University of Glasgow

Statistical challenges in estimating the long-term health impact of air pollution

Duncan Lee

Lancaster University - 17th June 2015

Acknowledgements

- This is joint work with Christophe Sarran from the UK Met Office and Sujit Sahu from the University of Southampton.

- The work is funded by the EPSRC grants EP/J017442/1 and EP/J017485/1.

- The main part of the work in this talk will appear in Environmetrics under the title *Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies*.

EPSRC
Engineering and Physical Sciences Research Council

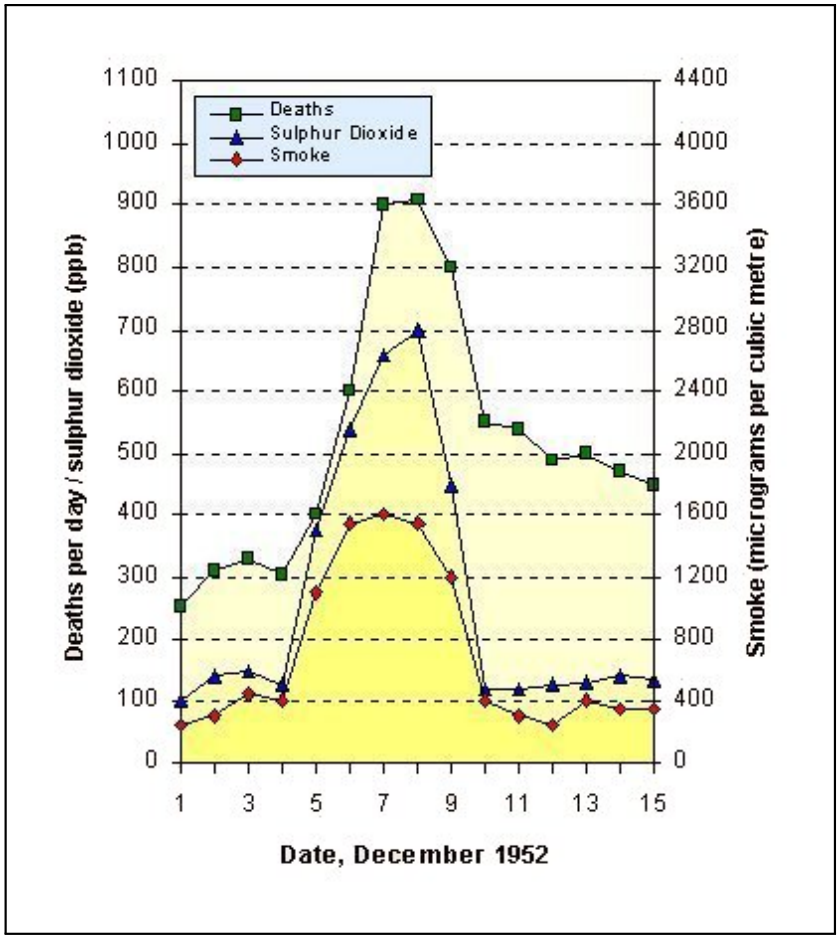
1. Background

- Air pollution has long been known to adversely affect public health, in both the developed and developing world.
- A recent report by the UK government estimates that particulate matter alone reduces life expectancy by 6 months, with a health cost of £19 billion per year.
- Epidemiological studies into the effects of air pollution have been conducted since the 1990s, with one of the first being that conducted by Schwartz and Marcus (1990) in London.
- Since 1990 a large number of studies have been conducted, which collectively have investigated the short-term and long-term health impact of air pollution.

A brief history

The relationship between air pollution exposure and mortality came to prominence during high air pollution episodes in:

- the Meuse Valley, Belgium in 1930;
- Donora, Pennsylvania in 1948; and
- London in December 1952.



High pollution

Government action

- Clean air acts in 1956, 1968 and 1993
  - Prohibited and regulated pollution sources.
  - Set up *'smoke control areas'* in which it was prohibited to emit smoke from buildings or chimneys.
- UK air quality strategy 1997, 2000, 2003 and 2007
  - Set target limits for annual or daily average concentrations for a number of common pollutants.
- Set up the Committee on the Medical Effects of Air Pollution (COMEAP).

Pollution is still a problem today

BBC Sign in News Sport Weather iPlayer TV Ra

NEWS UK

Home World UK England N. Ireland Scotland Wales Business Politics Health Education Sci/En

3 April 2014 Last updated at 22:15

Share

Air pollution: Forecasters hope for cleaner air on Friday

People with lung and heart problems have been advised to avoid strenuous outdoor activity

Pollution is still a problem today

UK faces £300m fine over failure to meet air pollution targets by 2010

European Commission to take legal action against Britain over high levels of dangerous gas

Study designs

Pollution legislation continues to be informed by epidemiological studies investigating both the short-term and long-term health effects of air pollution exposure.

Acute studies investigate the effects resulting from a few days of high exposure.

- e.g. NMMAPS in the USA, Dominici *et al.* (2002) and APHEA in Europe, Katsouyanni *et al.* (2001).

