THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #267

Preferential sampling in long term monitoring of air pollution: a case study

BY

GAVIN SHADDICK & JAMES V ZIDEK

September 2012

THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #267

Preferential sampling in long term monitoring of air pollution: a case study

BY

GAVIN SHADDICK & JAMES V ZIDEK

September 2012

Preferential sampling in long term monitoring of air pollution: a case study.

Gavin Shaddick[1] and James V Zidek[2]

[1]Department of Mathematical Sciences, University of Bath

[2]Department of Statistics, University of British Columbia.

Abstract

The potential effects of air pollution are a major concern both in terms of the environment and in relation to human health. In order to support environmental policy there is a need for accurate measurements of the concentrations of pollutants. Here we examine long term changes in concentrations of black smoke using data from an extended period (1966-1996) from a long established network in the UK. Over this period, the number of sites reduced dramatically and there is the possibility of selection bias if the monitoring sites are kept in polluted areas. Bayesian models were used to model concentrations over time and space and to assess explore the evidence of preferential sampling. In cases such as this, with large spatial datasets, inference using MCMC can be a challenge due to computational issues and here we perform 'approximate' Bayesian inference using Integrated Nested Laplace Approximations. For the spatial components of the models we employ methods that represent a Gaussian field with Matern covariance function as a Gaussian Markov Random Field through use of the Stochastic Partial Differential Equations. The results presented here give support to the hypothesis of preferential sampling which has largely been ignored in environmental risk analysis.

Keywords: preferential sampling; spatial-temporal modelling; INLA, air pollution.

1 Introduction

Air pollution has been a concern for many centuries: during the middle ages, monarchs in several countries tried to reduce air pollution by banning practices such as burning coal, and travellers in the seventeenth centuries commented on the poor air quality in many cities. Following the industrial revolution, problems associated with air pollution worsened in many areas of Europe. During the first half of the twentieth century major pollution episodes occurred in London, notably in 1952 an episode of fog, in which levels of black smoke exceeded 4,500 $\mu g m^{-3}$, was associated with 4000 excess deaths (Ministry of Health, 1954).

Other early episodes, which were caused by a combination of industrial pollution sources and adverse weather conditions, and resulted in large numbers of deaths among the surrounding populations include those in the Meuse valley (Firket, 1936) and the US (Ciocco and Thompson, 1961)).

Attempts to measure levels of air pollution in a regular and systematic way largely arose as a result of these episodes. Early air pollution control legislations were focused on setting restrictions on the use of smoke-producing fuels and smoke-producing equipment (Garner and Crow, 1969; Stern et al., 1973) and in 1961 the worlds first co-ordinated national air pollution monitoring network was established in the UK, the 'National Survey' which was used to monitor black smoke and sulphur dioxide at around 1000 sites (Clifton, 1964).

Since then all European countries have begun to establish monitoring networks, some of them run at the national level, others by local authorities or municipalities. Because of the different ways in which these have developed, and the different purposes for which they have been established, many of the networks vary in terms of which pollutants they measure, how they measure them, where monitoring sites are located, and how results are reported. In addition, over time many of the networks have changed; some growing, others shrinking, as attention has shifted to new pollutants and geographical areas. During much of the twentieth century, for example, the main concern was soot (or black smoke) and sulphur dioxide from industry and domestic fires. Most networks thus focused on measuring these pollutants, especially in industrial areas where concentrations were likely to be high. More recently, attention has moved to potential hazards associated with fine particulates and reactive gases such as nitrogen dioxide, volatile organic compounds and ozone, so monitoring networks for these have expanded.

Black smoke (BS) is one of a number of measures of fine particulate matter, other examples including the coefficient of haze (CoH), total suspended particulates (TSP), as well as direct measurements of PM_{10}, and $PM_{2.5}$. Attempts have been made to standardise the measures of pollution by converting the measurements into 'equivalent' amounts of PM_{10}, for example $PM_{10} \approx 0.55$ TSP, $PM_{10} \approx$ CoH/0.55, $PM_{10} \approx$ BS and $PM_{10} \approx PM_{2.5}/0.6$ (Dockery and Pope (1994)). BS is measured using the light reflectance of particles collected in filters to assess the blackness of the collected material. The method was originally developed to measure smoke from coal combustion.

Each of these measures of particulate matter have been associated with adverse health outcomes, for example PM_{10} (Samet et al., 2000), $PM_{2.5}$ (Goldberg et al., 2001), TSP (Lee et al., 2000), black smoke (Verhoeff et al., 1996), and CoH Gwynn et al. (2000). BS continues to be used in epidemiological studies, for example, Elliott et al. (2007); Hansell et al. (2011), and in several recent European studies, BS was found to be at least as predictive of negative health outcomes as PM_{10} or $PM_{2.5}$ (Hoek et al., 2000; Samoli et al., 2001). These findings indicate that black smoke, which is closely-related in the modern urban setting with diesel engine exhaust, could serve as a useful marker in epidemiological studies, perhaps even retrospective analyses using the historic data available in many European urban areas (WHO, 2003).

Black smoke has been measured in the UK since the early 1960s as part of the UK Smoke and Sulphur Dioxide network and it's predecessor the National Survey. The monitoring network, which measures both SO_2 and black smoke, was established in the early 1960s, and by 1971 included over 1200 sites. As levels of black smoke and SO_2 pollution have declined, the network has been progressively rationalised and reduced and by 1996 comprised approximately 220 sites.

Originally, the National Survey was designed on the basis of the classification of towns into categories according to factors that were thought to affect air pollution with the aim to use these 'representative towns' to assess the concentrations of BS and SO_2 throughout the country. Towns were categorised according to (i) domestic and (ii) industrial coal consumption per unit area and (iii) natural ventilation (the tendency for pollution to be trapped due to local topography) with each of these factors being graded as high, moderate and low. Clifton (1964). Within the resulting categories, locations were selected stratified by geographical region and population size. In addition to choosing a location that was deemed to be representative of the pollution experienced by a community, local authorities were also requested to locate a second monitoring site as far as possible from sources of pollution to give information about background levels over the country.

In later years, the network has been used to monitor compliance with the relevant EC Directives on sulphur dioxide and suspended particulate matter. The original Directive, 80/779/EEC1, was introduced in 1980 and has been updated in Daughter Directives for SO2 and suspended particulate matter. The standards for monitoring of black smoke remained in force until 2005 but more recent standards for suspended particulate relate to PM_{10} and not black smoke. Daily average black smoke has been shown to be a reasonable predictor of PM_{10}. In general, PM_{10} concentrations are usually higher than black smoke except during high episodes, and hence, if smoke exceeds the PM_{10} limit, it is likely that PM_{10} has also done so (Muir and Laxen, 1995).

