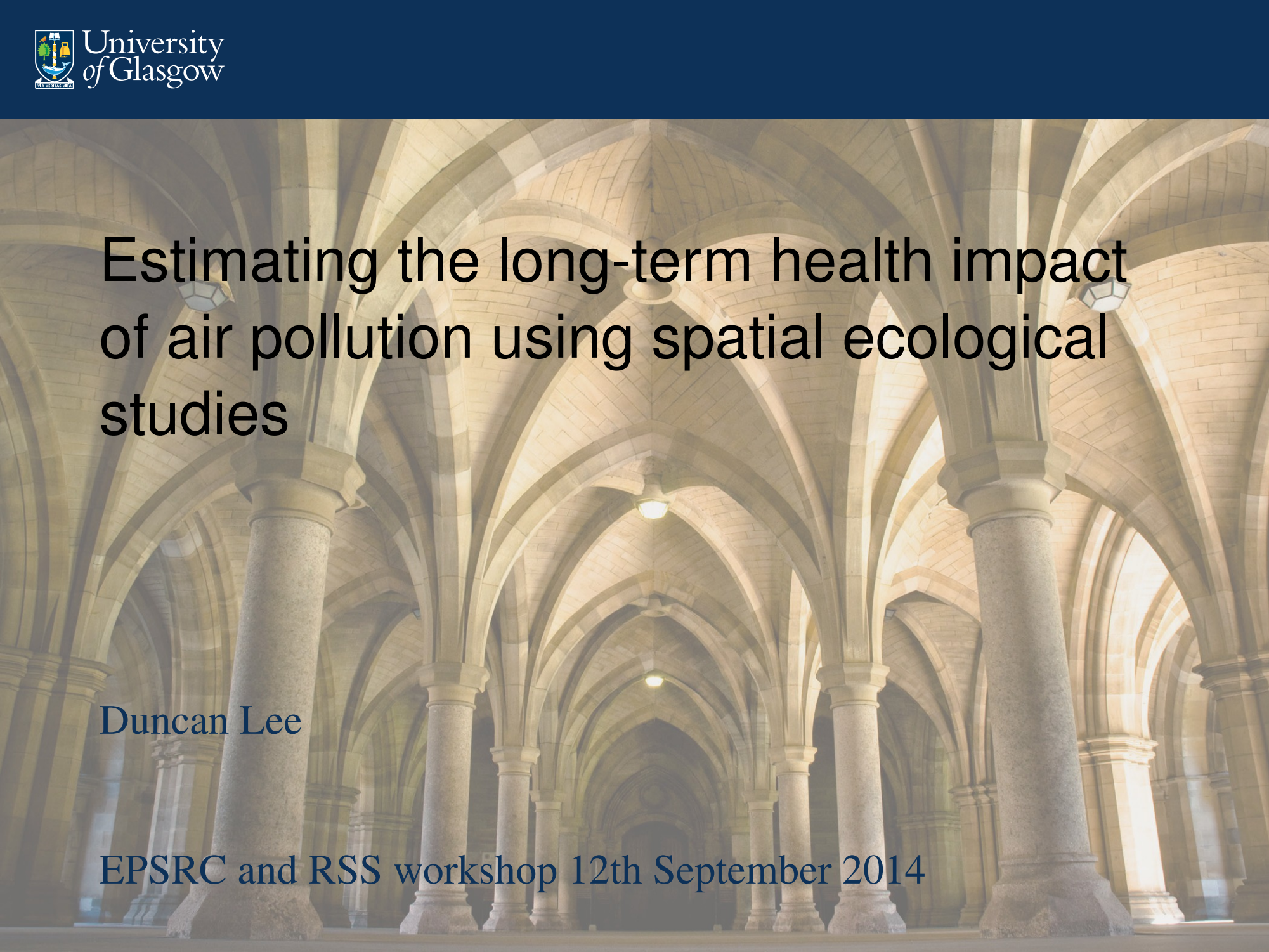


University of Glasgow

Estimating the long-term health impact of air pollution using spatial ecological studies

Duncan Lee

EPSRC and RSS workshop 12th September 2014

Acknowledgements

- This is joint work with Alastair Rushworth and Richard Mitchell from the University of Glasgow and Sujit Sahu and Sabyasachi Mukhopadhyay from the University of Southampton, and Paul Agnew, Christophe Sarran, Fiona O'Connor and Rachel McInnes from the Met Office.