Chronic studies investigate the cumulative effects of exposure over numerous years

- e.g. Dockery *et al.* (1993) in six US cities, and Elliot *et al.* (2007) in the UK.

Chronic studies

There are two main study designs when investigating the effects of long-term exposure to air pollution.

Cohort studies e.g. The Six Cities study by Dockery *et al.* (1993) and the American Cancer study by Pope *et al.* (2002), which relate average air pollution concentrations to the health status of a large pre-defined cohort of people.

Ecological studies e.g. Elliot *et al.*(2007) and Lee *et al.* (2009), which relate average air pollution concentrations in contiguous small areas (such as electoral wards), against yearly numbers of health events from the population living in that area.

Ecological study design

- Small area studies have an ecological design, because the data relate to populations living in a set of n non-overlapping areal units, rather than to individuals.
- Examples of such studies include Jerrett *et al.* (2005), Elliott *et al.* (2007), Lee *et al.* (2009) and Greven *et al.* (2011).
- The health data are denoted by $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{E} = (E_1, \ldots, E_n)$, which are the observed and expected numbers of disease cases in each areal unit over a year.
- The expected numbers of cases are computed using external standardisation, based on age and sex specific disease rates.

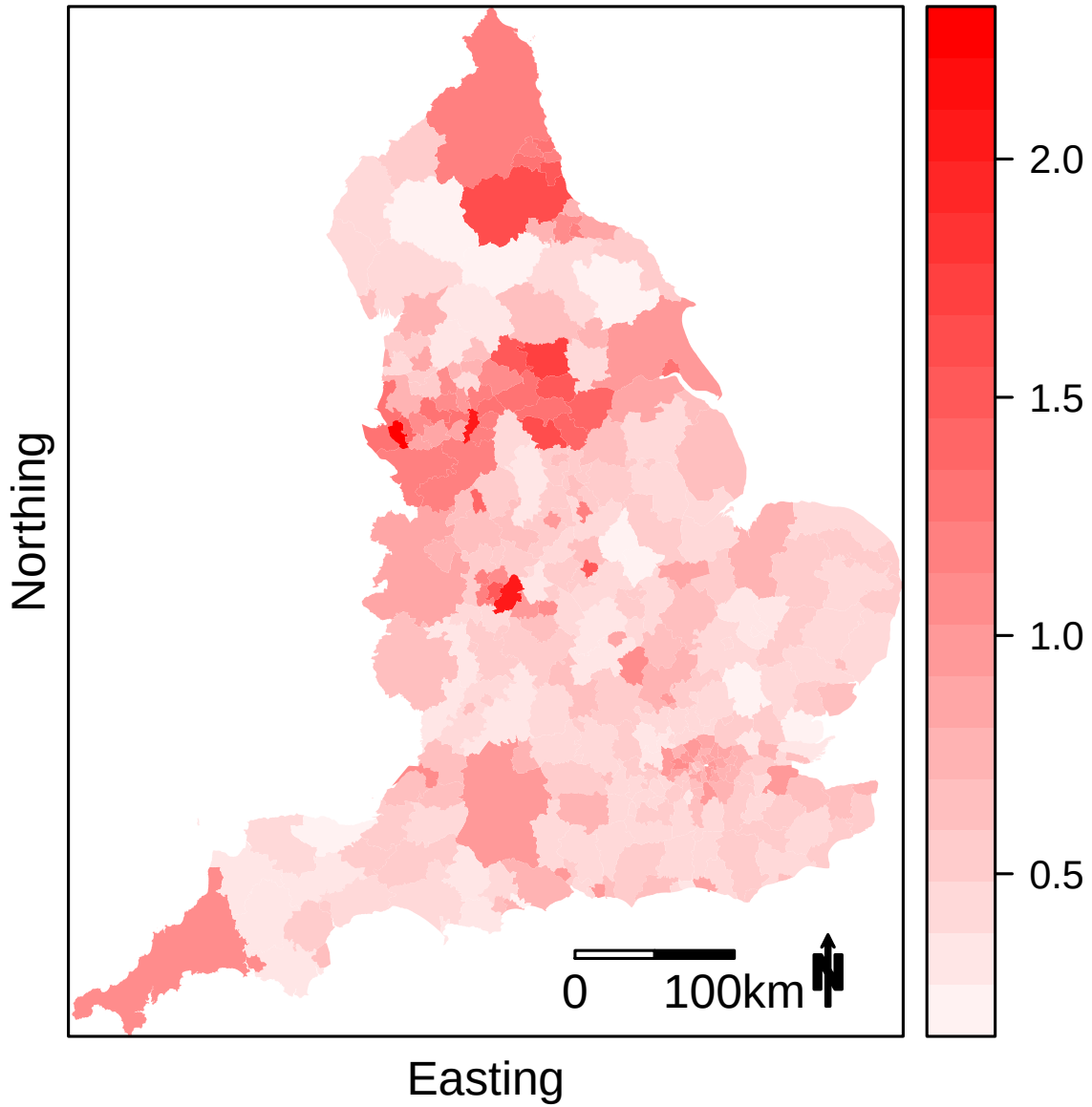
Motivating application

- Our study region is mainland England, which has been split into n=323 Local and Unitary Authorities (LUA).
- The disease data are counts of the number of emergency admissions to hospital due to respiratory disease in each LUA in 2010. The simplest measure of disease risk is the standardised morbidity ratio (SMR), which is given by

$$\text{SMR}_k = Y_k / E_k.$$

- The pollutants we consider are nitrogen dioxide (NO_2) and two measures of particulate matter, PM_{10} and $PM_{2.5}$.

