and includes a wide range of questions, plus physical measurements, such as weight and height. There are a lot of papers published based on this study. They can be found on the web site (http://www.mimas.ac.uk/surveys/ncds/ncds.html). Several are concerned with asthma. Kaplin and Mascie (1985) did a study on bio-social factors in the epidemiology of childhood asthma. McCready *et. al.* (1997) used it to study relationship between asthma and road.

3.1.1 Variables

The NCDS data is stored in SPSS form. Each NCDS survey dataset has hundreds variables. Some of them are related to the study of asthma, i.e. sex, ever-asthma, current-asthma, etc.

We will work out the pollution exposure from the house locations. Compared to children, adults move more frequently. After a person reach 18, he starts to change his address frequently i.e. he goes to university which may be in another city, finds a job in another place, he travels etc. The current house location in our data may not reflect the pollution exposure he has got in the past. We decide to use the asthma information obtained from the 1974 survey (when subjects are 16, they change their addresses less frequently).

Listed below are some variables in 1974 survey (when subjects are 16) related to our asthma study. (Details of their codes are listed in Appendix A).

IDNO	The unique id for every person
N2887	Child's smoking habits when they are 16
N259	Ever had an asthma attack -when 5
N2617	Ever asthma or wheezy bronchitis - when 16
N5770	Asthma/bronchitis since 16th Birthday
SEX	Sex
N2622	Most recent asthma attack
N2623	Frequency if attack in last 12months
CURAS16	Current asthma/wb at age 16
GR74	Grid Ref. in 1974
EVERASTH	Ever asthma - recoded n2617
EVAS16D	Ever asthma at 16 - derived from asth16
N2017	Ethnic group from features
N504021	Ever been told has asthma
N504024	Wheezing/asthma inhibited speech last 12m

N504023 Used inhaler etc over the past 12m

N504025 Admitted overnight for wheezing/asthma

N504026 No. of times hospitalised for wheeze/asth past 12m

3.1.2 Simple analyses and choice of variables

Using all of the variables will make the model too complicated. To reduce the number of variables in our spatial model, some simple screening analyses have been carried on.

We will choose variables from this survey. There are two primary response variables about subjects' asthma status. Curas16 is the current-asthma outcome of subjects at age 16. EVAS16D is whether subject has ever-asthma before 16 (include 16).
Tables 3.1 and 3.2 give the frequencies of these two variables.

Response	Frequency	Valid Percent
No	10863	96.2
Yes	427	3.8
Missing	7268	-
Total	18558	100.0

Table 3.1 Proportion of children having current asthma at age 16

Response	Frequency	valid Percent
No	8731	70.2
Yes	3711	29.8
Missing	6116	-
Total	18558	100.0

Table 3.2 Proportion of children ever having asthma up to age 16

Among other variables, genders of the subjects and their smoke habits always play an important part in asthma (McCready *et al.* 1997). N2887 is a subject's smoking habit when they are 16. Sex is the gender of a subject. Tables 3.3 and 3.4 show their frequencies.

Response	Frequency	Valid Percent
Don't smoke	7687	64.2
<1/week	381	3.2
1-9/week	684	5.7
10-19	544	4.5
20-29	594	5.0
30-39	524	4.4
40-49	477	4.0
50-59	380	3.2
60+	698	5.8
Missing	6589	-
Total	18558	100.0

Table 3.3 Proportion of children's smoking habits at age 16

	Frequency	Valid Percent
Male	9593	51.7
Female	8960	48.3
Missing	5	-
Total	18558	100.0

Table 3.4 Proportion of children's gender

As presented in Table 3.1, only 427 (3.8%) subjects have current-asthma (variable Curas16) at age 16. For reasons given in Chapter 4, we will carry on our main studies on two sub-national regions (the London region and the Birmingham region). In these two regions there are only 37 (3.5%) and 16 (4.0%) subjects respectively who have current-asthma. With such small number of cases, the test would be of lower power. In contrast, the number subjects of having ever-asthma (variable EVAS16D) nationally is 3711 (29.8%). In our two restricted areas (London and Birmingham) the number of children who have ever-asthma is 283 (27%) and 131 (32%) respectively. The large number of cases of asthma would enable a more powerful analysis than is possible for current-asthma outcome. Hence it was decided to use ever-asthma as the asthma outcome in this study.

Logistic regressions are used to study the relationships between gender (sex), smoking habit (N2887) and ever-asthma outcome. Let y be the outcome of a subject having or having not ever-asthma. The model is

$y \sim B(p(\eta))$

with $p(\eta) = \dfrac{1}{1+e^{-\eta}}$ and $\eta = \rho + \alpha_1\, sex + \alpha_2\, N2887$ (3.1)

in which $B(p_0)$ is a Bernoulli distribution, $P(y=1)= p_0$, $P(y=0)=1-p_0$.

We also study the sub-models

$\eta = \rho + \alpha_1\, sex$ (3.2)

$\eta = \rho + \alpha_2 N2887$ (3.3)

The estimated logistic regression parameters for (3.1), (3.2) and (3.3) (by SPSS) are shown in Tables 3.5, 3.6 and 3.7 respectively.

	Estimate	S.E
α_1	0.3209	0.0460
α_2	0.0092	0.0089
ρ	-0.8785	0.0759
Log-likelihood	-5695.62	-

Table 3.5 Estimates using logistic regression for ever-asthma, with gender and smoking habit as covariates

	Estimate	S.E
α_2	0.0144	0.0089
ρ	-1.0422	0.0325
Log-likelihood	-5696.16	-

Table 3.6 Estimates using logistic regression for ever-asthma, with smoking habit as covariate

	Estimate	S.E
α_1	0.3299	0.0395
ρ	-0.7027	0.0605
Log-likelihood	-5721.4	-

Table 3.7 Estimates using logistic regression for ever-asthma, with gender as covariate

Compare models (3.1) and (3.2), (3.3), the significant levels for α_1, α_2 are 0.0000 and 0.2989 respectively (Tables 3.5, 3.6, 3.7). That is, gender is highly significant ($p < 0.01$). The p-value for smoking habit is only 0.29. The reason for this is probably that children just start to smoke in their teens. Smoking does not affect their past histories, so it is not significant for ever-asthma (variable EVAS16D).

After these initial analyses, it has been decided to use ever-asthma as asthma outcome and include sex as a covariate in our model.

3.2 Bartholomew's digital map of Great Britain

Bartholomew's digital map of Great Britain is a layered vector map dataset comprising of point, line and area features. Separate datasets cover London, Britain, Europe and the rest of the world at various scales. MIMAS supports the Bartholomew data in the Arc/Info coverage format, permitting the data to be viewed, manipulated in ArcView and outputted as ASCII text file.

3.2.1 Roads map

Bartholomew's digital map of Great Britain is specifically for use within a Geographical Information System. Dataset is divided into 100km by 100km square based on the National Grid (Figure 3.1). It contains 18 computerized maps or 'coverages', for each of the 55 (100 km x 100 km) squares that make up the National Grid. The road coverage includes twenty-two road types which have been collapsed into four main road types according their traffic volumes i.e. motorway (Figure 3.2), trunk road (Figure 3.3), principal road (Figure 3.4) and minor road (Figure 3.5).
The following is the code list for the road layer called RD. The OBS_ACC_NO is the code used for each type of road.

OBS_ACC_NO	DESCRIPTION
235	MOTORWAY
245	MOTORWAY UNDER CONSTRUCTION

173287	MOTORWAY TUNNEL
173291	A ROAD PRIMARY TRUNK DUAL C/W
173292	A ROAD PRIMARY TRUNK SINGLE C/W
173293	A ROAD PRIMARY TRUNK PASSING PLACES
233	A ROAD PRIMARY NON-TRUNK DUAL C/W
226	A ROAD PRIMARY NON-TRUNK SINGLE C/W
173294	A ROAD PRIMARY NON-TRUNK PASSING PLACES
231	A ROAD NON-PRIMARY DUAL C/W
227	A ROAD NON-PRIMARY SINGLE C/W
222	A ROAD NON-PRIMARY PASSING PLACE
243	A ROAD DUAL C/W UNDER CONS (ALL)
241	A ROAD SINGLE C/W UNDER CONS (ALL)
232	B ROAD DUAL CARRIAGE WAY
230	B ROAD SINGLE CARRIAGE WAY
229	B ROAD WITH PASSING PLACES
244	B ROAD DUAL C/W UNDER CONS (ALL)
242	B ROAD SINGLE C/W UNDER CONS (ALL)
73	ROAD TUNNEL
130	MINOR ROAD
173295	PRIVATE ROAD

All of the 22 types of road have been put into four road categories: motorway (235,245,173278), trunk road (173291, 173292, 173293, 233, 226, 173294), principle road (231, 227, 222, 243, 241) and minor road (232, 230, 229, 244, 242, 73, 130, 173295).

The digitized road map is likely to be no more accurate than the post-code locations of subjects. Each road is stored as combined by line-segments, which has a start point and end point each.

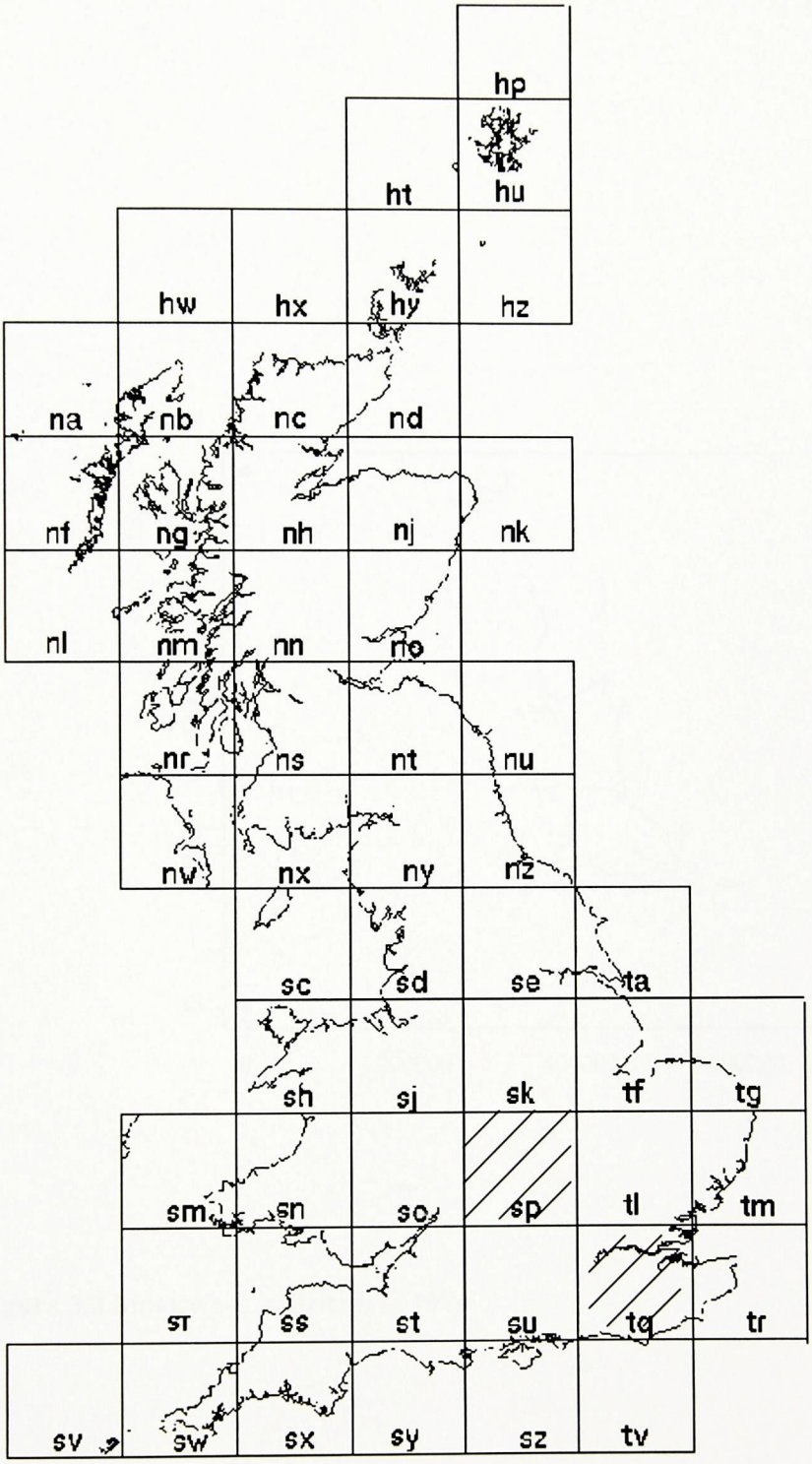


Figure 3.1 Regions in Britain. Square 'tq' is London, 'sp' is Birmingham

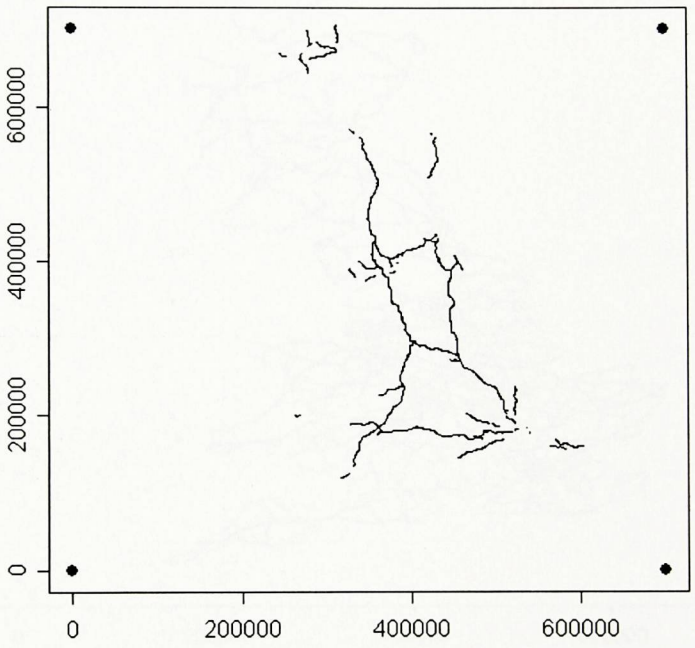


Figure 3.2 Motorways in Britain in 1974

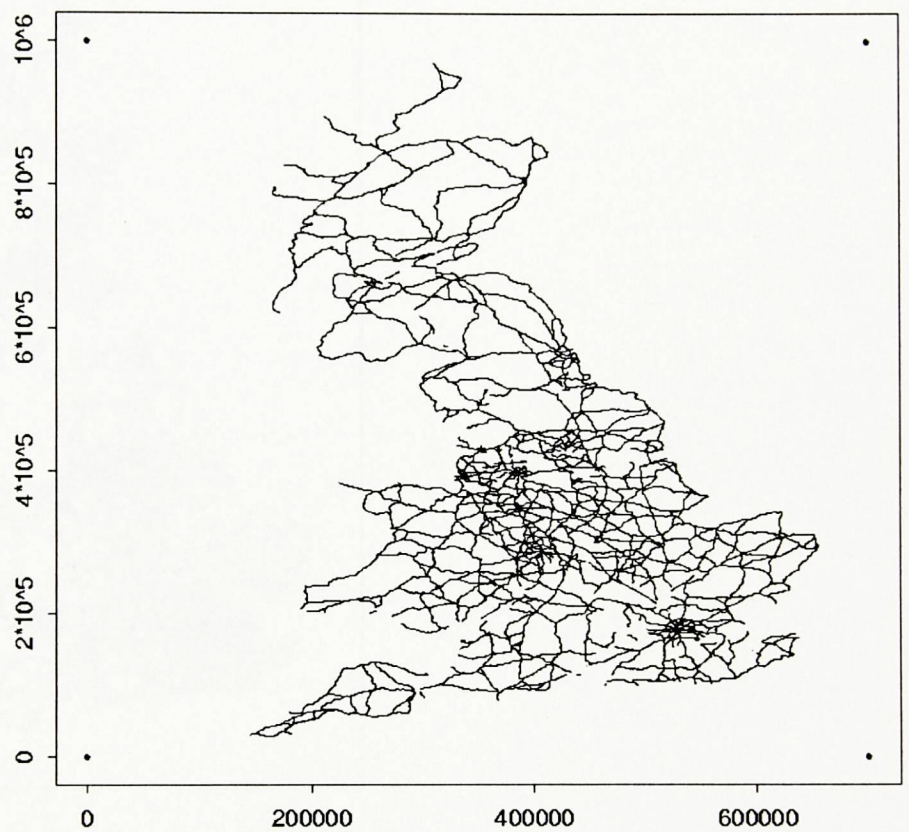


Figure 3.3 Trunk roads in Britain in 1974

Figure 3.4 Principal roads in Britain in 1974

Figure 3.5 Minor roads in Britain in 1974

In this study, results are reported for two 100km x 100km National Grid squares; one of these includes London (square 'tq' in Figure 3.1); the other includes Birmingham (square 'sp' in Figure 3.1). Each area has relative large number of subjects (1046 valid subjects in London and 402 valid subjects in Birmingham). Compared to the Britain, the variety of condition in each square is smaller than in whole Britain. The roads in the 'London region' (square 'tq') are illustrated in Figure 3.6, and the distribution of study subjects in a region near Woolwich is illustrated in Figure 3.7. The roads in the 'Birmingham region' (square 'sp') are illustrated in Figure 3.8.

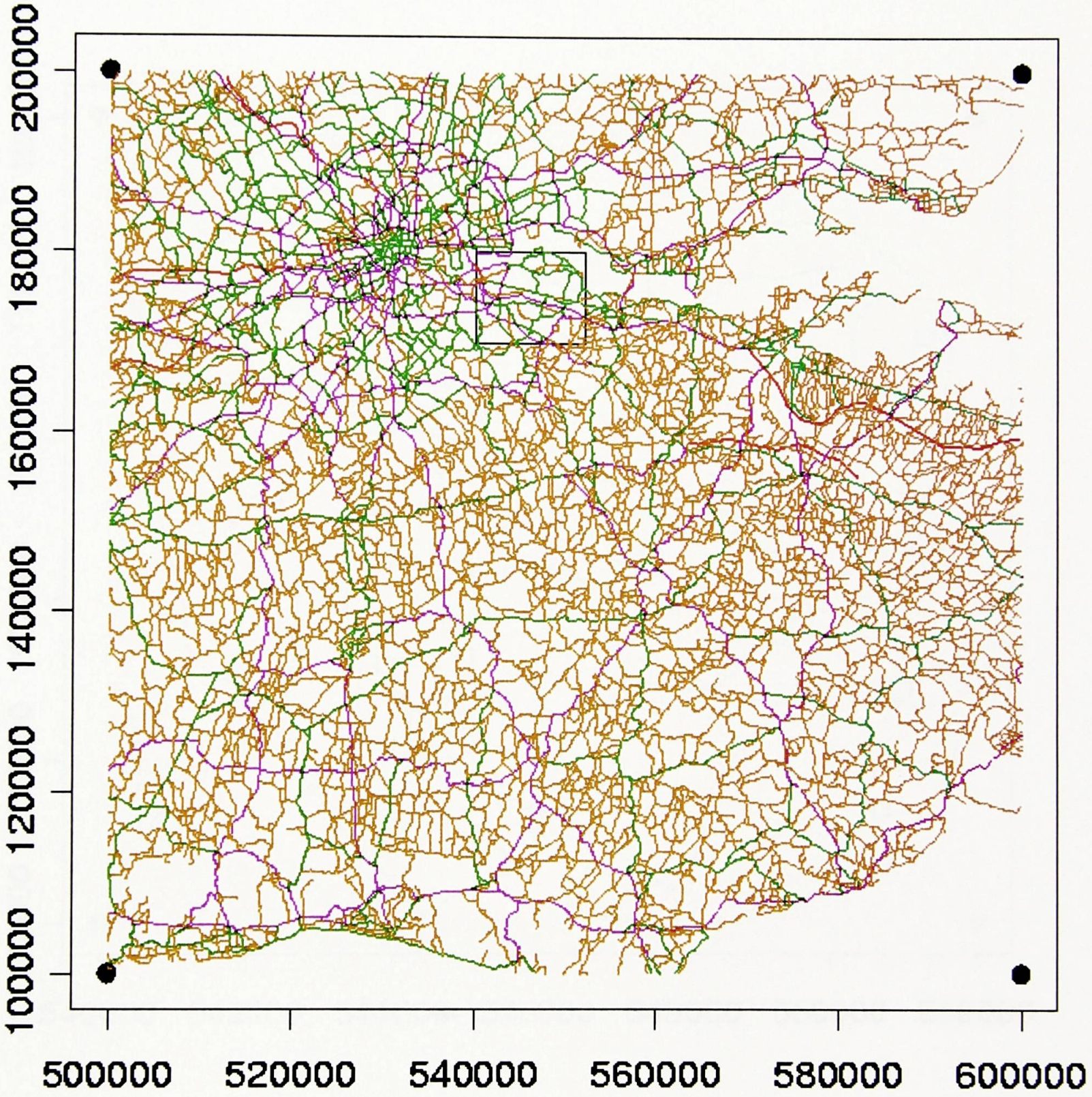


Figure 3.6 Roads in the 'London region' (square 'tq'). A region near Woolwich is indicated by a rectangle box. Code: red: motorway; pink: trunk road; green: principle road; yellow: minor road.

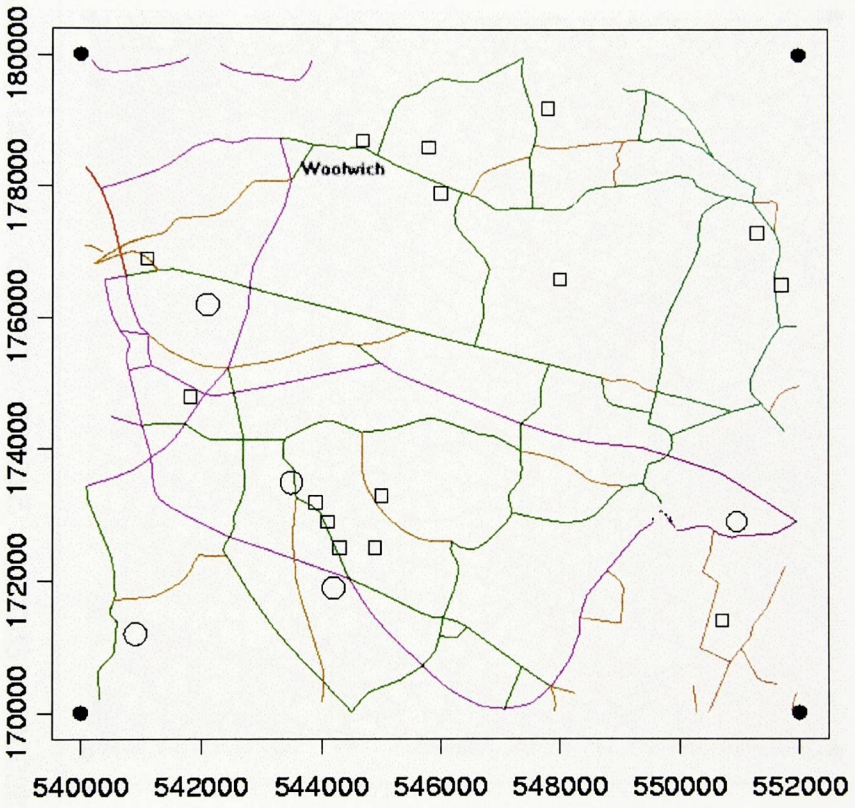


Figure 3.7 Spatial distributions of NCDS subjects in a region near Woolwich: Code: circle: asthmatic; square: not-asthmatic; red: motorway; pink: trunk road; green: principle road; yellow: minor road.

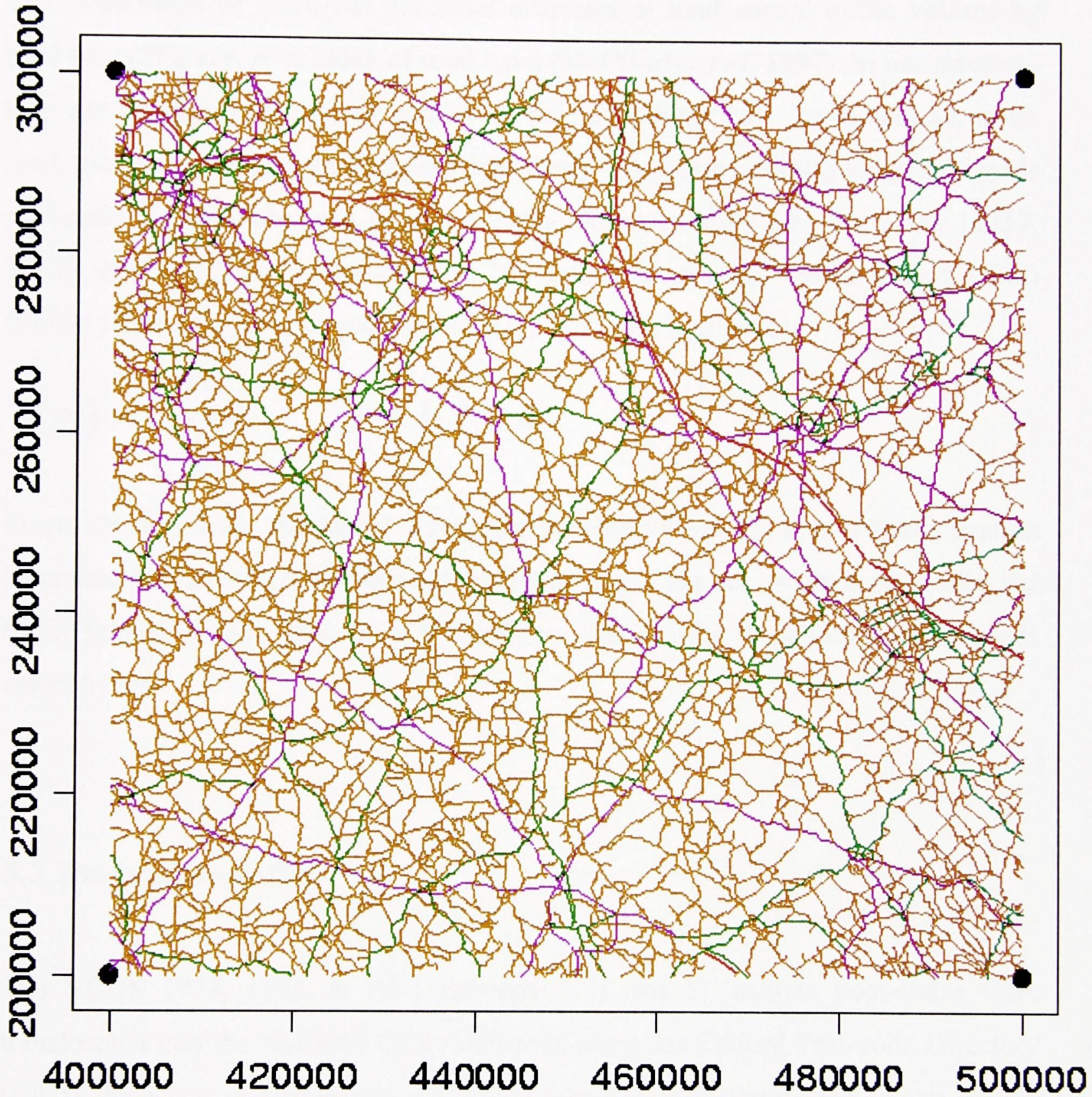


Figure 3.8 Roads in the Birmingham region. Code: red: motorway; pink: trunk road; green: principle road; yellow: minor road.

