A Spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London

Abstract

It has long been known that air pollution is harmful to human health, as many epidemiological studies have been conducted into its effects. Collectively, these studies have investigated both the acute and chronic effects of pollution, with the latter typically based on individual level cohort designs that can be expensive to implement. As a result of the increasing availability of small-area statistics, ecological spatio-temporal study designs are also being used, with which a key statistical problem is allowing for residual spatio-temporal autocorrelation that remains after the covariate effects have been removed. We present a new model for estimating the effects of air pollution on human health, which allows for residual spatio-temporal autocorrelation, and a study into the long-term effects of air pollution on human health in Greater London, England. The individual and joint effects of different pollutants are explored, via the use of single pollutant models and multiple pollutant indices.

Keywords: air pollution; Gaussian Markov random fields; respiratory disease; spatio-temporal autocorrelation

1. Introduction

Air pollution is well known to be detrimental to human health, and can exacerbate many respiratory problems. It is a much greater issue in highly urbanised environments compared with rural locations, because of elevated pollutant concentrations due to emissions from traffic and industry, and high population densities resulting in large populations at risk. The health impact of air pollution is known to be different over different exposure periods, with epidemiological studies having been conducted into the effects of exposure in both the short and the long term. Studies investigating the effects of short-term (acute) exposure are the most common, and utilise time series of health and pollution data recorded at daily or weekly intervals. One of the first such studies was [41], and more recent examples include [37], [40], [28] and [7]. Much less research has been conducted into the health impact of long-term (chronic) exposure to pollution, and individual level cohort studies investigating this problem include [15], [21] and [5]. However, cohort studies are both expensive and time consuming to implement, due to the need to follow up a large cohort of people over an extended period of time.

Recently, small area health and social statistics from government run repositories have been made publicly available, with examples being the Surveillance Epidemiology and End Results (SEER, http://seer.cancer.gov/) database in the USA, and the Health and Social Care Information Centre (HSCIC, https://indicators.ic.nhs.uk/webview/) indicator portal in the UK. These databases contain population level annual aggregated summaries of disease incidence and socio-economic status for a set of irregularly shaped administrative units, such as electoral wards or census tracts. Additionally, modelled yearly average pollution concentrations estimated by computer dispersion models on a regular grid have also become freely available in recent times, with such data for the UK being provided by the Department for the Environment, Food and Rural Affairs (DEFRA, http://uk-air.defra.gov.uk/data/pcm-data). These data sources have enabled researchers to estimate the long-term health impact of air pollution using small-area spatio-temporal study designs, which due to the easy availability of the data are quick and inexpensive to implement. While these studies cannot assess the causal health effects of air pollution due to their ecological design, their ease of implementation mean that they contribute to and independently corroborate the body of evidence about the long-term population level impact of air pollution.

Examples of such studies in a purely spatial context include [18], [33], [34], [24], [1], [45], [14] and [23], while spatio-temporal designs include [11], [17],[13] and [22]. Poisson log-linear models are typically used for the analysis, where the linear predictor includes pollution concentrations and measures of socio-economic deprivation as covariates. However, the disease data typically contain residual spatial or spatio-temporal autocorrelation after the covariate effects have been accounted for, which violates the assumption of statistical independence that often underpins the models. This autocorrelation may be caused by numerous factors, including unmeasured confounding, neighbourhood effects (where subjects behaviour is influenced by neighbouring subjects), grouping effects (where subjects choose

to be close to similar subjects), and the fact that disease counts in consecutive time periods come from largely the same susceptible population. The common solution to this problem in the spatial studies listed above is to add a set of autocorrelated random effects to the linear predictor to account for this residual spatial structure. A number of different approaches have been adopted, including the use of conditional autoregressive (CAR) models [24], simultaneous autoregressive models [18], and geographically weighted regression [45]. In contrast, [22] is one of the only studies that has allowed for residual spatio-temporal autocorrelation, because the data modelled by [11], [17] and [13] do not relate to contiguous areal units, and hence in the main they assume the observations are independent.

Therefore, this paper makes two main contributions. Firstly, we propose a new Poisson log-linear hierarchical model for estimating the effects of air pollution on human health, which allows for the residual spatio-temporal autocorrelation using a set of autocorrelated random effects. A new Gaussian Markov Random Field (GMRF) model is proposed to represent this autocorrelation, and inference is based in a Bayesian setting using Markov Chain Monte Carlo (MCMC) simulation. We note that the development of such GMRF models has been a rich area of research in the related literature of space-time disease mapping. For example, [3] used different linear terms to represent location-specific temporal trends; [32] extend this to non-linear temporal effects using region-specific smooth spline terms; [20] and [19] investigate different GMRF constructions for accounting for space time structure, upon which subsequent studies have been based such as [43]. Functional approaches using penalized splines, such as [44], as well as GMRFs, have also been used to account for spatio-temporal structure in disease mapping applications, and a comparison of the relative performance of GMRF and functional approaches can be found in [12]. However, in disease mapping the random effects are of primary interest, where as in the ecological regression context considered here they are nuisance parameters included purely to remove the residual autocorrelation. As a result we consider a less highly parameterised GMRF model here, which is an extension of that proposed by [43]. We also note that the majority of the models listed above are not accompanied by freely available software to allow others to apply them to their own data, which is a limitation we rectify in this paper.

The second contribution of this paper is a new study investigating the long-term effects of air pollution on human health in London, England, which has a long history of air pollution problems dating back to the Great Smog of December 1952. London has been the location for many acute air pollution and health studies, including one of the first that was ever conducted [41]. However, no long-term studies utilising a spatio-temporal ecological design have been conducted in London, which is a gap in the literature that this paper aims to fill. Moreover, few studies in this context consider the combined effects of multiple pollutants simultaneously, despite the availability of such data. Therefore, this paper also seeks to construct appropriate air quality indicators based on a Principal Components analysis, that will enable all available pollutant data to be combined and used to estimate the health impact of a proxy measure of the air we breathe.