SMR for England

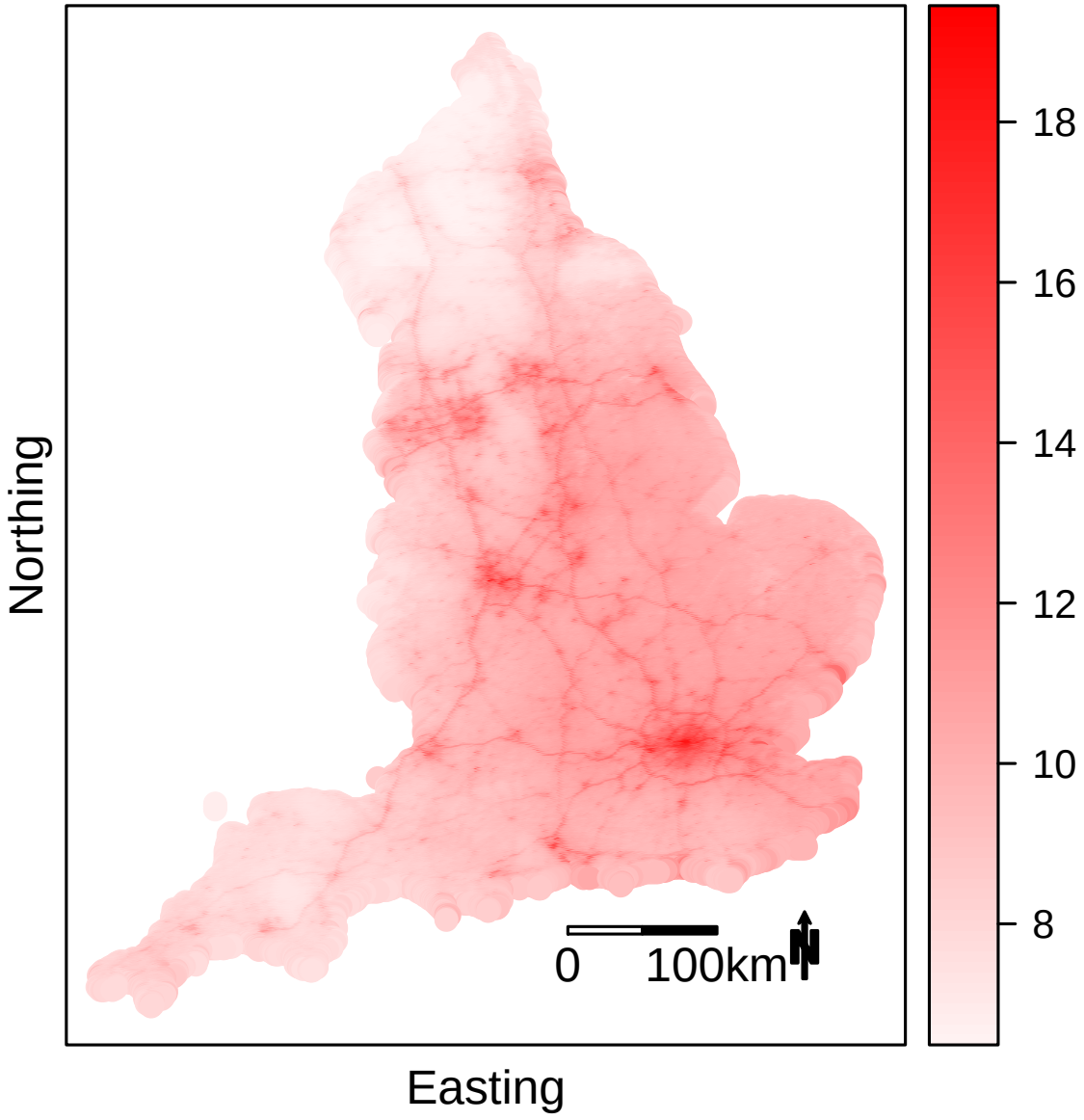


Covariate data

1. Annual average air pollution concentrations estimated from the Community Multi-Scale Air Quality (CMAQ) Model at a 1km squared resolution for all of England. Measured pollution data are not used because the network of monitoring sites is not dense at the small area scale. The vector of concentrations for the kth LUA are denoted by $(x_{k1}, \ldots, x_{kM_k})$

2. Measures of socio-economic deprivation that acts as a proxy for risk inducing behaviours such as smoking, and here we have the proportion of people in receipt of Job Seekers Allowance (JSA) and average property price. The vector of covariate values for the kth LUA are denoted by $\mathbf{v}_k$.