3.2.2 Traffic intensity

The Department of Transport publishes estimates of total annual traffic volume by road type (TV_m say, m is index of road type) (McCready *et al.* 1997). In our research, they are 60.95, 69.43, 127.06 and 74.20 (in 10^9 vehicle km) for motorway, trunk road, principle road and minor road respectively. The corresponding lengths of roads nationally (L_m say), obtained from the Bartholomew's digital map are 3100, 12357, 35629 and 16723 (km) respectively. The estimated average traffic intensity I(m) (traffic volume per km per year) for each type of road, is given by

$$I(m) = c\frac{TV_m}{L_m} \qquad m = 1,2,3,4$$

Since we only need the traffic intensities as relative values, c is a normalization parameter to make the intensities not too big or small. We use c=100 here. Hence, the traffic intensities are 1.966, 0.56, 0.35665, and 0.04532 for these four types of roads respectively.

3.3 The post-code file

The NCDS 1974, 1981 & 1991 (sweeps 3, 4 and 5), subject post-codes were transformed into the National Grid references using the Central Post-code Directory (CPD). This is a computerized file which contains every post-code in the Britain along with its corresponding grid reference accurate to within 100m. Normally a post-code covers a region along a road (or part of a road). The average number of delivery points per post-code is 14. Very few post-codes cover more than 50 delivery points (Raper *et al.* 1992). Grid references relate to the Southwest corner of the 100m square in which the first house (the first delivery point) of each post-code lies. (In our study, we change the grid reference to be centre of the square for reducing error and simplicity reason.) Hence, while the 'first' house of a post-code will have an

accuracy of ±50m, the others with the post-code will have an unknown accuracy level which may be greater than ±50m, depending on the extent of the post-code.

Chapter 4 Models of asthma in relation to traffic pollution

It is mentioned in Chapter 2 that most of studies on asthma and traffic pollution either study air pollution or road traffic. Research to study asthma and pollution exposure specified from road traffic needs to be carried out. And a more precise model for pollution also needs to be developed. However a physical model is very complicated. Physical models of the spread of particulate pollutants, either from 'one-shot' sources or continuously emitting sources, yield a range of models depending on how the vertical dimension, deposition, and drift are taken into account (Pasquill 1983). We have no access to such information. Another approach has to be adopted.

A concept of 'pollution exposure' is developed to measure the effect of pollution on a subject. Following several assumptions, a simple dispersal model for the pollution exposure is given. It measures the effect of pollution from a spatially distributed network source.

By using the conditional spatial location approach, non-linear logistic regression models are developed to describe the spatial distribution of asthma in relation to air pollution exposure due to road traffic. A logistic regression model is used as the asthma incidence model in which pollution exposures are covariates. Depending on how different road types are treated, several models are obtained. Maximum likelihood estimation (MLE) is used for estimation and the likelihood ratio test is used for inference.

The results are presented in this chapter. The limitations of our models are discussed, and how to deal with the measurement error issues, which will be studied in Chapter 6.

4.1. Pollution dispersal models

Observed spatial data on pollution level is not available in the context of this study. Hence physical models are not used in the analysis of NCDS asthma data, in relation to roads as a pollution source. Simulation of such physical models could make use of various information such as weather condition, wind speed and the intensity of pollutant at release in order to model the dispersion more accurately.

4.1.1 Physical models

Effective dispersion of gaseous or finely divided material released into the atmosphere near the ground depends on natural mixing processes on a variety of scales. In the main, this mixing is a direct consequence of turbulent and convective motions generated in the boundary layer itself.

Pasquill (1983) studied the dispersion of windborne material from industry and other sources. The dispersion was analysed in theory, experiment and real data.

The starting point of most mathematical treatments of diffusion from sources is a diffusion equation. Denoting by χ, the local concentration of mass per unit volume of fluid, the velocity of flow along the x, y, z axis by μ, ν, ω respectively, the diffusion equation is:

$$\frac{\partial \chi}{\partial t} = -[\frac{\partial(\mu\chi)}{\partial x} + \frac{\partial(\nu\chi)}{\partial y} + \frac{\partial(\omega\chi)}{\partial z}] \qquad (4.1)$$

Under different assumptions, the solutions of this equation can give models of the diffusion respectively.

In the simplest two-dimensional situation (no wind, no deposition, no height variable z), pollution from a point source at (0,0) is a Brownian motion, the distribution is a two-dimensional Gaussian distribution:

$$ce^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (4.2)$$

But the real situation is more complicated, since the pollution source (roads) and subjects are distributed over a large area during a long period. More information is needed for the use of physical models: models of time heterogeneity, local conditions (weather condition, wind speed, direction, height of local buildings, hills and the intensity of pollutant at release). Most of these are not available in the context. The complexity also will change the solution of the diffusion models. It makes the physical models difficult to use.

4.1.2 An Exposure model

4.1.2.1 Pollution exposure

The level of pollution at a particular point, due to a putative source of pollution at another point, is definable and in the term of the concentration of a particular pollutant gas (e.g. SO_2), the number of particles of soot pert unit volume or the particle size distribution of a particulate pollutant.

However, when we consider the effect of the pollution on an animal or human population, we are often interested in the occurrence of an adverse effect which may not be directly related to the local measure of pollution experience (either instantaneously or cumulatively). The particular aspect of pollution which is the most appropriate risk variable for the occurrence of the adverse event (which we term pollution exposure) will in general not be clearly known. This is the case with respect road pollution and its influence on the occurrence of asthma. However we do postulate that such a pollution exposure measure does exist, even we do not have means of deriving it from physical pollution measures.

Furthermore, in the investigation considered here, i.e. the effect of roads on childhood asthma, we have no measurements of pollution either at the road source, or over the environment over which it disperses. It follows therefore that we have no data/measurement of pollution exposure either at the road source, or in the spatial environment.

In accordance with the traditional principle of parsimony in statistical modelling we made some (null) assumptions of the way in which pollution exposure from a road is related to the type of road i.e. its traffic and the traffic intensity on the road. We adopt an assumption of proportionality between the pollution exposure at a point on a road and the traffic intensity on that road.

This may be regarded as our chosen definition of what we mean by 'pollution exposure' at a point on a road. The risk of occurrence of the adverse event in terms of the defined pollution exposure may take a simple form if this definition of pollution exposure is appropriate, or more complex forms if the dependency of pollution exposure/dose – at a road location were other than proportional to traffic intensity.

4.1.2.2 'Spread' of dispersal models

In the absence of pollution data, and information on local physical conditions we do not use physical models directly, but rather use statistical models which have some similarity with physical dispersal models. We make some assumptions about the pollution exposure dispersal:

1) There is no drift in the dispersal process and there is no wind, so that the distribution of pollution exposure will be spatially symmetric about the pollution source.

We considered several pollution exposure dispersal models from a point source:

(a) We use a symmetric bivariate normal distribution to model dispersion of pollution exposure. For a unit point source at $\mathbf{x}$, the pollution exposure at $\mathbf{x}_0$ is given by:

$$y' = s\frac{1}{2\pi\sigma^2}e^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|^2}{2\sigma^2}} \qquad (4.3)$$

where σ is a measure of the distance of spread of the pollutant from the source. s is the strength of the pollution source.

This model is similar to (4.2), the simplest pollution dispersal model.

(b) We have also considered models

$$y' = \frac{s}{\|\mathbf{x} - \mathbf{x}_0\|^2 + \sigma^2} \tag{4.4}$$

and

$$y' = se^{-c\|\mathbf{x}-\mathbf{x}_0\|} \tag{4.5}$$

Because (4.3) is easy for computing and has similarity with physical model (4.2), we select (4.3) as our pollution exposure dispersal model. The results we report are all from this model.

2) An additive model is adopted on the pollution exposure scale. Hence, the pollution exposure can be added or integrated.

It has been pointed out in (4.1.2.1) that the validity of such an assumed relationship is not known and no relevant data are available. Though such a simple assumption provides a simple null model framework for statistical analysis, it may be noted that it has some logical/mathematical limitations when compared to physical models.

Next, from assumptions 1) and 2) we calculate the pollution exposures from more complicate sources.

4.1.2.3 Combining the exposures from separate point sources

The natural way of combining pollution exposures at two or more point sources is additivity. It is the assumption 2).

Hence, if there are sources at locations $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k$, the strengths of sources are $s_1,\ldots, s_k$ respectively, the combined pollution exposure, y', for the subject at $\mathbf{x}_0$ is:

$$y' = \sum_{j=1}^{k} s_j \frac{1}{2\pi\sigma_j^2} e^{-\frac{1}{2\sigma_j^2}\|\mathbf{x}_0-\mathbf{x}_j\|^2} \tag{4.6}$$

4.1.2.4 Exposure from a road line-segment

Take $\mathbf{x}_0$ as the position of a subject, and ($\mathbf{x}_1$, $\mathbf{x}_2$) as a linear road-segment (see Figure 4.1).

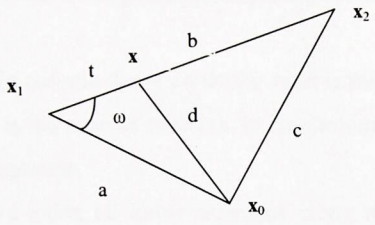


Figure 4.1 Generic subject (at x_0) in relation to a generic road-segment ($\mathbf{x}_1$, $\mathbf{x}_2$)

$t = \|\mathbf{x} - \mathbf{x}_1\|$, $a = \|\mathbf{x}_1 - \mathbf{x}_0\|, b = \|\mathbf{x}_1 - \mathbf{x}_2\|$, $c = \|\mathbf{x}_2 - \mathbf{x}_0\|$, $d = \|\mathbf{x} - \mathbf{x}_0\|$,

where $\| \|$ is the Gaussian distance and ω is the angle between ($\mathbf{x}_2$, $\mathbf{x}_1$) and ($\mathbf{x}_1$, $\mathbf{x}_0$),

Assume there is unit traffic intensity along ($\mathbf{x}_1$, $\mathbf{x}_2$). A very small length of road can be treated as a point pollution source, The strength of point $\mathbf{x}$ is d$\mathbf{x}$ (d$\mathbf{x}$ is the length of the small segment). If point pollution sources are on same type of road, then we may argue that the traffics they have are similar. The physical pollution produced from these points may be assumed similar. Ignoring difference in local conditions, the dispersal from the points is similar. Hence, the exposure models at all source points on the roads have same dispersal parameter. We integrate the pollution exposure from the elements of the road-segment to give the total exposure from the line-segment.

$$T = \int_{(\mathbf{x}_1, \mathbf{x}_2)} \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|^2}{2\sigma^2}} d\mathbf{x} \qquad (4.7)$$

where $\mathbf{x}$ is a point on the line-segment ($\mathbf{x}_1$, $\mathbf{x}_2$), and integration is from one end of the line-segment to the other. (for the integration of (4.7) refer to page 57, 58).

The exposure from a point on a road is assumed to be proportion to the traffic intensity at that point. If the traffic intensity along ($\mathbf{x}_1$, $\mathbf{x}_2$) is I, the pollution exposure will be IT.

4.1.2.5 Total pollution exposure from a road network

We assume that a particular road is made of linear segments along that road, at least it is the case of our data. In Bartholomew's Digital map, each road is made of linear segments.

We index all linear segments along roads of type m by j. Then the total pollution exposure from roads of type m say is obtained by summing over line-segments indexed by j. If the line segment j has traffic intensity $I_j(m)$, Then the total exposure y'_m is proportional to

$$y'_m = \sum_j I_j(m) T_j(m) \tag{4.8}$$

where $T_j(m)$ is calculated using parameter σ_m assuming unit traffic intensity. Parameter σ_m is the dispersal parameter for road type m. Here we made an approximation by using same dispersal parameter for same road type.

Finally, the total pollution exposure for a subject from all road types is given by

$$y' = \sum_{m=1}^{4} y'_m \tag{4.9}$$

same as

$$y' = \sum_{m=1}^{4} \sum_j I_j(m) T_j(m) \tag{4.10}$$

However we do not have the traffic intensity $I_j(m)$ of each road linear-segment. Only mean traffic intensities of each type of road are available, no detached information on particular roads is available. So we replace $I_j(m)$ by *I(m)* (*I(m)* is the average traffic intensity of road type m) as an approximation. If $I_j(m)$ were available, more detailed modelling of pollution exposure would be possible, and it does not increase the number of numerical calculations.

$$y' = \sum_{m=1}^{4} \sum_{j} I(m) T_j(m) \tag{4.11}$$

In this model, two approximations are made. One is using same dispersal parameter σ_m for same type of road. Second is using same traffic intensity *I(m)* for same type of road. The local conditions can be very different in the whole Britain. A smaller region has similar local condition. So London (square 'tq') and Birmingham (square 'sp') region are chosen for detailed study and the whole Britain region is used for simple study.

4.2 Models of pollution-exposure/asthma-incidence (PEAI)

From simple analyses in Section 3.1.2, outcome variable is whether the subjects have ever-asthma and sex is a covariate needed to be included. Denote $\mathbf{x}_i$ as the home location of subject *i*. The occurrences of asthma of subject *i* in the study are taken to be independent Bernoulli variables (Diggle and Rowlingson 1994), conditionally on the locations $\{\mathbf{x}_i\}_{i=1,\ldots,n}$, of the subjects. Let $p(\mathbf{y}(\mathbf{x}_i), \mathbf{y}'(\mathbf{x}_i))$ be the probability of subject *i* having asthma, where $\mathbf{y}(\mathbf{x}_i)$ is a vector of covariates (it is the gender variable in this study) and $\mathbf{y}'(\mathbf{x}_i) = (y_1'(\mathbf{x}_i), \cdots, y_4'(\mathbf{x}_i))$ is the vector of pollution exposures at $\mathbf{x}_i$ from the four road-types.

Four nested logistic models are formulated, which allow the differing road types to have differing effects on asthma incidence. Each uses only the gender variable as covariate.

Model 0: $\text{logit}(p) = \rho + \gamma\ sex$ (4.12)

Model 1: $\text{logit}(p) = \rho + \alpha y' + \gamma\ sex$ (4.13)

Model 2: $\text{logit}(p) = \rho + \alpha \sum_{m=1}^{4} y'_m + \gamma\ sex$ (4.14)

Model 3: $\text{logit}(p) = \rho + \sum_{m=1}^{4} \alpha_m y'_m + \gamma\ sex$ (4.15)

In these models, Model 1 uses same dispersal parameter σ for all road types whereas Models 2 and 3 allow four different σ's, i.e. fit σ_m for road type m where m = 1,...,4. Model 3 also allow different types of roads have different effects (α_m) on asthma.

4.3 The log-likelihood function; estimation and inference

4.3.1 Log-likelihood

Conditionally on the locations $\{\mathbf{x}_i\}_{i=1,\ldots,n}$, of the subjects, the log-likelihood function is:

$$L(\rho, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \gamma) = \sum_{i \in A} \log p(\mathbf{y}(x_i), \mathbf{y}'(x_i)) + \sum_{i \notin A} \log\left(1 - p(\mathbf{y}(x_i), \mathbf{y}'(x_i))\right) \quad (4.16)$$

where A={i: subject i has asthma}

Equation (4.16) is in general form. We give some detailed formulas for integration and differential.

For equation (4.7), denote $\mathbf{x}_0$ is the position of a person, ($\mathbf{x}_1$, $\mathbf{x}_2$) is a road segment, a, b, ω as in Figure 4.1, σ is the dispersal parameter of the road segment. We have

$$T = \int_{(\mathbf{x}_1, \mathbf{x}_2)} \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|^2}{2\sigma^2}} d\mathbf{x}$$

$$= \oint_{(\mathbf{x}_1, \mathbf{x}_1 + (\mathbf{x}_2 - \mathbf{x}_1))} \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1) - \mathbf{x}_0\|^2}{2\sigma^2}} d(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1))$$

$$= \int_0^b \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(a^2 + t^2 - 2at\cos\omega)} dt \tag{4.17}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{a^2 \sin^2 \omega}{2\sigma^2}} \int_0^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t - a\cos\omega)^2}{2\sigma^2}} dt$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{a^2 \sin^2 \omega}{2\sigma^2}} \left(\Phi\left(\frac{b - a\cos\omega}{\sigma}\right) - \Phi\left(-\frac{a\cos\omega}{\sigma}\right) \right) \tag{4.18}$$

$$\frac{\partial T}{\partial \sigma}$$

$$= \frac{a^2 \sin^2 \omega}{\sigma^3} T - \frac{1}{\sigma} T$$

$$- \frac{1}{2\pi} \frac{b - a\cos\omega}{\sigma^3} e^{-\frac{c^2}{2\sigma^2}} - \frac{1}{2\pi} \frac{a\cos\omega}{\sigma^3} e^{-\frac{a^2}{2\sigma^2}} \tag{4.19}$$

The log-likelihood function is

$$\mathsf{L} = \sum_{i \in A} \ln \frac{e^{f_i}}{1 + e^{f_i}} + \sum_{i \notin A} \ln \frac{1}{1 + e^{f_i}}$$

$$= \sum_{i \in A} f_i - \sum_i \ln(1 + e^{f_i}) \tag{4.20}$$

in which

A={i: subject i has asthma}

For model 1 and 2

$$f_i = \rho + \alpha \left(I(1) \sum_j T_{i1j} + I(2) \sum_j T_{i2j} + I(3) \sum_j T_{i3j} + I(4) \sum_j T_{i4j} \right) + \gamma \, sex(i)$$

For model 3,

$$f_i = \rho + \alpha_1 I(1)\sum_j T_{i1j} + \alpha_2 I(2)\sum_j T_{i2j} + \alpha_3 I(3)\sum_j T_{i3j} + \alpha_4 I(4)\sum_j T_{i4j} + \gamma\, sex(i)$$

in which, T_{imj} is the pollution exposure of subject i gets from road segment j of road type m, (all the line-segments on all the roads of type m is indexed by j).

4.3.2 Estimation

Maximum likelihood estimation (MLE) is used for parameter estimation. For a log-likelihood function $\mathrm{L}(x|\theta)$, there is a property of MLE (McCullagh and Nelder 1989).

In multi-dimensional situation:

The consistent solution $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_n)$ of the likelihood equations

$$\frac{\partial \mathrm{L}(x|\theta)}{\partial \theta_j} = 0 \qquad j = 1,2,\cdots n$$

under certain regularity conditions , has a joint distribution of $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_n)$ which is asymptotically normal with covariance matrix (v_{rs})

$$(v_{rs})^{-1} = \left(-E\frac{\partial^2 \mathrm{L}}{\partial\theta_i \partial\theta_j} \right)$$

To get the expectation of second derivative is difficult. We use the second derivative value $(\frac{\partial^2 \mathrm{L}}{\partial\hat{\theta}_i \partial\hat{\theta}_j})$ at the estimated parameters from our data as approximations.

For model 3, we give the first and second derivatives.

The derivatives of the log-likelihood function (4.20) are

$$\frac{\partial \mathrm{L}}{\partial} = \sum_{i\in A} \frac{\partial f_i}{\partial} - \sum_i \frac{e^{f_i}}{1+e^{f_i}} \frac{\partial f_i}{\partial} \qquad (4.21)$$

in which

$$\frac{\partial f_i}{\partial \rho} = 1$$

$$\frac{\partial f_i}{\partial \alpha_m} = I(m)\sum_j T_{imj} \qquad m=1,2,3,4$$

$$\frac{\partial f_i}{\partial \sigma_m} = \alpha_m \cdot I(m) \cdot \sum_j \frac{\partial T_{imj}}{\partial \sigma_m} \qquad m=1,2,3,4$$

$$\frac{\partial f_i}{\partial \gamma} = sex(i)$$

The second derivatives of the log-likelihood function are:

$$\frac{\partial^2 \mathrm{L}}{\partial x_1 x_2} = \sum_{i \in A} \frac{\partial^2 f_i}{\partial x_1 \partial x_2} - \sum_i \frac{e^{f_i}}{(1+e^{f_i})^2} \frac{\partial f_i}{\partial x_1} \frac{\partial f_i}{\partial x_2} - \sum_i \frac{e^{f_i}}{1+e^{f_i}} \frac{\partial^2 f_i}{\partial x_1 \partial x_2} \qquad (4.22)$$

in which

1) $\frac{\partial^2 f_i}{\partial x \partial \rho} = 0 \qquad x = \rho, \alpha_m, \gamma, \sigma_m \qquad m=1,2,3,4$

2) $\frac{\partial^2 f_i}{\partial x \partial \alpha_m} = 0 \qquad x = \rho, \alpha_m, \gamma \qquad m=1,2,3,4$

$$\frac{\partial^2 f_i}{\partial \sigma_k \partial \alpha_m} = \begin{cases} 0 & m \neq k \\ I_k \sum_j \frac{\partial T_{ikj}}{\partial \sigma_k} & m = k \end{cases}$$

3) $\frac{\partial^2 f_i}{\partial \rho \partial \sigma_m} = 0 \qquad x = \rho, \gamma \qquad m=1,2,3,4$

$$\frac{\partial^2 f_i}{\partial \alpha_m \partial \sigma_k} = \begin{cases} 0 & m \neq k \\ I_k \sum_j \frac{\partial T_{ikj}}{\partial \sigma_k} & m = k \end{cases}$$

$$\frac{\partial^2 f_i}{\partial \sigma_l \partial \sigma_k} = \begin{cases} 0 & l \neq k \\ \alpha_k I_k \sum_j \frac{\partial^2 T_{ikj}}{\partial^2 \sigma_k} & l=k \end{cases}$$

4) $\frac{\partial^2 f_i}{\partial x \partial \gamma} = 0 \qquad x = \rho, \alpha_m, \gamma, \sigma_m \qquad m=1,2,3,4$

4.3.3 Inference

The Likelihood Ratio Test is used for inference.

For model 1 a likelihood ratio test of $\alpha = 0$ would involve a comparison between the test statistic $D = 2\{L(\hat{\rho},\hat{\alpha},\hat{\sigma},\hat{\gamma}) - L_0(\hat{\rho},\hat{\gamma})\}$ and critical value of the χ^2-distribution on 2 degrees of freedom.

For model 2 a likelihood ratio test of $\alpha = 0$ would involve a comparison between the test statistic $D = 2\{L(\hat{\rho},\hat{\alpha},\hat{\sigma}_1,\hat{\sigma}_2,\hat{\sigma}_3,\hat{\sigma}_4,\hat{\gamma}) - L_0(\hat{\rho},\hat{\gamma})\}$ and critical value of the χ^2-distribution on 5 degrees of freedom.

For model 3 a likelihood ratio test of $\alpha_1 = 0$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\alpha_4 = 0$ would involve a comparison between the test statistic:

$D = 2\{L(\hat{\rho},\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4,\hat{\sigma}_1,\hat{\sigma}_2,\hat{\sigma}_3,\hat{\sigma}_4,\hat{\gamma}) - L_0(\hat{\rho},\hat{\gamma})\}$ and critical value of the χ^2-distribution on 8 degrees of freedom. Critical values for χ^2 distribution are shown in Appendix B.

4.4. The Software environment

The computers we used for computing in our research are several SUN workstations (SPARC 5 or ULTRA 2). The operator system is UNIX (i.e. Solaris). They are public machines in the University of Greenwich. Sometime multi-tasks are running on them. In this study, various software are used. The road map (Bartholomew's Digital map of Great Britain) is stored as GIS information. ARC/INFO is used to access, display and transfer road map data. Several C programmes are written for file format transformation. S-plus is used to draw figures. SPSS is used for some simple analyses presented in Chapter 3. FORTRAN is used to do the numerical calculation during estimation and inference.

We choose Fortran to do main computation because its speed on float calculation and its rich libraries. We use NAG (Numerical Algorithms Group) to do most of difficult numerical calculations. NAG is a mathematical library written in FORTRAN. It is developed by the Numerical Algorithms Group (NAG), based in Oxford, UK, as a library to solve numerical and statistical problems.

Subroutine E04KAF is used to maximize the likelihood function. It optimises a function which has several variables with simple bounds using a quasi-Newton

algorithm. It requires the first derivative, which we supply. E04XAF is used to calculate the second derivative / Hessian matrix of a function at a particular point. It does not require the first derivative.

4.5. Results

4.5.1 Study areas and parameter constraints

Numerical difficulties

The parameters of the models are estimated by maximizing the log-likelihood function using NAG. The standard error of the estimated parameter can be calculated from the inverse of Hessian matrix. Numerical approach is used to calculate second derivatives. NAG routine E04XAF uses numerical algorithm to calculate the second derivatives from log-likelihood function. However, the data in whole Britain is too large. This algorithm does not converge on Britain data. Standard errors cannot be obtained in this case.