The remainder of this paper is organised as follows. Section 2 describes the study design and the data available for the Greater London study presented in this paper, and provides some exploratory analyses that motivate the use of the complex methods developed in this paper. Section 3 describes the Bayesian hierarchical modelling framework proposed here for estimating the health effects of air pollution, which accounts for non-separable spatio-temporal residual autocorrelation. In section 4 the results of the Greater London study are presented, including estimates of the health impact of both individual pollutants and a composite index. Finally, Section 5 concludes the paper by summarising the main findings of the work, and includes some avenues for future research.

2. Epidemiological study

The methodology developed in this manuscript was motivated by a new epidemiological study of air pollution concentrations and respiratory hospital admissions in London, UK. London has a long history of air pollution problems, with very high levels of smoke and SO_2 being observed since the industrial revolution. Famous pollution events include the 'Great Smog' of 1952, in which London was shrouded in a thick layer of airborne pollutants, predominantly originating from coal smoke. Many studies of the health impacts of this smog event have been undertaken, including [2], who estimate that it resulted in 12000 excess deaths between December 1952 and February 1953. As a result of this event, greater regulation on black smoke and coal burning was introduced in the Clean Air Act of 1956, and air pollution was vastly reduced over subsequent years as a result. However, air pollution in London remains a critical public health issue today, with an estimated 4000 deaths a year attributable to poor air quality alone [35]. In addition to being the most populous city in the European Union, London is also one of the most economically diverse, often with localised clusters of very affluent neighbourhoods bordering some of the most deprived. It is for these reasons that a spatio-temporal autocorrelation model is likely to prove vital, so that the residual spatio-temporal autocorrelation driven by observed and unobserved confounding factors is accounted for and does not distort the estimated effects of pollution exposure on health.

2.1. Health data

The data analysed in this paper consist of a set of annualised counts (Y_{ij}) of the numbers of hospital admissions for respiratory disease (International Classification of Diseases codes J00-J99) for each of the 624 electoral wards that make up Greater London (indexed by i), and for each of the years spanning 2003 to 2009 (indexed by j). Although each of the electoral wards have approximately similar population sizes they are not identical, and the observed numbers of admissions will depend on these differences as well as the demographic structure therein. Therefore, the expected numbers (E_{ij}) of admissions were calculated by external standardisation, using age and gender specific respiratory admissions rates for the UK. Insight into the spatial distribution of risk can be obtained by looking at maps of the Standardised Incidence Ratio (SIR) defined as $SIR_{ij} = Y_{ij}/E_{ij}$, which is shown for 2005 in the top panel of Figure 1. Spatially, the highest values of the SIR appear to be concentrated around the east of Central London, and on the western periphery around Heathrow Airport and the M1 motorway. These are persistent features during the time period for which data are available, and largely correspond to socio-economic deprivation across the city (not shown). Overall, there appears to be little change in the SIR over the 7 year period as the median values for each year vary between 0.75 and 0.8.

2.2. Pollutant and covariate data

One of the most important contributors to respiratory disease and ill-health in general is socio-economic deprivation, which is multi-factorial and cannot be measured directly. One approach is to use a deprivation score such as the English indices of deprivation (https://www.gov.uk/government/organisations/department-for-communities-and-local-government/series/english-indices-of-deprivation) provided by the Department for Communities and Local Government. However, such indices are typically unavailable at the electoral ward level for the temporal extent for which the respiratory health data are available. As a result, proxy measures of deprivation and socio-economic status are available, such as median house price (`Price`) in each area and the proportion of the population in each electoral ward that are in receipt of Job Seekers Allowance (`JSA`). The `JSA` data are available for each time period and electoral ward, while `Price` is only available at Local Authority level (32 Local Authority areas make up Greater London). Although the impact on respiratory disease of smoking prevalence at local area level will dwarf that of air pollution, smoking prevalence data are unavailable at the ward level and for the 7 year period of this study and are not included in the analysis. However, London borough-level smoking prevalence data are available for 2009 [30], and this exhibited a linear relationship with `JSA` (Pearson's correlation correlation coefficient of 0.67), which suggests that socio-economic proxy variables can control for the majority of the health effects due to smoking.

Ideally, pollution data from a network of ground monitoring stations would be used to represent population level exposure. However, the network available is too sparse to give a full spatial picture of air pollution concentrations in each of the 624 electoral wards across Greater London. Therefore background pollution maps based on dispersion models and provided by DEFRA (http://www.uk-air.defra.gov.uk) were used, which contain modelled annual mean concentrations in $\mu\mathrm{gm}^{-3}$ for each of carbon monoxide (CO); nitrogen dioxide (NO_2); the total of nitrogen monoxide and nitrogen dioxide (NO_x); particulate matter less than 2.5 micrometers in diameter ($PM_{2.5}$); particulate matter less than 10 micrometers in diameter (PM_{10}) and sulpher dioxide (SO_2) each on a 1km$\times$1km grid. Each pollution variable was lagged by one year relative to the respiratory admission data, so that the pollution exposure occurs before the health events. Ideally, lags extending beyond a single year would be investigated in order to understand the impact of historic air pollution exposures, and 'distributed lag' models could be used to investigate how these effects can accumulate with time. However, each additional lag introduced requires a reduction in the number of years of data available for the study, and a one year lag was used to limit this data loss. In order to align the pollution grids to the electoral ward scale, the median was calculated for each pollutant in each electoral ward of the modelled pollution tiles that were in that area. These concentrations are shown in Figure 1 for PM_{10} in 2005, ad the concentration of each pollution variable are summarised in Table 1. While the median levels of CO, NO_2, NO_x and SO_2 appear to be on an overall downwards trajectory, it is less clear whether PM_{10} and $PM_{2.5}$ are rising or falling during this period. Median house price (`Price`) has been transformed by dividing by 1000 and taking the natural log, while the proportion of a ward claiming job seekers allowance is on the percentage scale - both of these conventions will be maintained for the remainder of the paper.