Modelled $PM_{2.5}$ concentrations



Standard model

A Poisson Generalised Linear Mixed Model is given by:

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k R_k), \\ R_k &= \exp(\mathbf{v}_k^{\mathrm{T}} \boldsymbol{\beta} + \theta_k + x_k \beta_x), \end{aligned}$$

- R_k quantifies disease risk in area k, so $R_k = 1.2$ means a 20% increased risk of disease.
- x_k is a representative pollution concentration for the kth LUA such as $x_k = \frac{1}{M_k} \sum_{i=1}^{M_k} x_{ki}$.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are random effects to model residual spatial autocorrelation not captured by the covariates.

A Bayesian approach is adopted, using McMC simulation, which can be implemented using the *CARBayes* software in R.

Modelling spatial correlation

Conditional Autoregressive (CAR) models are typically specified to capture the spatial autocorrelation, and can be written as a set of n univariate full conditional distributions $f(\theta_k|\boldsymbol{\theta}_{-k})$ for $k = 1, \ldots, n$. Here we use the model proposed by Leroux *et al.* (1999) which is:

$$\theta_k|\boldsymbol{\theta}_{-k}, \tau^2, \rho, \mathbf{W} \sim \mathrm{N}\left(\frac{\rho\sum_{i=1}^{n} w_{ki}\theta_i}{\rho\sum_{k=1}^{n} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{k=1}^{n} w_{ki} + 1 - \rho}\right)$$

Here $\mathbf{W} = (w_{ki})$ is a binary $n \times n$ neighbourhood matrix, with $w_{ki} = 1$ if areal units (k, i) share a common border and $w_{ki} = 0$ otherwise.