On the other hand, it was considered that the very different environmental conditions in the different regions of the Britain would mean that different local parameterisations would be appropriate in different regions. It is therefore considered most appropriate to analyse the data separately in a number of sub-regions. Two regions were selected, each with size 100km x 100km, based around London and Birmingham. In these two cases, Hessian matrix and standard errors can be obtained.

Asthma prevalence

Ever-asthma prevalence in these regions is illustrated in Tables 4.1 and 4.2 and is about 30% in both regions, with the clear sex difference in both regions.

	Male	Female	Total
Ever-asthma	164	119	283
Non-asthma	381	382	763
Total	545	501	1046

Table 4.1 Ever-asthma prevalence in London area

	Male	Female	Total
Ever-asthma	68	63	131
Non-asthma	124	147	271
Total	192	210	402

Table 4.2 Ever-asthma prevalence in Birmingham area

	Male	Female	Total
Ever-asthma	2142	1569	3711
Non-asthma	4325	4406	8731
Total	6467	5975	12442

Table 4.3 Ever-asthma prevalence in Britain

Parameter constraints

When fitting the models for these regions another numerical problem was encountered. The optimization routines generally assigned one of the dispersal distance parameters to 'essentially' zero, corresponding to a massive spike in the effect of one type of road. The reason for this was that the large number of subjects, and the large road network, usually resulted in just a few subjects who had asthma being assigned a location which was essentially coincident with a road. Because the levels of accuracy of both the post-code and road locations not being more than ±50m, the coincidence of locations is leading to a spurious effect. To avoid this

outcome the optimizations were conducted with a constraint that the dispersal distance parameter exceeded 100m. On the other hand, we assume 5km is the maximum value for dispersal parameter. So σ_m is between 0.1 and 5 (unit is km). The α_m parameters were constrained to be greater than or equal to zero.

4.5.2 Estimated models in London and Birmingham

The estimated parameters of the fitted models, and their standard errors, are shown in Tables 4.4 and 4.5 for the London and Birmingham regions respectively.

It can be seen from the very small increase in log-likelihood from the null model, model 0, to the simplest of road pollution models, model 1, with a pollution dispersal parameter of 0.98km, that this simple model of road pollution effect has no asthma explanatory power. Figure 4.2(a) shows the distribution of exposure values over all subjects in the London region, using the fitted parameter values for model 1. Figures 4.2(b) and 4.2(c) show the distributions of exposures obtained if the dispersal parameter is halved, and doubled, respectively, the other parameters being kept fixed. It can be seen that the distribution of exposures is relatively insensitive to variations in the pollution dispersal parameter. This is likely to be part of the reason that the fitted models are so poor. Figure 4.3(a) and 4.3(b) show the distributions of estimated exposures, under the fitted model 1 in the London region, for those who have ever-asthma, and those who have not. It would seem that those having asthma have lower exposure levels than those who do not have asthma. It is therefore not surprising that the pollution effect models do not have a high explanatory power for asthma occurrence.

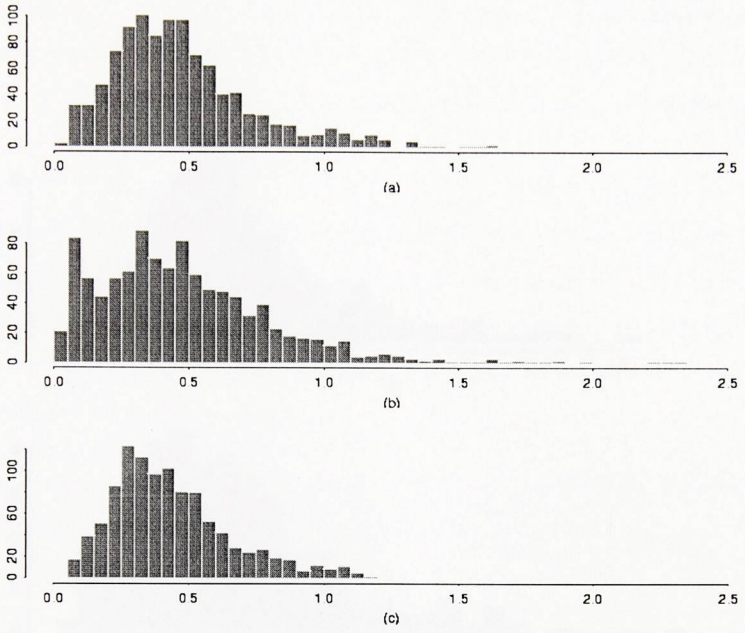


Figure 4.2 Frequency distribution (in the form of histogram) of exposure values over NCDS subjects in the London region, using model 1: (a) σ=0.98km; (b) σ=0.49km; (c) σ=1.96km.

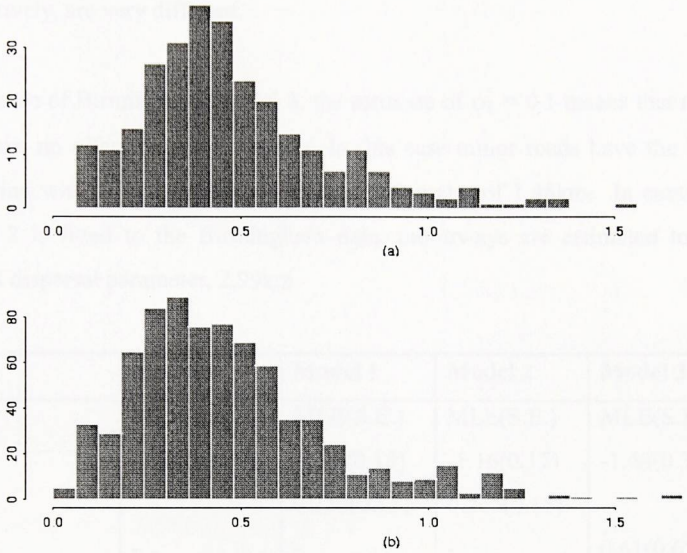


Figure 4.3 Frequency distribution of exposure values over NCDS subjects in the London region, (a) asthmatics, (b) not-asthmatics

Parameter estimates on a constraint boundary do not have asymptotic standard errors. In such cases, (London, model 2, σ_2 = 0.1; London, model 3, α_2 =0; Birmingham, model 3, σ_1 = 0.1) the standard errors of the other parameters are obtained from the profile likelihood with respect to those other parameters.

In the case of London dataset, for model 2, σ_2 equals to the minimal constraint value of 100m (0.1km). This corresponds to a very small dispersal parameter for pollution exposure from trunk roads and to an effect on relatively few of the subjects in the study. This may be contrasted with the dispersal parameter estimates for the other types of roads, which range between 0.1 and 1 km. For the London dataset, model 3,

α_2 equals zero, so that again trunk roads will have no pollution exposure effect on the subjects in the study. The relatively high values for α_3 and α_4 mean that principal and minor roads have relatively larger effect than motorways and trunk roads, though the estimated dispersal parameters for principal and minor roads, 2.34km and 0.8km respectively, are very different.

In the case of Birmingham, model 3, the estimate of $\sigma_1 = 0.1$ means that motorways will have no effect on most subjects. In this case minor roads have the highest α-weighting with a high estimated dispersal parameter of 1.48km. In contrast, when model 2 is fitted to the Birmingham data, motorways are estimated to have the highest dispersal parameter, 2.99km.

	Model 0	**Model 1**	**Model 2**	**Model 3**
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.84(0.009)	-0.96(0.15)	-1.16(0.17)	-1.48(0.30)
α	-	0.258(0.15)	0.265(0.15)	-
α_1	-	-	-	0.61(0.63)
α_2	-	-	-	0
α_3	-	-	-	2.29(1.3)
α_4	-	-	-	2.34(1.4)
σ	-	0.98(1.16)	-	-
σ_1	-	-	1.15(0.71)	1.14(0.79)
σ_2	-	-	0.1	1.07
σ_3	-	-	1.92(1.48)	2.34(1.58)
σ_4	-	-	1.08(1.67)	0.8(0.55)
γ	-0.32(0.02)	-0.33(0.14)	-0.34(0.14)	-0.35(0.14)
Log-likelihood	-607.99	-607.52	-605.20	-604

Table 4.4 Analysis result for London using different models

	Model 0	Model 1	Model 2	Model 3
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.6(0.023)	-0.86(0.18)	-1.05(0.26)	-1.06(0.45)
α	-	0.59(0.4)	0.97(0.76)	-
α_1	-	-	-	1.07(0.88)
α_2	-	-	-	0.53(0.25)
α_3	-	-	-	1.32(0.75)
α_4	-	-	-	3.27(1.05)
σ	-	0.72(0.63)	-	-
σ_1	-	-	2.99(1.8)	0.1
σ_2	-	-	1.1(0.87)	0.87(0.61)
σ_3	-	-	0.33(0.19)	0.38(0.22)
σ_4	-	-	1.31(0.91)	1.48(0.96)
γ	-0.247(0.05)	-0.224(0.17)	-0.199(0.14)	-0.19(0.16)
Log-likelihood	-253.87	-251.9	-250.6	-249.6

Table 4.5 Analysis results for Birmingham using different models

All sub-models of model 3, in which subsets of road types are assigned a zero effect, were fitted as well (Table 4.6). The best model of this form was that in which only principal roads featured, with parameter estimates:

$\rho = -1.117\ (0.168)$, $\alpha_3 = 1.36(0.83)$, $\sigma_3 = 1.98km\ (1.82km)$ and $\gamma = -0.33$

However the p-value for the test of this model against the model 0 is 0.15.

We use SPSS to do the Hosmer and Lemeshow goodness-of-fit test (Hosmer and Lemeshow 1989, divided all subjects into several groups, used chi-square test to test the goodness-of-fit by fitting observed case numbers with expected case numbers under the models) for models 1, 2, 3 on London data. The significance levels are 0.6958, 0.6596, 0.6387 respectively.

	road1,3,4	road 3,4	road1,3	Road1,4	road 3	road 4
	MLE	MLE	MLE	MLE	MLE	MLE
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
ρ	-1.49	-1.50(0.28)	-1.13	-0.89	-1.12(0.17)	-0.87
α						
α_1	0.62		0.71	0.79		
α_2						
α_3	2.30	2.54(0.98)	1.26		1.36(0.83)	
α_4		6.30(1.18)		0.59		0.7
σ						
σ_1	1.14		1.19	1.24		
σ_2						
σ_3	2.35	2.58(0.15)	1.75		1.98(1.82)	
σ_4		0.80(0.05)		0.30		0.3
γ	-0.36	-0.35(0.14)	-0.34	-0.36	-0.33(0.14)	-0.32
Log-likelihood	-604.2	-604.60	-605.25	-607.04	-605.98	-607.9

Table 4.6 Analysis results for London using different sub-models

4.5.3 Estimated models in Britain

We also fitted the models 0, 1, 2, 3 for whole Britain (Table 4.7). As we mentioned early, standard errors can not be obtained due to convergence problem. The p-value for model 1 is just above 0.05 (Appendix B). Though they are still not significant, the significance level is higher than models fitted for London and Birmingham area. Figure 4.4(a) and 4.4(b) show the distribution of estimated exposures, under the fitted model 1 for Britain region, for those who have ever-asthma, and those who have not.

	Model 0	**Model 1**	**Model 2**	**Model 3**
	MLE (S.E.)	MLE	MLE	MLE
ρ	-0.87(0.04)	-0.99	-0.98	-0.97
α	-	0.45	0.6	
α_1	-	-		0.97
α_2	-	-		0.18
α_3	-	-		1.05
α_4	-	-		0
σ	-	2.92		
σ_1	-	-	3.04	4.23
σ_2	-	-	5	1.56
σ_3	-	-	3.69	5
σ_4	-	-	5	2.26
γ	-0.35(0.07)	-0.35	-0.35	-0.35
Log-likelihood	-4319	-4316.2	-4315.9	-4315.5

Table 4.7 Analysis results for Britain

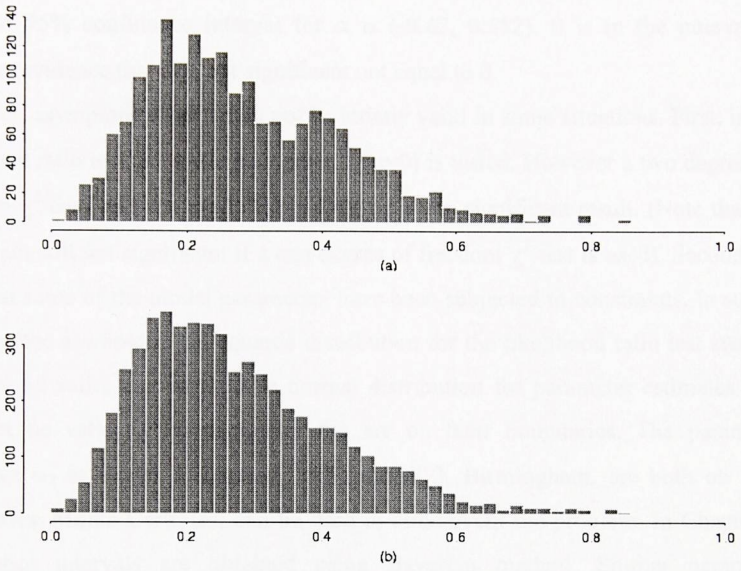


Figure 4.4 Frequency distribution of exposure values over NCDS subjects in Britain (a) asthmatics, (b)not-asthmatics

4.6 Discussion

The p-values from comparing any of the more complex models with the null spatial effect model, model 0 in different areas (London, Birmingham, Britain), are all greater than 0.05. Hence we cannot say, for either the London, Birmingham regions or Britain, there is strong evidence of a relationship between ever-asthma at age 16 and pollution caused by traffic.

The test of α=0 can be achieved in two ways. One is using the Likelihood Ratio Test; one is using the confidence interval (for 5% test, the interval is $\alpha \pm 1.96 \cdot s.e.$). The results from these two methods are consistent. For example for model 1 for London area (Table 4.4), the likelihood ratio test is not significant. The standard error of α is 0.15, so 95% confidence interval for α is (-0.42, 0.552). 0 is in the interval. It provides evidence that α is not significant not equal to 0.

However, asymptotic theory may not be strictly valid in some situations. First, in the likelihood ratio test, only one parameter α (α=0) is tested. However a two degrees of freedom χ^2 test has been used (Table 4.4), with non significant result. (Note that the results are still not significant if a one degree of freedom χ^2 test is used). Second, we note that some of the model parameters have been subjected to constraints. In such a situation the asymptotic chi-squared distribution for the likelihood ratio test statistic may not be valid, and asymptotic normal distribution for parameter estimates may also not be valid, when parameter fits are on their boundaries. The parameter estimates α_2 in model 3, London, σ_1 in model 3, Birmingham, are both on their boundaries. Monte Carlo test can be used to circumvent the problem. In Chapter 7, confidence intervals are obtained using Bayesian method. Similar parameter estimates are obtained and no significance is found there.

The test in Britain is more significant than tests in the London and Birmingham region. This leads us to think whether the number of subjects is large enough for the test to pick the significance. Chapter 5 will calculate the power for the test.

There are measurement errors in the variables in our models, such as the locations of houses and roads. The locations of homes of subjects living are derived from their post-code. The post-code covers an area grid references relates to the Southwest corner of the 100m square in which the first house of each post-code lies. Each house location we use has measurement error. Also the road map is likely to be no more accurate than the post-code locations of subjects. In Chapter 6, a measurement error model is developed to tackle these issues.

Chapter 5 Power analysis

From the statistical analyses of the data in London and Birmingham regions, with models ((4.13), (4.14), (4.15)) in Chapter 4, no significant effect of road traffic pollution on asthma incidence was found.

The nature of significance test is such that it is possible for an 'important' actual effect to be undetected on use of a hypothesis test (of size α) of a given sample size, if the sample size is not large enough. Hence it is important to know the probability of detecting an important actual effect (i.e. the alternative H_1) with a significance test (i.e. the power of the test). If the power is too small, then more samples are needed.

The general theory of power (Hodges and Lehmann 1970) is summarised:

If the null hypothesis is H_0, alternative hypothesis is H_1, a decision is made according to: if a test statistic $T \in \Omega_1$ (the critical region), we reject H_0; otherwise, we accept H_0. The significance level of the test is $P_{H_0}(T \in \Omega_1)$. It is the probability of the error (called type I error) that H_0 is rejected while H_0 is true. If we only worry about minimizing this error, we could simply decide to always accept the hypothesis H_0. This will reduce the type I error to zero.

There is another kind of error (type II error), which is the error that H_0 is mistakenly accepted while it is false. The probability of the error is $\{1 - P_\theta(T \in \Omega_1)\}$, where θ is actual parameter value. It is also a serious error which needs to be controlled. Type II error often happens when the number of samples is not enough. For each particular alternative A, we can calculate the probability of type II error, β. $1-\beta$ is called the *power* of the test against alternative A. Denote it by π. It is the probability of rejecting the null hypothesis when alternative A is true, and hence of correctly detecting the null hypothesis H_0 to be false.

The method used here to calculate the power of a test for alternative A is as follows: Assume an alternative A is true, simulate the outcomes of samples a large number of

times (let us say n), and then apply the test to the simulation results. Hence the number of rejecting the null hypothesis H_0 is obtained, let say m. Then m/n is an approximation for the power of the test of H_0, against the alternative A.

If the power for the entire set of possible alternative A's is obtained, a 'power curve' can be drawn, which plots the estimated power against the actual value of the parameter.

Power analysis, through examination of the power curve, can be used to determine the efficiency of a test (Waller and Lawson 1995) or determine whether the number of cases is large enough.

Because the process is very time consuming, the power analysis is only done for model 1 in London, and Britain.

5.1 Power analysis for London

For the London region, model 1 (4.13) is

$$\text{logit}(p) = \rho + \alpha\, y'(\sigma) \quad + \quad \gamma\, sex \qquad (5.1)$$

The estimated values of parameters are ρ= -0.96, α=0.258, σ=0.98, γ= -0.33 (from Table 4.4).

To study the power of our analyses of the relationship between asthma and road traffic pollution is to study the power against α under null hypothesis H_0: α=0.

Assuming model 1 is the true model (i.e. the estimates of ρ, σ, and γ are real parameter values), the power against α is the 'conditional power' under estimated values of ρ, σ, γ.

Let α change from 0.0 to 1.5 (α=0 corresponds to no association between road pollution and asthma; α=1.5 corresponds to strong association). α is incremented by 0.01. The state of each person (asthmatic or non-asthmatics) can be simulated using (5.1). Then the likelihood ratio test is applied to test whether H_0 is significant (5% or 1%). For each α, 1000 realizations of the asthma states of the population are simulated. The proportions of the 1000 simulations for which the null hypothesis (α=0) is rejected at a 5% and 1% test are calculated. The result is shown in Figure 5.1

(square for 5%, dot for 1%). When α is 1.0, power of a 5% test is 99%. In our model $\alpha = 0.258$. Hence the approximate power for a 5% level test is 20%, and for 1% level test is 10%. This power value is very small. The 5% test is very unlikely to detect an association between asthma and road traffic pollution, with $\alpha = 0.258$.

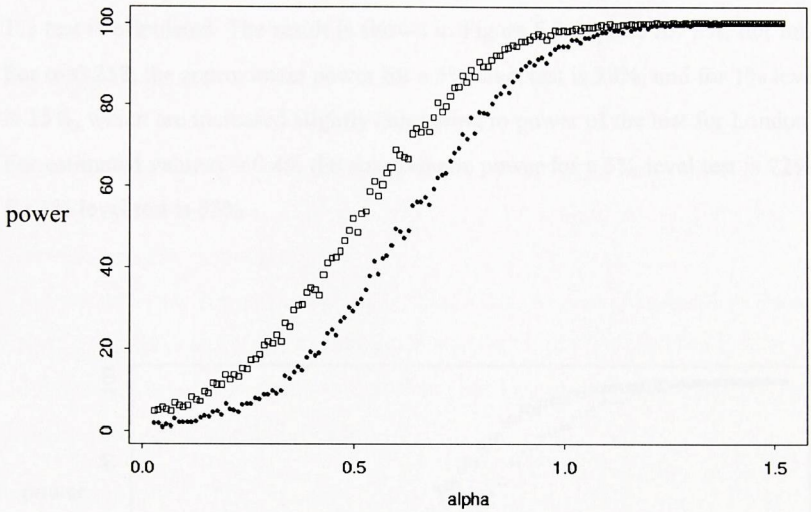


Figure 5.1 Power analysis of London area using model 1

Key: Square – 5% test

Dot – 1% test

The number of subjects in London is 1046. It is possible that model 1 is realistic and the parameter values might be real, actual, but that the number of subjects is too small to detect $\alpha = 0.258$ as significant with a 5% test.

Hence, the power of the test on an area with larger number of subjects: the whole Britain area is also studied.

5.2 Power analysis for Britain

Though the same estimated values of ρ, σ, γ (ρ= -0.96, α=0.258, σ=0.98, γ= -0.33, which were estimated from the London area) could be used in the whole Britain, the estimates from Britain (ρ= -0.99, α=0.45, σ=2.92, γ= -0.3527) are used.

We let α change from 0.0 to 1.0, and incremented by 0.01. For each α, 1000 realizations of the asthma status of the population are simulated, then the proportion of the 1000 simulations for which the null hypothesis (α=0) is rejected with a 5% or 1% test is calculated. The result is shown in Figure 5.2 (square for 5%, dot for 1%). For α=0.258, the approximate power for a 5% level test is 28%, and for 1% level test is 15%, which are increased slightly comparing to power of the test for London data. For estimated value α = 0.45, the approximate power for a 5% level test is 72%, and for 1% level test is 53%.

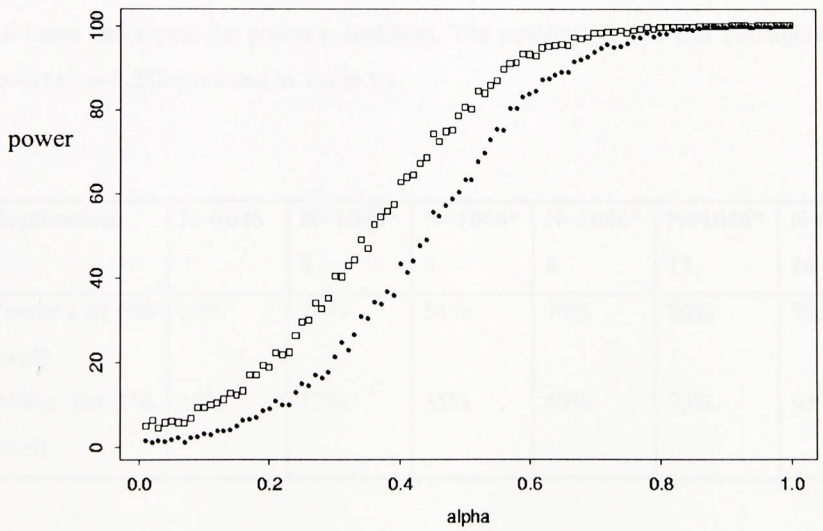


Figure 5.2 Power analysis of Britain area using model 1

Key: Square – 5% test

Dot – 1% test

On moving from the London area to Britain, the power of the test increases as expected. Considering the wide variation of environment in different areas in Britain, we conjecture that if the increased numbers of subjects (as for Britain) were observed in London, then the increase of power should be bigger than in this study. Thus, we suggest that a survey with a larger number of subjects and less variables comparing to NCDS, is needed for a more conclusive study.

5.3 Sample size

Heieh (1989) and Whitemore (1981) developed a method to estimate sample size for simple logistic regression. Their method cannot be applied to our non-linear model due to complexity of our model.

To determine how large the sample size should be to achieve satisfactory power, the London data is used for an experiment. We replicate the London data 2, 4, 8, 12 and 16 times and repeat the power calculation. The powers (for 5% and 1% significant level) at α=0.258 are listed in Table 5.1.

Replications	**N=1046**	**N=1046* 2**	**N=1046* 4**	**N=1046* 8**	**N=1046* 12**	**N=1046* 16**
Power (at 5% level)	20%	27%	54%	74%	86%	99.5%
Power (at 1% level)	10%	17%	35%	59%	73%	93%

Table 5.1 Power for different sample sizes, for London; α=0.258.

From Table 5.1, to achieve a high power of a 5% test for model 1 in the London area (>95%), the sample size need to be about 16,000. The power curve is plotted when n=1046*16=16736 in Figure 5.3.

It is noted that in chapter 4, the p-value for the test of H_0: $\alpha=0$ are 0.35 and 0.44 for models 2 and 3 with a 5% test. Since they are lower than the corresponding p-value (0.65) for model 1 for London. As analysed in this chapter, to achieve satisfactory power, a lower sample size would be required with model 2 and 3 than model 1.

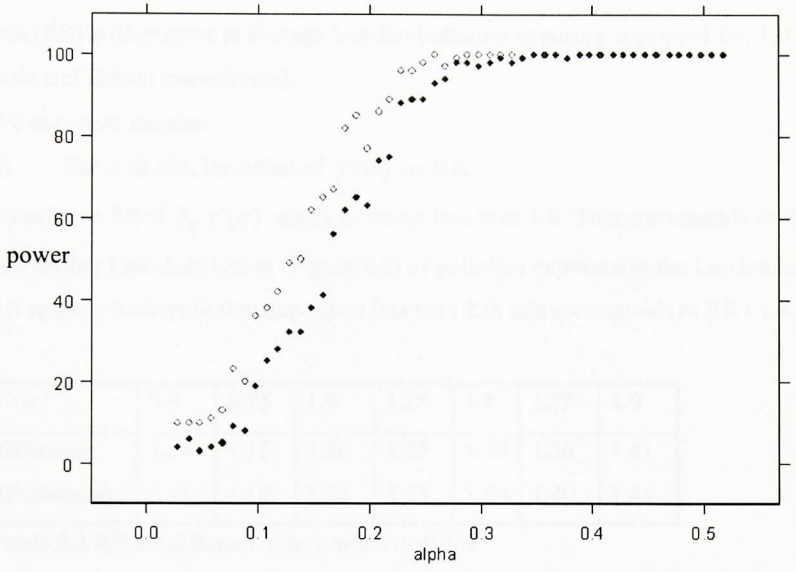


Figure 5.3 Power analysis of repeated London data

Key: Circle – 5% test

Dot – 1% test

5.4 Relative risk

Finally, we need to consider what value of α, (in model 1), would be epidemiologically important to detect using a significance test.