Over time, many sites have been moved or replaced in order to reflect changing patterns and levels of pollution, and to reduce redundancy in the network. Therefore there is the possibility of selection bias if the monitoring sites are kept in polluted areas. This may occur for example, if the locations of sites remaining in the network were chosen to assess whether guidelines and policies are being adhered to. This will lead to *preferential sampling* which occurs when the process that determines the locations of the monitoring sites and the process being modelled (concentrations) are in some ways dependent Diggle et al.

(2010). In the context of air pollution and health in epidemiological analyses, Guttorp and Sampson (2010) state that air pollution monitoring sites may be intentionally located for a number of reasons, including to measure: (i) background levels outside of urban areas; (ii) levels in residential areas; and (iii) levels near pollutant sources. Standard geostatistical methods which assume sampling is non-preferential are often employed despite the presence of a preferential sampling scheme. Ignoring preferential sampling may lead to incorrect inferences and biased estimates of pollution concentrations.

In this paper we aim to examine patterns in both the concentrations of BS over an extended period (1966-1996) and to investigate the possible effects of the reduction in size of the network on estimates of levels of air pollution and the evidence for preferential sampling. The remainder of this paper is as follows, Section 2 gives details of black smoke concentrations measured in the UK. Section 3 provides details of the proposed statistical model and details the methods for inference and Section 4 presents the results of applying the models. Finally, Section 5 contains a discussion and details of potential future research.

2 Data

The data on annual concentrations of BS were obtained from the UK National Air Quality Information Archive. This is a routinely maintained data set prepared by the National Environmental Technology Centre (NETCEN) on behalf of the UK Department of the Environment, Food and Rural Affairs (DEFRA), which collates information from all the national air quality monitoring networks. Sites are classified into one of eighteen types in terms of their local environment. Air pollution statistics for each site are reported in pollution years (April-March) and include mean concentration, standard deviation and number of valid reporting days. Site locations (at a 10 metre resolution) and annual average concentrations of BS (μgm^3) for monitoring sites operating between April 1966 and March 1996 from the Great Britain Air Quality Archive (www.airquality.co.uk). The minimum data capture requirement used is 75% (as stated in the EC Directive), equivalent to 273 days a year, and only sites reaching this requirement are considered in the analyses presented here. The locations of the sites can be seen in Figure 1. Concentrations of black smoke were typically highest in areas where the use of coal for domestic heating was relatively widespread, such as parts of Yorkshire and Northern Ireland, and also at some sites in large cities. The locations of the monitoring sites were also classified as either urban or rural based on CORINE land cover data using discriminate analysis as detailed in Beelen et al. (2009).

In total there were 1466 sites operational within the period 1966 to 1996 of which 35 *consistent sites* were operating throughout the entire period. Figure 2 shows a schematic of when sites were operational within the study period. It clearly shows the reduction in the size of the network from the original set of sites in 1966 and the introduction of a smaller number of sites over time. Also evident is the marked reduction of the network in 1981 when there was a dramatic reduction of almost 50% as the network was reorganized owing to falling urban concentrations and to comply with EC directive 80/779/EEC (Colls, 2002). Figure 3 shows the mean concentrations over all sites by year for the set of consistent sites and *non-consistent sites*, i.e. those that weren't operational throughout the entire period. There is a marked decline in the level of BS over the this period, with a significant difference between the levels recorded in the two groups.

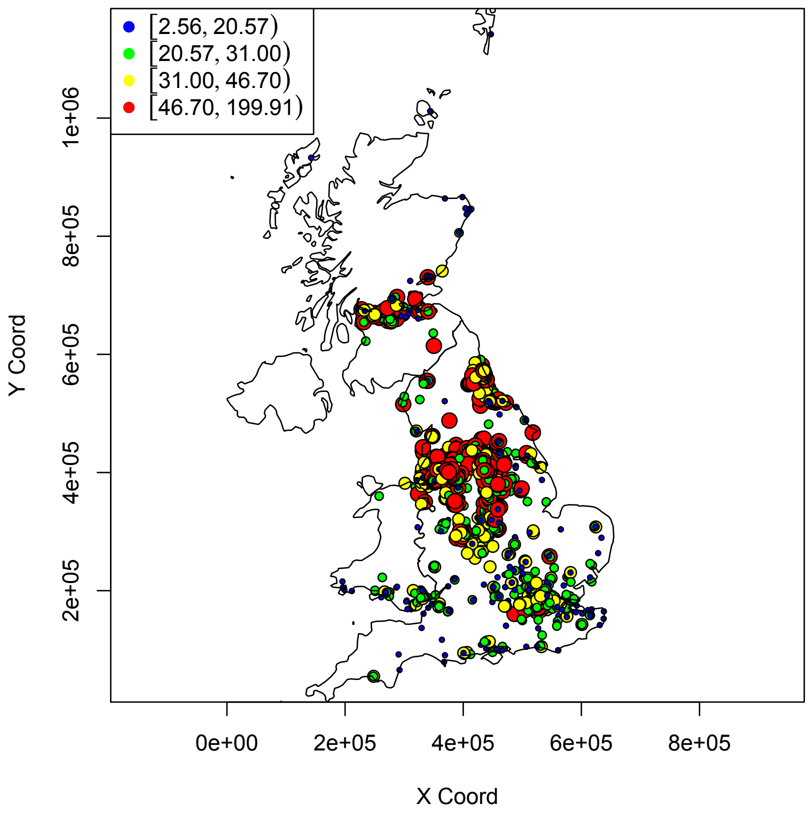

Figure 1: Locations of black smoke monitoring sites together with annual mean of daily concentrations at those sites for the times they were operational within the study period, 1966-1996