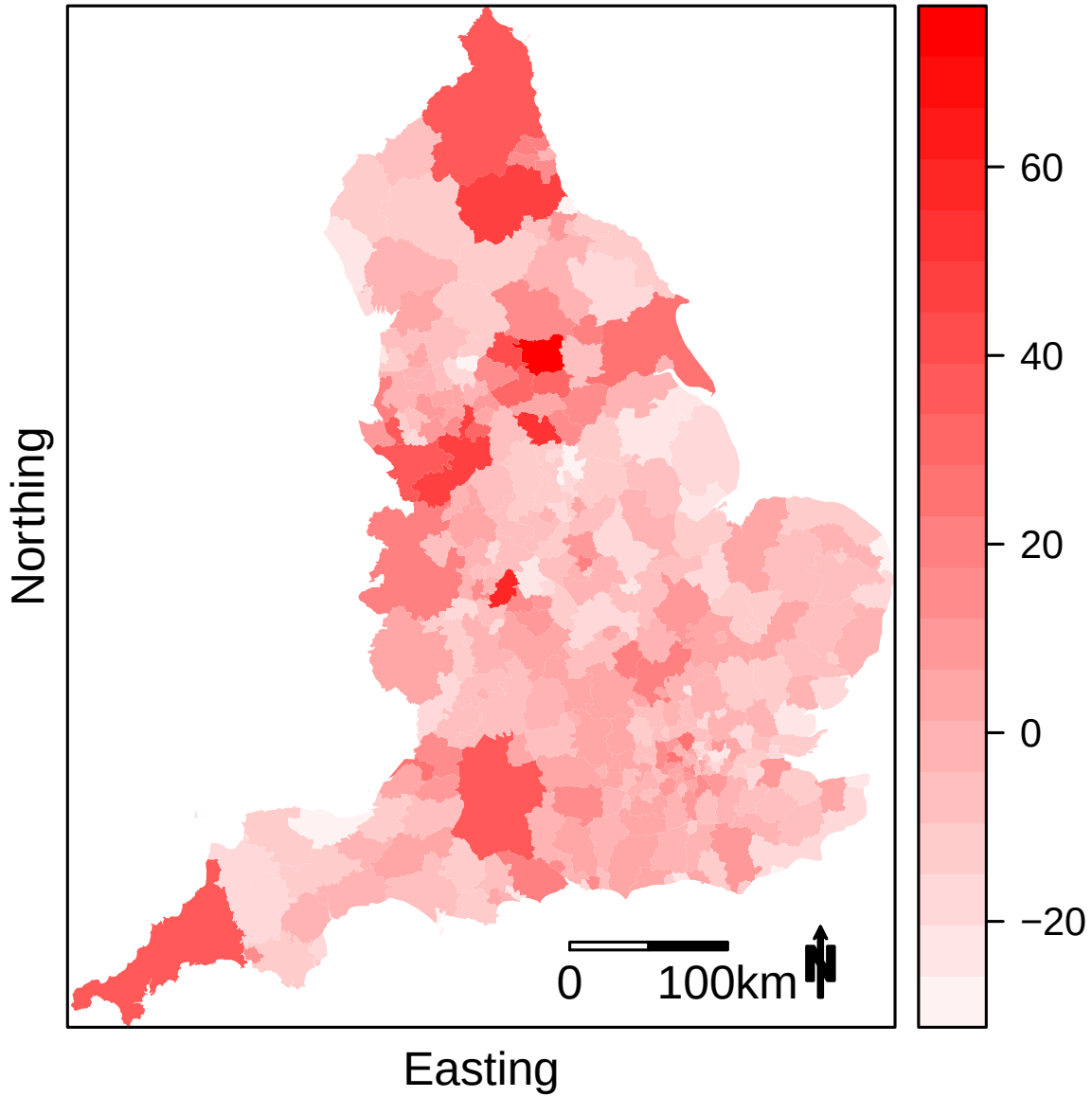
Limitations - random effects

The random effects model forces a single level of spatial smoothness on the random effects surface which is controlled by the spatial dependence parameter ρ. The partial autocorrelation between (θ_k, θ_i) implied by this prior is given by:

$$\text{Corr}[\theta_k, \theta_i | \boldsymbol{\theta}_{-ki}] = \frac{\rho w_{ki}}{\sqrt{(\rho \sum_{j=1}^{n} w_{kj} + 1 - \rho)(\rho \sum_{l=1}^{n} w_{il} + 1 - \rho)}}.$$

So the strength of the dependence between all pairs of adjacent LUA is controlled by ρ. Is this realistic?

Residuals from a covariate only model



Localised spatial autocorrelation

- The previous figure suggests the unmeasured spatial structure captured by the random effects exhibits localised spatial dependence (smoothness), which is strong between some pairs of adjacent LUA but weak between others.
- In fact some pairs of adjacent LUA appear to exhibit large changes (step changes) in the values of this unmeasured structure, potentially driven by an unmeasured confouder that exhibits this pattern.
- Additionally, Reich *et al.* (2006) showed there is the potential for collinearity between any covariate that is spatially smooth such as air pollution and globally smooth CAR random effects. This led to the idea of orthogonal smoothing.

Orthogonal smoothing

Hughes and Haran (2013) proposed replacing the random effects $\boldsymbol{\theta}$ with a regression component $\mathbf{M}\boldsymbol{\delta}$, where the columns of the design matrix $\mathbf{M}$ are eigenvectors from the matrix product $\mathbf{PWP}$ where

$$\mathbf{P} = \mathbf{I}_n - \tilde{\mathbf{V}}(\tilde{\mathbf{V}}^{\mathrm{T}}\tilde{\mathbf{V}})^{-1}\tilde{\mathbf{V}}^{\mathrm{T}},$$

where $\tilde{\mathbf{V}} = (\mathbf{x}, \mathbf{V})$ is the complete covariate matrix. The columns of $\mathbf{M}$ correspond to all possible mutually distinct patterns of spatial autocorrelation orthogonal to the covariates.

Limitations - Pollution data

The use of modelled pollution data relating to 1km grid squares has the following limitations:

1. They are assumed to be true known measurements where as they are in fact subject to error and potential biases.
2. The uncertainty due to them being estimates rather than known data should be accounted for in the model.
3. They are estimated on a regular grid and have a different spatial support compared with the irregularly shaped LUA, which is known as the *change of support problem*. Therefore there is spatial variation within an LUA in the pollution concentrations.

2. Methodology - motivation

This talk:

1. Proposes an alternate likelihood model based on aggregating an idealised smaller group level model (where each small group has no within group variation in the pollution concentrations) to the LUA scale which accounts for the spatial variation in the pollution concentrations within a LUA.

2. Proposes an extension to the random effects model that allows for localised spatial autocorrelation and step changes in the random effects surface.

A smaller group-level model

Suppose there are M_k modelled pollution concentrations $(x_{k1}, \ldots, x_{kM_k})$ in the kth LUA, and that each member of the population in the kth LUA is exposed to one of these M_k concentrations. Then let $(Y_{k1}, \ldots, Y_{kM_k})$ denote the (unobserved) disease counts relative to the population exposed to each pollution concentration. Then following Wakefield and Shaddick (2006) an appropriate model is

$$\begin{aligned} Y_{ki} &\sim \text{Poisson}(E_{ki}R_{ki}), \\ R_{ki} &= \exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\beta} + \theta_k + x_{ki}\beta_I), \end{aligned}$$

Here $E_k = \sum_{i=1}^{M_k} E_{ki}$. In this model there is no within group (i.e. within i) variation in pollution concentrations, as M_k could be in theory be arbitrarily large.

Aggregating this model

However, $(Y_{k1}, \ldots, Y_{kM_k})$ are unknown, and only $Y_k = \sum_{i=1}^{M_k} Y_{ki}$ is observed. Therefore assuming conditional independence between the M_k groups and using the additive property of Poisson distributions yields an aggregated model of the form:

$$\begin{aligned} Y_k | E_k, R_k &\sim \text{Poisson}(E_k R_k) \\ R_k &= \exp(\mathbf{v}_k^{\mathrm{T}} \boldsymbol{\beta} + \theta_k) \sum_{i=1}^{M_k} E_{ki}^* \exp(x_{ki} \beta_I), \end{aligned}$$

where $E_{ki}^* = E_{ki}/E_k$ and $\sum_{i=1}^{M_k} E_{ki}^* = 1$.