2.3. Exploratory analysis using a generalised linear model

Initially, a Poisson generalised linear model of the form

$$\begin{aligned} Y_{ij} &\sim \mathrm{Poisson}(E_{ij}R_{ij}) \\ \log(R_{ij}) &= \beta_0 + \beta_1 \mathtt{JSA}_{ij} + \beta_2 \mathtt{Pollutant}_{i,j-1} + \beta_3 \mathtt{Price}_{ij} \end{aligned} \tag{1}$$

was fitted to the data with just the covariates, to assess the presence of residual spatio-temporal correlation. Here, R_{ij} represents disease risk in area i at time j. However, the residuals that result from this analysis exhibit strong spatio-temporal autocorrelation, with an associated Moran's I statistic [36] of 0.3434 and an associated p-value of 0.0099, suggesting that some source of variation in respiratory disease risk has not been adequately accounted for. This autocorrelation is particularly evident from the plot of the spatial residuals shown in the top panel of Figure 2, in which high levels of spatial smoothness are visible. Furthermore, there is evidence of non-separable space-time smoothness in the spatial residuals, which can be seen in the bottom panel of Figure 2. The figure displays the difference between two successive years residuals, and under the assumption of a separable structure a 'flat' overall surface would be expected. However, the surface exhibits substantial spatial variation, suggesting the presence of non-separable space-time structure. Therefore in the next section we propose a model that allows for non-separable space-time residual autocorrelation in an intuitive manner.

3. Modelling

The study region is partitioned into a set of N non-overlapping areal units indexed by $i \in \{1, \ldots, N\}$, and data are observed for each of these units for $j \in \{1, \ldots, T\}$ consecutive time periods. A Bayesian hierarchical model is proposed for these data, with inference based on MCMC simulation. The first level of the hierarchical model is given by

$$\begin{aligned} Y_{ij}|E_{ij}, R_{ij} &\sim \text{Poisson}(E_{ij}R_{ij}), \\ \ln(R_{ij}) &= \mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta} + \phi_{ij}, \\ \beta_k &\sim \text{N}\,(0, 1000) \qquad\qquad k \in \{1, \ldots, p\}, \end{aligned} \tag{2}$$

where Y_{ij} and E_{ij} are the observed and expected numbers of disease cases in areal unit i during time period j, and are described in Section 2.1. Here $\mathbf{x}_{ij}^{\text{T}}$ is a $p \times 1$ vector of covariates relating to areal unit i during time period j, while $\boldsymbol{\beta}$ is the associated $p \times 1$ vector of regression parameters. For the epidemiological study discussed in Section 2, the covariate component is given by $\beta_0 + \beta_1 \texttt{JSA}_{ij} + \beta_2 \texttt{Pollutant}_{i,j-1} + \beta_3 \texttt{Price}_{ij}$, where $\texttt{Pollutant}_{ij}$ is generic notation for one of the pollutants summarised in Section 2.2.

The random effects ϕ_{ij} are included in (2) to allow for any residual spatio-temporal autocorrelation in the data after the covariate effects have been removed, and are represented by a GMRF prior distribution. Numerous GMRF priors have been proposed for spatio-temporal random effects relating to areal unit data in the related field of disease mapping, although their application in long-term air pollution and health studies is rare ([22] being one such example). Both separable [20] and non-separable [19] spatio-temporal structures have been proposed, and as the former makes the restrictive assumption that the residual spatial structure is the same for all time periods which from Section 2.3 is unlikely to be realistic, we consider a non-separable model here. The non-separable model of [19] contains both spatial and temporal main effects and a non-separable interaction term, and is thus appropriate when the aim of the analysis is to identify these constituent parts of the spatio-temporal structure in the data. However, in the ecological regression context considered here the random effects are nuisance parameters included to account for any residual spatio-temporal autocorrelation in the data, and are not of direct interest. Therefore here we follow the less highly parameterised model proposed by [43], and decompose the single set of random effects $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T)$ as

$$f(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T) \sim f(\boldsymbol{\phi}_1) \prod_{j=2}^{T} f(\boldsymbol{\phi}_j|\boldsymbol{\phi}_{j-1}), \tag{3}$$

where $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{Nj})$ denotes the vector of random effects for time period j. This decomposition induces temporal autocorrelation by explicitly allowing $\boldsymbol{\phi}_j$ to depend on $\boldsymbol{\phi}_{j-1}$, while $\boldsymbol{\phi}_1$ is specified marginally as $\boldsymbol{\phi}_0$ does not exist. The GMRF prior specified for $f(\boldsymbol{\phi}_1)$ induces spatial autocorrelation into the random effects at time period 1 by means of a binary $N \times N$ adjacency matrix $W = (w_{ik})$, which is based on the contiguity structure of the N areal units. Element $w_{ik} = 1 \iff$ areal unit i shares a border with areal unit k, otherwise $w_{ik} = 0$, and also $w_{ii} = 0\ \forall i$. The joint prior distribution for $\boldsymbol{\phi}_1$ is given by $\boldsymbol{\phi}_1 \sim \text{N}\left(\mathbf{0}, \tau^2 Q(\rho, W)^{-1}\right)$, where spatial autocorrelation is induced by the precision matrix $Q(\rho, W)$. A number of GMRF specifications have been used in the spatial modelling literature for $Q(\rho, W)$, most of which are special cases of conditional autoregressive (CAR) models. The one we adopt here was proposed by [29] and is given by $Q(\rho, W) = \rho(\text{diag}(W\mathbf{1}) - W) + (1-\rho)I$, where I is the $N \times N$ identity matrix and $\mathbf{1}$ is an $N \times 1$ vector of ones. This matrix is proper if $\rho \in [0, 1)$, and the spatial structure amongst $\boldsymbol{\phi}_1$ can be observed more clearly from the univariate full conditional distributions which are given by

$$\phi_{i1}|\boldsymbol{\phi}_{-i1} \sim \mathrm{N}\left(\frac{\rho\sum_{k=1}^{N} w_{ik}\phi_{k1}}{\rho\sum_{k=1}^{N} w_{ik} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{k=1}^{N} w_{ik} + 1 - \rho}\right). \tag{4}$$