Model 1 may be written as

$$p = \frac{1}{1+e^{-(\rho+\alpha y'(\sigma)+\gamma sex)}}$$

where p is the probability has asthma.

The average value of $y'(\sigma)$ for the model 1 estimates is 0.41 with parameter estimates ρ= -0.96, α=0.258, γ= -0.33. This means that the probability of an individual with average London pollution exposure having asthma is p_1=0.298, 0.234 (for male and female respectively). This to be compared with probabilities p_2= 0.277, 0.215 (for male and female respectively), if α=0, (or $y'(\sigma)$ =0). Hence the relative risk (RR) with respect to average London pollution exposure is p_1/p_2=1.07, 1.09 (for male and female respectively).

We may now enquire:

(i) For α=0.258, the effect of $y'(\sigma)$ on RR.

To achieve RR=1.2, $y'(\sigma)$ needs to be no less than 1.0. This corresponds to the 95 percentile of the distribution (Figure 4.2) of pollution exposure in the London region. All subjects have pollution exposures less than 2.0; this corresponds to RR =1.4.

$y'(\sigma)$	**0.5**	**0.75**	**1.0**	**1.25**	**1.5**	**1.75**	**2.0**
RR(male)	1.09	1.15	1.20	1.25	1.30	1.36	1.41
RR(female)	1.11	1.16	1.22	1.28	1.34	1.40	1.46

Table 5.2 RR for different $y'(\sigma)$ when α=0.258

(ii) For estimated $y'(\sigma)$ for London, the effect of α on RR.

α	**0.5**	**1.0**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**
RR(male)	1.15	1.32	1.50	1.68	1.86	2.05	2.83
RR(female)	1.17	1.35	1.56	1.78	2.01	2.25	3.35

Table 5.3 RR for different α for estimated $y'(\sigma)$

For the Britain data, average $y'(\sigma)$ is 0.236 (Figure 4.4) with ρ= -0. 99, α=0. 45, γ= -0. 35, the RR can be calculated. It is 1.08, 1.09 (for male and female respectively). It is similar to the result from London, though the estimated parameters (α, σ) for London and Britain have large differences. The reason is that $\alpha\, \overline{y}'(\sigma)$ is similar in

both data (0.106, 0.108 for London and Britain respectively). It is expected if the model fit the data.

Chapter 6 Measurement error models

Subjects do not live entirely at locations corresponding to their home locations. They spend a lot of time in school and other places. In addition they are also exposed to traffic pollution on the road between school and home. But no such information is currently available in the NCDS.

In NCDS data, the locations where subjects are living are defined by their houses. The house locations are derived from their post-code. A post-code covers an area referenced from the Southwest corner of the 100m square in which the first house of the post-code lies. So, each house location has 'measurement error'. This chapter introduces the error-in-variable model (Cochran 1968, Cox and Dolby 1977, Dolby 1976, Hoschel 1989) and investigates the effect of inaccuracy of the house location in detail. A measurement error model is developed to tackle the inaccuracy of the house locations. Also two additional methods are developed to approximate the effect of the diffused subject location on estimation and inference. One uses expectation of pollution exposures in the regions where the true house locations could be as the true pollution exposures. The other simulates true locations and estimates parameters for each simulation. It is concluded that the effect of home location measurement error on estimation and inference in this study is small. The effect can be omitted. Another measurement error (i.e. inaccuracies in the locations of roads) is also suggested to be omitted.

6.1 Review of error-in-variable models

Variables used in statistical analysis are often taken as if they were exactly determined. In fact, they are often measured with error. Such error can be classified as either 'measurement error' or 'intrasubject error'. Measurement error can be due to

instrument imprecision or human inability to accurately read an instrument. Intrasubject error can rise from diurnal or other short-term variability when the measured average or other simple value is used. Both of these types of variability will be referred to as 'variable error'.

Variable error in the covariate has long been known to actually change the functional relationship between output and observed variables (Adcock 1878). There are many papers that discuss the effect of the error on inference, estimation and how to deal with this kind of error (Cochran 1968, Cox and Dolby 1977, Dolby 1976). Hoschel (1989) reviewed research results in this area.

The simplest case is a simple linear regression model.

$$y_i = \alpha + \beta\xi_i + \varepsilon_i \tag{6.1}$$

$$x_i = \xi_i + \delta_i \tag{6.2}$$

in which ε_i is referred to 'equation error'. The ε_i is i.i.d. such that $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$. ξ_i is i.i.d. random variables, $\xi_i \sim N(v,\sigma_\xi^2)$. δ_i is referred to as 'measurement error'; they are i.i.d. such that $\delta_i \sim N(0,\sigma_\delta^2)$. x_i is the measured variable with error.

Reformatting (6.1) and (6.2), the representation is obtained:

$$y_i = \alpha + \beta x_i + (\varepsilon_i - \beta\delta_i) \tag{6.3}$$

Error $\varepsilon_i - \beta\delta_i$ depends on parameter β. If ignoring the form of the error $\varepsilon_i - \beta\delta_i$, the ordinary least squares estimator gives approximated estimates:

$$\hat{\beta} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sum_i (x_i - \bar{x})^2$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

It can be proved (Kendal and Stuart 1979), when number of subjects goes to infinity,

$$\hat{\beta} \to \beta\sigma_\xi^2 / (\sigma_\delta^2 + \sigma_\xi^2) \tag{6.4}$$

$$\hat{\alpha} \to \alpha + \beta v\sigma_\delta^2 / (\sigma_\delta^2 + \sigma_\xi^2) \tag{6.5}$$

The estimate $\hat{\beta}$ is smaller than true value. $\hat{\alpha}$ is also biased.

In the case of replication of observations, the model is

$$y_{ij} = \alpha + \beta\xi_{ij} + \varepsilon_{ij} \tag{6.6}$$

$$x_{ij} = \xi_{ij} + \delta_{ij} \tag{6.7}$$

in which ε_{ij} is i.i.d. with distribution $N(0,\sigma_\varepsilon^2)$, ξ_{ij} is i.i.d. with distribution $N(v_i,\sigma_\xi^2)$, δ_{ij} is i.i.d. with distribution $N(0,\sigma_\delta^2)$, $i=1,\cdots,n, j=1,\cdots,m_i$.

Cochran (1968) pointed out that in this model when variable errors exist, ordinary ANCOVA estimates the coefficient to be (when number of subjects goes to infinity):

$$\hat{\beta} \to \beta\sigma_\xi^2/(\sigma_\delta^2+\sigma_\xi^2) < \beta \qquad (6.8)$$

Maximum Likelihood Estimation (MLE) is also affected by errors in variables. The effects are in two ways. If the measurement error is ignored, the estimate will be biased, though in some situations the estimate error is not large e.g. when the measurement error δ_{ij} is small compared to equation error ε_{ij}.

On the other hand, when the measurement error is considered, MLE is changed from the non-measurement error situation.

In the regression model (6.6), (6.7), by combining (6.6) and (6.7) we get:

$$y_{ij} = \alpha + \beta x_{ij} + (\varepsilon_{ij} - \beta\delta_{ij}) \qquad (6.9)$$

The error $\varepsilon_{ij} - \beta\delta_{ij}$, has a normal distribution, $N(0,\sigma_\varepsilon^2+\beta^2\sigma_\delta^2)$

Denote $z_{ij} = [x_{ij}, y_{ij}]$ and $u_i = [v_i, \alpha+\beta v_i]$.

z_{ij} has covariance

$$\Sigma = \begin{pmatrix} \sigma_\delta^2+\sigma_\xi^2 & \beta\sigma_\xi^2 \\ \beta\sigma_\xi^2 & \sigma_\varepsilon^2+\beta^2\sigma_\xi^2 \end{pmatrix}$$

Then the log-likelihood function is

$$L(\alpha,\beta) = -(m/2)\log\det[\Sigma] - \sum_i\sum_j (z_{ij}-u_i)'\Sigma^{-1}(z_{ij}-u_i)/2 \qquad (6.10)$$

For replicated observations, e.g. m_i >1, a solution can be obtained (Hoschel 1989, Cox and Dolby 1977). For models without replicated observations, e.g. m_i =1, it has no general solution. Dolby (1976) discussed it in several situations, such as, $\sigma_\xi^2/\sigma_\delta^2$ is known.

The general model with error-in-variables is

$$\eta_i = H(\xi_i),$$

$$x_i = \xi_i + \delta_i,$$

$$y_i = \eta_i + \varepsilon_i \qquad (6.11)$$

in which ξ_i is a d-dimension variable. x_i is the observed d-dimensional variable. y_i is the c-dimensional output. H: $R^d \rightarrow R^c$ is the functional relation. δ_i, ε_i are measurement and equation errors respectively.

For this general model, closed formulas can no longer be obtained. Numerical methods are needed to calculate the maximum likelihood estimates.

6.2 The pollution-exposure/asthma-incidence (PEAI) error-in-variable model

The grid references of post-codes given with NCDS are different from the house locations. The grid references can be treated as the observed house locations with errors. Hence an error-in-variable model can be adopted.

Assume the true position of a house is $\mathbf{x}^h$, $\mathbf{x}^h$ follows a distribution $h(\mathbf{x}^h)$. $h(\mathbf{x}^h)$ can be treated as a uniform distribution in a certain area. The grid reference of post-code is $\mathbf{x}$, the pollution exposure at $\mathbf{x}^h$ is y'. Suppose that

$$\mathbf{x} \sim g(\mathbf{x}^h) \qquad (6.12)$$

The conditional distribution of $\mathbf{x}^h$ based on $\mathbf{x}$ can be obtained:

$$\mathbf{x}^h = \mathbf{x} + \boldsymbol{\varepsilon} \qquad (6.13)$$

in which $\boldsymbol{\varepsilon}$ is a location measurement error based on $\mathbf{x}$. $\boldsymbol{\varepsilon}$ is independent of $\mathbf{x}$, and i.i.d. Its' detailed distribution will be discussed later in Section 6.3.

Given grid reference of post-code $\mathbf{x}_1, \cdots, \mathbf{x}_n$, denote the true positions of houses by $\mathbf{x}_1^h, \cdots, \mathbf{x}_n^h$.

Then the full hierarchical model for pollution-exposure/asthma-incidence model (see in Chapter 4) is

$$\mathbf{x}_i^{\mathrm{h}} = \mathbf{x}_i + \varepsilon_i \tag{6.14}$$

$$y_i' = y'(\mathbf{x}_i^{\mathrm{h}}) \tag{6.15}$$

$$\mathrm{logit}(\mathrm{p}_i) = \rho + \alpha y_i' + \gamma\, sex(i) \tag{6.16}$$

where y_i' is the pollution exposure for subject i, y' denotes the road pollution exposure function, $sex(i)$ is the gender of subject i, p_i is the probability of subject i has asthma.

For model 2 and 3 in Chapter 4, similar hierarchical measurement models can be obtained.

The marginal likelihood function of parameters $\rho, \alpha, \sigma, \gamma$ on ε_i is

$$L(\rho,\alpha,\sigma,\gamma) = \prod_{\text{suject i has asthma}} \int \frac{\exp\left(\rho + \alpha y'(\mathbf{x}_i + \boldsymbol{\varepsilon}) + \gamma\, sex(i)\right)}{1 + \exp\left(\rho + \alpha y'(\mathbf{x}_i + \boldsymbol{\varepsilon}) + \gamma\, sex(i)\right)} f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon}$$
$$\cdot \prod_{\text{suject i has no asthma}} \int \frac{1}{1 + \exp\left(\rho + \alpha y'(\mathbf{x}_i + \boldsymbol{\varepsilon}) + \gamma\, sex(i)\right)} f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} \tag{6.17}$$

in which $f(\boldsymbol{\varepsilon})$ is the density function of error $\boldsymbol{\varepsilon}$ in (6.13).

Early in this study, (6.17) is considered too time-consuming to be explicitly maximized by MLE, due to huge amount of data and the integration involving the complex pollution exposure dispersal function y'. To simplify the computation, two numerical approaches were developed as approximations to MLE.

Numerical approach 1: Expected Exposure Method (EEM)

To simplify the computation, we use expectation of pollution exposure over possible house location $\mathrm{E}\, y_i' = \mathrm{E}(y'(\mathbf{x}_i + \boldsymbol{\varepsilon}))$ as an estimation of y_i'. This method is called Expected Exposure Method (EEM). This method reflects the fact that a subject gets the pollution exposure not only at his house, but also from the area about his house because of his daily activity.

We can use MLE to do parameter estimation and use likelihood ratio test for inference.

Numerical approach 2: Location simulation Method (LSM)

We simulate the true positions $\mathbf{x}^h$ of houses from the relevant post-code regions. Estimations and inferences are conditioned on these simulations. After we repeat these processes many times, we can investigate the estimates from all simulations. This method is called Location Simulation Model (LSM). This method can give us an idea of how large of the estimated parameter is influenced by the house location measurement error.

Later in this study, it was realised that the integration in (6.17) can be performed approximately by suitable numerical technique, summing values on several points. Then MLE is used to maximize (6.17).

Results from all three methods are reported (MLE in Section 6.4, EEM in Section 6.5 and LSM in Section 6.6).

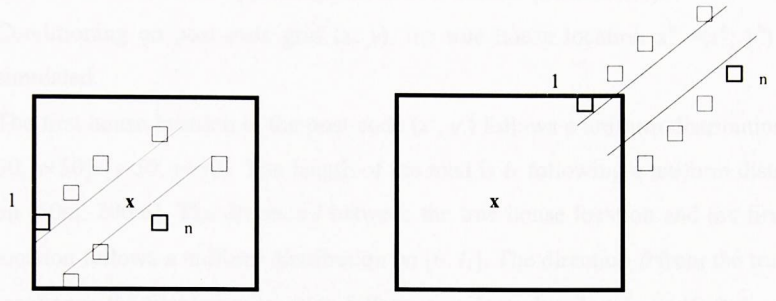
6.3 Measurement errors of house locations

The measurement error in (6.13) is to be investigated.

Normally a post-code covers a region along a road (or part of a road). The average number of delivery points per post-code is 14. Very few post-codes cover more than 50 delivery points (Raper *et al.* 1992). The grid reference (x, y) of a post-code is the grid reference of the southwest corner of the 100m square in which the first house (the first delivery point) of the post-code lies. In order to correct this Southwest corner bias to some extent, the grid reference is changed to $\mathbf{x}=(x+50,y+50)$ in this research. From now on, the grid reference is the centre of the 100m square in which the first house of the post-code lies.

The extent of particular post-codes can be obtained from some commercial software (Raper *et al.* 1992), but they are not available in our research. Hence, we used some simple assumptions about the extent of a post-code. We assume a particular post-code covers a road segment whose length is between 50m and 200m (Raper *et al*, 1992). The house No. 1 is the first house in this post-code region. It lies in a (100m x

100m) square whose centre is **x**. Suppose that the subject under study lives in house No. n. Then the distance between the house and post-code grid **x** can vary from 0m to 270m (Figure 6.1), in which $270 \approx \sqrt{2} * 50 + 200$.



(a) **(b)**

Figure 6.1 An illustration of cases for which the distance between the subject's house and the post-code grid is (a) 0m and (b) 270m

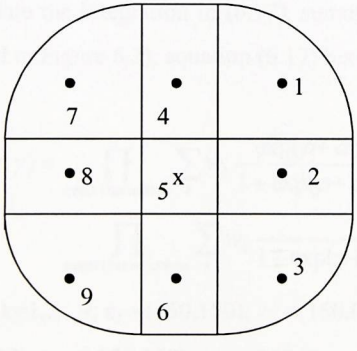


Figure 6.2 An illustration of the region which true house location in

It is assumed that the road orientation is isotropic. It is also assumed that the house numbering is in an arbitrary direction and independent of road orientation.

Let the true location of the house $\mathbf{x}^h = \mathbf{x} + \varepsilon$, ε is the measurement error conditioned on $\mathbf{x}$ and ε follows a distribution which is isotropic. In this example with assumed road length (in a post-code area) of between 50m and 200m. The true house location lies in an area illustrated in the Figure 6.2. The area has been divided into nine areas, which will be used for approximation in next section (Section 6.4).

Conditioning on post-code grid (x, y), the true house location $\mathbf{x}^h = (x^h, y^h)$ can be simulated.

The first house location in the post-code (x', y') follows a uniform distribution on $[x-50, x+50]x[y-50, y+50]$. The length of the road is l_1 following a uniform distribution on [50m, 200m], The distance l between the true house location and the first house location follows a uniform distribution on $[0, l_1]$. The direction θ from the true house location to the first house location follows a uniform distribution on $[0, 2\pi]$.

The value of (x', y'), l_1, l, θ can be simulated. Therefore the true house location $(x^h, y^h) = (x' + l^*\cos\theta, y' + l^*\sin\theta)$ can be simulated.

6.4 Results using MLE

To calculate the integration in (6.17), summing (weighted) values at nine points (as illustrated in Figure 6.2), equation (6.17) becomes

$$L(\rho,\alpha,\sigma,\gamma) = \prod_{\text{suject i has asthma}} \sum_k w_k \frac{\exp(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i))}{1+\exp(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i))} \cdot \prod_{\text{suject i has no asthma}} \sum_k w_k \frac{1}{1+\exp(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i))} \tag{6.18}$$

in which k=1,…,9, ε_1 =(150,150), ε_2 =(150,0), ε_3 =(150,-150), ε_4 =(0,150), ε_5 =(0,0), ε_6 =(0,-150), ε_7 =(-150,150), ε_8 =(-150,0), ε_9 =(-150,-150).

w_i are obtained with Monte-Carlo method explained in Section 6.3 to simulate the probability of true house location lying in each region (Figure 6.2). w_1= 0.04154, w_2= 0.10515, w_3= 0.04154, w_4= 0.10515, w_5= 0.41324, w_6= 0.10515 w_7= 0.04154, w_8= 0.10515, w_9=0.04154.

6.4.1 Models and inference

Models and inference are similar with those in Chapter 4. Only some equations are changed.

The log-likelihood function is

$$\begin{aligned} \mathrm{L}(\rho,\alpha,\sigma,\gamma) = &\sum_{i\in A}\log\left(\sum_k w_k \frac{\exp\left(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i)\right)}{1+\exp\left(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i)\right)}\right) \\ &+\sum_{i\notin A}\log\left(\sum_k w_k \frac{1}{1+\exp\left(\rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i)\right)}\right) \end{aligned} \tag{6.19}$$

in which, A={i: subject i has asthma}.

The derivative of log-likelihood function is

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial} = &\sum_{i\in A}\left(\frac{\exp(f_{ik})}{\sum_k w_k \frac{1}{1+\exp(f_{ik})}}\sum_k w_k \frac{\exp(f_{ik})}{(1+\exp(f_{ik}))^2}\frac{\partial f_{ik}}{\partial}\right) \\ &-\sum_{i\notin A}\left(\frac{1}{\sum_k w_k \frac{1}{1+\exp(f_{ik})}}\sum_k w_k \frac{\exp(f_{ik})}{(1+\exp(f_{ik}))^2}\frac{\partial f_{ik}}{\partial}\right) \end{aligned} \tag{6.20}$$

in which $f_{ik} = \rho+\alpha y'(\mathbf{x}_i+\boldsymbol{\varepsilon}_k)+\gamma\, sex(i)$

$\frac{\partial f_{ik}}{\partial}$ has the same form as $\frac{\partial f_{ik}}{\partial}$ in section 4.3.2, in which the location $\mathbf{x}_i$ is replaced by $\mathbf{x}_i+\boldsymbol{\varepsilon}_k$.

6.4.2 Estimation and inference

The method is applied on the data from the London region and the Birmingham region. The results for the London region are shown in Table 6.1, for the Birmingham are shown in Table 6.2.

	Model 0	Model 1	Model 2	Model 3
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.84(0.009)	-0.96(0.16)	-1.08(0.19)	-1.49(0.31)
α	-	0.26(0.20)	0.64(0.58)	-
α_1	-	-	-	0.62(0.63)
α_2	-	-	-	0
α_3	-	-	-	2.30(1.52)
α_4	-	-	-	6.02 (5.85)
σ	-	0.98(1.49)	-	-
σ_1	-	-	1.13(0.72)	1.14(0.88)
σ_2	-	-	5	-
σ_3	-	-	0.51(0.36)	2.35(1.71)
σ_4	-	-	0.13(0.44)	0.80(0.57)
γ	-0.32(0.02)	-0.33(0.16)	-0.33(0.14)	-0.36(0.14)
Log-likelihood	-607.99	- 607.52	-605.81	-604.01

Table 6.1 Analysis results for London area from measurement error model using approximate MLE

	Model 0	**Model 1**	**Model 2**	**Model 3**
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.6(0.023)	-0.87(0.19)	-1.05(0.25)	-0.99(0.44)
α	-	0.597(0.44)	0.97(0.79)	-
α_1	-	-	-	2.03(2.05)
α_2	-	-	-	0(-)
α_3	-	-	-	1.186(0.94)
α_4	-	-	-	5.196(5.15)
σ	-	0.72(0.61)	-	-
σ_1	-	-	2.99(1.09)	-5(-)
σ_2	-	-	1.09(0.76)	-
σ_3	-	-	0.33(0.21)	0.37(0.20)
σ_4	-	-	1.31(0.91)	3.55(1.87)
γ	-0.25(0.05)	-0.23(0.18)	-0.20(0.14)	-0.23(0.17)
Log-likelihood	-253.87	-251.91	-250.62	-250.58

Table 6.2 Analysis results for Birmingham area from measurement error model using approximate MLE

Overall the results are similar to the results from the original models without measurement error. This may be because most subjects are far from the main road. The measurement errors of the house locations are small compared to the mean distances to main roads.

6.5 Expected exposure method (EEM)

In this method the average pollution exposure in the post-code region is used as the true pollution exposure.

Let square A with corner co-ordinates $((x_1, y_1), (x_2, y_1), (x_2, y_2), (x_1, y_2))$, denote the vicinity we want to average over, and line $((x_0, y_0), (x_0+a, y_0+b))$ denote a linear road-segment. Also let |A| denote the area of square A and l denote the length of $((x_0, y_0), (x_0+a, y_0+b))$, i.e. $l = \sqrt{a^2 + b^2}$.

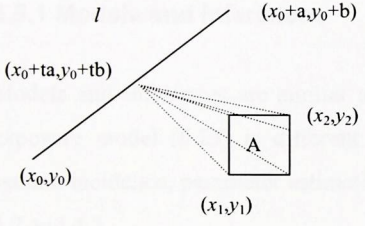


Figure 6.3 Pollution exposure on the square from a road-segment

The assumptions about pollution exposure dispersal are the same as in Chapter 4.

$$y' = s\frac{1}{2\pi\sigma^2}e^{-\frac{\|\mathbf{x}-\mathbf{x_0}\|^2}{2\sigma^2}} \qquad (6.21)$$

From Section 6.3 the location error ε follows a distribution over an area (illustrated in Figure 6.2). In (6.14) the true location of the house is $\mathbf{x}^h = \mathbf{x} + \varepsilon$. To simplify the computation, a uniform distribution over [-200,200]x[-200,200] is used.

Hence the Expected pollution exposure of the subject will be the averaged pollution exposure in the vicinity over $[x\text{-}200, x\text{+}200]\text{x}[y\text{-}200, y\text{+}200]$.

$$T=\frac{1}{|A|}\iint_{(x,y)inA}\int_0^1 \frac{l}{2\pi\sigma^2}\exp\left(-\frac{(x-x-ta)^2+(y-y_0-tb)^2}{2\sigma^2}\right)dt\ \ dxdy \qquad (6.22)$$

$$=\frac{1}{|A|}\int_0^1 \iint_{(x,y)inA}\frac{l}{2\pi\sigma^2}\exp\left(-\frac{(x-x-ta)^2+(y-y_0-tb)^2}{2\sigma^2}\right)dxdy\ \ dt$$

$$=\frac{l}{|A|}\int_0^1\left[\Phi(\frac{x_2-x_0-ta}{\sigma})-\Phi(\frac{x_1-x_0-ta}{\sigma})\right]*\left[\Phi(\frac{y_2-y_0-tb}{\sigma})-\Phi(\frac{y_1-y_0-tb}{\sigma})\right]dt$$

(6.23)

The integration in (6.23) is obtained by sum of values at t=0.0, 0.2, ..., 1.0 in our study. Φ is calculated using NAG routine S15ABF.

6.5.1 Models and inferences

Models and inferences are similar as those in Section 4.2, and 4.3. The pollution exposure model (6.23) is different because of the averaging, but the models of asthma incidence, parameter estimation and inference are similar to those in Section 4.2 and 4.3.

Maximum likelihood estimation is used to do parameter estimation. All the formulas for first derivatives are still the same as those in Section 4.3.2, except the derivative of T for σ (formula (6.23)) becomes:

$$\frac{\partial T}{\partial \sigma}=$$

$$\frac{l}{A}\int_0^1\left(-\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_2-x_0-ta)^2}{2\sigma^2})\frac{x_2-x_0-ta}{\sigma^2}+\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_1-x_0-ta)^2}{2\sigma^2})\frac{x_1-x_0-ta}{\sigma^2}\right)$$
$$*\left[\Phi(\frac{y_2-y_0-tb}{\sigma})-\Phi(\frac{y_1-y_0-tb}{\sigma})\right]$$
$$+\left(-\frac{1}{\sqrt{2\pi}}\exp(-\frac{(y_2-y_0-ta)^2}{2\sigma^2})\frac{y_2-y_0-ta}{\sigma^2}+\frac{1}{\sqrt{2\pi}}\exp(-\frac{(y_1-y_0-ta)^2}{2\sigma^2})\frac{y_1-y_0-ta}{\sigma^2}\right)$$
$$*\left[\Phi(\frac{x_2-x_0-tb}{\sigma})-\Phi(\frac{x_1-x_0-tb}{\sigma})\right]dt$$

(6.24)

The integration in (6.24) is obtained by sum of values at t=0.0, 0.2, …, 1.0 in our study.