Aggregate and Ecological models

The key difference is in the pollution components:

- Naive ecological model - $\exp\left(\frac{1}{M_k}\sum_{i=1}^{M_k} x_{ki}\beta_x\right)$
- Aggregate model - $\sum_{i=1}^{M_k} E^*_{ki}\exp(x_{ki}\beta_I)$

So aside from differential weighting due to $(E^*_{k1},\ldots,E^*_{kM_k})$ the difference is that you average the exponentiated risks and not evaluate the risk at the average pollution concentration.

Therefore $\beta_x \neq \beta_I$ in general. Using the first of these assuming it gives you an individual level effect is known as *ecological bias*.

How much bias do you get?

- If there is no within LUA variation, that is $x_{ki} = x_k$ then there is no bias and $\beta_x = \beta_I$.
- If the mean and standard deviation of $(x_{k1}, \ldots, x_{kM_k})$ are independent over the set of n LUA (i.e. over k) then there is no bias and $\beta_x = \beta_I$.
- If $x_{k1}, \ldots, x_{kM_k} \sim \mathrm{N}(x_k, \sigma_k^2)$ where $\hat{\sigma}_k^2 = a + bx_k$, then it can be shown that

$$\beta_x = \beta_I + 0.5b\beta_I^2$$

so that the ecological model produces biased effect estimates.

Localised residual spatial smoothness

We account for potentially localised spatial smoothness (dependence) in the random effects by augmenting the mean model to

$$\begin{aligned} Y_k|E_k, R_k &\sim \text{Poisson}(E_k R_k) \\ R_k &= \exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\beta} + \theta_k + \lambda_{Z_k}) \sum_{i=1}^{M_k} E_{ki}^* \exp(x_{ki}\beta_I), \end{aligned}$$

where:

- θ_k comes from the same CAR model as before and captures globally smooth patterns.
- λ_{Z_k} is a piecewise constant intercept surface and captures step changes between adjacent areal units.

Piecewise constant intercept surface

The piecewise constant intercept surface $(\lambda_{Z_1}, \ldots, \lambda_{Z_n})$ comprises at most G distinct values $\boldsymbol{\lambda} = (\lambda_1 < \lambda_2 \ldots < \lambda_G)$ which are ordered to mitigate against label switching via the prior

$$\lambda_i \sim \text{Uniform}(\lambda_{i-1}, \lambda_{i+1})$$

where $\lambda_0 = -\infty$ and $\lambda_{G+1} = \infty$. Here $Z_k \in \{1, \ldots, G\}$ controls the allocation of the kth LUA to one of the G different intercept terms.

Choosing G

- Here G is the maximum number of different intercept terms.
- We propose a shrinkage based approach that is a discrete random variable analogue of ridge regression, which fixes G to be larger than necessary and uses a penalty prior to encourage each Z_k towards the middle class.
- This middle class is $G^* = (G+1)/2$ if G is odd and $G^* = G/2$ if G is even, and this penalty ensures that Z_k is only estimated to be in one of the extreme classes if supported by the data.
- Thus $\mathbf{Z} = (Z_1, \ldots, Z_n)$ are allowed to take values in the set $\{1, \ldots, G\}$ but are not forced to completely cover the set.

Ridge regression style penalty

The shrinkage prior we propose for $\mathbf{Z}$ is a discrete analogue of ridge regression and is given by:

$$f(\mathbf{Z}) = \prod_{k=1}^{n} f(Z_k)$$

where

$$\begin{aligned} f(Z_k) &= \frac{\exp(-\delta(Z_k - G^*)^2)}{\sum_{r=1}^{G} \exp(-\delta(r - G^*)^2)}, \\ \delta &\sim \text{Uniform}(0, M = 100), \end{aligned}$$

So δ controls the amount of shrinkage. This model can also be implemented in the *CARBayes* package in R.

3. Simulation study

This simulation study has two main goals:

1. Assess the impact of allowing for and ignoring within LUA variation in the pollution concentrations.
2. Quantify the performance of a number of different approaches to allowing for residual spatial autocorrelation.

These goals are addressed by two different studies, and in each case all results quoted are based on five hundred data sets.

Study 1 - variation in pollution within an LUA

- Disease data are generated for the $n = 323$ LUA that make up mainland England from a Poisson likelihood, and the mean pollution concentrations in each LUA are based on those observed from the CMAQ model (for PM_{10}).
- The number of pollution concentrations observed within each LUA are taken from the real data, and range between 11 and 4889 with a median value of 215.
- The distribution of within LUA concentrations is generated from a Gaussian distribution.

Scenarios considered

Three different aspects of the data generation mechanism are changed to observe their impact on model performance.