In the above equation, $\boldsymbol{\phi}_{-i1}$ denotes the vector of random effects for time period 1 except for ϕ_{i1}. From (4) it is clear that ρ controls the spatial autocorrelation structure, with $\rho = 1$ corresponding to the intrinsic CAR prior [4] for strong spatial autocorrelation, where the conditional expectation is the mean of the random effects in geographically adjacency areal units. In contrast, $\rho = 0$ corresponds to independent random effects with constant mean and variance. Temporal autocorrelation is induced into the random effects by the conditional specifications $f(\boldsymbol{\phi}_j|\boldsymbol{\phi}_{j-1})$, which are given by

$$\boldsymbol{\phi}_j|\boldsymbol{\phi}_{j-1} \sim \mathrm{N}\left(\alpha\boldsymbol{\phi}_{j-1}, \tau^2 Q(\rho, W)^{-1}\right) \qquad j \in \{2, \ldots, T\}, \tag{5}$$

where the precision matrix $Q(\rho, W)$ is as defined above. This model thus induces temporal autocorrelation through the conditional expectation, while spatial autocorrelation is induced via the precision matrix. The level of temporal autocorrelation is controlled by α, with $\alpha = 0$ corresponding to temporal independence, while $\alpha = 1$ corresponds to strong temporal autocorrelation and is a first order random walk model. We specify weakly informative hyperpriors for the parameters (τ^2, ρ, α) as

$$\begin{aligned} \tau &\sim U[0, 1000], \\ \alpha &\sim U[0, 1], \\ \rho &\sim U[0, 1], \end{aligned}$$

which allows their values to be informed by the data. Values of (ρ, α) equal to one correspond to non-stationary processes in space and time, while values in the interval $[0, 1)$ lead to stationary specifications. The random effects as modelled above are non-separable in space and time, as the spatial structure at time j is equal to a proportion of the spatial structure at time $j - 1$ plus error. The spatial structure thus evolves through time, with the magnitude and strength of this evolution being controlled by the hyperparameters (τ^2, ρ, α). Our model is a straightforward extension of that proposed by [43], which makes the restriction of strong temporal dependence by setting $\alpha = 1$.

To obtain posterior summaries for the model parameters $\Theta = (\boldsymbol{\beta}, \boldsymbol{\phi}, \tau^2, \rho, \alpha)$, samples were drawn from the posterior distributions using MCMC simulation, based on a mixture of Gibbs sampling and Metropolis-Hastings steps. The analyses presented in Section 4 are based on running the model for a burn in period of 10,000 iterations, after which visual diagnostics suggested that the Markov chains had converged. Inference in all cases was then based on a further 50,000 samples. Spatio-temporal models of this type are very computationally intensive to simulate form, due to the large number of random effects and their complex spatio-temporal autocorrelation structure. As a result, we have implemented our MCMC algorithm using an efficient `C++` script written using the R package `Rcpp` [10], [9]. All of the software developed was implemented within the R [38] statistical programming language, and is available for download with this paper along with the hospital admissions, $PM_{2.5}$, `JSA` and `Price` data.

4. Results

4.1. Investigating α

We first investigate whether the inclusion of the temporal smoothness parameter α was required, by fitting the model described in Section 3 to the data with α fixed at the values 0 and 1, as well as allowing it to be selected by the data. Each of the three different models were fitted to the London data, and for each the Deviance Information Criterion (DIC) [42] was calculated to give an indication of how well each model fits the data. The lowest DIC value of 36073 is associated with the most flexible scenario in which α is estimated from the data, and is substantially lower than those associated with the scenarios in which α is held fixed (36986 for $\alpha = 0$ and 36125 for $\alpha = 1$ respectively). This illustrates that if the strength of temporal dependence is assumed fixed in advance, the model does not fit the data as well. The estimated posterior median was $\hat{\alpha} = 0.85$, which suggests that while the residual temporal autocorrelation in the London data is strong, the temporal evolution of the random effects is stationary.

4.2. Estimation of the health impact of air pollution and other covariates

The main aim of this study is to estimate the human health impacts of different types of air pollution, and to investigate techniques for estimating the joint effects of numerous air pollutants simultaneously. For the former, each pollutant can be included separately in the spatio-temporal model described in Section 3, in order to identify its relative impact on respiratory health. The principle underlying the latter is that the air we breathe is a complex

mixture of different pollutants, and thus its health effects may be different to those of individual pollutants. However, air pollution measurements exhibit strong linear correlations, because they may be generated by common processes or be driven by similar factors such as meteorology. This means that it is inappropriate to include a number of different pollutants in a single model as they are collinear, and thus their individual effects would not be well estimated. Therefore, we propose constructing an Air Quality Indicator (AQI) composed of an appropriately chosen linear combination of the air pollution measures available.

A first approach might be to construct an AQI based on the average of the 6 air pollutants in time and space. This approach has been taken by a number of studies, and was further developed by [25] who recognise that the uncertainty present in the individual pollutant measures should be adequately accounted for in the final AQI. Even so, the restrictive assumption that each of the pollutants contributes equally to the resulting AQI is made, and does not take account of the correlation between the contributing pollutants. The UK Met Office calculate a Daily Air Quality Index (DAQI) by assigning each measured air pollutant a score between 1 and 10, and the overall DAQI is defined as the maximum of these scores (www.metoffice.gov.uk/guide/weather/air-quality). However, for the DAQI the assignment of scores and the thresholds that define them are based on daily pollution levels and it is unclear whether this type of AQI is appropriate for the annual data considered here. An alternative approach is to implement a dimension reduction analysis such as Principal Components Analysis (PCA), in order to identify a small number of orthogonal indicators based on different linear combinations of the pollutants that can account for the maximum level of variability in the multivariate pollution data. The principal component loadings that result from this analysis are shown in Table 2, which are the weights each pollutant is multiplied by to create the composite index.
The first principal component (PC) shown in Table 2 shows similar positive loadings across all pollutants, indicating that 56% of the variation in air pollution is due to the overall amount of air pollution, and this PC has a similar interpretation to the average pollution AQIs discussed earlier. The second PC describes a contrast between the effects of large particulate matter as measured by PM_{10}, and SO_2. The third PC contrasts the effects of Nitrogen Oxides with small particulate matter as measured by $PM_{2.5}$, and SO_2. Since the first three principal components account for 91% of the total variation in the air pollution data, only these three will be used as air quality indicators.