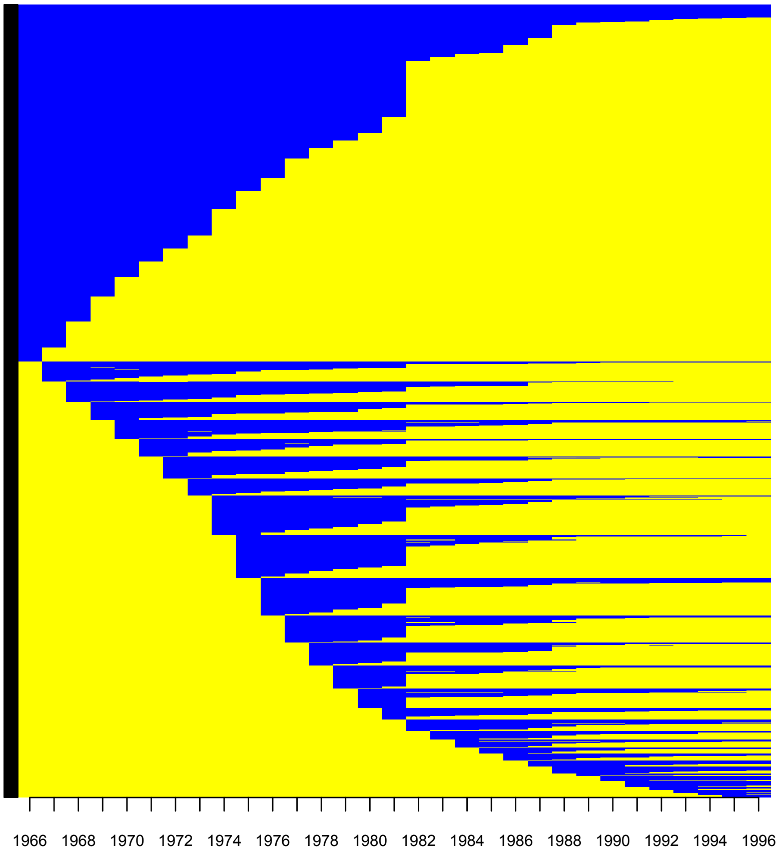

Figure 2: Schematic showing the years for which black smoke monitoring sites were operational (blue lines) and those when they were not (yellow) in the UK Smoke and Sulphur Dioxide network, 1966-1996.

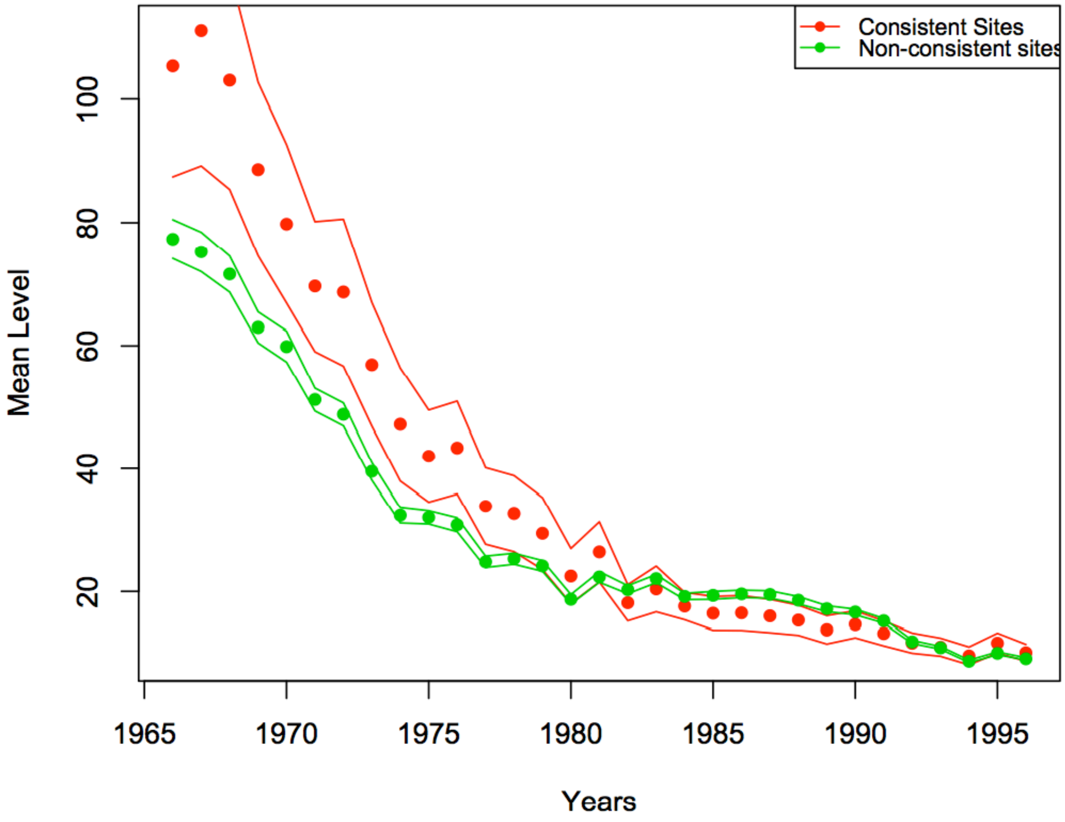

Figure 3: Annual means of black smoke concentrations by year. Dots denote the yearly mean value with associated 95% confidence intervals denoted by lines. Values are given for consistent (red) and non-consistent (green) sites from year 1966 to 1996.

3 Statistical Modelling

Let Z_{it} be the concentration of black smoke measured at location, i, at time, t. Ott (1990) has suggested that a log transformation is appropriate for modelling pollution concentrations, because in addition to the desirable properties of right-skew and non-negativity, there is justification in terms of the physical explanation of atmospheric chemistry. In general, as transcendental functions, neither of the functions $\exp(x)$ nor $\log(x)$ can be applied when x is a measurement, meaning in particular, that neither x nor the two functions can be meaningfully have units of measurement. In short x must be a number. To quote Monk and Munro (2010):

> "We cannot take the logarithm of anything except a number. Therefore, a logarithm term will never have units because it is merely a dimensionless number."

To nondimensionalize our measurements, we divide them by 78 (units), roughly the level of black smoke concentrations in 1961. The unitless ratio, now represents the number of baseline units of decline in that particulate concentration since that time. For example 1/2 would represent a 50% decline.

Here then $Y_{it} = \log(Z_{it}/78)$ with,

$$Y_{it} = (\beta_0 + \beta_{0i}) + (\beta_x + \beta_{x_i})X_{it} + (\beta_{x2} + \beta_{x2_i})X_{it}^2 + \beta_c C_i + \beta_u U_i + \epsilon_{it} \qquad (1)$$

where $i = 1, \cdots, N$ denotes the site and $t = 1, \cdots, T$ the year. The model includes both linear and quadratic effects, β_x and β_{x2} of time reflecting the shapes of decline in the decline in levels of BS observed in the data. Site specific random effects, β_{xi} and β_{x2i} and β_{0i}, are assigned to the slopes of the linear, quadratic and intercept components respectively. These are contained to sum to zero, around fixed effects, β_0, β_x and $\beta_2 x$ respectively. Whether a site is consistent or not is represented by the indicator C_i and whether the location is classified as urban or rural by the indicator U_i. The effects of being a consistent site and being located in a rural area are represented by β_c and β_u respectively. The ϵ_{it} is a random error term, which is assumed to be Normally distributed, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$.

In addition, we consider interactions between the slope terms (linear and quadratic) and the consistent site indicator, which allows for both a shift in overall levels between the two groups (consistent and non-consistent) and different rates of decline over time.

In the simplest case, the random effects terms can be assumed to be independent, i.e. $\beta_{\mathbf{s}} \sim MVN(0, \sigma_s I)$, however there is likely to be residual spatial auto-correlation in the data after allowing for the effects of time and so this is relaxed to the assumption that they are multivariate normally distributed,

$\beta_{\mathbf{s}} \sim MVN(0, \sigma_s \Sigma)$, with the structure of the covariance reflecting any spatial auto-correlation.