1. The standard deviation of the concentrations within each LUA is either 1 or 10.
2. The relationship between the mean and variance of the Gaussian exposure distribution within each LUA is either linear or independent.
3. The true health impact of pollution β_I is either of standard size or large (relative risks of 1.05 or 1.5 for a $2\mu gm^{-3}$ increase).

Results

Risk	Pollution	Bias		RMSE		Coverage	
		E	A	E	A	E	A
1.05	SD=1, Indep	-0.31	-0.29	6.06	6.08	95.3	95.5
1.05	SD=1, Linear	-0.16	-0.22	5.85	5.83	96.2	96.4
1.05	SD=10, Indep	0.11	0.10	5.69	5.67	95.7	95.5
1.05	SD=10, Linear	0.45	-0.17	5.81	5.74	94.8	95.6
1.5	SD=1, Indep	-0.22	-0.34	14.06	13.92	95.4	94.4
1.5	SD=1, Linear	17.96	0.47	23.00	10.56	73.4	95.4
1.5	SD=10, Indep	-1.17	-0.64	7.87	5.35	92.6	94.3
1.5	SD=10, Linear	20.72	-0.09	23.57	2.37	44.6	95.9

Message - For realistic sized pollution-health relationships the ecological model (E) performs as well as the more theoretically appropriate aggregate (A) model.

Study 2 - Models for spatial confounding

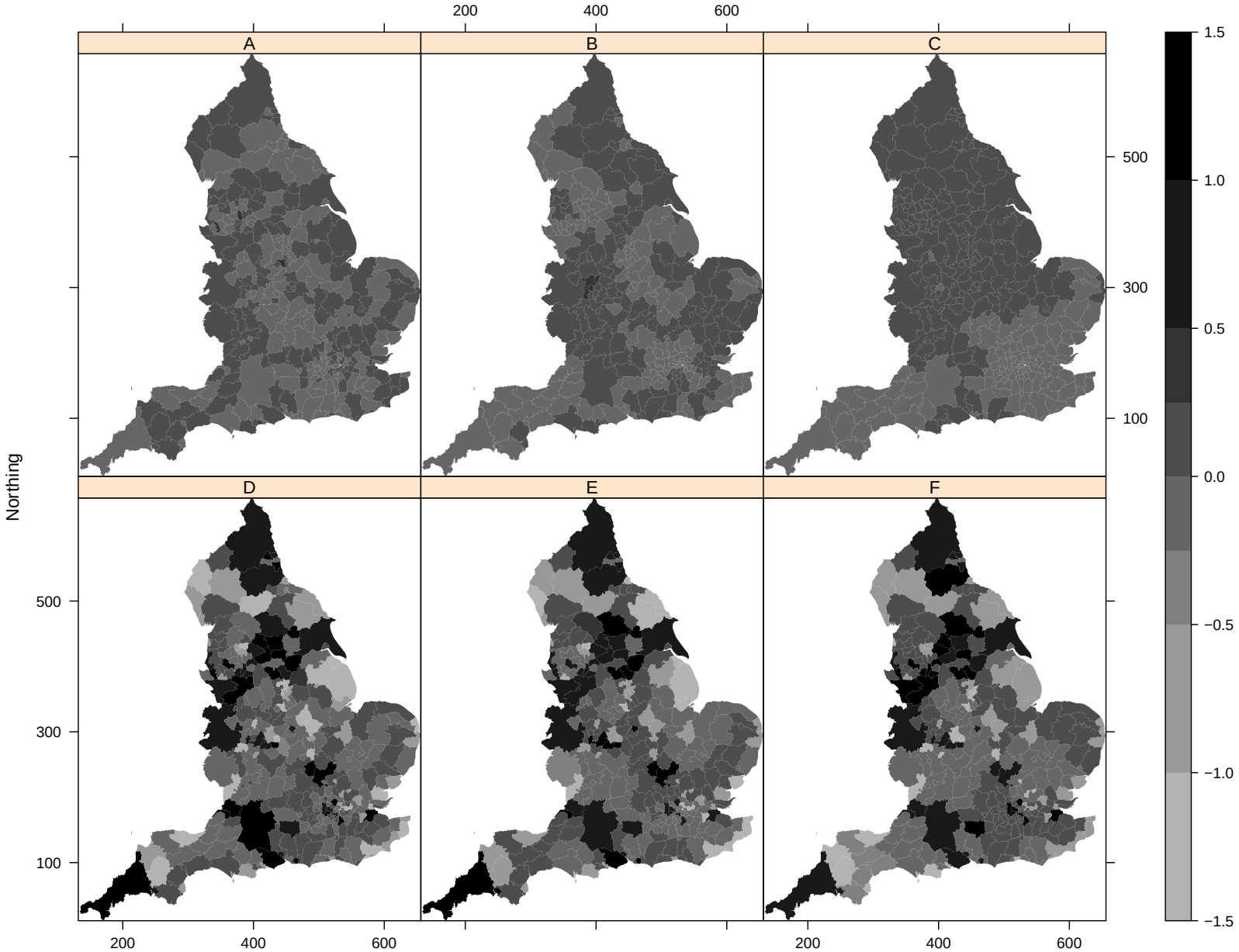
We compared the following models for allowing for residual spatial autocorrelation:

- A simple generalised linear model that ignores this autocorrelation.
- A standard CAR model for globally smooth autocorrelation.
- The CAR model with a piecewise constant intercept term proposed here for localised autocorrelation.
- The orthogonal spatial autocorrelation model proposed by Hughes and Haran (2013).