For each pollutant, Table 3 presents the relative risk associated with a 1 standard deviation increase with it's associated 95% credible interval (CI). All of the estimated relative risks associated with air pollutants are greater than 1, indicating median increases in respiratory admissions of between 0.7% (SO_2) and 2.7% ($PM_{2.5}$) should be expected for 1 standard deviation increases in pollution. The adverse health effects associated with PM_{10}, $PM_{2.5}$ and CO are substantial at the 95% level, although all of the other pollutant's relative risks had lower levels of the credible intervals that were less than one. The strongest overall effect was associated with $PM_{2.5}$, with annual increases in respiratory hospital admissions of between 0.9% and 4.4% expected for a 2.03 $\mu\mathrm{gm}^{-3}$ increase in small particulate matter concentrations.

Of the three PCs fitted to account for the joint impact of air pollution, the first and the third were estimated as having substantial impacts on health. These risks were 2.6% for PC_1 and 1.7% for PC_3. Since these variables are orthogonal it is appropriate to include them in the same model simultaneously, and is a more efficient use of the air pollution data. PC_1 is a roughly even weighted average of all pollutants, and thus increasing overall pollution levels leads to a substantial increase in respiratory disease risk. PC_2 exhibits no relationship to respiratory ill health, which means that changing the composition of air pollution to include relatively more coarse particles (PM_{10}) at the expense of SO_2, is not harmful to health. In contrast, PC_3 does exhibit a substantial health impact, which means that a relative reduction in CO, NO_2 and NO_x and increase in fine particulate matter ($PM_{2.5}$) results in substantial health effects.

It can also be seen from Table 3 that for a 6.6% increase in the proportion of a ward population claiming JSA, an increase in admissions of 20.7% would be expected, indicating that JSA is a very informative covariate and that increasing JSA, as a proxy for smoking has a much stronger impact than air pollution. The House price variable was also found to be strongly associated with decreased hospital admissions, with a decrease in risk of 5% associated with a 0.31 increase in the log House price. For example, this is the difference in admissions that would be expected between two electoral wards in which the average house prices are £200,000 and £270,000 on the original scale.

5. Discussion

In this paper we have proposed a novel spatio-temporal modelling approach for estimating the long-term health impact of air pollution while allowing for non-separable residual spatio-temporal autocorrelation, and have applied this methodology to a new epidemiological study of the effects of air pollution on respiratory ill health in Greater London. While London has been the scene for many short-term time series studies in the past, this is the first study

of the longer-term impact of pollution using a small-area spatio-temporal design. In addition, we are one of very few studies to take a multifactorial view of the health impact of pollution, as we have considered both single pollutant analysis, as well as composite indicators of air quality generated using principal components analysis.

The main results of the study are that after accounting for socio-economic deprivation, air pollution concentrations are associated with an increase in the incidence of respiratory hospital admissions. In particular, one standard deviation increases in $PM_{2.5}$ and CO were found to significantly increase the rate of respiratory hospital admissions, by around 1.8% and 2.7% respectively. Furthermore, an increase in PC_1, which represents a roughly equal weighted average of all pollutants, is associated with a 2.6% increase in respiratory ill health, while an increase in PC_3 is associated with a 1.7% increase. The latter represents a change in the composition of pollution, obtained by increasing the amounts of $PM_{2.5}$ and SO_2 relative to CO, NO_2 and NO_x. This result thus re-enforces the significant association observed between $PM_{2.5}$ and ill health in the single pollutant analysis, and suggests that it is fine particles that may pose the greatest risk to respiratory ill health. Therefore air pollution still has a substantial population level health impact in London, even at the relatively low concentrations observed in recent years.

The limitations of data availability mean that the results of this study are subject to the following caveats. Firstly, this study assumes that impact of smoking can be accounted for by socio-economic variables, and that this relationship is linear. In Section 2, JSA was found to increase linearly with smoking prevalence at Borough level in 2009, which is consistent with these assumptions. However, interaction with other factors such as ethnic composition could mean that this specification is too simplistic. For example, [6] find that ethnicity affects smoking rates even after controlling for wealth, and so incorporating ethnic composition data should be considered in future studies. In addition, population transience has the potential to dilute the area-specific impact of an exposure, if a large enough proportion of the population in each area have re-located in a short time period. For London boroughs, these movements can be particularly large, for example [16] indicates that in 2008, both Hammersmith and Fulham, and Islington experienced population turnovers of over 27%. However, the impact these might have on air pollution and health studies is unclear, since [8] show that a large proportion of UK internal population movement is due to residents in their late teens or early twenties, who in turn are at low risk of respiratory problems and are unlikely to make up the high-risk subpopulation from which the respiratory admissions are drawn. It is therefore not expected that the effects of population transience would diminish the results presented, however we note that this is an issue for all small-area studies.

One of the key motivators of this work was to develop a model that captured the residual spatio-temporal autocorrelation in the respiratory disease data after the covariate effects have been removed, which if ignored can bias the estimated health effects of air pollution. A class of models based on GMRF priors was used for this purpose, which was a simple extension of that proposed by [43] in the related field of disease mapping. Where our model differs from theirs is that we consider varying degrees of temporal autocorrelation in the random effects structure, and this flexibility was found to be necessary for the Greater London data. However, in a purely spatial setting [39] has shown that there is potential for collinearity between spatially smooth covariates and the globally spatially smooth random effects, which may lead to unstable fixed effects estimation. A related issue is that the residual spatio-temporal autocorrelation that the random effects are designed to model may not be globally smooth, as two bordering areal units might contain communities that have very different characteristics. This feature calls into question the use of border sharing as a proxy measure of similarity, with which a smooth residual structure is defined. Some recent developments try to address this issue in a purely spatial setting by treating the neighbourhood structure as a quantity that must be estimated as part of the modelling process. Approaches include [31], [26] and [27], and for each, attempts are made to avoid overparameterisation and reduce computational complexity. Therefore, an avenue of future work will be to investigate these phenomena in the spatio-temporal context considered here, to see what impact they may have on the estimated pollution-health relationships.