3.0.1 Spatial Process

If there is spatial correlation between sites (after allowing for the effect of time) then Σ will be determined by the form of the relationship between correlation and distance. We assume that the spatial effects represent a stationary spatial process, meaning that the correlation between the sites is dependent only on the distance between sites and not their actual location. A common class of models used to model such relationships is the Matern Class, where the spatial covariance between two points (u, v) function takes the form

$$r(\mathbf{u}, \mathbf{v}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa || \ \mathbf{v} - \ \mathbf{u}||)^{\nu} K_{\nu}(\kappa||\mathbf{v} - \mathbf{u} \ ||). \qquad (2)$$

where K_ν is a modified Bessel function of the second kind, σ^2 is the overall variance and (ν, κ) are parameters that control the smoothness and strength of the distance–correlation relationship respectively. The limiting case of the Matern class of models, when $\nu \rightarrow \infty$, is the Gaussian model and the commonly used exponential model is a special case with $\nu = 1/2$. In addition to acknowledging spatial correlation in the intercept and slope terms of model (1), the spatial component of the models allow the prediction of measurements at locations for times where there is no monitoring site.

3.1 Inference

A number of studies have incorporated spatial modelling of air pollution within a Bayesian framework, for example (Shaddick and Wakefield, 2002; Sahu et al., 2006; Molitor et al., 2007; Lee and Shaddick, 2010). Commonly, inference has been performed using Markov Chain Monte Carlo (MCMC) often using software packages such as WinBUGS (Lunn et al., 2000). The main constraint of this approach, particularly when using large spatial datasets, is its demanding computational requirements. This can be both because of the requirement to manipulate large matrices within each simulation of the MCMC and also in convergence of parameters in complex models (Finley et al., 2007).

An alternative approach is that of marginalisation (Finley et al. (2007), Banerjee et al. (2008)) in which the spatial effects are marginalised out which reduces the parameter space and thus lessens the computational burden. This means however, that estimates of the spatial effects, which are required for prediction, are not available as they cannot be sampled. In the Gaussian case they can be reconstructed in a posterior predictive fashion (Banerjee et al., 2008).

Here we use recently developed techniques which perform approximate Bayesian inference based on integrated nested Laplace approximations (INLA) and thus do not require full MCMC sampling to be performed (Rue et al., 2009). INLA has been developed as a computationally attractive, practical alternative to the MCMC. The increasing size and complexity of experiments and the databases they generate together has outpaced the speed of readily available computational hardware. This has forced the development of of practical alternatives to MCMC algorithms, which although approximations to their MCMC counterparts are better than no solution at all.

The dataset considered here contains a large number of missing values as can be seen in Figure 2 although we consider the entire set of potentially 31 measurements over time (years) for all 1466 monitoring sites under consideration. Due to both the size of the spatial component of the model and the number of missing values, for which predictions will be required, it is computationally impractical to run this model using WinBUGS or bespoke MCMC in any straightforward fashion. However, it is easily possible to fit models to datasets of this size using INLA. Until recently, in terms of spatial models, such methods were generally used for areal level, rather than point level spatial data as considered here, but here we use a recent update which representing a Gaussian field (GF) with Matern covariance function as a Gaussian Markov Random Field (GMRF) through use of the Stochastic Partial Differential Equations (SPDE) (Lindgren et al., 2011). We now present a brief review of the INLA method and the extension which allows spatial modelling of point level spatial data.

Development of INLA began with GMRF models. Central to that development is a Laplace approximation (Rue et al., 2009; Shrödle and Held, 2011) to the posterior distribution of a hyper–parameter vector θ given the measurement vector $\mathbf{y}$, $\tilde{\pi}(\theta|\mathbf{y})$. The measurements constitute the observations of a response vector, which depends stochastically on a latent random field $\mathbf{x}$ that is indexed by spatial locations or areas, i. This random vector will include all the Gaussian components of the model including the model parameters that have a Gaussian prior distribution. Moreover, its joint distribution including that of the model parameters in $\mathbf{x}$, is that of a GMRF. The parameter vector θ would then include the hyperparameters for those parameters, including for example, the prior variances which index the covariance matrices.

Often inference will be also be required about $\mathbf{x}_i$ conditional on the data through the predictive distribution

$$\pi(\theta|\mathbf{y}) = \int \pi(x_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta.$$

The Laplace approximation is given by

$$\tilde{\pi}(\theta|\mathbf{y}) = \left. \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_G((\mathbf{x}, \theta|\mathbf{y})} \right|_{x=x^*(\theta)}$$

where $\tilde{\pi}_G$ is a Gaussian approximation at the mode $x^*(\theta)$ of the conditional distribution of $\mathbf{x}$ given θ. Note that in the spatio–temporal context, θ may contain parameters that are indexed by i while $\mathbf{y}$ includes the responses over time as well as over space.

Shrödle and Held (2011) address the problem of mapping disease over a relatively small number of geographical areas and illustrates the use of INLA in a spatio–temporal context. In this example, the responses are counts over time that depend stochastically on a latent GMRF, which is determined for any one area by a linear model whose parameters and residual terms have conditional distributions specified in terms of its surrounding areas.

However, in the case presented in this paper involves a Gaussian field and responses measured with error over time at a large number of point–referenced locations. The field itself has no natural Markov random field structure, although conditional on certain hyperparameters, the field and model parameters may realistically be supposed to have a joint Gaussian distribution. Thus INLA as originally developed does not apply directly as it does in Shrödle and Held (2011).

Instead we use the SPDE approach described presented by Lindgren et al. (2011), which starts with a GF over a continuous domain of arbitrary dimension and induces from it and its joint distribution, a GMRF to which INLA does apply. The method involves a number of key elements. First is its restriction to the class of GFs, which are thus characterized by their second order properties. Next it requires that field have a Matern covariance structure. Basic theory then implies that such a field must then be the solution of an SPDE. Then it approximates that solution using a finite element method, whose elements are triangles over the field's domain. The induced Gaussian random weights attached to its vertices now determine the joint distribution of the induced GMRF representation of the original GF. Finally the precision matrix for that GMRF is approximated by a sparse precision matrix $\mathbf{Q}$ to achieve computational simplicity, one that represents the covariance Σ of the GF well, i.e with $\mathbf{Q}^{-1}$ close to Σ. The result is a GF model for the process but a GMRF for doing the computations that would be hard to do with the GF itself. The resulting algorithm is represented in R-INLA (www.r-inla.org), an extensive R library of programs which accesses the core INLA computational engine. This implementation of the methodology allows problems of enormous size to be tackled, well beyond what can be tackled with MCMC as demonstrated in this paper.