6.5.2 Results

The EEM method is applied on the data from the London region and the Birmingham region. The results for the London region are shown in Table 6.3, for the Birmingham region are shown in Table 6.4.

	Model 0	**Model 1**	**Model 2**	**Model 3**
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.84(0.009)	-0.96 (0.16)	-1.08 (0.18)	-1.49(0.31)
α	-	0.26 (0.19)	0.64 (0.55)	-
α_1	-	-	-	0.64 (0.58)
α_2	-	-	-	0
α_3	-	-	-	2.29 (1.49)
α_4	-	-	-	6.00 (4.65)
σ	-	0.99 (1.48)	-	-
σ_1	-	-	1.11 (0.68)	1.13(0.83)
σ_2	-	-	5	-
σ_3	-	-	0.54(0.34)	2.34(1.65)
σ_4	-	-	0.52(1.74)	0.80(0.55)
γ	-0.32(0.02)	-0.33(0.16)	-0.33(0.14)	-0.35(0.14)
Log-likelihood	-607.99	-607.51	-605.87	-603.99

Table 6.3 Analysis results for London area using EEM

	Model 0	Model 1	Model 2	Model 3
	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)	MLE(S.E.)
ρ	-0.6(0.023)	-0.86(0.18)	-1.03(0.26)	-1.10(0.45)
α	-	0.58(0.43)	1.02(0.76)	-
α_1	-	-	-	1.17(0.95)
α_2	-	-	-	0.56(0.27)
α_3	-	-	-	1.34(0.76)
α_4	-	-	-	3.67(1.15)
σ	-	0.71(0.63)	-	-
σ_1	-	-	0.20(0.18)	0.19(0.17)
σ_2	-	-	1.19(0.87)	0.84(0.60)
σ_3	-	-	0.32(0.19)	0.36(0.21)
σ_4	-	-	1.32(0.92)	1.49(0.96)
γ	-0.247(0.05)	-0.226(0.18)	-0.19(0.13)	-0.20(0.16)
Log-likelihood	-253.87	-251.96	-249.73	-249.4

Table 6.4 Analysis results for Birmingham area using EEM

Overall the results are similar to the results from the original models without averaging. Since most subjects are far away from the main road, the measurement errors of the house locations are small compared to the mean distances to main roads, so the averaging does not change the pollution level much.

Comparing measurement models with the original models in Chapter 4, the significance levels are smaller and estimation errors are bigger because the averaging of the pollution makes the difference between the pollution exposure of different subjects smaller (see Figures 6.4 and 6.5, comparing to Figure 4.3). Hence no significant relationship between asthma and pollution exposure is found.

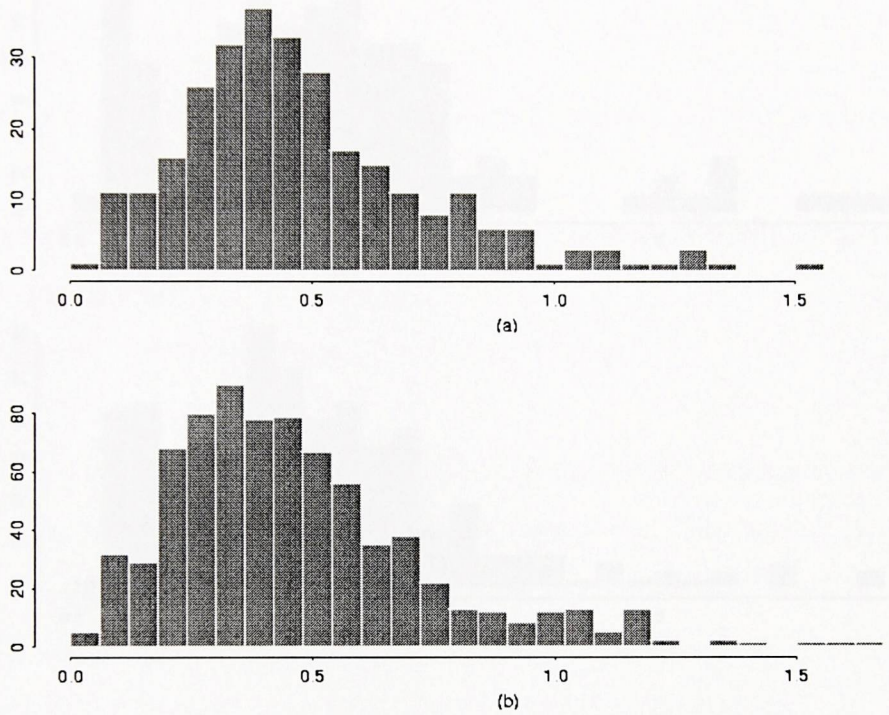


Figure 6.4 Histograms of expected pollution exposures on (a) cases and (b) controls in London

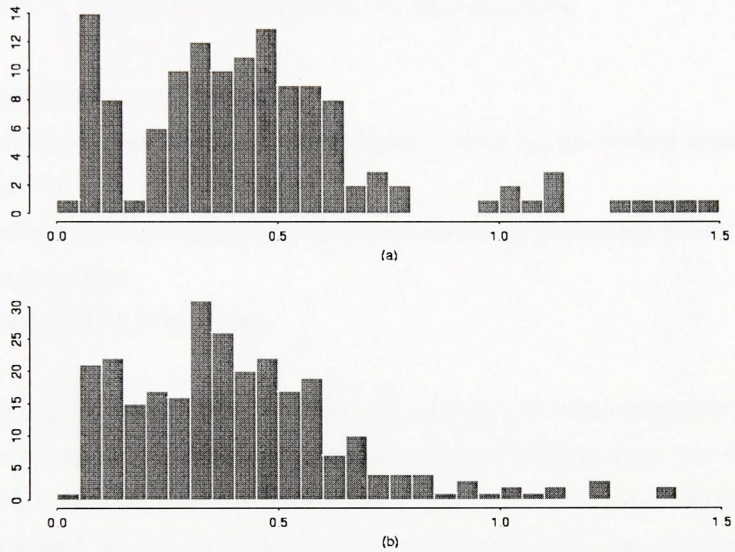


Figure 6.5 Histograms of expected pollution exposures on (a) cases and (b) controls in Birmingham

6.6 Location simulation method (LSM)

In this method, we simulate the true house locations (as the method explained in Section 6.3), then do parameter estimation for each simulation.

This method works as following:

1) Simulate the true locations of all subjects' houses (as the method explained in Section 6.3).

2) Maximize the log-likelihood function and estimate the parameters, standard errors for this simulation.

3) Repeat steps 1), 2) 100 times.

Steps 1), 2) are repeated 100 times. Since one repetition is independent with another, we can run parallel repetitions on different computers. This process is very slow. For model 1, it takes one week for four SUN workstations to run 1) and 2) 100 times. For model 2 and 3, since there are more parameters to be estimated, they take much more time to run. Hence only the model 1 for London area is simulated.

The histograms of parameter estimates from 100 simulations are shown in Figure 6.6

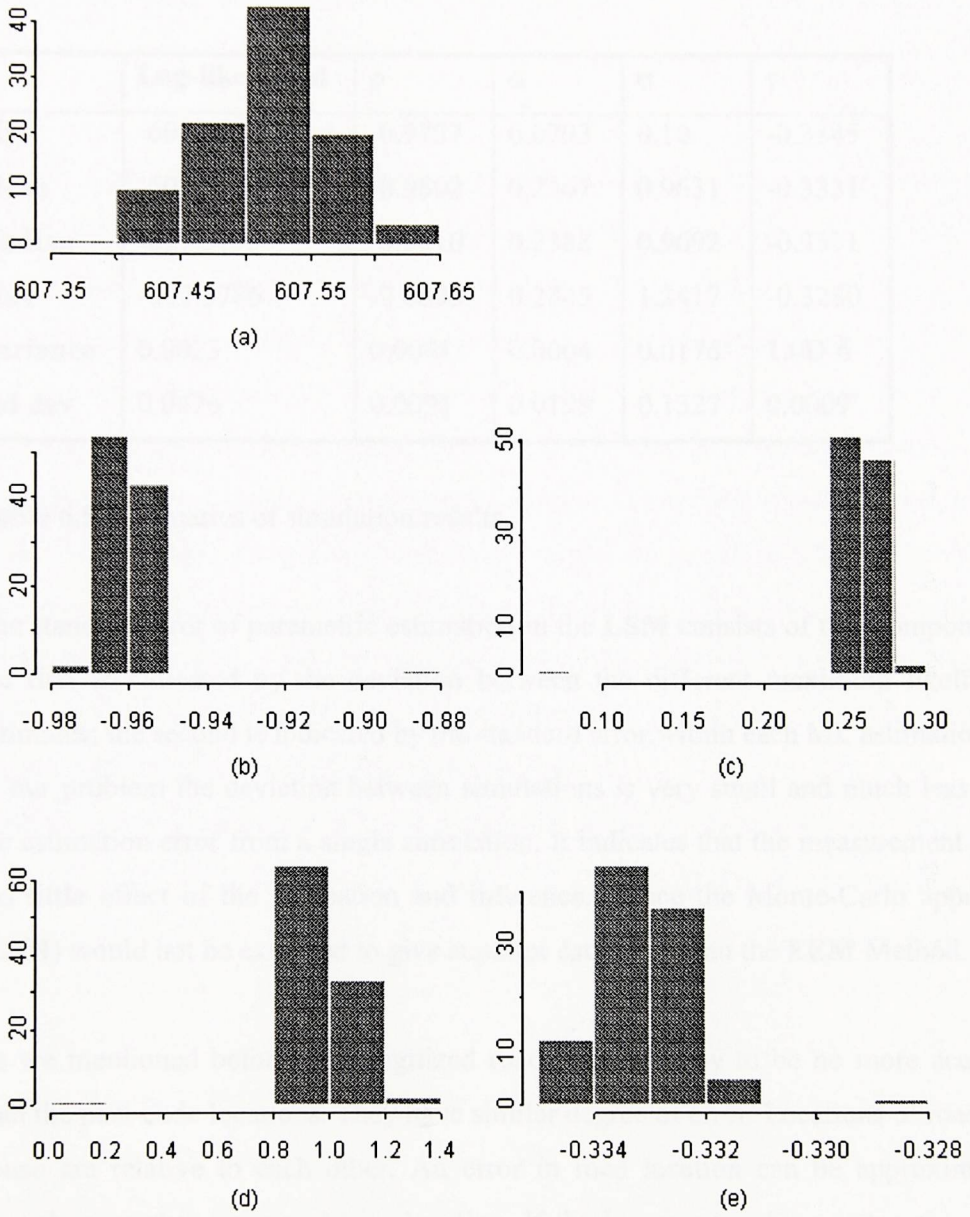


Figure 6.6 Histograms of simulation results (a) log-likelihood, (b) ρ, (c) α, (d) σ, and (e) γ

The statistical summaries of the estimated parameter distributions are shown in Table 6.5.

	Log-likelihood	ρ	α	σ	γ
Min	-607.6188	-0.9737	0.0793	0.10	-0.3345
Mean	-607.5170	-0.9602	0.2567	0.9631	-0.3331
Median	-607.5195	-0.9610	0.2588	0.9692	-0.3331
Max	-607.3786	-0.8803	0.2845	1.2417	-0.3280
Variance	0.0023	0.0001	0.0004	0.0176	1.0D-6
Std dev	0.0476	0.0091	0.0199	0.1327	0.0009

Table 6.5 Summaries of simulation results

The standard error of parametric estimation in the **LSM** consists of two components: the first is indicated by the deviation between the different maximum likelihood estimates; the second is indicated by the standard error within each ML estimation. In our problem the deviation between simulations is very small and much less than the estimation error from a single simulation. It indicates that the measurement error has little effect of the estimation and inference. Hence the Monte-Carlo approach (**LSM**) would not be expected to give superior estimates than the **EEM** Method.

As we mentioned before, The digitized road map is likely to be no more accurate than the post-code locations. They have similar degree of error. Locations of road and house are relative to each other. An error in road location can be approximately treated as another error on house location. If the inaccuracy of house location does not affect our study, the inaccuracy of road map will have no more effect.

Chapter 7 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a very useful tool in the study of posterior probability densities in Bayesian analysis. It enables a complex joint density to be used without it having to be normalized. Since the invention of MCMC method, Bayesian statistical methodology has become a mainstream statistical method (Robert and Casella 1999, Smith and Roberts 1993). MCMC naturally finds its use in spatial statistics (Besag and Green 1993, Lawson 1995). But MCMC is computationally intensive, requires careful design and evaluation. It uses the density function very many times and creates a small fraction of samples for use. If the density function is very complex, then each evaluation is time consuming. If Bayesian approach is to be adopted in this study, each posterior density/likelihood-function (for model 1 for the London region) takes seven minutes to evaluate on a SUN workstation. In such a situation the method is practically impossible to apply.

In this chapter a version of MCMC method is developed, which allows the MCMC method to be used with a complex posterior density/likelihood function. It can be used on such posterior density which can be divided into two components. One component should be such that the bulk of computation depends on small number of the model's parameters (assumed to be continuous). During the calculation of a density function, most of computations are spent on evaluating this component. Our method involves treating these parameters as discrete. Hence values of this component in the posterior density can be calculated prior to the MCMC and used later from a look-up table without the need to calculate them again. In this way, the computing time is much reduced, speed is improved and MCMC can be implemented practically for complex models on a large dataset.

This discretization method is implemented in our asthma study. It is pointed out that a similar approach can be used in a wide area of applications.

7.1 Review of the literature of MCMC

Markov Chain Monte Carlo (MCMC) can be used for sampling from a distribution (Robert and Casella 1999). The basic idea involves building a Markov chain whose equilibrium distribution is the target distribution. Then after a long period, the chain's distribution converges to the target distribution. Diagnostic methods are also developed to determine when the chain converges (Robert and Casella 1999, Gelfand and Smith 1990, Raftery and Lewis 1992, Gelman and Rubin 1992). Several methods have been developed to improve the speed of convergence (Robert and Casella 1999, Prop and Wilson 1996, Geyer 1991, Marinari and Parisi 1992, Charles and Thompson 1995).

7.1.1 Method

There are two basic MCMC methods: the Gibbs sampler and the Metropolis-Hastings Algorithm. They can be used in different situations.

7.1.1.1 The Gibbs sampler

The Gibbs sampler was developed by Geman and Geman (1984) for distributions on a lattice. It was generalized by Gelfand and Smith (1990) for a continuous distribution. The notation in Smith and Roberts (1993) is used here to explain the Gibbs sampler.

For the target distribution $\pi(\mathbf{x})= \pi(x_1, x_2,\ldots, x_k)$, we construct a Markov chain whose equilibrium distribution is $\pi(\mathbf{x})$.

Start from $\mathbf{x}^0 = (x_1^0, \cdots, x_k^0)$

Draw $x_1^1 \sim \pi(x_1 \mid x_2^0, \cdots, x_k^0)$

$x_2^1 \sim \pi(x_2 \mid x_1^1, x_3^0, \cdots, x_k^0)$

$x_3^1 \sim \pi(x_3 \mid x_1^1, x_2^1, x_4^0, \cdots, x_k^0)$

$\vdots$

$x_k^1 \sim \pi(x_k \mid x_1^1, \cdots, x_{k-1}^1)$

So we get $\mathbf{x}^1$, repeating, we get

$\mathbf{x}^0, \mathbf{x}^1, \cdots, \mathbf{x}^t$.

$\mathbf{x}^0, \mathbf{x}^1, \cdots, \mathbf{x}^t, \cdots$ form a Markov Chain. It can be proved that the equilibrium distribution of the Markov Chain is $\pi(\mathbf{x})$ (Geman and Geman 1984), i.e.

$$\lim_{t \to \infty} P(\mathbf{x}^t = \mathbf{x} \mid \mathbf{x}^0) = \pi(\mathbf{x})$$

7.1.1.2 The Metropolis-Hastings algorithm

MCMC in the form of the Metropolis-Hastings algorithm was developed by Metropolis *et al.* (1953) and generalized by Hastings (1970). We now describe the Metropolis-Hastings algorithm using the notation as in Smith and Roberts (1993).

Assume the target distribution is $\pi(x)$. We construct a Markov chain $x^0, x^1, \cdots, x^t, \cdots$.

The change from x^t to x^{t+1} is made as follows.

Let $q(x, x')$ denote a transition probability function from x to x'. If $x^t = x$, x' drawn from $q(x, x')$ is considered as a possible value for x^{t+1}. However a further randomization takes place. With probability $\alpha(x, x')$ (given in (7.1)) $x^{t+1} = x'$ is accepted. Otherwise, it is rejected, we set $x^{t+1} = x$.

$$\alpha(x, x') = \begin{cases} \min\{\dfrac{\pi(x')q(x', x)}{\pi(x)q(x, x')}, 1\} & \text{if } \pi(x)q(x, x') > 0 \\ 1 & \text{if } \pi(x)q(x, x') = 0 \end{cases} \tag{7.1}$$

Provided $q(x, x')$ is chosen be irreducible and aperiodic on a suitable state space, $\pi(x)$ is the equilibrium distribution, i.e.

$$\lim_{t \to \infty} P(x^t = x \mid x^0) = \pi(x)$$

The Gibbs sampler can only be implemented when conditional distributions are easy to simulate. Theoretically, the Metropolis-Hastings algorithm can be used in any situation. In general, the speed of (weak) convergence of Gibbs sampler is faster than

the Metropolis-Hastings algorithm, though in some cases, the Metropolis-Hastings algorithm is faster (Frigessi *et al.* 1991).

In practice, the first part of the Markov chain will be thrown out. This is called 'burn in' process. After a sufficient large 'burn in', the chain starts to converge i.e. to stabilize. Among each M steps (for a suitable choice of M), a value is chosen to use as a sample. MCMC takes a long time to converge. The choice of jump step between x and x' influences convergence and acceptance rate, and the speed with which the chain covers the state space (this speed is commonly referred to as 'mixing'). If the jump step is large, the acceptance probability will be small. The chain hardly changes for a long time. If the jump step is small, it is difficult to cover the entire state space. Several methods have been developed to improve convergence speed and mixing of the MCMC chain.

7.1.1.3 Simulated tempering

When a distribution is very complicated and has sharp shape. A MCMC is hard to mix. To make the processes faster, more distributions can be added.
Geyer (1991) developed a method called Metropolis-coupled MCMC (MCMCMC). It improves the mixing by changing between parallel simulations of different but similar distributions. Marinari and Parisi (1992) developed a method called 'simulated tempering'. Charles and Thompson (1995) later developed a similar method called Annealing Markov Chain Monte Carlo. They have advantages over MCMCMC, in that there is only one chain instead of parallel chains, so the chain uses less storage and also mixes better.

Simulated tempering is based on an optimization method called 'simulated annealing' (Kirkpatrick *et. al.* 1983). But it does not 'cool' like simulated annealing. It uses a one-parameter family of probability distribution indexed by a parameter called 'temperature', ranging from the distribution of interest as the 'coldest' temperature to a 'hottest' distribution that is much easier to simulate.

Suppose a distribution $h_1(x)$ is the target distribution. A sequence of m distributions $h_i(x)$, $i = 1,\cdots,m$ (such as, for example, $h_i(x) = h(x)^{1/i}$) can be simulated parallel instead. We call $h_1(x)$ is the 'cold' distribution, $h_m(x)$ the 'hot' distribution. Sometime all distributions are of interest (such as later in our study), usually only the cold distribution is of interest, and the rest only to increase the mixing.

The joint distribution of x and i forms a new distribution. A Markov chain whose state is (x, i) can be created. It starts from the hot distribution (i.e. i = m), then x and i are updated separately. The detail is as follows:

1.Update x using a Metropolis-Hastings or Gibbs update for h_i.

2. Update i by the following step.

1) Set $j = i \pm 1$ according to probabilities $q_{i,j}$, where $q_{i,j}$ is

$$q_{1,2} = q_{m,m-1} = 1 \text{ and } q_{i,i+1} = q_{i,i-1} = \frac{1}{2} \text{ if } 1 < i < m .$$

2). Calculate the Hastings ratio

$$r = \frac{h_j(x)}{h_i(x)} \frac{q_{j,i}}{q_{i,j}}$$

and accept the transition (set i to j) or reject it according to the Metropolis rule: accept with probability min(r,1).

Repeat these steps, a Markov Chain of (x, i) is created.

The stationary distribution of (x, i) is $c\,h_i(x)$, where c is normalizing constant (Marinari and Parisi, 1992). If only the samples (x, i) whose i is 1 is chosen, the samples x follow a distribution proportional to $h_1(x)$.

It was also suggested by Charles and Thompson (1995) that use of the regeneration procedure would improve the estimation (Riply, 1987). Regeneration can be achieved by updating x with an independent sample from h_m when i=m. It can simplify estimation of Monte Carlo error.

7.1.1.4 Exact MCMC

For the MCMC method, the Markov chain converges after a large number of steps. It is usually quite difficult to determine how large the number should be.

Prop and Wilson (1996) developed an exact sampling method. Using this method, the Markov chain can decide when to stop and the distribution at stopping time is exactly the target distribution. This method samples from a finite set of objects in accordance with some distribution. This method runs two coupled Markov Chains and the running stops when two chains reach the same states. This is an effective method although it only works for a finite state space case and the state space must have a partial order. Murdoch and Green (1998) extended this method to continuous state space case for some circumstances.

7.1.2 Diagnosis

A Markov chain has to run a large number of steps to converge to the target distribution. Several methods have been developed to diagnose whether the chain has converged. Gelfand and Smith (1990) proposed the use of the Q-Q graph, in which the last n values of sample outputs are plotted against the penultimate n iterations, to test whether the chain has converged. Raftery & Lewis (1992) proposed the use of the estimated transition matrix to decide how long the chain should be run. However Gelman & Rubin (1992) argued it is difficult to obtain sufficient diagnosis from one chain. A single chain easily gets stuck in one region, giving the wrong impression of convergence. They argued that running several chains with different starting points could avoid this problem. They developed an index with which to compare variation of the samples within and between the different chains. If the variations are similar, the chains have converged to stability. If the variations have large difference, the chains have not converged. Brooks & Gelman (1998) extended this method to use covariance as well as variance within and between chains. Zellner and Min (1995) also proposed other convergence criteria.

7.2 Discretized MCMC

7.2.1 The Computational 'wall'

In the use of MCMC, after a long 'burn in' process, a single sample is selected in each of several samples. To get enough samples a very long chain needs to be simulated.

In our study, a posterior density is likelihood function multiplied by prior distribution. During calculation of the likelihood function (4.16) in Chapter 4, pollution exposure needs to be calculated for each subject (In Britain there are about 18,000 subjects in total, while in London there are about 1,000 subjects). The pollution exposure for the simplest model in Chapter 4 (model 1) is calculated as below:

$$y' = \sum_{m=1}^{4} \sum_{j} I_j(m) T_j(m) \tag{7.2}$$

$$T = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{a^2 \sin^2 \omega}{2\sigma^2}} \left(\Phi\left(\frac{b - a\cos\omega}{\sigma}\right) - \Phi\left(-\frac{a\cos\omega}{\sigma}\right) \right) \tag{7.3}$$

where notations are as defined in Chapter 4.

Pollution exposure of a subject is the sum of the pollution exposures from each road segment. There are nearly 10,000 road segments of motorway, 140,000 segments of trunk road, 170,000 segments of principle road and 1,000,000 segments of minor road (though not all the segments are used when a small region is studied). For each pollution exposure, some complicated functions need to be calculated, i.e. sin, exp and Φ (a NAG routine S15ABF is called to get its value).

It takes seven minutes to calculate one posterior function value (likelihood function multiplied by the prior distribution) of the model for the London region. Using MCMC method, if each sample is selected among 1000 steps. To get 1000 samples, the MCMC chain has to run 1,000,000 steps. That takes 4862 days on a SUN workstation. This does not even take the 'burn in' process in account. Therefore it is impossible to simulate enough samples for MCMC be usable.

7.2.2 Discretization method

From the previous section, concerned with computational difficulty of using MCMC on complex spatial models, it is clear that if the time required calculating the likelihood function is reduced, MCMC still can be implemented.

During the evaluation of likelihood function (4.16), getting the value of y′ (pollution exposure) takes most of the time. Calculation (for the simplest model) of y′ only involves parameter σ ((4.11)). If the values of σ are restricted to a limited set of values, then it is possible to calculate y′ for all σ-values and to save them in a look-up table. Hence, during the calculation of likelihood function, values of y′ only needed to be obtained from the look-up table. The method is simple to use, very fast, and hence MCMC becomes applicable in our pollution-asthma model analysis. This method is called Discretized MCMC (DMCMC).

Assume a posterior function Po (i.e. likelihood function multiplied by the prior distribution) has this form:

$$Po(\theta,\alpha) = f_1(\theta,\alpha)\prod_l f(x_l \mid \theta,\alpha) \tag{7.4}$$

where $f(x_l \mid \theta,\alpha) = g(h(x_l \mid \theta) \mid \alpha)$ (7.5)

θ, α can be one dimensional or multi-dimensional parameters, and θ takes values from a region in $\mathbf{R}^M$. $h(x_l \mid \theta)$ is a complex function, which is very time-consuming to evaluate. Function g is a simple function which is not time-consuming.

Using MCMC, in each simulation step, the value of $h(x_l \mid \theta)$ needs to be calculated (see (7.5)). So each evaluation takes a long time.

If θ is restricted to take discrete values: $\{v_1, v_2, v_3, \ldots v_N\}$ (Notation v_i is used, instead of θ_i to avoid confusion with θ_i as a component of vector θ), N is large and the values $\{v_1, v_2, v_3, \ldots, v_N\}$ cover the relevant region of θ sufficiently finely, then the results of continuous-θ and discrete-θ MCMC will not be materially different.

In this case the posterior function becomes

$$Po = f_1(\theta,\alpha)\prod_l f(x_l \mid \theta,\alpha) = f_1(\theta,\alpha)\prod_l g(h(x_l \mid \theta) \mid \alpha) \qquad (7.6)$$

$\theta \in \{v_1, v_2, v_3, \ldots v_N\}$

In the beginning of MCMC simulation, all the $h(x_l \mid \theta)$ values for θ in $\{v_1, v_2, v_3, \ldots v_N\}$. $h(x_l \mid \theta = v_i)$ can be calculated and stored in the computer memory. During MCMC, only simple computing of function g need be calculated.