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

EPSRC
Engineering and Physical Sciences Research Council

1. Introduction

- 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.

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 effects of cumulative (long-term) exposure over months and 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.

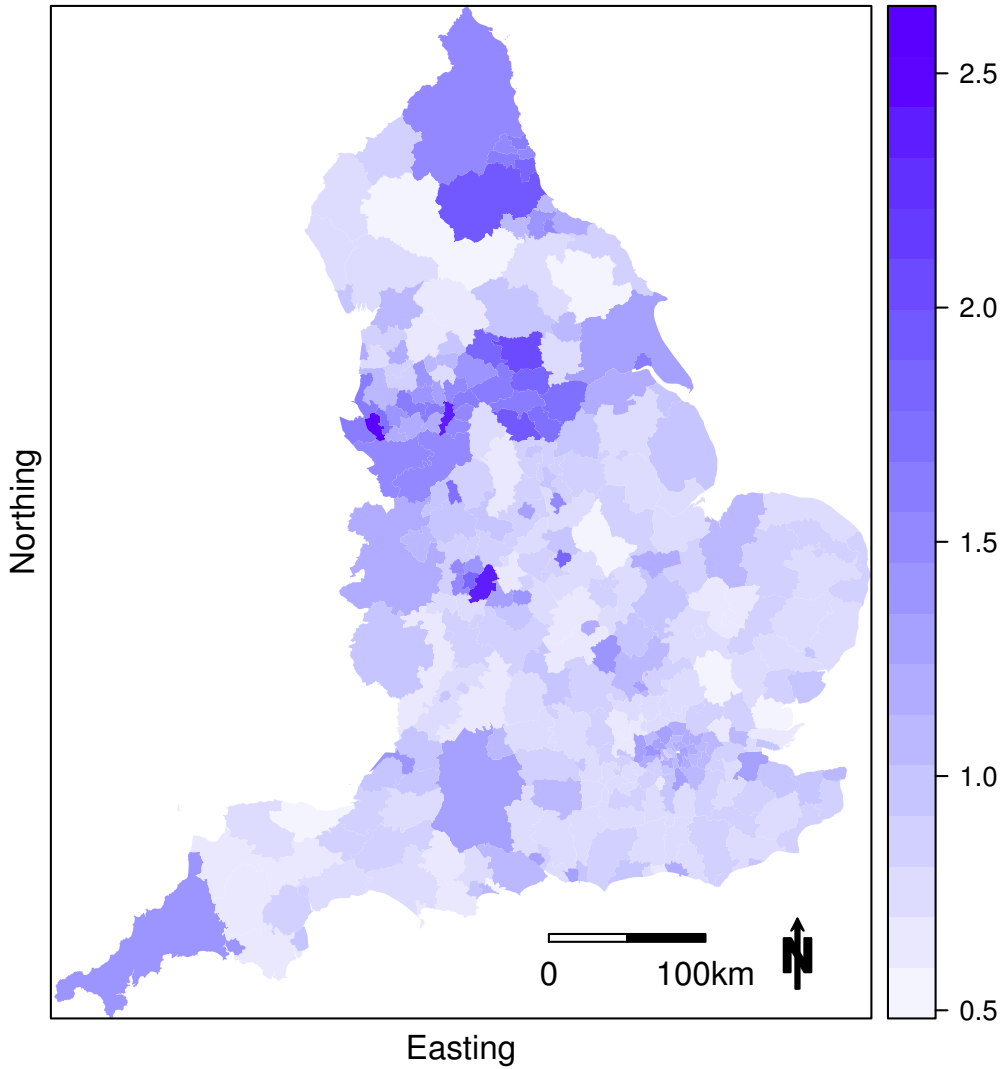
Spatial ecological studies e.g. Elliot et al (2007) and Lee et al (2009), which relate yearly average air pollution concentrations in a set of contiguous areas (such as electoral wards), against yearly numbers of health events from the population living in that area.