To give further details of the SPDE – GRMF approximation of INLA, we follow Lindgren et al. (2011). INLA assumes the GF $\{x(\mathbf{u}); \mathbf{u} \in \mathcal{R}^d\}$ has Matern spatial covariance field as given by (2) which must then be the solution of the SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2} x(\mathbf{u}) = \mathcal{W}(\mathbf{u}),\ \ \mathbf{u}\ \in \mathcal{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \qquad (3)$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo − differential operator, Δ is the Laplacian, $\mathcal{W}$ is spatial white noise with unit variance. To complete the specification, assume the process's marginal variance is given by

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}.$$

Representing the process in this way is key to the developments that follow. For it provides the bridge over which we can cross from the GF to the GMRF via an approximate solution to the SPDE.

An infinite dimensional solution x of the SPDE over its domain $\{\mathcal{D}\}$ is characterized by the requirement that for all members of an appropriate class of test functions $\{\phi\}$

$$\int \phi_j(\mathbf{u})(\kappa^2 - \Delta)^{\alpha/2} x(\mathbf{u}) d\mathbf{u} = \int \phi_j(\mathbf{u})\mathcal{W}(\mathbf{u}) d\mathbf{u}. \tag{4}$$

However in practice, only approximate solutions are available. In R - INLA, the R implementation of INLA, that approximate solution is obtained by the conventional finite element approach, which uses a Delauney triangulation (DT) over $\{\mathcal{D}\}$. Initially DT's triangles are formed with vertices at the points of the sparse network where observations are available. Additional triangles are judiciously added until $\{\mathcal{D}\}$ is covered and an irregular array of locations (vertices) is obtained.

Considering the example of the black smoke network, Figure 4 shows the mesh that was constructed using DT for the locations of the BS monitors. In this case, there are 3799 edges and the mesh was constructed using triangles that have minimum angles of 26 and a maximum edge length of 100km. There are 1466 monitoring locations being considered over the period of study and these are highlighted in red. This lattice underlies the GRMF at the other end of the SPDE bridge. It simultaneously gives a finite element representation of the solution of (3):

$$x(\mathbf{u}\ ) = \sum_{k=1} \psi_k(\mathbf{u}) w_k. \tag{5}$$

Here n is the number of vertices in DT, the $\{w_k\}$ are Gaussian weights while $\psi_k(u)$ is piecewise linear in each triangle, 1 at vertex k but 0 at all other vertices. It remains to link the $\psi_k(u)$ to the class of test functions and in that way obtain an approximate solution to the SPDE.

To use this in practice, R-INLA takes this approximation one step further by requiring just n test functions to get a finite dimensional approximation to the SPDE. Lindgren et al. (2011) provides the details, but for example, $\phi_k = (\kappa^2 - \Delta)^{\alpha/2}\psi_k$ is used when $\alpha = 1$. Then substituting these test functions into (4) along with the approximation (5), gives a set of n equations which may be solved. These equations characterize the elements of that approximation, including a sparse precision matrix for the GMRF distributed over the vertices of the irregular lattice and given by the random Gaussian weights $\{w_k\}$ located there. This is all performed within R − INLA once the initial distributions of the GF have been specified. As in Shrödle and Held (2011), we can then handle spatio–temporal problems, but now based on point–referenced measurements of the field.

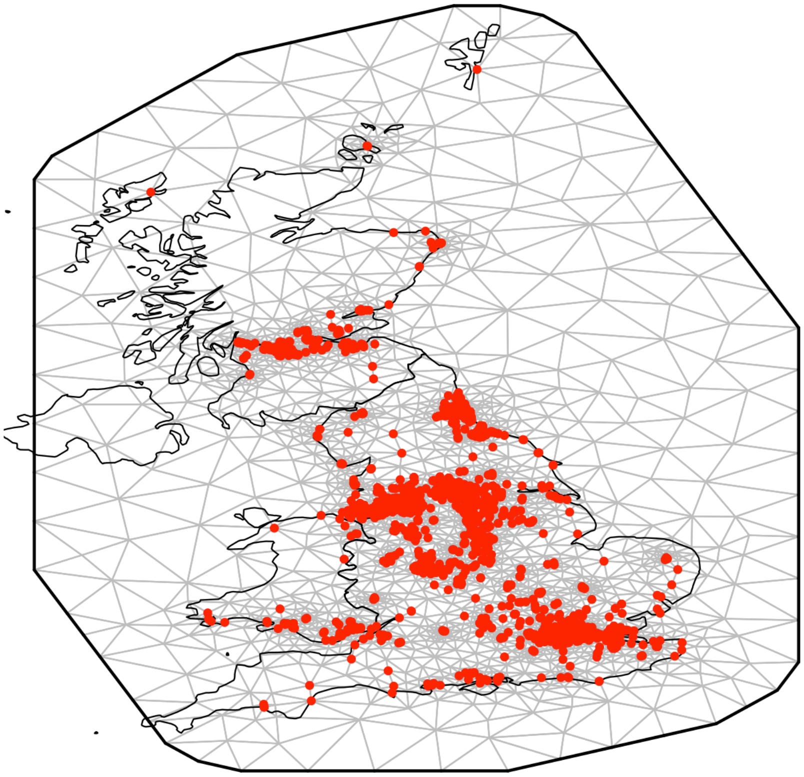

Figure 4: Triangulation for the UK. The mesh comprises of 3799 edges and was constructed using triangles that have minimum angles of 26 and a maximum edge length of 100km. The 1466 monitoring locations being considered over the period of study are highlighted in red.

4 Results

A series of models based on that shown in (1) were fit to the data on black smoke concentrations over the period of study (1966-1996). The models differed both in the covariates that were used and in whether spatial random effects were used for the slope parameters (all models presented and discussed here had spatial random effects for the intercept terms). Covariates were used to model the decreasing pattern over time (represented by year and year^2 terms), the effect of a monitoring site being located within an urban or rural area (ur) and whether a site was consistently operational throughout the period of study (cons). Interactions between time and period of operation were also considered (denote by int(cons, year, year^2)). Model choice was informed by the deviance information criteria (DIC) (Spiegelhalter et al., 2002) which can be computed using INLA, for details see Rue et al. (2009). The model with the lowest DIC provides the best trade-off between fit and model complexity.

	Spatial random effects	
Covariates	Intercept	Intercept and slope
year	5825.2	2875.7
year+year^2	5161.1	1900.8
year+year^2+ ur	5161.1	1823.2
year+year^2+ ur + cons	5160.8	1823.8
year+year^2+ ur +int(cons, year, year^2)	5103.1	1786.6

Table 1: Deviance information criteria for a series of models incorporating different sets of covariates. Covariates are linear and quadratic effects of time (year, year^2), whether the location of the monitoring site was urban or rural (ur) and whether a site was consistently operational throughout the period of study (cons), see text for details. Interaction between operational status and time are denoted by int(cons, year, year^2). Left hand column gives results for models with spatial random effects for the slopes and fixed slope, right hand gives results for models with spatial random effects for both slope and intercept terms.