There is a special case which yields even more time saving. For a multi-dimensional parameter $\theta=(\theta_1,\theta_2,\theta_3,\ldots,\theta_M)$, If $h(x|\theta)$ has the form

$$h(x \mid \theta) = k(h_1(x \mid \theta_1), h_2(x \mid \theta_2), \ldots, h_M(x \mid \theta_M)) \qquad (7.7)$$

in which the most computing costuming part is function h_j, in which $k(\cdot)$ is a simple function, then we can change all θ_j to discrete values $\{v_1^j, v_2^j, \ldots v_{N_j}^j\}$, (the discrete values can be different for each θ_j).

Then all $h_j(x_l \mid v_i^j)$ for x_l (x value for subject l) and discrete values of θ_j, can be calculated and stored in a look-up table.

7.2.2.1 Algorithm

We now implement MCMC on the new discretized likelihood function. However we have not stated how this should be done. There are several possible ways. The basic approach is to treat the whole posterior distribution as one distribution. This is the usual approach to MCMC.

Another approach is related to the 'simulated tempering' method (Marinari and Parisi, 1992, see 7.1.1.3). Different discrete values v_i are treated as different 'temperatures'. The reason to choose this approach is that the posterior distributions conditioning on different v_i have similar shapes. In this situation, 'simulated tempering' method has faster mixing time compare to ordinary MCMC method (Marinari and Parisi, 1992). Also, the simulated tempering approach is algorithmically simpler than the ordinary MCMC.

Since in our case there are a much larger number of temperatures than in usual simulated tempering, we suggest the jump step between temperatures should be much larger. The chain not only jumps to the two adjacent temperatures but to 2K adjacent temperatures (K is a suitable number).

We rearrange our notation. After discretization of θ to $\{v_i : i=1,\ldots, N\}$, the posterior density function (7.4) is

$Po(\alpha, v_i)$, $i=1,\ldots, N$.

We also denote function $Po(\alpha, v_i)$ by $Po_i(\alpha)$.

The joint distribution of α and i forms a new distribution. A Markov chain whose state is (α, i) can be updated by the following steps. It includes updating α and 'temperature' i. Let K be the width of jump between 'temperatures'.

1.Update α using a Metropolis-Hastings or Gibbs update for $Po_i(\alpha)$.

2.Update i,

1) Set

$$j = i + k \begin{cases} k = -i+1,\ldots, K & \text{if} \quad i \le K \\ k = -K,\ldots, K & \text{if} \quad K < i < N-K+1 \\ k = -K,\ldots, N-i & \text{if} \quad i \ge N-K+1 \end{cases}$$

according to probabilities $q_{i,j}$, where $q_{i,j}$ is

$$q_{i,i+k} = \begin{cases} 1/(i+K) & i \le K \\ 1/(2K+1) & K < i < N-K+1 \\ 1/(K+N-i+1) & i \ge N-K+1 \end{cases}$$

2). Calculate the Hastings ratio

$$r = \frac{Po_j(\alpha)}{Po_i(\alpha)} \frac{q_{j,i}}{q_{i,j}}$$

and accept the transition (set i to j) or reject it according to the Metropolis rule: accept with probability $\min(r,1)$.

If more than one variable is discretized, we can call the set of discretized parameters a 'multi-dimension temperature'. When we update the temperature, we can update each dimension one by one.

The posterior density function is:

$\mathrm{Po}(\alpha, \mathrm{v}_{i_1}^1, \cdots, \mathrm{v}_{i_m}^m, \cdots, \mathrm{v}_{i_M}^M)$, $i_m = 1, \cdots, n_m$.

Let $\mathrm{Po}_{i_1,\ldots,i_M}(\alpha) = \mathrm{Po}(\alpha, \mathrm{v}_{i_1}^1, \cdots, \mathrm{v}_{i_m}^m, \cdots, \mathrm{v}_{i_M}^M)$,

$(\alpha, i_1, ..., i_M)$ forms a new MCMC chain by updating it using the following steps:

1.Update α using a Metropolis-Hastings or Gibbs update for $\mathrm{Po}_{i_1,\ldots,i_M}(\alpha)$.

2. Update i_m, m=1,…, M one by one. For i_m the rule is:

1) Set

$$j_m = i_m + k_m \quad \begin{cases} k_m = -i_m + 1, ..., K_m & \text{if} \quad i_m \le K_m \\ k_m = -K_m, ..., K_m & \text{if} \quad \mathrm{K_m} < i_m < N_m - K_m + 1 \\ k_m = -K_m, ..., N_m - i & \text{if} \quad i_m \ge N_m - K_m + 1 \end{cases}$$

according to probabilities q_{i_m, j_m}^m, where q_{i_m, j_m}^m is

$$q_{i_m, i_m + k_m}^m = \begin{cases} 1/(i_m + K_m) & i_m \le K_m \\ 1/(2K_m + 1) & \mathrm{K_m} < i_m < N_m - K_m + 1 \\ 1/(K_m + N_m - i + 1) & i_m \ge N_m - K_m + 1 \end{cases}$$

2) Calculate the Hastings ratio

$$r = \frac{Po_{i_1,\cdots,i_{m-1},j_m,i_{m+1},\cdots i_M}(\alpha)}{Po_{i_1,\cdots,i_{m-1},i_m,i_{m+1},\cdots i_M}(\alpha)} \frac{q_{j_m,i_m}^m}{q_{i_m,j_m}^m}$$

and accept the transition (set i_m to j_m) or reject it according to the Metropolis rule: accept with probability min(r,1)

It can be prove that the equilibrium distribution of the Markov Chain of $(\alpha, i_1, ..., i_M)$ (or (α, i) in the one dimension case) is $\mathrm{cPo}(\alpha, i_1, ..., i_M)$. The proof is shown in Appendix D.

7.2.2.2 Comparisons with continuous parameter MCMC

The main computational effort of the discretized method is in preparing the pre-calculated look-up table. The time depends on how fine the discretization is. Let N be

the number of v_i. For continuous MCMC, the speed depends how many samples are needed, say n, how many steps the burn in process takes, say k, and from how many steps one sample is selected, say m. Then the total the number of evaluations of likelihood function will be $nm + k$. If $N << nm + k$, the discretized method is useful. If $N \geq nm + k$, then discretized MCMC should not be used.

From the computing view, the smaller N is the better. But when N is too small, the θ-interval will be too big. It is then more likely that the discretization would not be acceptable as an approximation of the continuous-θ likelihood function.

7.2.3 Application of the DMCMC to the roads-asthma models

If the likelihood function is $L(\cdot)$, the prior distribution for parameters is $pr(\cdot)$, the posterior distribution density is proportional to $L(\cdot)pr(\cdot)$.

We have three models (4.13), (4.14) (4.15). Their likelihood functions are:

$L(\rho,\alpha,\sigma,\gamma)$,

$L(\rho,\alpha,\sigma_1, \sigma_2 ,\sigma_3 ,\sigma_4,\gamma)$,

$L(\rho,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\sigma_1,\sigma_2,\sigma_3,\sigma_4,\gamma)$.

where σ, σ_1, σ_2, σ_3, σ_4 are the dispersal parameters for pollution exposure. Parameters α, α_1, α_2, α_3, α_4 capture the association between asthma and pollution exposure and γ captures the association between asthma and gender.

Since there is no prior knowledge about the parameters available, uniform distributions are used as their prior distributions.

Parameter ρ follows prior which is continuous uniform distribution on [-100,100].

Parameter γ follows prior which is continuous uniform distribution on [-100,100].

Dispersal parameters, σ, σ_1, σ_1, σ_1, σ_1 follow continuous uniform distributions on [0.1, 5]. The lower limit has been chosen to be a small positive value, to avoid the consequences of the situation where a few asthma cases have home locations which coincide with a road location. In this situation it was found that the estimate of σ

was zero! This is clearly not appropriate since we know that the pollution affects from roads does disperse. The upper limit of this vague prior distribution for σ has been chosen at 5kmfor two reasons: (i) the calls of plausibility; it is not considered plausible that roads would have a measurable effect on pollution exposure at distances of over 5km, particularly since we have assumed a symmetric normally distributed pollution exposure dispersal model. Further, we would not expect there to be associated asthma events at such large distances. (ii) Since the road networks in the (residential) areas we have considered are relatively dense, and of similar density in all areas, a pollution dispersal model extending over 5 km would result in a fairy common level of exposure of the subjects. This effect can be seen in Fig. 4.2, where the s.d. of the exposure distribution decreases as σ increases. Furthermore, increasing the value of σ beyond 5km has a relatively small effect on the distribution of the exposure. We would therefore expect that the estimated value of σ would not achieve this boundary value if the model of asthma causation is appropriate to the asthma incidence data. We shall see later that this expectation is largely not fulfilled.

Since the hypothesis is increasing pollution causes increasing asthma occurrence, we let the prior of α (models 1 and 2) α_1, α_2, α_3, α_4 (model 3) be uniform distributions on [0,100].

The form of prior distribution can affect the posterior distribution, hence estimation and inference. Other forms of prior distribution of α (or α_1, α_2, α_3, α_4) have been considered. Only α will be mentioned in this paragraph for simplicity. Since α only is greater than (or equal to) zero, the prior distribution of α should take values from regions greater than 0 (including 0). The Gamma distribution is often used as a prior distribution in this situation. However it is not suitable for our study because $\alpha=0$ is the H_0. If prior is a Gamma distribution (whose density is zero at $\alpha=0$), the marginal posterior distribution of α will always be zero at $\alpha=0$. Hence the test of $\alpha=0$ will always be rejected (see illustration in Figure 7.1). So the Gamma distribution is not an appropriate prior distribution.

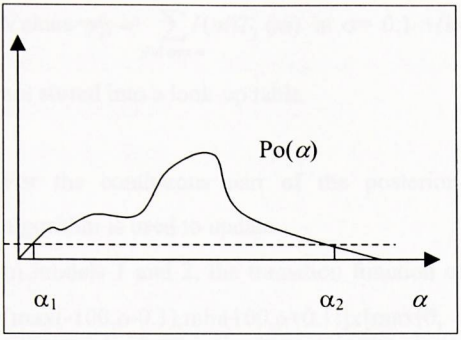


Figure 7.1 An illustration of the fact that $\alpha=0$ is outside of the 'confidence interval'. Po(α) is the posterior distribution.

A prior distribution over region (0, U) (U is a suitable value larger than 0) whose density is greater than zero when $\alpha=0$ is more appropriate. An exponential distribution also is considered. Since it has high density at 0, the posterior density will have higher value than posterior density obtained using uniform distribution. The confidence interval will have lower left side. If $\alpha=0$ is accepted by using uniform distribution, it will be accepted by using exponential distribution. Since there is no information available on the values of parameter, a uniform distribution is a reasonable choice for the study.

The discretized MCMC method has been applied on these three models to fit London region data. Samples from posterior distribution will be used to obtain confidence interval to test if $\alpha=0$.

Pollution exposure only involves parameter σ (model 1), or σ_1, σ_2, σ_3, σ_4 (model 2, 3), they are changed to be discrete values.

$\sigma_i \in [0.1, 5]$

let $\sigma = 0.1 + (k-1)\, 0.01$. $0.1 \le \sigma \le 5$, $k=1,\ldots,491$.

There are 491 values of σ, so there are 491 temperatures. After some experimentation, the width of jump is set as 20 in the updating steps.

Values $y'_m = \sum_{j \text{ of type } m} I(m)T_j(m)$ at $\sigma = 0.1 + (k-1)\ 0.01$, (k=1,…,491) for each subject are stored into a look-up table.

For the continuous part of the posterior distribution, the Metropolis-Hastings algorithm is used to update.

In models 1 and 2, the transition function of (ρ, α, γ), is a uniform distribution on $[\max(-100,\rho-0.1),\min(100,\rho+0.1)]x[\max(0, \alpha-10),\min(100, \alpha+10)]x[\max(-100, \gamma-0.1),\min(100, \gamma+0.1)]$.

For model 3, the transition function of $(\rho,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\gamma)$, is a uniform distribution on

$[\max(-100,\rho-0.1),\min(100,\rho+0.1)]x[\max(0, \alpha_1-10),\min(100, \alpha_1+10)]x$

$[\max(0, \alpha_2-10),\min(100, \alpha_2+10)]x[\max(0, \alpha_3-10),\min(100, \alpha_3+10)]x$

$[\max(0, \alpha_4-10),\min(100, \alpha_4+10)]x\ [\max(-100, \gamma-0.1),\min(100, \gamma+0.1)]$.

The choices of values of intervals (0.1, 10 etc) are based on experience and data from previous chapters to achieve high speed of convergence. If small change of a parameter causes large change on posterior function, the interval is set smaller, otherwise larger.

Several chains from different start points are simulated. In each chain, the first 100,000 steps are thrown away as 'burn in' process. A sample is obtained from each 1,000 steps thereafter. 1,000 samples are obtained from each chain.

Two methods are used to diagnose convergence: (1) a comparison of variances and means from different chains, to see if the summaries of posterior distributions are the same. (2) Draw Q-Q plot for parameters from one chain against parameters from another chain.

7.2.3.1 Model 1

For model 1, eight Markov chains whose equilibrium distributions are the posterior distribution have been created using DMCMC. Each starts from different starting points.

For four parameters, the variances, means, minimum values and maximum values from each chain are compared. For each parameter, the summaries from eight chains are presented in one table (Tables 7.1, 7.2, 7.3, and 7.4). Also summaries of log-likelihood functions from eight chains are presented in Table 7.5.

	1	2	3	4	5	6	7	8
Min	-1.63	-1.63	-1.47	-1.47	-1.45	-1.44	-1.49	-1.47
1st Qu	-1.09	-1.09	-1.07	-1.08	-1.08	-1.08	-1.08	-1.08
Mean	-0.9997	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99
Median	-0.99	-0.99	-0.98	-0.98	-0.98	-0.98	-0.98	-0.97
3rd Qu	-0.903	-0.90	-0.89	-0.89	-0.89	-0.89	-0.89	-0.89
Max:	-0.63	-0.63	-0.65	-0.64	-0.59	-0.64	-0.65	-0.55
std Dev	0.139	0.139	0.133	0.133	0.134	0.137	0.137	0.142

Table 7.1 Summaries of samples of ρ from eight chains, model 1

	1	2	3	4	5	6	7	8
Min	0.0010	0.0007	0.0002	0.0009	0.0002	0.0022	0.0019	0.0012
1st Qu	0.2264	0.2124	0.2163	0.2329	0.2219	0.2341	0.2103	0.2242
Mean	0.5000	0.4963	0.4988	0.5100	0.5081	0.5400	0.4967	0.5105
Median	0.4359	0.4050	0.4207	0.4136	0.4244	0.4525	0.4168	0.4290
3rd Qu	0.6819	0.6777	0.6998	0.7064	0.6956	0.7414	0.6873	0.7046
Max:	2.2725	2.0821	2.3665	2.5549	2.1545	2.4865	2.2778	2.3574
std Dev	0.3728	0.3845	0.3786	0.3875	0.3864	0.4123	0.3810	0.3882

Table 7.2 Summaries of samples of α from eight chains, model 1

	1	2	3	4	5	6	7	8
Min	0.12	0.21	0.10	0.10	0.14	0.11	0.12	0.11
1st Qu	1.76	1.76	1.87	1.73	1.98	1.84	1.70	1.87
Mean	2.90	2.93	2.98	2.94	3.07	3.01	2.94	3.01
Median	3.06	3.10	3.16	3.08	3.28	3.18	3.07	3.20
3rd Qu	4.09	4.09	4.17	4.21	4.26	4.19	4.25	4.21
Max:	4.96	4.99	4.99	4.99	4.99	4.99	4.99	4.99
std Dev	1.34	1.32	1.34	1.38	1.32	1.34	1.37	1.34

Table 7.3 Summaries of samples of σ from eight chains, model 1

	1	2	3	4	5	6	7	8
Min	-0.77	-0.72	-0.85	-0.82	-0.813	-0.77	-0.72	-0.77
1st Qu	-0.41	-0.43	-0.43	-0.43	-0.425	-0.42	-0.42	-0.44
Mean	-0.32	-0.33	-0.33	-0.34	-0.33	-0.33	-0.33	-0.34
Median	-0.32	-0.33	-0.33	-0.34	-0.32	-0.33	-0.33	-0.34
3rd Qu	-0.22	-0.24	-0.24	-0.25	-0.23	-0.24	-0.25	-0.24
Max:	0.14	0.08	0.10	0.09	0.062	0.11	0.17	0.03
std Dev	0.14	0.14	0.14	0.14	0.137	0.14	0.14	0.14

Table 7.4 Summaries of samples of γ from eight chains, model 1

	1	2	3	4	5	6	7	8
Min	-619.44	-614.19	-614.76	-615.16	-615.45	-616.29	-615.02	-615.13
1st Qu	-609.62	-609.57	-609.58	-609.61	-609.53	-609.75	-609.53	-609.64
Mean	-609.15	-609.04	-609.06	-609.09	-609.07	-609.16	-609.017	-609.15
Median	-608.81	-608.72	-608.74	-608.81	-608.76	-608.82	-608.75	-608.85
3rd Qu	-608.23	-608.22	-608.25	-608.23	-608.25	-608.25	-608.22	-608.27
Max	-607.56	-607.53	-607.55	-607.59	-607.58	-607.58	-607.53	-607.57
std Dev	1.27	1.12	1.14	1.14	1.14	1.19	1.08	1.21

Table 7.5 Summaries of log-likelihood from eight chains, model 1

In these tables, the means and variations from different chains are similar. It indicates these Markov chains have 'converged'.

Q-Q graphs (Gelfand & Smith, 1990) are also used to diagnose the convergence. Instead of comparing samples from a single chain, samples from two different chains are compared. The Q-Q graph is obtained by samples from the first chain against samples from the second chain. For each parameter and log-likelihood function, the Q-Q plots are presented in Figure 7.2. They are all close to y=x, that is an evidence of convergence.

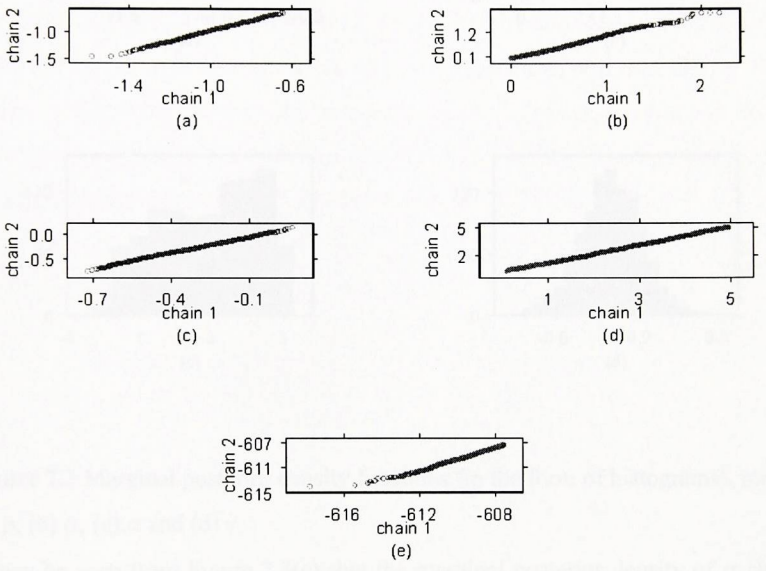


Figure 7.2 Q-Q plots of samples of parameters from chain 1 against chain 2, model 1. (a) ρ, (b) α, (c) γ, (d) σ and (e) log-likelihood

The marginal posterior densities of the four parameters are plotted as histograms (Figure 7.3).

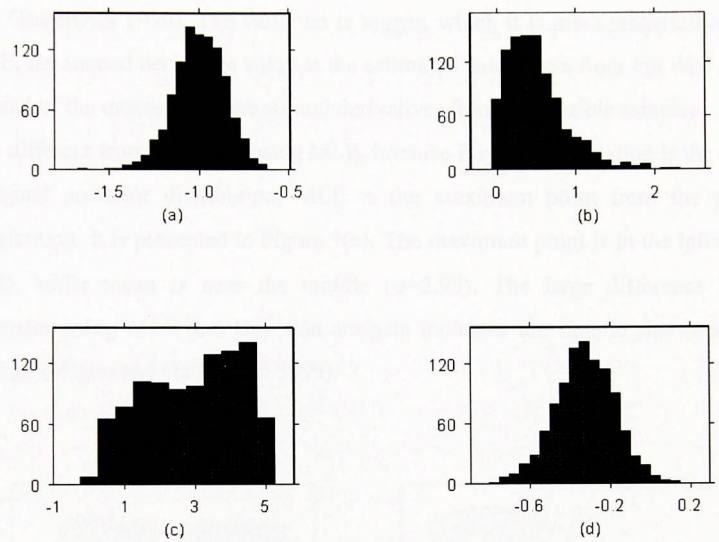


Figure 7.3 Marginal posterior density functions (in the form of histograms), model 1. (a) ρ, (b) α, (c) σ and (d) γ

It may be seen from Figure 7.3(c) that the marginal posterior density of σ shows a cut-off at the 5km value, heavily reflecting the form of prior used. Hence the Bayesian estimates of σ are negatively biased by this boundary effect. A further research might therefore either consider the use of a prior distribution which is not uniform, or the use of a much larger upper limit of the uniform prior. This work has not been done as part of this study. In spite of this we proceed.

The estimates from the marginal posterior density are: ρ = -0.9997 (-1.27, -0.75), α = 0.496(0, 1.28), σ = 2.93(0.55, 4.93), γ = -0.33(-0.61, -0.07). (The interval inside the bracket is the 95% confidence interval).

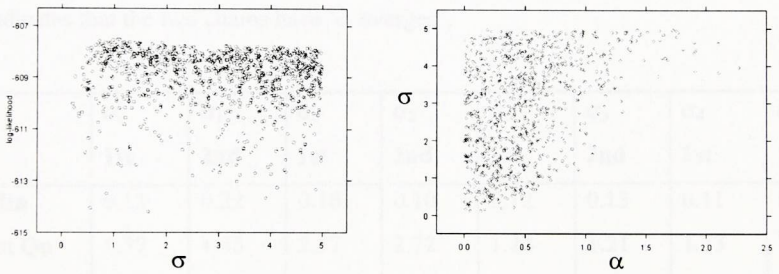
95% confidence interval is defined as HPD (highest posterior density, Migon and Gamerman 1999) which is a region A, which satisfies:

$Po(\theta_1) \geq Po(\theta_2)$, $\theta_1 \in A$, $\theta_2 \notin A$ and $\int_{\theta \in A} Po(\theta)dx = 0.95$.

Po(θ) is the posterior density function of θ.

Most of the estimates are similar to the estimates using MLE from chapter 4 (Migon and Gamerman 1999). The variation is bigger, which it is more realistic because in MLE, the second derivative value at the estimated parameters from our data are used instead of the means of all the second derivatives from all possible samples.

σ is different from estimates using MLE, because Bayesian estimation is the mean of marginal posterior distribution, MLE is the maximum point from the posterior distribution. It is presented in Figure 7(a). The maximum point is in the left side (σ= 0.98), while mean is near the middle (σ=2.99). The large difference between estimates using MLE and Bayesian analysis indicates the sample size is not large enough (Migon and Gamerman 1999).



(a) (b)

Figure 7.4 (a) Plot of log-likelihood function value against σ. (b) Plot of α against σ

From the samples, correlation between α and σ can also be calculated, it is 0.41. The correlation between these two parameters can be seen from Figure 7.4 (b). σ is the dispersal parameter. When σ goes larger (after a certain extent), the mean pollution exposures $\overline{y}'(\sigma)$ which subjects get is smaller. To make $\alpha\,\overline{y}'(\sigma)$ not changed much (to fit the data), α will be bigger.

The test for H_0: $\alpha=0$ is to see if 0 is inside the 95% confidence interval of marginal posterior distribution of α. If outside the interval, it is rejected. If inside the interval it is accepted.

The marginal posterior density has a high value at $\alpha=0$ (Figure 7.3, b), so the left side of confidence interval is 0. The 95% confidence interval of marginal posterior distribution of α is (0, 1.28). α is inside the interval, so $\alpha=0$ is accepted, i.e. there is no significant relationship between asthma and road pollution.

7.2.3.2 Model 2

For model 2, samples from posterior distribution are obtained using DMCMC. Two Markov chains are created. Each starts from different starting points.

The comparison of summaries of parameter estimates and log-likelihood values are presented in Tables 7.6 and 7.7, The summaries from two chains are similar. It indicates that the two chains have 'converged'.

	σ_1 **1st**	σ_1 **2nd**	σ_2 **1st**	σ_2 **2nd**	σ_3 **1st**	σ_3 **2nd**	σ_4 **1st**	σ_4 **2nd**
Min	0.13	0.22	0.16	0.10	0.12	0.13	0.11	0.11
1st Qu	1.37	1.35	2.57	2.72	1.24	1.21	1.13	1.03
Mean	2.52	2.50	3.37	3.43	2.37	2.34	2.42	2.34
Median	2.38	2.42	3.65	3.72	2.33	2.22	2.40	2.35
3rd Qu	3.75	3.68	4.41	4.42	3.48	3.39	3.70	3.52
Max	4.99	4.99	4.99	4.99	4.98	4.99	4.98	4.99
Std Dev	1.34	1.32	1.22	1.21	1.33	1.34	1.44	1.45

Table 7.6 Summaries of samples of σ_1, σ_2, σ_3, σ_4 from two Markov chains, model 2 ('1st' means the first chain, '2nd' means the second chain)

	ρ 1st	ρ 2nd	α 1st	α 2nd	γ 1st	γ 2nd	Log-like 1st	Log-like 2nd
Min	-1.58	-1.53	0.0003	0.0004	-0.82	-0.79	-614.87	-615.36
1st Qu	-1.11	-1.11	0.23	0.24	-0.43	-0.44	-609.35	-609.31
Mean	-1.02	-1.02	0.51	0.52	-0.33	-0.33	-608.75	-608.72
Median	-1.01	-1.01	0.45	0.46	-0.33	-0.33	-608.48	-608.42
3rd Qu	-0.91	-0.91	0.71	0.73	-0.23	-0.24	-607.89	-607.86
Max	-0.63	-0.54	2.08	2.28	0.18	0.18	-606.07	-606.41
Std Dev	0.15	0.15	0.36	0.35	0.14	0.14	1.27	1.28

Table 7.7 Summaries of samples of ρ, α, γ and the log-likelihood function from two Markov chains, model 2 ('1st' means the first chain, '2nd' means the second chain)

Q-Q plots are drawn to compare parameters from these two chains (Figures 7.5 and 7.6). The Q-Q plots are all close to y=x. which gives further evidence that these two chains have converged.