Six spatial confounding scenarios

- A - The spatial confounding had no spatial structure.
- B - The spatial confounding was spatially correlated but less smooth than the pollution covariate.
- C - The spatial confounding was spatially correlated and as smooth as the pollution covariate.
- D - The spatial confounding had no spatial structure and included step changes in values.
- E - The spatial confounding was spatially correlated but less smooth than the pollution covariate and included step changes in values.
- F - The spatial confounding was spatially correlated and as smooth as the pollution covariate and included step changes in values.

Example confounding surfaces



Results 1 - RMSE

Scenario	Model			
	Model-GLM	Model-CAR	Model-Local	Model-HH
A	6.37	6.43	6.45	6.43
B	17.71	15.23	15.28	17.64
C	26.70	18.73	18.91	26.44
D	53.49	45.42	7.60	47.23
E	59.71	49.87	16.88	53.18
F	60.71	50.32	22.64	54.20

Results 2 - Coverage

Scenario	Model			
	Model-GLM	**Model-CAR**	**Model-Local**	**Model-HH**
A	94.8	96.6	94.8	84.8
B	52.8	85.4	85.6	42.8
C	35.0	77.4	77.0	25.8
D	67.8	91.2	94.4	16.4
E	62.6	89.0	76.2	10.6
F	63.6	87.4	63.2	17.0

4. England study

- We now apply all models to the England respiratory hospitalisations data in 2010, where the pollutants considered are concentrations of NO_2, $PM_{2.5}$ and PM_{10} from the CMAQ model.
- Socio-economic deprivation was controlled for using two the percentage of working age people in receipt of job seekers allowance, and median property price in each LUA.
- The residuals from a covariate only model exhibited substantial spatial autocorrelation, with a Moran's I statistic of 0.282 (p-value 0.00001).
- All results are based on 100,000 McMC samples obtained from 5 Markov chains following an appropriate burn-in period.

Results

Model	Pollutant		
	NO_2	$PM_{2.5}$	PM_{10}
GLM	1.085 (1.052, 1.118)	1.032 (1.005, 1.060)	1.008 (0.989, 1.027)
CAR	1.094 (1.055, 1.133)	1.055 (1.022, 1.094)	1.037 (1.014, 1.062)
Local	1.089 (1.071, 1.104)	1.032 (1.017, 1.047)	1.013 (1.003, 1.023)
Local-Agg	1.086 (1.072, 1.100)	1.035 (1.021, 1.054)	1.010 (1.001, 1.019)
HH	1.088 (1.086, 1.091)	1.046 (1.044, 1.047)	1.019 (1.017, 1.020)

The relative risks are for the following increases in the yearly average concentrations: NO_2 ($5.0\mu gm^{-3}$), $PM_{2.5}$ ($1\mu gm^{-3}$), PM_{10} ($1\mu gm^{-3}$).

Here we specified $G = 5$. We tried other values of G and obtained almost identical results.

5. Conclusions

1. We have proposed an integrated modelling framework for estimating the long-term effects of air pollution on human health accounting for localised spatial autocorrelation and spatial misalignment between the exposure and the response.

2. Inappropriate control for residual (i.e. after the effects of known covariates have been removed) spatial autocorrelation in the disease data can result in incorrect fixed effects estimation, in terms of both point estimation and uncertainty quantification.

3. However, these studies where the effect size is small, within area variation in the exposure can be solved by simply averaging the exposures within an area.

- Spatial ecological studies are inexpensive and quick to implement, due to the now routine availability of the required data. Thus, while they cannot provide individual-level evidence on cause and effect, they provide important corroborating evidence of health effects.
- Air pollution still appears to exhibit substantial health risks, as the uncertainty intervals from all models and pollutants (except 1) show evidence of a relationship.
- In future we aim to investigate fusion modelling to combine both modelled and monitored pollution data to produce improved areal level pollution estimates.

Sensitivity to G - RMSE

Scenario	Value of G			
	$G = 3$	$G = 4$	$G = 5$	$G = 7$
A	6.41	6.37	6.45	6.45
B	15.43	15.51	15.28	15.48
C	19.22	19.90	18.91	19.17
D	7.34	8.00	7.60	7.86
E	16.45	17.50	16.88	16.96
F	21.55	22.35	22.64	23.37

Sensitivity to G - Coverage

Scenario	**Value of G**			
	$G = 3$	$G = 4$	$G = 5$	$G = 7$
A	95.4	95.4	94.8	96.0
B	84.4	81.6	85.6	83.8
C	73.6	70.0	77.0	73.2
D	95.4	92.8	94.4	91.0
E	79.6	77.6	76.2	74.0
F	67.2	63.6	63.2	58.8