The DICs for the series of models can be seen in Table 1 where the left hand side shows the results for models with just the intercepts having spatial structure and the right hand side additionally allowing spatial auto-correlation in the slope terms rather than just a fixed slope. There is a marked improvement in the fit of models which allow spatial structure in both slopes and intercepts and thus can allow different sites to have different relationships between concentrations and time. Using the simpler model (with fixed slopes) also has another side effect which can be seen when comparing maps of the spatial effects from models with and without this flexibility for the slope terms. This can be seen by comparing Figures 5 and 6 which show maps of the spatial effects from models with fixed and random slopes respectively. In the first case, there is much less spatial smoothing reflecting the fact that fitting a fixed slope is essentially just subtracting a mean term (over time) with the random effects for the intercepts just reflecting the concentrations at individual sites at the beginning of the study period. In contrast, the map shown in Figure 6 shows clear smoothing over space and indicates that it would be much more suitable for predictions at times for which there was no recorded measurements at a particular time.

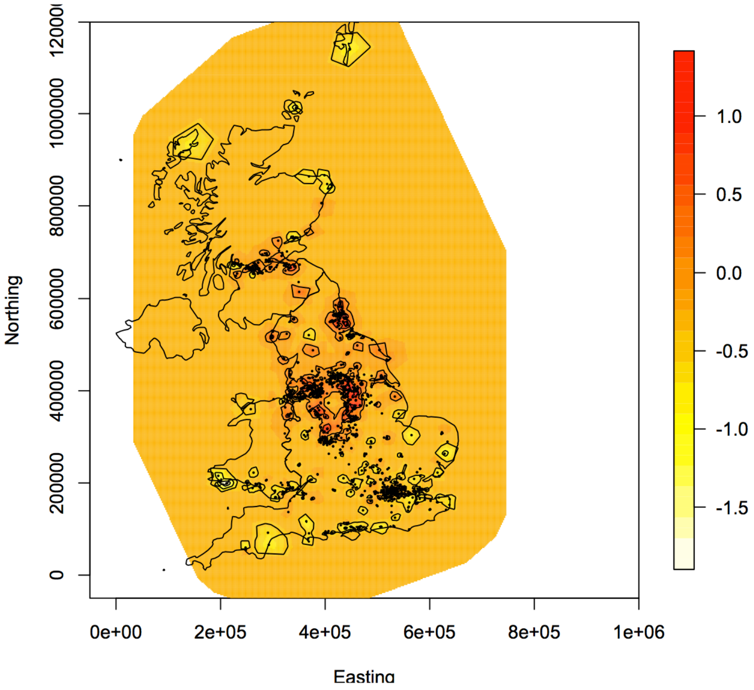

Figure 5: Map of the means of posterior predicted distributions black smoke concentrations on the logarithmic scale from a model with spatial structure on the intercepts and fixed slopes.

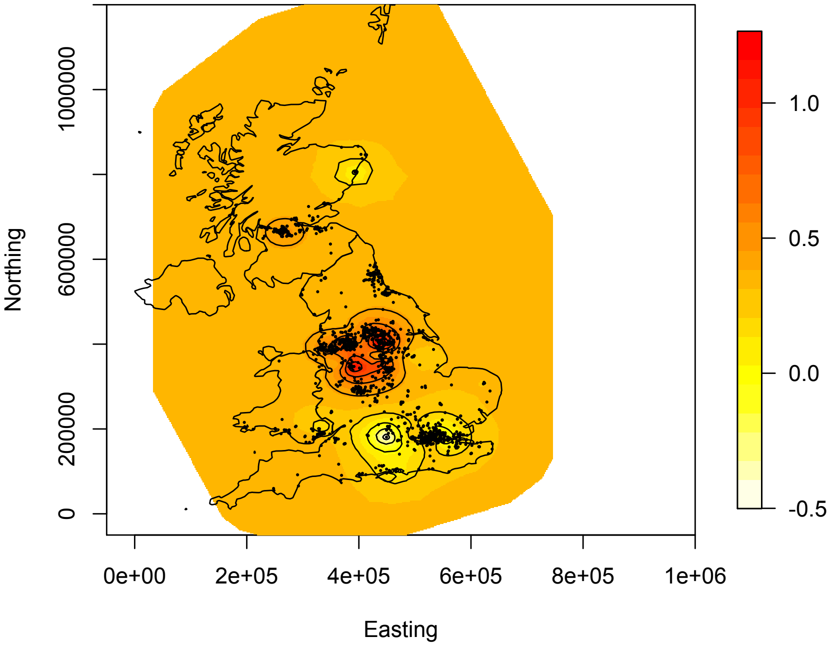

Figure 6: Map of the means of posterior predicted distributions on the logarithmic scale from a model with spatial structure on the intercepts and slopes

Estimates of the parameters for the model with the lowest DIC can be seen in Table 2. Significant effects are seen for both the linear and quadratic effects of year reflecting the overall shape of the declines seen over time. The difference between urban and rural locations was also significant with, as might be expected, lower concentrations in rural areas than urban ones. The consistent set of sites were also found to have significantly higher concentrations than the non-consistent sites with the interaction terms indicating that the decline over the period of study is greater for this set of sites. It should be noted that the values presented are unitless and for the log transformed, rescaled measurements. For example the coefficient associated with the urban-rural indicator, which signifies the lower concentrations in rural as compared to urban areas, is -0.0856 on this scale. On the original scale of the concentrations the effect is therefore $\exp(-0.0856) \times 78 = 71.6$ indicating an overall reduction of 9.5% (7.4) from the baseline of 78 μgm^{-3}. Similarly, the overall difference between consistent and non-consistent sites is 13%, (95% CI 4-24%). This quantifies what was observed in Figure 3 and provides evidence of preferential sampling.

	Median	2.5%	97.5%
Intercept	-0.0413	-0.1355	0.0522
ur	-0.0856	-0.1102	-0.0611
year	-0.1087	-0.1186	-0.0988
year^2	0.0006	0.0002	0.0009
cons	0.1182	0.0425	0.1939
year:cons	-0.0077	-0.0167	0.0013
year^2:cons	0.0005	0.0002	0.0008
σ^2_ϵ	0.0604	0.0592	0.0612
κ_1	0.1795	0.1649	0.1912
τ_1	2.5996	2.4528	2.7751
κ_2	0.0869	0.0753	0.1047
τ_2	79.3	68.6	88.5
κ_3	0.1175	0.0938	0.1714
τ_3	1647.1	1341.0	1901.78

Table 2: Estimates of the parameters for the model with linear and quadratic terms for time, indicators for urban-rural status of the monitor locations and whether a site has been operational throughout the study period and interaction between time and operational status. Medians of the posterior distributions are given together with 95% credible intervals for parameters of fixed (above the horizontal line) and random (below the line) effects, see text for details. It should be noted that the values presented are unitless and for the log transformed, rescaled measurements (see text for details).