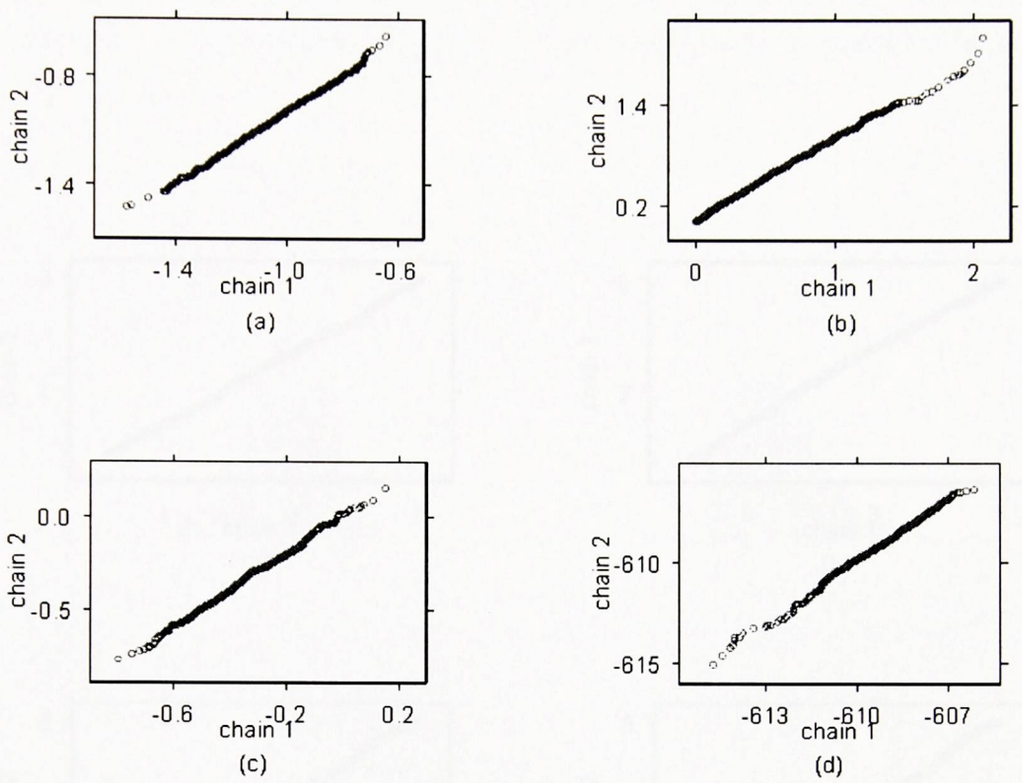


Figure 7.5 Q-Q plots of samples of parameters from chain 1 against chain 2, model 2. (a) ρ, (b) α, (c) γ and (d) log-likelihood function

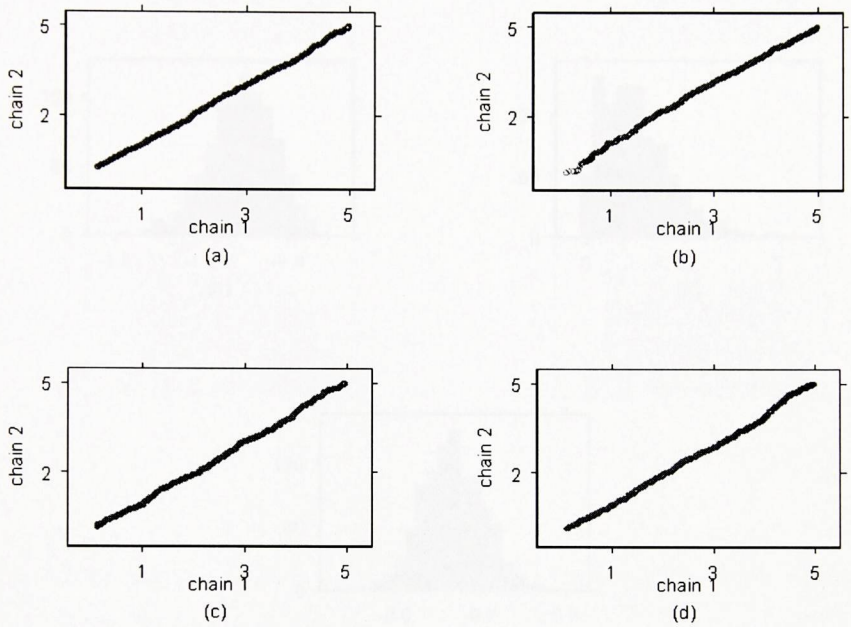


Figure 7.6 Q-Q plots of samples of parameters from chain 1 against chain 2, model 2. (a) σ_1, (b) σ_2, (c) σ_3 and (d) σ_4

The marginal posterior densities of parameters are plotted in Figures 7.7 and 7.8.

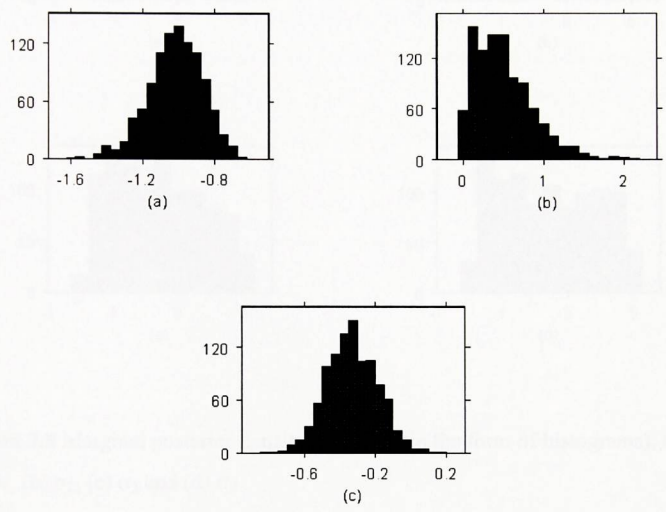


Figure 7.7 Marginal posterior density functions (in the form of histograms), model 2. (a) ρ, (b) α and (c) γ

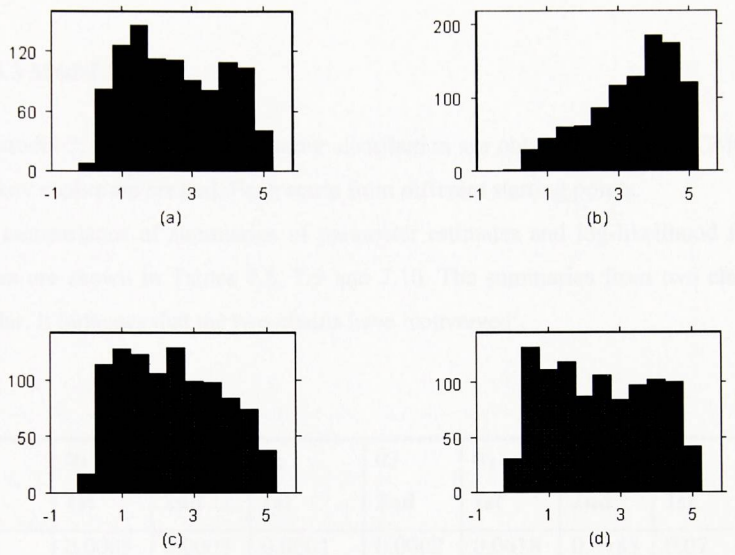


Figure 7.8 Marginal posterior density functions (in the form of histograms), model 2. (a) σ_1, (b) σ_2, (c) σ_3 and (d) σ_4

We note that the same comments about Figure 7.8 can be made as were made of Figure 7.3(c).

The Bayesian estimates and confidence intervals are

ρ = -1.0176 (-1.35, -0.74),

α = 0.5114 (0, 1.21),

γ = -0.3318 (-0.60, -0.05),

σ_1 = 2.5210(0.41, 4.88),

σ_2 = 3.3733(0.92, 5),

σ_3 = 2.3772(0.27, 4.84),

σ_4 = 2.4266 (0.22, 4.84).

The marginal posterior density has a high value at α=0 (Figure 7.7 b), so the left side of the confidence interval is 0. The 95% confidence interval of marginal posterior

distribution of α is (0, 1.21), α=0 is inside the confidence interval. α=0 is accepted, i.e. no significant relationship between road pollution and asthma is found.

7.2.3.3 Model 3

For model 3, samples from posterior distribution are obtained using DMCMC. Two Markov chains are created. Each starts from different starting points.

The comparisons of summaries of parameter estimates and log-likelihood function values are shown in Tables 7.8, 7.9 and 7.10. The summaries from two chains are similar. It indicates that the two chains have 'converged'.

	α_1 **1st**	α_1 **2nd**	α_2 **1st**	α_2 **2nd**	α_3 **1st**	α_3 **2nd**	α_4 **1st**	α_4 **2nd**
Min	0.0005	0.0008	0.0001	0.0002	0.0618	0.0185	0.07	0.0085
1st Qu	0.3728	0.3605	0.1089	0.1130	1.6311	1.55	3.83	3.94
Mean	1.0007	0.9633	0.4170	0.3870	2.8256	2.74	8.29	8.09
Median	0.7853	0.7840	0.2757	0.2590	2.5995	2.44	7.03	7.04
3rdQu	1.3273	1.3570	0.5622	0.5303	3.7251	3.65	11.30	11.01
Max	5.9501	4.8078	3.1621	2.5119	9.7735	10.62	41.38	48.41
Std Dev	0.8779	0.7997	0.4517	0.3956	1.5747	1.63	6.15	5.94

Table 7.8 Summaries of samples of α_1, α_2, α_3 and α_4 from two chains, model 3 ('1st' means the first chain, '2nd' means the second chain)

	σ_1 1st	σ_1 2nd	σ_2 1st	σ_2 2nd	σ_3 1st	σ_3 2nd	σ_4 1st	σ_4 2nd
Min	0.19	0.11	0.11	0.12	0.22	0.33	0.14	0.19
1st Qu	1.54	1.51	2.13	2.14	2.36	2.30	1.11	1.14
Mean	2.72	2.72	3.20	3.13	3.29	3.25	2.38	2.41
Median	2.63	2.65	3.52	3.40	3.42	3.39	2.05	2.17
3rdQu	3.96	3.94	4.33	4.27	4.33	4.31	3.73	3.74
Max	4.99	4.99	4.99	5.00	4.99	5.00	4.99	4.99
Std Dev	1.33	1.34	1.34	1.32	1.16	1.19	1.43	1.43

Table 7.9 Summaries of samples of σ_1, σ_2, σ_3, σ_4 from two chains, model 3 ('1st' means the first chain, '2nd' means the second chain)

	ρ 1st	ρ 2nd	γ 1st	γ 2nd	Log-like 1st	Log-like 2nd
Min	-2.35	-2.45	-0.84	-0.74	-617.99	-618.90
1st Qu	-1.65	-1.6295	-0.45	-0.44	-609.57	-609.44
Mean	-1.47	-1.46	-0.35	-0.35	-608.57	-608.47
Median	-1.45	-1.43	-0.35	-0.35	-608.24	-608.19
3rdQu	-1.28	-1.277	-0.25	-0.26	-607.15	-607.19
Max	-0.85	-0.81	0.12	0.12	-605.14	-604.47
Std Dev	0.26	0.26	0.147	0.14	1.926	1.854

Table 7.10 Summaries of samples of ρ, γ, log-likelihood function value from two chains, model 3 ('1st' means the first chain, '2nd' means the second chain)

Q-Q plots are drawn to compare parameter estimates from these two chains (Figures 7.9, 7.10 and 7.11). The Q-Q plots are close to y=x. It gives further evidence that these two Markov chains have 'converged'.

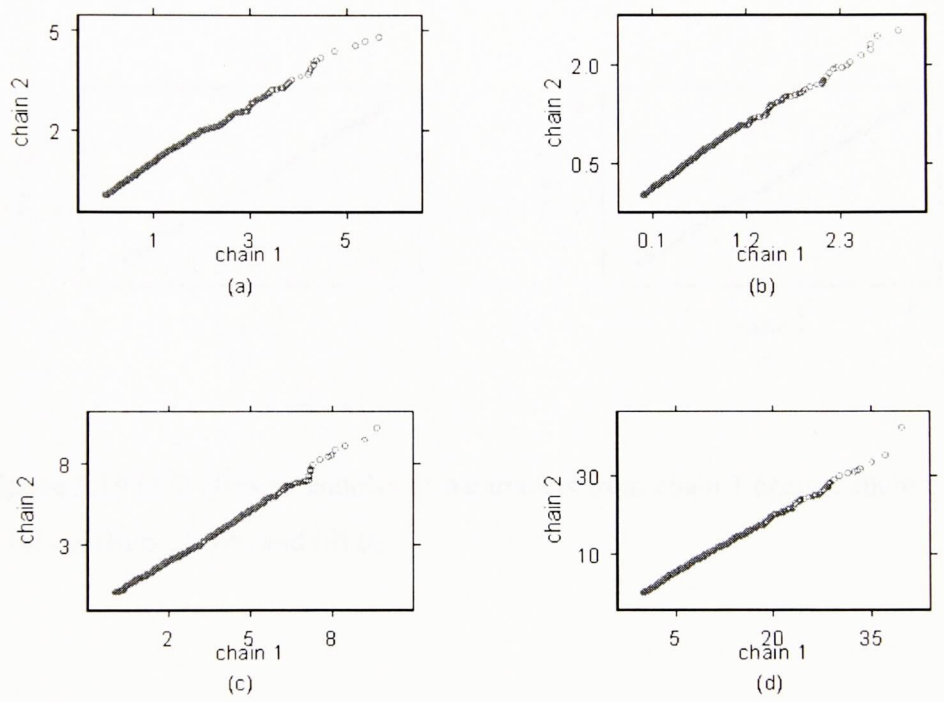


Figure 7.9 Q-Q plots of samples of parameters from chain 1 against chain 2, model 3. (a) α_1, (b) α_2, (c) α_3 and (d) α_4

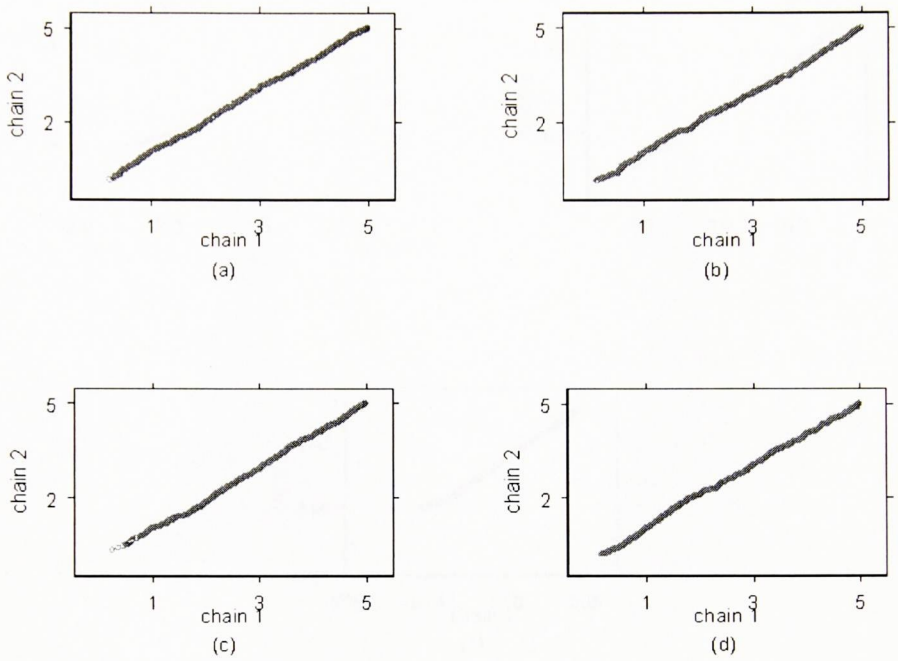


Figure 7.10 Q-Q plots of samples of parameters from chain 1 against chain 2, model 3. (a) σ_1, (b) σ_2, (c) σ_3 and (d) σ_4

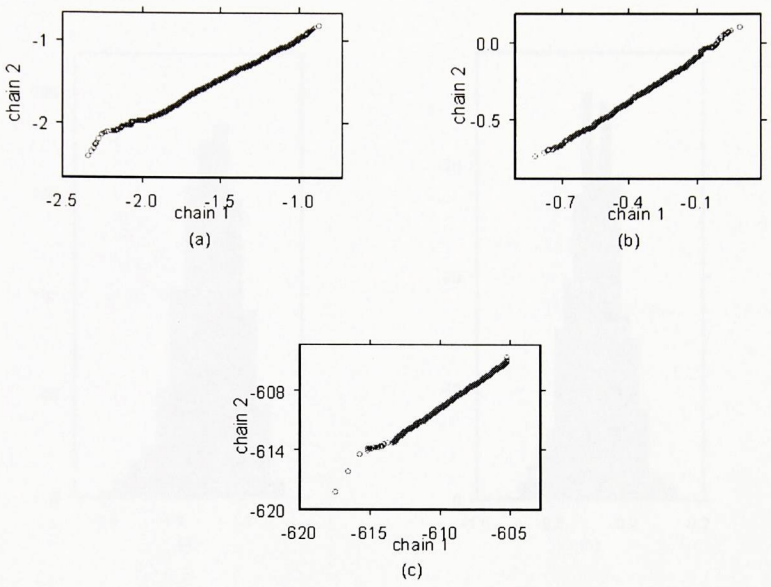


Figure 7.11 Q-Q plots of samples of parameters from chain 1 against chain 2, model 3. (a) ρ, (b) γ and (c) log-likelihood function

The marginal posterior densities of parameters are shown in Figures 7.12, 7.13 and 7.14 in the form of histograms.

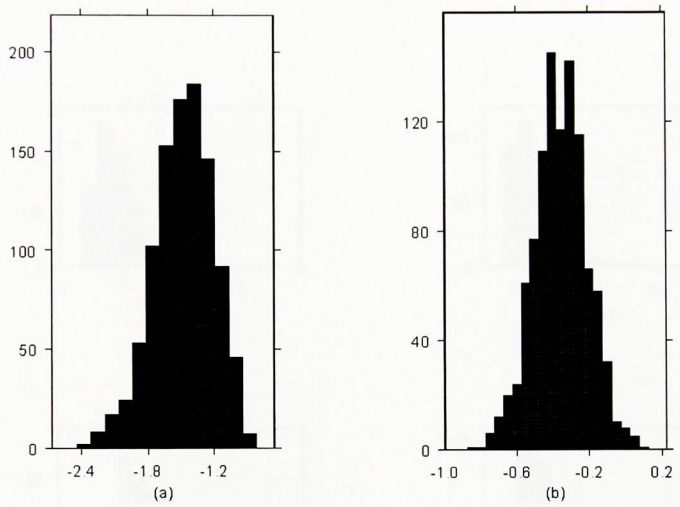


Figure 7.12 Marginal posterior density functions (in the form of histograms), model 3. (a) ρ and (b) γ.

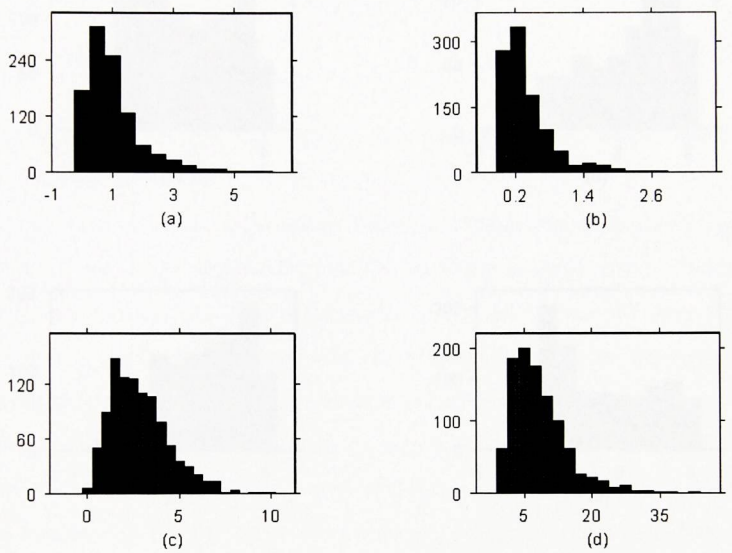


Figure 7.13 Marginal posterior density functions (in the form of histograms), model 3. (a) α_1, (b) α_2, (c) α_3 and (d) α_4.

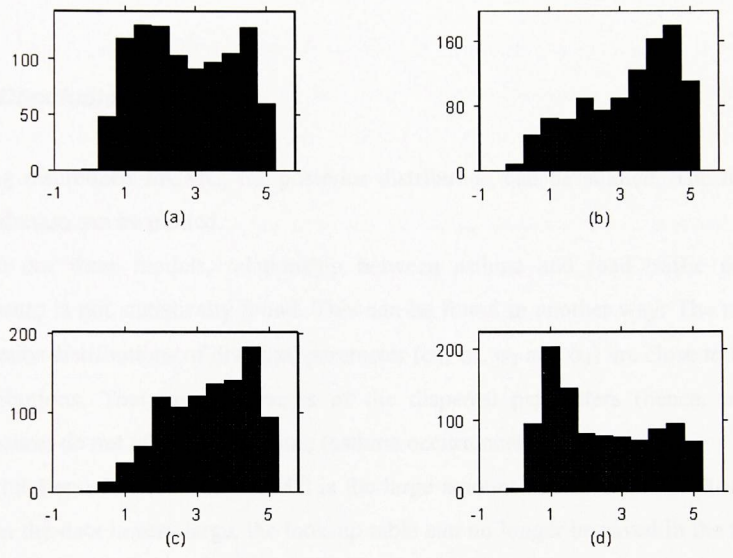


Figure 7.14. Marginal posterior density functions (in the form of histograms), model 3. (a) σ_1, (b) σ_2, (c) σ_3 and (d) σ_4.

We note that the same comments about Figure 7.14 can be made as were made of Figure 7.3(c).

The estimates and confidence intervals of parameters are

ρ = -1.4743 (-2.07, -1.01), γ = -0.3557 (-0.66, -0.08),

α_1 = 1.0007 (0, 2.84), α_2 = 0.4170 (0, 1.42),

α_3 = 2.8256 (0, 5.87), α_4 = 8.2925 (0, 20.16),

σ_1 = 2.7212 (0.58, 4.88), σ_2 = 3.2044 (0.73, 5),

σ_3 = 3.2919 (1.23, 5), σ_4 = 2.3888 (0.42, 4.92).

α_i=0 (i=1,2,3,4) are inside the confidence intervals. But α_3 is very close to the 95% interval, it is outside the 90% interval. That is consistent with the finding using sub-model with only road type 3 included (Table 4.6).

Our conclusion is α_i=0 (i=1, 2, 3, 4). There is no evidence of relationship between asthma and traffic pollution exposure from four types of roads.

7.3 Conclusion

Using discretized MCMC, the posterior distribution can be studied. The marginal distribution can be plotted.

From our three models, relationship between asthma and road traffic pollution exposure is not statistically found. This can be found in another way: The marginal posterior distributions of dispersal parameter (σ_1, σ_2, σ_3 and σ_4) are close to uniform distributions. That means changes of the dispersal parameters (hence, pollution exposure) do not affect the outcome (asthma occurrences).

A limitation of discretized MCMC is the large amount of data in the look-up table. When the data is very large, the look-up table can no longer be saved in the memory of the computer. It has to be read from files or swap memory on hard disk which is much slower than read directly from memory. This method will become slow. In our London data, the look-up table occupies 10M in the computer memory. Our computer can handle it in memory. But for data in Britain, the look-up table takes nearly 200M. It is larger than the SUN workstations which we use can handle. To implement discretized MCMC for such large amount of data, a less fine discretization has to be used.

The discretized MCMC has two parts: discretization and simulate tempering. Simulate tempering is used to implement MCMC. Discretization and the look-up table are used to save the computation time.

The way to use discretization and look-up table can also be used in other situations, such as Monte-Carlo integration of high dimensional function.

Chapter 8 Discussion and Conclusions

The Substantive results

The result of an extensive and fairly sophisticated modelling exercise on some fairly large data sets has failed to clearly demonstrate the influence of road traffic pollution on asthma incidence. This is rather unfortunate since this means that the research project has really added relatively little to the debate on the effects of traffic pollution on childhood asthma. No confirmation of the conclusions of the Committee on the medical effect of air pollution (1995), following their review of the relevant literature ("*there is a consistent, though modest, association between exposure to traffic and asthma prevalence in children*") has been provided.

It has to be admitted that a very simple view of the available data has been selected at earlier stages of this research project. That is, that only the data at age 16 has been used, only ever-asthma has been analysed in great detail, gender has been the only covariate used in the presented analysis. The reasons for all of these choices have been presented in the early chapters of this thesis, and are still considered to be justified. Also the putative source of pollution exposure has been taken to be a fixed road system with traffic intensities set for only four classes of road. Clearly this representation of the pollution source can only reflect an average level of exposure experienced by individuals. Our data contains no information on the extremes of pollution exposure that occur over relatively short time intervals, and similarly our response variable, being ever-asthma does not say much about the occurrence of acute asthma attacks. We would therefore not expect the models and analyses used in this thesis to reveal links between extreme levels of road pollution and acute attacks of asthma. It is exactly in this area of asthma aetiology that the anecdotal cases and the prior research studies have suggested the relationship exists. Consequently we should not be surprised or disappointed by our non-conclusive substantive results.