Part of the focus of this study was to investigate the impacts of different pollutants both individually and jointly, to determine if the air we breathe has a larger health impact than individual pollutants. The PCA undertaken for this purpose successfully yielded two transformed variables that were both found to be strongly related to respiratory admissions, indicating that additional information can be unlocked by taking a multifactorial view of pollution. However, this analysis has ignored two types of uncertainty, which should be allowed to propagate through the PCA. First a measure of variability should be associated with the modelled air pollutant concentrations themselves, and secondly the factor loadings obtained from the PCA have also been estimated and are thus subject to error. In future work, approaches to account more fully for these sources of uncertainty in the pollution data will be pursued.

6. Funding

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC), via grant EP/J017442/1.

7. Supplementary materials

The supplementary materials include annual $PM_{2.5}$ concentrations for each London ward between 2002 and 2008, and respiratory hospital admissions, `JSA` and `Price` for each ward between 2003 and 2009. The materials also include the `C++` functions and an R script describing how to fit the spatio-temporal random effects model described in Section 3.

8. Acknowledgements

The authors gratefully acknowledge the editor and a reviewer whose thoughtful comments have improved the content and presentation of this paper.

References

[1] Barceló, M. A., M. Saez, and C. Saurina (2009). Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the barcelona metropolitan region, spain. *Science of the Total Environment 407*(21), 5501–5523.

[2] Bell, M. L. and D. L. Davis (2001). Reassessment of the lethal london fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environmental health perspectives 109*(Suppl 3), 389.

[3] Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini (1995). Bayesian analysis of spacetime variation in disease risk. *Statistics in Medicine 14*(21-22), 2433–2443.

[4] Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*(1), 1–20.

[5] Beverland, I., C. Robertson, C. Yap, M. Heal, G. Cohen, D. E. J. Henderson, C. Hart, and R. Agius (2012). Comparison of models for estimation of long-term exposure to air pollution in cohort studies. *Atmospheric Environment*.

[6] Bhopal, R., A. Vettini, S. Hunt, S. Wiebe, L. Hanna, and A. Amos (2004). Review of prevalence data in, and evaluation of methods for cross cultural adaptation of, uk surveys on tobacco and alcohol in ethnic minority groups. *BMJ: British Medical Journal 328*(7431), 76.

[7] Chang, H. H., R. D. Peng, and F. Dominici (2011). Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. *Biostatistics 12*(4), 637–652.

[8] Dennett, A. and J. Stillwell (2008). Population turnover and churn: enhancing understanding of internal migration in britain through measures of stability. *Population trends 134*, 24–41.

[9] Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York: Springer. ISBN 978-1-4614-6867-7.

[10] Eddelbuettel, D. and R. François (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*(8), 1–18.

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

[12] Goicoa, T., M. D. Ugarte, J. Etxeberria, and A. Militino (2012). Comparing car and p-spline models in spatial disease mapping. *Environmental and Ecological Statistics 19*(4), 573–599.

[13] Greven, S., F. Dominici, and S. Zeger (2011). An approach to the estimation of chronic air pollution effects using spatio-temporal information. *Journal of the American Statistical Association 106*(494), 396–406.

[14] Haining, R., G. Li, R. Maheswaran, M. Blangiardo, J. Law, N. Best, and S. Richardson (2010). Inference from ecological models: estimating the relative risk of stroke from air pollution exposure using small area data. *Spatial and Spatio-temporal Epidemiology 1*(2), 123–131.

[15] Hoek, G., B. Brunekreef, S. Goldbohm, P. Fischer, and P. A. van den Brandt (2002). Association between mortality and indicators of traffic-related air pollution in the netherlands: a cohort study. *The Lancet 360*(9341), 1203–1209.

[16] Hollis, J. (2010). Focus on london: Population and migration.

[17] Janes, H., F. Dominici, and S. L. Zeger (2007). Trends in air pollution and mortality: an approach to the assessment of unmeasured confounding. *Epidemiology 18*(4), 416–423.

[18] Jerrett, M., M. Buzzelli, R. T. Burnett, and P. F. DeLuca (2005). Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Social Science & Medicine 60*(12), 2845–2863.

[19] Knorr-Held, L. (1999). Bayesian modelling of inseparable space-time variation in disease risk.

[20] Knorr-Held, L. and J. Besag (1998). Modelling risk from a disease in time and space. *Statistics in medicine 17*(18), 2045–2060.

[21] Laden, F., J. Schwartz, F. E. Speizer, and D. W. Dockery (2006). Reduction in fine particulate air pollution and mortality: extended follow-up of the harvard six cities study. *American Journal of Respiratory and Critical Care Medicine 173*(6), 667.

[22] Lawson, A. B., J. Choi, B. Cai, M. Hossain, R. S. Kirby, and J. Liu (2012). Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. *Journal of agricultural, biological, and environmental statistics 17*(3), 417–441.

[23] Lee, D. (2012). Using spline models to estimate the varying health risks from air pollution across scotland. *Statistics in Medicine 31*(27), 3366–3378.

[24] Lee, D., C. Ferguson, and R. Mitchell (2009). Air pollution and health in scotland: a multicity study. *Biostatistics 10*(3), 409–423.

[25] Lee, D., C. Ferguson, and E. M. Scott (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(1), 109–126.

[26] Lee, D. and R. Mitchell (2012). Boundary detection in disease mapping studies. *Biostatistics 13*(3), 415–426.

[27] Lee, D. and R. Mitchell (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(4), 593–608.

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

[29] Leroux, B. G., X. Lei, and N. Breslow (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 179–191. Springer.

[30] London Knowledge and Intelligence Team at Public Health England (2013). Local tobacco control profiles for england.

[31] Ma, H., B. P. Carlin, and S. Banerjee (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics 66*(2), 355–364.