2. Spatial ecological studies

- Spatial ecological studies 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.

Example - 323 local authorities in England

Respiratory hospitalisation risk in 2010 - $\text{SIR}_k = Y_k/E_k$.



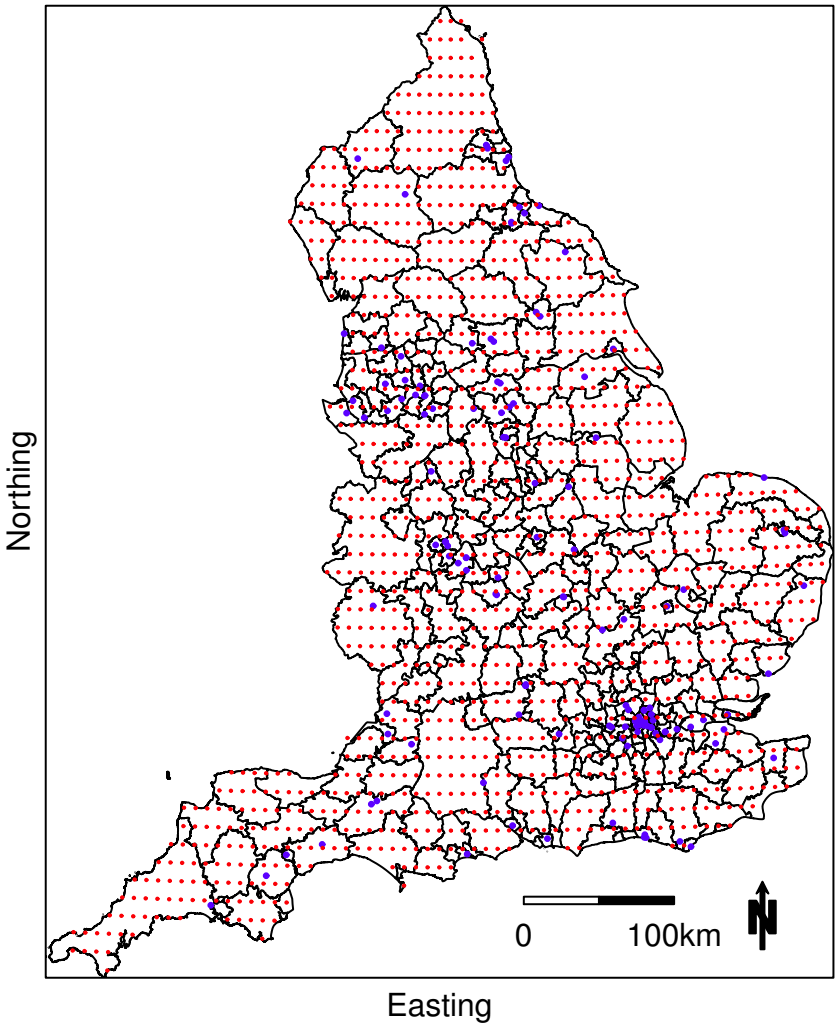
Pollution data

Pollution data comes from two distinct sources.

- Observed data from the AURN network at single fixed geographical points located throughout the study region. These data are known to be measured with little error but do not provide complete spatial coverage of England.
- Estimated background concentrations over 12 Kilometre grid cells from the Air Quality in the Unified Model (AQUM) run by the Met Office. These model estimates provide complete spatial coverage of the study region, but are known to contain biases and are less accurate then the monitoring data.