Considering the parameters of the Matern spatial terms, κ_k and $\tau_k, i = 1, 2, 3$ correspond to the parameters for the intercept, slope of linear and quadratic effects of time respectively. On the basis of the empirically derived definition given in Lindgren et al. (2011) the posterior mean range at which correlation falls to approximately 0.1, ρ_k, is equal to $\rho = \frac{\sqrt{8\nu_k}}{\kappa_k}$, where here $\nu_k = 1$. For the slope term this equates to 16km, with the results for the slope terms showing more correlation over distance, for example correlation falls to 0.1 at 33km for the slope of the linear term of time. Also from Lindgren et al. (2011), the spatial variance, σ_{s_k} is given by $\frac{1}{4\pi\kappa_k^2\tau_k^2}$ which allows us to compare the spatial variation with that is left unexplained (represented by the random error term σ_ϵ^2. Here the value of σ_{s_1} for the spatial effects assigned to the intercepts is 0.3655 indicating that the more variation is explained by the spatial term rather than by the measurement error.

By combining the temporal and spatial components of the model, predictions can be made at locations where there are missing values for certain years. This can be seen in Figure 7 in which the observed and predicted concentrations over time are shown at a selection of sites. The top centre panel shows the increased uncertainly associated with predictions that are made further away from times where data is observed. In this case, the site (Caerleon 1) was not particularly close to other sites and so there was little possibility of borrowing information over space. The effect of predicting for a site (Newport 26) that is in close proximity to others can be see in the bottom right panel for which there were nearby sites (notably Newport 22) providing information for the period for which it was not operational.

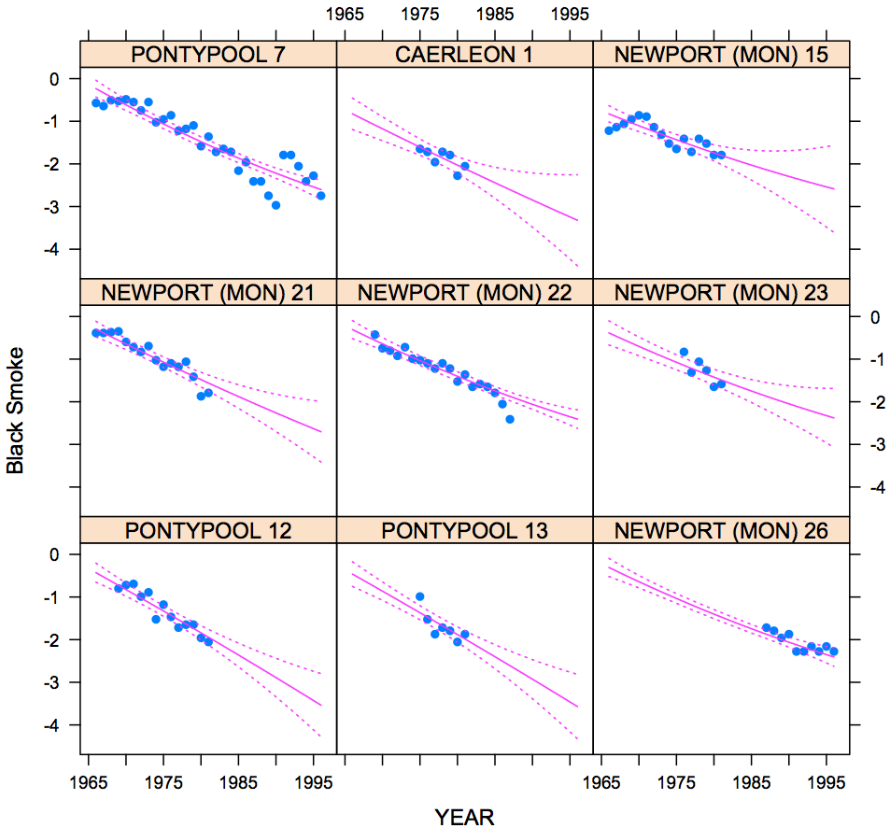

Figure 7: Observed and predicted concentrations of black smoke over time at a selection of sites. Blue dots are measured values with pink lines denoting the predictions from a model with spatial structure for the intercepts and slopes. The model includes a quadratic effect of time and an interaction with consistent sites (see text for details). Pink dotted lines show 95% credible intervals around the predicted values. It should be noted that the values presented are unitless and for the log transformed, rescaled measurements (see text for details)

5 Discussion

In this paper we have examined patterns in concentrations of BS over an extended period (1966-1996) and investigated the possible effects of the reduction in size of the network on estimates of levels of air pollution and the evidence for preferential sampling. The level of BS declined over the period of the study in accordance with regulatory guidelines and since 1982 there have been no exceedences of the EU limit of 68 μgm^{-3} (European Commission, 1980).

Originally, the network was designed to produce a sample of locations that would provide representative data for the entire country. However, over time many sites have been moved or replaced in order to reflect changing patterns and levels of pollution, and to reduce redundancy in the network. A particularly dramatic change occurred in 1981 when the network was reorganised owing to falling urban concentrations and to comply with EC directive 80/779/EEC (Colls, 2002). The analyses presented in this paper compared the levels and patterns over time for two groups of sites. Those retained over the whole period (consistent sites) had distinctly higher values than those that were terminated (the non-consistent sites). On the original scale of the data, the overall difference between consistent and non-consistent sites was 13%, (95% CI 4-24%). This quantifies what was observed in Figure 3 and provides evidence of preferential sampling.

The models considered here were fit using INLA with the SPDE approach being used to allow point referenced spatial components to be incorporated. An alternative might have been to perform inference using MCMC but in an initial attempt it turned out to be computationally prohibitive. This is often the case when dealing with large spatial datasets, both because of the requirement to manipulate large matrixes within each simulation of the MCMC and with convergence of parameters in complex models. In terms of prediction at a very high number of locations techniques, such as INLA, which perform 'approximate' Bayesian inference and thus do not require full MCMC sampling provide an extremely appealing approach. As shown in Lindgren et al. (2011), INLA can be extended to cover other situations, including non–stationary random GFs. However, as was pointed out in the discussion of Lindgren et al. (2011), it does not allow for the commonly used exponential model, where $\nu = 1/2$, or other noninteger values of ν. However, there are details in the authors reply to the discussion that the GMRF construction can be extended into a more general class of continuous domain Markov models, which contains close approximators of Matern models with fractional ν. Overall, the implantation of the INLA and SPDE approaches in this paper demonstrate how the methods can provide a remarkably fast computational algorithm for application over large domains when standard computational methods might fail.