[32] MacNab, Y. C. and C. Dean (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics 57*(3), 949–956.
[33] Maheswaran, R., R. P. Haining, P. Brindley, J. Law, T. Pearson, P. R. Fryers, S. Wise, and M. J. Campbell (2005). Outdoor air pollution and stroke in sheffield, united kingdom a small-area level geographical study. *Stroke 36*(2), 239–243.
[34] Maheswaran, R., R. P. Haining, T. Pearson, J. Law, P. Brindley, and N. G. Best (2006). Outdoor nox and stroke mortality: adjusting for small area level smoking prevalence using a bayesian approach. *Statistical methods in medical research 15*(5), 499–516.
[35] Miller, B. G. (2010). Report on estimation of mortality impacts of particulate air pollution in london. *Institute of Occupational Medicine*. [Online; accessed 29-July-2013].
[36] Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.
[37] Peters, A., D. Dockery, J. Heinrich, and H. Wichmann (1997). Short-term effects of particulate air pollution on respiratory morbidity in asthmatic children. *European Respiratory Journal 10*(4), 872–879.
[38] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
[39] Reich, B. J., J. S. Hodges, and V. Zadnik (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics 62*(4), 1197–1206.
[40] Schwartz, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology 11*(3), 320–326.
[41] Schwartz, J. and A. Marcus (1990). Mortality and air pollution j london: a time series analysis. *American journal of epidemiology 131*(1), 185–194.
[42] Spiegelhalter, D. J., N. G. Best, B. P. 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.
[43] Ugarte, M. D., J. Etxeberria, T. Goicoa, and E. Ardanaz (2012). Gender-specific spatio-temporal patterns of colorectal cancer incidence in navarre, spain (1990–2005). *Cancer Epidemiology 36*(3), 254–262.
[44] Ugarte, M. D., T. Goicoa, and A. Militino (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics 21*(3-4), 270–289.
[45] Young, L. J., C. A. Gotway, J. Yang, G. Kearney, and C. DuClos (2009). Linking health and environmental data in geographical analysis: Its so much more than centroids. *Spatial and Spatio-temporal Epidemiology 1*(1), 73–84.

Pollutant	2002	2003	2004	2005	2006	2007	2008
CO	372.00	372.00	343.00	338.00	264.00	251.00	229.00
NO_2	34.80	36.70	31.90	33.70	32.40	34.40	30.30
NO_x	59.50	62.00	53.70	56.30	53.10	57.80	50.10
$PM_{2.5}$	11.50	17.10	16.70	14.90	15.10	13.50	14.10
PM_{10}	17.60	20.20	24.60	23.40	22.70	23.80	20.30
SO_2	3.75	6.02	3.10	3.02	3.08	3.16	2.37

Table 1: Median annual air pollution concentrations of London wards between 2002 and 2008 inclusive.

Pollutant	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
CO	0.43	-0.31	-0.19	-0.75	0.34	-0.02
NO_2	0.51	-0.04	-0.33	0.30	-0.23	-0.70
NO_x	0.51	-0.04	-0.34	0.28	-0.20	0.71
$PM_{2.5}$	0.38	0.27	0.64	-0.30	-0.54	0.01
PM_{10}	0.31	0.68	0.13	0.15	0.63	-0.01
SO_2	0.24	-0.61	0.57	0.39	0.32	0.00
Cumulative proportion of variance	0.56	0.77	0.91	0.97	1.00	1.00

Table 2: Loadings corresponding to a PCA performed on the air pollution data and the cumulative proportion of the variance explained.

Pollutant	RR	95% CI	St.Dev
CO	1.023	(1.000, 1.039)	75.83
NO_2	1.013	(0.995, 1.030)	6.99
NO_x	1.009	(0.987, 1.031)	14.67
$PM_{2.5}$	1.027	(1.009, 1.044)	2.03
PM_{10}	1.018	(1.001, 1.038)	2.85
SO_2	1.007	(0.996, 1.019)	1.38
PC_1	1.026	(1.006, 1.044)	1.82
PC_2	1.003	(0.987, 1.025)	1.12
PC_3	1.017	(1.001, 1.033)	1.12
JSA	1.207	(1.193, 1.220)	6.64
Price	0.950	(0.925, 0.973)	0.31

Table 3: A summary of the parameter estimates from the spatio-temporal model. The estimated covariate effects are relative risks (RR) for a one-standard deviation (Std. dev) increase in each variable's value, shown in the final column