The well known fact that failure to demonstrate the significance of a conjectured effect does not prove that the effect is zero leaves the substantive conclusions of this research project to be rather weak. In such a situation it is still possible that an important effect really does exist, but that either the data set used is of insufficient size, or the methods of analysis are insufficiently sensitive.

In Chapter 5 the question of power has been considered. The power of the basic analysis method/model (of Chapter 4) has been investigated and the size of samples required to demonstrate the significance of a real effect equal to the estimated effect have been estimated. In summary, the power of the present study, using the models of Chapter 4 are of the order of 30%, (for a test of level 0.05). It has been estimated that for real effects equal to the size of those estimated for the NCDS population in London area, a sample of 16 times that available would have been required in order to demonstrate significance. This means that the data available from NCDS for the whole country might be regarded as of about the required size. However the differences between regions of the asthma relationships, as demonstrated between the London and Birmingham areas, (in Chapter 4), would mean that the rather simple model formulations of the models in Chapter 4 should either be regarded as inappropriate, or that they have a higher degree of residual variation. This latter position means that a larger sample than that were suggested in our power calculations would be required to demonstrate the asthma effects.

New data collection?

Access to such larger data sets was not a feasible alternative in this study, but it is suggested that it might be useful to consider the development of new datasets on the issue of asthma and many other conditions. The computerisation of most Medical General Practices in the UK, and the inclusion into such records of more than just age and sex means that there is a vast potential data resource for environmental epidemiology. The widely use of Internet means the data could be easily accessed. However, there are clearly major issues of confidentiality and organisation and such population based demographic/epidemiological studies can really only be envisaged

in the context of an MRC/NHS funded or managed project with considerable funding over long periods.

POSSIBLE IMPROVEMENTS in DATA QUALITY and MODEL SOPHISTICATION

1) Better road data.

The Bartholomew Digital Road Map used has a limited resolution. Even though it lists 22 different classes of roads it still does not include many of the small access roads in residential areas. Since the home location is surrounded by the roads of the residential area in which the home is located this might be reason for the insensitivity of the analysis presented in this thesis. There are more up to date digitised road maps currently available, but their use is part of future studies.

An alternative source of road information would be from satellite imagery. However, manual digitisation of the roads would be very time consuming. Automatic recognition and digitisation of roads is a difficult pattern recognition task which will be subject to considerable error.

Another of the weaknesses of the data used is the traffic intensity data. At the time when this study was initiated the only available data on traffic intensities was that from the Department of Transport. This was only an estimate average intensity, by FOUR major road types, over the WHOLE of the UK. Such data is clearly inadequate for the estimation of individual traffic pollution exposures, and this must certainly be one of the contributory factors to the study obtaining only null results.

It is conceivable that satellite imagery of roads, possibly in the night and early mornings, would provide local measures of traffic intensity on unlit roads, through the measurement of the headlight images. However there would be substantial problems in image analysis, and it is unlikely that such methods would be applicable to the small roads of the urban areas which are most important to asthma studies.

Furthermore, different intensities for diesel vehicles and petrol vehicles could be used to represent the difference of diesel and petrol pollution (McCready *et al.* 1997).

Finally, it is clear that the road maps are themselves subject to location errors. The magnitude of these errors is not known and has not been considered in this thesis. However, it is considered that effects on analysis and inference of indeterminacy of road locations would be roughly equivalent to slightly increased measurement error of home locations, (or an increase in the size of the individual utilisation distributions).

2) Space utilisation data

It has been assumed (as a simplification for modelling purposes) that subjects live entirely at spatial position corresponding to their home location. This is, of course untrue. In Chapter 6 this assumption has been relaxed by the consideration of ***measurement error models*** linked to the indeterminacy of the post-code of the home locations.

Within the NCDS dataset there is no data which can be used to give information about subject movement patterns about their home locations. Such as locations of schools and the routines used by a subject from home to school. Otherwise it would be possible to get a more appropriate measure of subjects' spatial utilisation. Hence it would be possible to derive an index of road traffic pollution exposure based on such an utilisation distribution and the local road network (using the GIS roads map data).

It is clear that space utilisation modelling would require extensive further data collection, which is outside of the scope of this project.

3) *Mixed effect models of pollution dispersion*

In the pollution dispersal models of Chapter 4, it has been assumed that the pollution dispersal range parameter, σ, is the same for all roads of the same type. Clearly this is far from the truth. The dispersal of traffic pollution must in fact depend on local

conditions such as exposure/cover from wind, and this will be influence by both the local topography and the building environment. Data on such variation is dispersal characteristics, is certainly not generally available over large regions and it would seem impractical to collect it. However, it would be possible to treat the dispersal range (σ) as a ***random effect*** (Zeger and Karim 1991) and to vary from site to site. A Bayesian mixed model approach could be adopted. More simply, different dispersal range parameters may be used for geographic strata, such as urban and rural. This would need a classification of road (segments) as urban or rural, a task which is not entirely obvious.

4) Use of spatio-temporal data and models

The nature of the NCDS data suggests that spatio-temporal models could be developed in which asthma incidence is related to (weighted) cumulative lifetime pollution exposure. However, the required spatial data on home locations prior to age sixteen is not available for the NCDS study. The repeated measures spatio-temporal models would be considerably more complex than the models considered in this thesis.

5) New survey

Considering the above discussions (especially 2) and 4)), it is suggested that, for a conclusive study, a new survey is needed. It should be conducted in a region such as the 'London region' on large number of subjects (such as all the children born within two or three certain months, which will give us around 20,000 subjects. It should be a sufficient number of subjects according the power analysis in Chapter 5). Unlike NCDS has hundreds variables, in the new survey only several variables are needed, such as gender, smoking habit, house location, school location, previous school and house locations, the frequency of asthma attacks.

A subject's cumulative lifetime pollution exposure can be obtained by summing of all the pollution exposures getting from current and previous home locations and school locations. For a subject has k previous addresses (include home and school), t_k

is the approximation time staying in that address, the pollution exposure for the subject is:

$$y' = \sum_k t_k \sum_{m=1}^{4} \sum_{j \text{ of } type\ m} I_j(m) T_j(m)$$

6) Different response variables

Instead of just using the binary ever-asthma variable as outcome, the degree of asthma severity could be used. For example the frequency of asthma attack, and a measure of their severity could be adopted as outcomes. This would require more sophisticated risk models than the simple logistic regression that has been used in this study.

7) Generalised Logistic Regression

In a logistic model, a subject will have asthma with certainty when pollution is very large. This may be not true in reality: some persons may be immune to asthma and they do not have asthma not matter how large the pollution is. A risk model with threshold may be more suitable. Other risk models can also be considered.

8) Survival models

If we treat the first time a subject has asthma attack as the study variable, then survival analysis models can be adopted. In this way, the research concentrates on whether pollution is the cause of asthma without been affected by the triggering of pollution on asthma (Committee on the medical effect of air pollution 1995).

Conclusion on methodology

The general methodology of this study may be of use in the analysis of other heath outcomes from a network pollution source. The methodology could also be extended

to deal with complicated sources which have continuous spatial distributions in two (such as radiation from a region), or possibly three dimensions. Our method put the spatial variable in a risk model (logistic regression in our case, can be extended to other models). In this way complicated pollution sources can be used.

Since complicated pollution source is used, the non-linear regression model becomes very complicated, the log-likelihood function takes long time to evaluate. MCMC method can not be implemented. Our discretized MCMC can be used in this situation. For discretized MCMC, a look-up table is created to store pre-calculated values. The limitation of this method is the large memory the look-up table occupies. Following the development of more powerful computer and cheaper memory, this method can be used in applications in wider areas.

References

Adcock, R. J. (1878), A problem in least squares. *Analyst* 5, 52-53.

Barnett, V. (1997), Statistical analyses of pollution problems, In *Statistics for the environment 3: pollution assessment and control.* pp.3-41, Edited by V.Barnett and K. Feridum Turkman, John Wiley & Sons.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice system, *JRSS B,* 36: 192.

Besag, J. and Green, P. (1993), Spatial statistics and Bayesian computation, *JRSS, B,* Vol. 55, No. 1, pp. 25-37.

Besag, J. and Newell, J. (1991), The detection of clusters in rare diseases, *JRSS A*, 154, pp.143-155.

Baxter, P. J., Ing, R., Falk, H. and Plikaytis, B. (1983), Mount St Helens eruption: the acute respiratory effect of volcanic ash in a North American community, *Arch Environ Health* 1983; 38, 138-143.

Brooks, S. and Gelman, A. (1998), General methods for monitoring convergence of iterative simulations, *Journal Of computational and graphical statistics*, 1998, Vol.7, No.4, pp.434-455.

Burney, P. G. J. (1992), Epidemiology. *Asthma,* edited by Clark TJH, Godfrey S, Lee TH, London, Chapman & Hall medical, 254-308.

Busse, W. W., Calhoun, W. F., and Sedgwick, J. D. (1993), Mechanism of airway inflammation in asthma, *Am rev Respir Dis* 1993, 147, pp520-524.

Butler, N. R. and Bonham, D. G. (1963), Perinatal mortality, Edinburgh, E & S Livingstone.

Cochran, W. G. (1968), Errors of measurement in statistics, *Technometrics* 10, 637-666.

Committee on the medical effect of air pollution (1995), Asthma and outdoor air pollution, London: HMSO,

Crager, M. R. (1987), Analysis of covariance in parallel-group clinical trails with pretreatment baselines, *Biometrics*, 43, 895-901.

Cox, N. R. (1976), The linear structural relation for several groups of data, *Biometrika,* 63, 231-237.

Cox, N. R., and Dolby, G. R. (1977), Corrections and amendments (to Cox(1976) and Dolby(1976)), *Biometrika*, 64, 427

Department of Health, (1991), Asthma: an epidemiological overview, Central health monitoring unit epidemiological overview series, London: HMSO, 1995.

Department of Health, (1992), Advisory Group on the medical aspects of air pollution episodes. Second report: sulphur dioxide, acid aerosols and particulates, London: HMSO,1992.

Department of Transport, (1978), *Transport statistics: Great Britain. London*: HMSO.

Diggle, P. (1990), A point process modelling approach to raised incidence of rare phenomenon in the vicinity of a prespecified point, *J.RSS. A.*153, 349-362.

Diggle, P. (1993), Point process modelling in environmental epidemiology, In *Statistics for the environment.* pp. 89-110. Edited by Vic Barnett and Turkman. John Wiley & Son Ltd.

Diggle, P. and Rowlingson, B. (1994), A conditional approach to point process modelling of elevated risk. *JRSS, A,* 157, 433-440.

Dolby, G. R. (1976), The ultrastructural relation a synthesis of the functional and structural relation, *Biometrika*, 63, 39-50.

Doob, J. L. (1953), Stochastic process, pp.190-218, John Wiley & Son Ltd.

Elizabeth, M. (1998), Concise Medical Dictionary, pp.726, Oxford University Press.

Elliot, P., Martuzzi, M. and Shaddick, G. (1995), Spatial statistical methods in environmental epidemiology: a critique, *Statistical methods in medical research,* 4: 137-159.

ESRI. (1989), *Understanding GIS: the ARC/INFO method.* Redlands, CA, USA.

Frigessi, A., Hwang, C. H., Stefano, P, and Sheu, A. J. (1991), Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics, *JRSS*, B, 55, No.1, pp. 205-219.

Gelfand, A. and Smith, A (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American statistical association,* 85: 398-409.

Gelman, A. and Rubin, D. (1992), A single series from the Gibbs sampler provides a false sense of security, in J. Bernardo, J. Berger, A. P. Dawid and A. Smith (eds), *Bayesian Statistics 4*, OUP, pp. 625-631.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and Bayesian restoration of images, *IEEE on pattern analysis and machine intelligence* PAMI-6(6):721-741.

Geyer, C. (1991), Markov chain Monte Carlo maximum likelihood, in E. Keramidis (ed.), Computing Science and Statistics: *Proceedings of the 23rd Symposium on the interface*, Interface Foundation, Fairfax Station, pp.156-163.

Geyer, C. and Thompson, E. (1995), Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American statistical association.* 1995, Vol.90, No.431, 909-920.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* 1970, 57, 97-109.

Hoschel, H. P. (1987), Models-in- variables. In *nonlinear regression, functional relations and robust method,* pp. 214-294, Edited by Helga Bunke and Olaf Bunk. John Wiley & Son Ltd.

Heieh, F. Y. (1989), Sample size tables for logistic regression, *Statistics in medicine*, Vol. 8, 795-802.

Hodges. L. Jr and Lehmann, E. L. (1970), Basic concepts of probability and statistics, pp. 370-411, University of California, Berkeley. Holden-Day Inc.

Hosmer, D. W. and Lemeshow, S. (1989), Applied logistic regression, New York; Chichester, Wiley, 1989.

Ishizaki, T., Koizumi, K., Ikemori, R., Ishiyama, Y., Kushibiki, E. (1987), Studies of prevalence of Japanese cedar pollinosis among the residents in a densely cultivated area, *Ann. allergy.* 1987, 58:265-270.

Kaplin, B., and Mascie, C. (1985), Biosocial factors in the epidemiology of childhood asthma in a British national sample, *Journal of Epidemiology and Community Health,* Vol. 39, pp 152-156.

Kendall, M. and Stuart, A. (1979), The advanced theory of statistics, Vol.2: Inference and relationship, 4th ed. London, Griffin, 1979.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983), Optimization by simulated annealing, *Science*, 220, 671-680.

Knoke, D. J. (1991), Nonparametric analysis of covariance for comparing change in randomized studies with baseline values subject to error, *Biometrics,* 47, 523-533.

Kobayashi, T., Shinozaki, Y. (1992), Induction of transient airway hyperresponsiveness by exposure to 4 ppm nitrogen dioxide in guinea pigs, *J Toxicol Environ Health*, 1992, 37, 451-461.

Lawson, A. B. (1993), On the analysis of mortality events associated with a prespecified fixed point, *JRSS, A,* 156, 363-377.

Lawson, A. B. (1995), MCMC methods for putative pollution source problems in environmental epidemiology, *Statistics in medicine*, Vol. 14, 2473-2485.

Lawson, A. B., Biggeri, A. and Williams, F. L. R. (1999), A review of modelling approaches in health risk assessment around putative sources. pp. 231-245 in Lawson A. B. *et al* (editors) *Disease Mapping and risk assessment for public health.* Wiley, New York.

Lawson, A. B. and Williams, F. L. R. (1994), Armadale: a case-study in environmental epidemiology. *JRSS, A,* 157,285-298.

Lawson, A. B. and Waller, L. A. (1996), A review of point pattern methods for spatial modelling of events around sources of pollution, *Environmetrics*, Vol. 7, 471-487.

Marinari, E. and Parisi, G.(1992), Simulated tempering: a new Monte Carlo scheme, *Europhysics Letters*, 19, 451-458.

McCready, M., Patel, S., and Rennolls, K. (1997), An investigation of the effect of road traffic pollution on asthma, using geographical information systems, In *Statistics for the environment 3: pollution assessment and control*, pp. 287-299. Edited by V. Barnett and K. Feridum Turkman, John Wiley & Sons.

McCullagh, P. and Nelder, J. (1989), Generalized linear models. 2nd ed. London, Chapman and Hall, 1989.

Metropolis, N., Rosenbluth, A., Rosenbluth, M. and Teller, E. (1953), Equations of state calculations by fast computing machine, *Journal of chemical physics*, 21, 1087-1091.

Migon, H. S. and Gamerman, D. (1999), Statistical inference: an integrated approach, Arnold.

Murdoch, D. J., Green, P. J. (1998), Exact sampling from a continuous state space, *Scandinavian journal of statistics*, 1998, Vol.25, No.3, pp.483- 502.

National Children's Bureau (1991), National child development study composite file including sweeps one to four, distributed by ESRC data archive, University of Essex, Colchester.

Ninan, T. K., Russell G. (1992), Respiratory symptoms in children at schoolchildren: evidence from two surveys 25 years apart, *BMJ* 1992; 304:873-875.

Pasquill, F. and Smith, F. B. (1983), Atmospheric diffusion, 3rd ed, Chichester: Ellis Horwood.

Propp, J. G., Wilson, D. B. (1996), Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random structures & algorithms*, 1996, Vol.9, No.1-2, pp.223-252.

Riply, B. D. (1987), *Stochastic simulation*, New York: John Wiley.

Raftery, A. & Lewis, S. (1992) How many iterations in the Gibbs sampler? *Bayesian statistics 4*, OUP, Oxford, pp. 763-773.

Raper, J. F., Rhind D. W. and Shepgerd J. W. (1992) Postcode: The new geography, Longman scientific & technical.

Rennolls, K., Zhang, Y., Ramesh, N. I., Chen, A. and Patel, S. (1999), Spatial Models of Pollution Exposure from a Road Network: The effect on Asthma, *Statistics for Environment 4:Pollution Assessment and Control*, Edited by V. Barnett, A. Stein and K. Feridun Turkman. 1999. John Wiley & Sons Ltd.

Robert, C. P. and Casella, G. (1999), Monte Carlo Statistical Methods. 1999. Springer Verlag.

Shy. C. M., Creason, J. P., Pearlman, M. E., McClain, K. E., Benson, B. F. and Young, M. M. (1970), The Chattanooga school children study: effects of community exposure to nitrogen dioxide. I: Method, description of pollutant exposure and results of ventilatory function testing, *JAPCA* 1970,20, 539-545.

Smith, A. and Roberts, G. (1993), Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods, *JRSS, B.* 55, No. 1, pp. 3-23.

Smith, D. and Neutra, R. (1993), Approaches to disease cluster investigations in a state health department, . *Statistics in medicine*, 12, 1757-1762.

Spinaci, S., Arossa, W., Bugiani, M., Natale, P., Bucca, C. and Candussio, G. (1985), The effects of air pollution on the respiratory health of children: a cross-sectional study, Pediatr Pulmonol, 1985; 1:262-266.

Strachan, D. P. (1995), Epidemiology, *Childhood asthma and other wheezing disorders*, edited by Silverman M. Chapman & Hall medical.

Stone, R. A. (1988), Investigations of excess environmental risks around putative sources: statistical problems and a proposed test, *Statistics in medicine*, 7, 649-660.

Vaughan, F. (1996), Pumping cleaner air into our city streets, *The Times* 16 August 1997 p.97 8.

Viel, J. F. , Pobel, D. and Carré, A. (1995), Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis, *Statistics in medicine*, Vol. 14, 2459-2472.

Waller, A. W., Turnbull, W. B., Clark, C. L and Nasca, P. (1994), Spatial pattern analysis to detect rare disease clusters, *Case studies in biometry*, Edited by Nicholas Lange, Louise Ryan, Lynne Billard, David Brillinger, Loveday conquest, and Joel Greenhouse. John Wiley & Sons, Inc.

Waller, L. A. and Lawson, A. B. (1995), The power of focused tests to detect disease clustering, Statistics in medicine, Vol. 14, 2291-2308.

Weiland, S. K. , Mundt, K. A., Ruchmann, A. and Keil U. (1994), Self-reported wheezing and allergic rhinitis in children and traffic density on street of residence, *Ann. Epidemiol.* 1994, 4: 79-83.

Whittemore, A. (1981), Sample size for logistic regression with small response probability, *Journal of the American statistical association*. March 1981, Vol. 76, Number 373, pp27-32.

Wjst, M., Reitman, P., Dold, S., Wulff, A., Nicolai, T., von Loeffelholz-Colberg, E., von Mutius, E. (1993), Road traffic and adverse effects on respiratory health in children, *BMJ* 1993; 307: 596-600.

Zeger, S. L. and Karim, M. R. (1991), Generalized linear models with random effects; a Gibbs sampling approach, , *Journal of the American statistical association,* March 1991, Vol. 86, No. 413, pp. 79-85.

Zellner, A. and Min, C. K. (1995), Gibbs sampler convergence criteria, *Journal of the American statistical association*, September 1995, Vol. 90, No. 431, pp. 921-927.

Appendix A Codes of some variables related to asthma study in the NCDS

IDNO the unique id for every person

N2887 Child's smoking habits when they are 16

Value	Label
1	Don't smoke
2	<1/week
3	1-9/week
4	10-19
5	20-29
6	30-39
7	40-49
8	50-59
9	60+

N259 Ever had an asthma attack -when 5

1	Don't Know
2	Yes
3	No

N2617 Ever asthma or wheezy bronchitis - when 16

1 Yes

2 No

3 Don't know

N5770 Asthma/bronchitis since 16th Birthday

1 Yes

2 No

8 Don't know

9 Don't know

SEX Sex

0 Male

1 Female

N2622 Most recent asthma attack

7 In past 12 months

N2623 Frequency if attack in last 12months

CURAS16 Current asthma/wb at age 16

0 No

1 Yes

GR74 Grid Ref. in 1974

0 No

1 Yes

EVERASTH Ever asthma - recoded n2617

1 No

2 Yes

EVAS16D Ever asthma at 16 - derived from asth16

0 No

1 Yes

N2017 Ethnic group from features

1 Euro-caucasian

2 African-negroid

3 Indian-Pakistan

4 Other asian

5 Mixed race

6 other/unsure

N504021 Ever been told has asthma

0 No

1 Yes

N504024 Wheezing/asthma inhibited speech last 12m

1 Yes

2 No

GR91 Grid ref status in 1991

Missing Values: 9

0 No GR

1 GR

9 M Orks/Shets

N504023 Used inhaler etc over the past 12m

1 Yes

2 No

N504025 Admitted overnight for wheezing/asthma

1 Yes

2 No

N504026 No. of times hospitalised for wheeze/asth past 12m

SMOKE33 SMOKING AT 33

0 No

1 <1/day

2 1-9/day

3 10-19/day

4 20-29/day

5 >30/day

CURRASTH Asthma status at 33

1 Yes

2 No

Appendix B Percentage points for the χ^2-distrituion

n-freedom	5%	1%
1	3.84	6.63
2	5.991	9.21
3	7.815	11.34
4	9.488	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.507	20.28
9	16.919	21.67
10	18.307	23.21

Appendix C The density function of the heterogeneous Poisson process

Suppose $\{\mathbf{x}_i \in A, i=1, \ldots, n\}$ is part of a realization of a heterogeneous Poisson process with intensity function $\lambda(\mathbf{x})$.

Since the number of points in an area A, N(A) follows a Poisson distribution with mean $u = \int_A \lambda(\mathbf{x}) d\mathbf{x}$.

The probability of N(A) equals n is

$$P(N(A) = n) = \frac{u^n e^{-u}}{n!} \tag{1}$$

Given N(A) = n, the n points in A form independent random samples from the distribution on A with PDF proportional to $\lambda(\mathbf{x})$, $\mathbf{x} \in A$, denote $\mathbf{X}_i$ as a random variable on A whose distribution to $\lambda(\mathbf{x})$, then

$$f(X_1 = \mathbf{x}_1, \cdots, X_n = \mathbf{x}_n \mid N(A) = n) = \left(\prod_{i=1}^{n} \lambda(\mathbf{x}_i) \right) u^{-n} \tag{2}$$

Since $X_1, \ldots, X_n$ can be $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in any order. There are $n!$ possible orders. The density function of the n point is $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$f(\mathbf{x}_1, \cdots, \mathbf{x}_n \mid N(A) = n) = n! \left(\prod_{i=1}^{n} \lambda(\mathbf{x}_i) \right) u^{-n} \tag{3}$$

Combining (3) and (1), The density function of there are n point ($\{\mathbf{x}_i, i=1, \ldots, n\}$) in A, is

$$L = P(N(A) = n) f(\mathbf{x}_1, \cdots, \mathbf{x}_n \mid N(A) = n) = \left(\prod_{i=1}^{n} \lambda(\mathbf{x}_i) \right) \cdot \exp(-u) \tag{4}$$

Appendix D Stationary distribution of discretized MCMC

We only need to prove the multi-dimension situation, since one dimension is a special case of it.

First we will show that $\mathrm{cPo}(\alpha, i_1, ..., i_M)$ is stationary probability distribution.

Denote the initial distribution of the Markov chain is $\mathrm{cPo}(\alpha, i_1, ..., i_M)$.

In step 1, since the update method is Gibbs or Metropolis-Hastings algorithm, $\mathrm{cPo}(\alpha, i_1, ..., i_M \mid i_1, ..., i_M)$ is equilibrium distribution for Gibbs and Metropolis-Hastings algorithm. Denote the transition density as $\mathrm{p}(\alpha, \alpha', i_1, ..., i_M \mid i_1, ..., i_M)$, we have

$$\int \mathrm{cPo}(\alpha, i_1, ..., i_M \mid i_1, ..., i_M)\, \mathrm{p}(\alpha, \alpha', i_1, ..., i_M \mid i_1, ..., i_M) d\alpha = \mathrm{cPo}(\alpha', i_1, ..., i_M \mid i_1, ..., i_M)$$

It is same as

$$\int \mathrm{cPo}(\alpha, i_1, ..., i_M)\, \mathrm{p}(\alpha, \alpha', i_1, ..., i_M) d\alpha = \mathrm{cPo}(\alpha', i_1, ..., i_M).$$

It means after step 1 the distribution is still $\mathrm{cPo}(\alpha, i_1, ..., i_M)$.

Also in Step 2, each update of i_m is a Metropolis-Hastings algorithm, so $\mathrm{Po}(\alpha, i_1, ..., i_M \mid \alpha, i_1, ., i_{m-1}, i_{m+1}.., i_M)$ is the stationary distribution, for the same reason of in the step 1, after a update in step 2, the distribution is $\mathrm{cPo}(\alpha, i_1, ..., i_M)$.

So after step, 1, 2, the distribution of the Markov chain is still $\mathrm{cPo}(\alpha, i_1, ..., i_M)$, i.e. $\mathrm{cPo}(\alpha, i_1, ..., i_M)$ is the stationery distribution.

Since this Markov chain is irreducible and aperiodic, it can be proved that there is only one stationary distribution, and this Markov chain converges to $\mathrm{cPo}(\alpha, i_1, ..., i_M)$ (Doob 1953).