The results presented here give support to the hypothesis of preferential sampling which has largely been ignored in environmental risk analysis. This may have implications if information from the network is used as a basis for measures of concentrations for the entire country. As such, if such information used as a basis for exposures experienced by entire populations in health studies then bias may be introduced. In addition to an assessment of the extent of the effect of preferential sampling, there is therefore a need for research into its potential for bias in health studies and policy guidance. Details of a number of possible approaches to correcting such bias can be found in Zidek and Shaddick (2012) which also gives some suggestions for possible implementations of the methods in practical settings. If such bias can be successfully corrected it would mean that estimates of pollution concentrations would be better suited for use in applications such as health studies and policy guidance.

Acknowledgement. The genesis of the work described in this paper was the 2009-10 thematic program on spatial analysis for environmental mapping, epidemiology and climate change developed by the Statistical and Applied Mathematics Institute on spatial statistics. We are grateful to members of its Working Group on preferential sampling, of which the second author was a member, for discussions that stimulated our interest in that topic. The authors would like to thank members of the INLA team and particularly Finn Lindgren and Harvard Rue for their invaluable assistance.

References

Banerjee, S., A. Gelfand, A. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Beelen, R., G. Hoek, E. Pebesma, D. Vienneau, K. de Hoogh, and D. Briggs (2009). Mapping of background air pollution at a fine spatial scale across the european union. *Science of the Total Environment 407*(6), 1852–1867.

Ciocco, A. and D. Thompson (1961). A follow-up of donora ten years after: methodology and findings. *Am J Public Health Nations Health 51*, 155–164.

Clifton, M. (1964). Air pollution. *Proceedings of the Royal Society of Medicine 57*(7), 615.

Colls, J. (2002). *Air pollution, modelling, and mitigation.* Abingdon, Oxford: Routledge.

Diggle, P., R. Menezes, and T. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Dockery, D. and C. I. Pope (1994). Acute respiratory effects of particulate air pollution. *Annu. Rev. Public Health 15*, 107–132.

Elliott, P., G. Shaddick, J. Wakefield, C. de Hoogh, and D. Briggs (2007). Long-term associations of outdoor air pollution with mortality in great britain. *Thorax 62*(12), 1088–1094.

European Commission (1980). Council directive 80/779/eec of 15 july 1980 on air quality limit values and guide values for sulphur dioxide and suspended particulates.

Finley, A., S. Banerjee, and B. Carlin (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software 19*(4), 1–24.

Firket, M. (1936). Fog along the meuse valley. *Trans. Faraday Soc. 32*, 1192–1197.

Garner, J. and R. Crow (1969). *Clean Air-Law and Practice.* Shaw and Sons Ltd.

Goldberg, M., R. Burnett, J. Bailar 3rd, J. Brook, Y. Bonvalot, R. Tamblyn, R. Singh, M. Valois, et al. (2001). The association between daily mortality and ambient air particle pollution in montreal, quebec. 1. nonaccidental mortality. *Environmental research 86*(1), 12.

Guttorp, P. and P. Sampson (2010). Discussion of Geostatistical inference under preferential sampling by Diggle, P.J., Menezes, R. and Su, T. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Gwynn, R., R. Burnett, and G. Thurston (2000). A time-series analysis of acidic particulate matter and daily mortality and morbidity in the buffalo, new york, region. *Environmental health perspectives 108*(2), 125.

Hansell, A., M. Blangiardo, C. Morris, D. Vienneau, J. Gulliver, K. Lee, and D. Briggs (2011). Association between black smoke and so2 air pollution exposures in 1971 and mortality 1972–2007 in great britain. *Epidemiology 22*(1), S29.

Hoek, G., B. Brunekreef, A. Verhoeff, J. van Wijnen, and P. Fischer (2000). Daily mortality and air pollution in the netherlands. *Journal of the Air & Waste Management Association 50*(8), 1380–1389.

Lee, D. and G. Shaddick (2010). Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics 66*(4), 1238–1246.

Lee, J., H. Kim, Y. Hong, H. Kwon, J. Schwartz, D. Christiani, et al. (2000). Air pollution and daily mortality in seven major cities of korea, 1991-1997. *Environmental research 84*(3), 247.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10*(4), 325–337.

Ministry of Health (1954). Mortality and morbidity during the london fog of december, 1962. *H.M.S.O. London*.

Molitor, J., M. Jerrett, C. Chang, N. Molitor, J. Gauderman, K. Berhane, R. McConnell, F. Lurmann, J. Wu, A. Winer, et al. (2007). Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environmental Health Perspectives 115*(8), 1147.

Monk, P. and L. Munro (2010). *Maths for chemistry: a chemist's toolkit of calculations*. OUP Oxford.

Muir, D. and D. Laxen (1995). Black smoke as a surrogate for pm¡ sub¿ 10¡/sub¿ in health studies? *Atmospheric Environment 29*(8), 959–962.

Ott, W. (1990). A Physical Explanation of the Lognormality of Pollutant Concentrations. *Journal of the Air Waste Management Association 40*, 1378–1383.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Sahu, S., A. Gelfand, and D. Holland (2006). Spatio-temporal modeling of fine particulate matter. *Journal of agricultural, biological, and environmental statistics 11*(1), 61–86.

Samet, J., F. Dominici, F. Curriero, I. Coursac, and S. Zeger (2000). Fine particulate air pollution and mortality in 20 us cities, 1987–1994. *New England journal of medicine 343*(24), 1742–1749.

Samoli, E., J. Schwartz, B. Wojtyniak, G. Touloumi, C. Spix, F. Balducci, S. Medina, G. Rossi, J. Sunyer, L. Bacharova, et al. (2001). Investigating regional differences in short-term effects of air pollution on daily mortality in the aphea project: a sensitivity analysis for controlling long-term trends and seasonality. *Environmental health perspectives 109*(4), 349.

Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 51*(3), 351–372.

Shrödle, B. and L. Held (2011). Spatio-temporal disease mapping using inla. *Environmetrics 22*, 725–734.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Stern, A., H. Wohlers, R. Boubel, and W. Lowry (1973). *Fundamentals of Air Pollution.* Academic Press.

Verhoeff, A., G. Hoek, J. Schwartz, and J. van Wijnen (1996). Air pollution and daily mortality in amsterdam. *Epidemiology 7*(3), 225–230.

WHO (2003). Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. *Report on a WHO Working Group, EUR/03/5042688*.

Zidek, J. and G. Shaddick (2012). Unbiasing estimates from preferentially sampled spatial data. Technical Report 268, Department of Statistics, University of British Columbia.