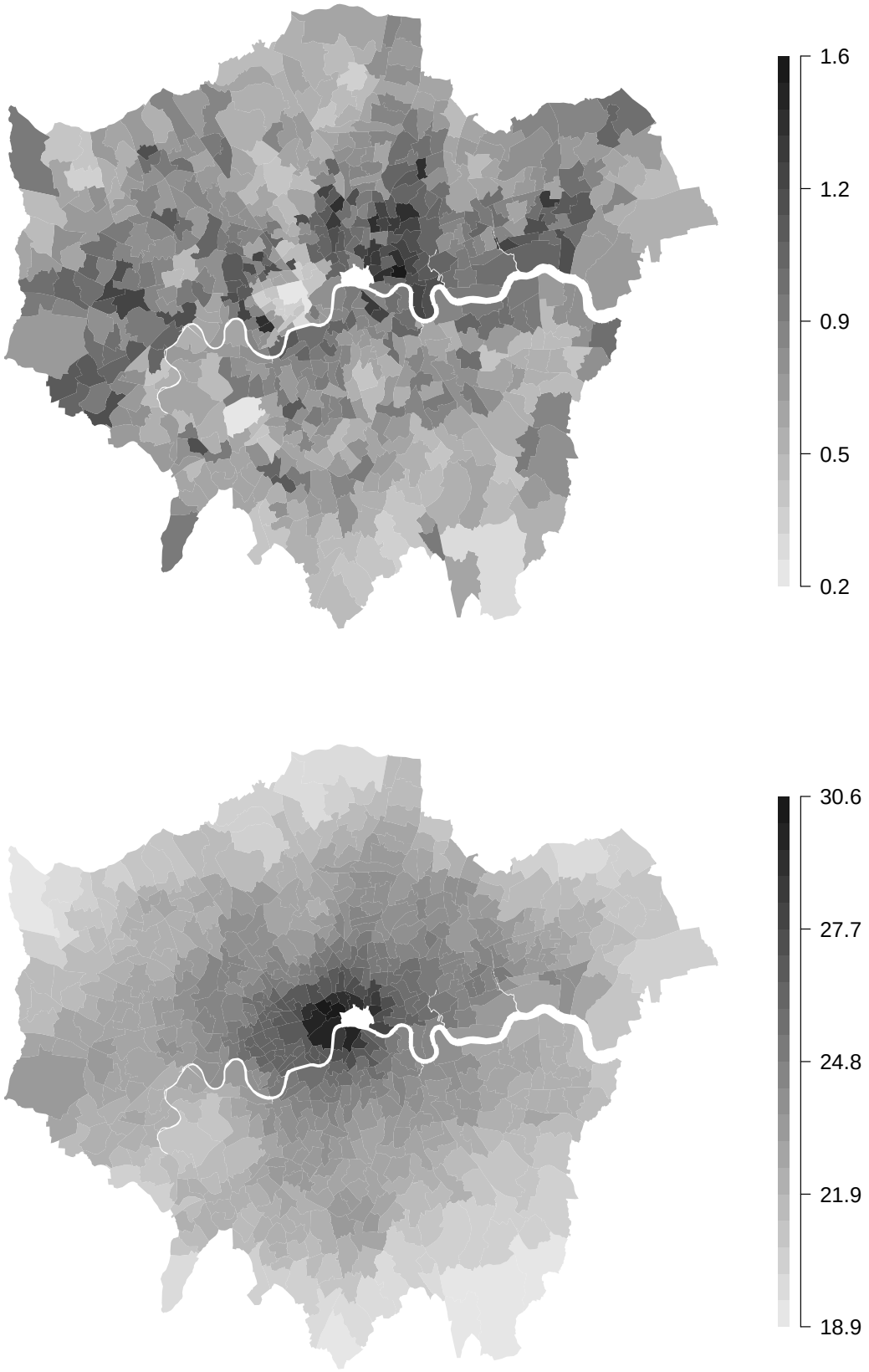

Figure 1: (Top panel) Standard incidence ratio (SIR) for respiratory hospital admissions across Greater London for 2005. (Bottom panel) Average PM_{10} pollution concentrations (μgm^{-3}) across Greater London for 2005

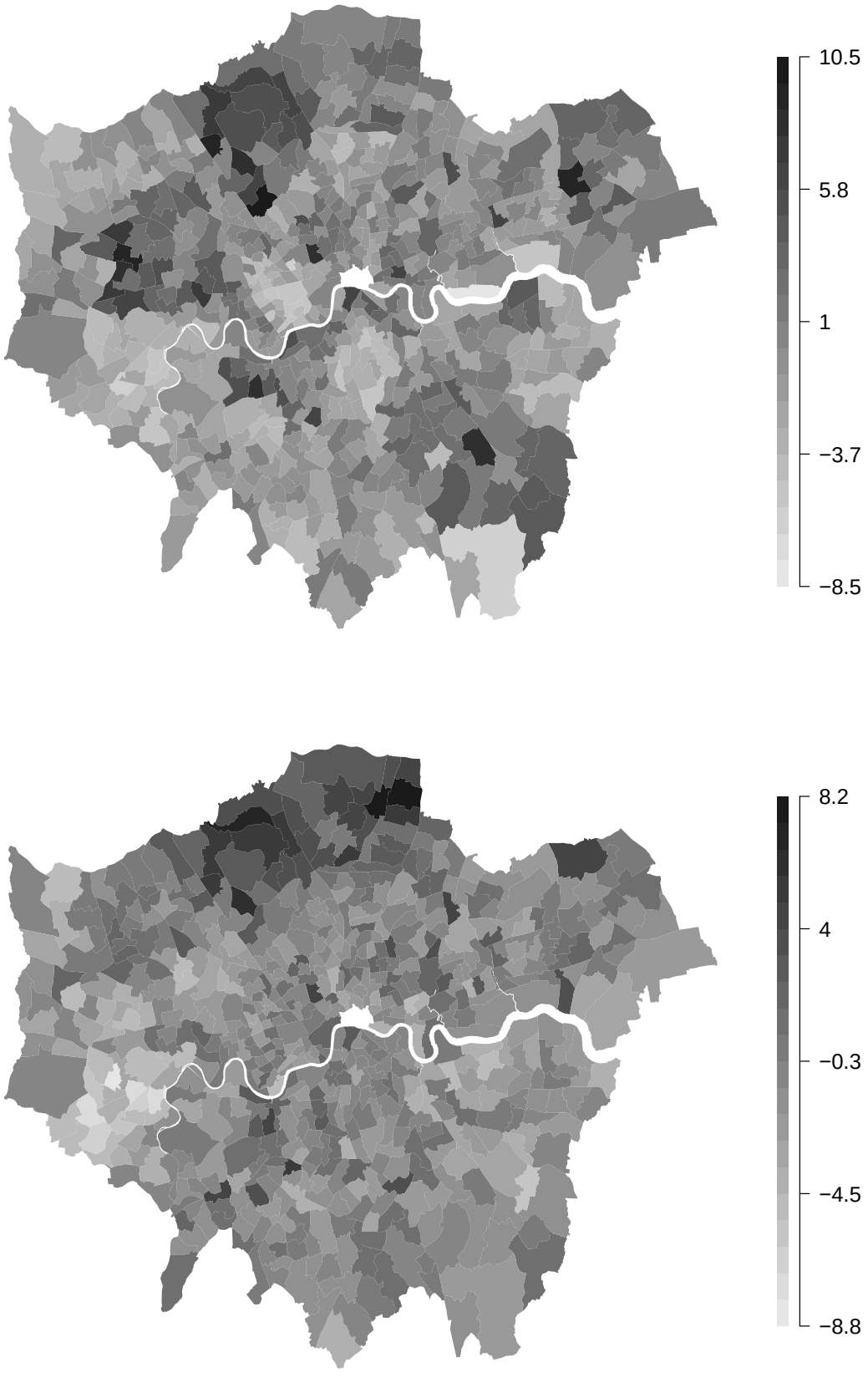

Figure 2: (Top panel) Standardised residuals plot for 2005 after fitting a generalised linear model, where clear spatial autocorrelation is evident. (Bottom panel) Plot showing the difference between the residuals for the years 2004 and 2005, which suggests that the spatiotemporal structure is non-separable.