Both data sets can be available at an hourly resolution.

$PM_{2.5}$ in local authorities in England



How best to combine these two sources of data to estimate annual average pollution levels in each local authority?

A common statistical model for these data

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

where

- R_k quantifies disease risk in area k, so $R_k = 1.2$ means a 20% increased risk of disease.
- $\mathbf{z}_k$ is a vector of other covariates influencing health risk in the kth areal unit, such as measures of deprivation.
- x_k is the estimated annual average pollution concentrations in the kth areal unit, and β_x is the log-risk of air pollution on health.
- $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are random effects to model residual spatial autocorrelation not captured by the covariates.

Residual spatial autocorrelation

- Spatial autocorrelation occurs when observations geographically close are more similar than those further apart.
- The residuals from fitting a regression model to the health data with just $(\mathbf{z}_k^{\mathrm{T}}, x_k)$ are typically spatially autocorrelated, requiring the random effect ϕ_k to account for it.
- The autocorrelation is typically caused by unmeasured confounding, namely the presence of important spatially smooth risk factors that have been omitted from the regression model.
- Ignoring this autocorrelation is known to bias β_x.

Modelling spatial correlation

Conditional Autoregressive (CAR, *Besag et al. (1991)*) models are typically specified to capture the spatial autocorrelation in ϕ, and can be written as a set of n univariate full conditional distributions $f(\phi_k|\boldsymbol{\phi}_{-k})$ for $k = 1, \ldots, n$ as:

$$\phi_k|\boldsymbol{\phi}_{-k}, \tau^2, W \sim \mathrm{N}\left(\frac{\sum_{i=1}^{n} w_{ki}\phi_i}{\sum_{i=1}^{n} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{n} w_{ki}}\right).$$

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

Interpreting β_x

The relative risk for a $1\mu gm^{-3}$ increase in pollution concentrations measures the proportional increase in health risk from increasing pollution by $1\mu gm^{-3}$, and is calculated as

$$\text{RR}(\beta_x, 1) = \frac{E_k \exp(\mathbf{z}_k^{\text{T}}\boldsymbol{\beta}_z + (x_k + 1)\beta_x + \phi_k)}{E_k \exp(\mathbf{z}_k^{\text{T}}\boldsymbol{\beta}_z + x_k\beta_x + \phi_k)} = \exp(1 \times \beta_x).$$

Hence a relative risk of 1.03 means a 3% increase in disease risk when the pollution level increases by $1\mu gm^{-3}$.

Limitations - spatial autocorrelation

The CAR prior forces the random effects $(\phi_1, \ldots, \phi_n)$ to be globally spatially smooth everywhere. This causes two problems:

- Collinearity with covariates that are also globally smooth such as air pollution, as was illustrated by Clayton *et al.* (1993) and Hughes and Haran (2013).
- The spatial autocorrelation in the data remaining after accounting for the covariates is unlikely to be globally spatially smooth, because the disease data (e.g. the SIR) are not globally smooth so the residuals after removing covariate effects are also unlikely to be.

Both these may result in biased estimates of β_x.

Limitations - Pollution data

The pollution concentration for area k, x_k is typically taken to be the average value from the set of modelled concentrations lying in area k. However, this has the following limitations:

1. The modelled AQUM data are known to contain biases, so the estimate of the average pollution concentration in each unit may be biased. In contrast, the AURN monitoring data are likely to measure with little error, but do not cover the entire study region.

2. The concentration for area k, x_k is assumed to be a true known measurement when estimating its health effect. However, the true average is unknown and x_k is an estimate and is subject to error and uncertainty.

Ignoring these two issues may result in biased estimates of β_x.

3. Modelling

This project aims to address all of these statistical shortcomings, by:

1. Developing a spatial regression model linking the monitoring and modelled pollution data, thus allowing average pollution concentrations x_k to be predicted for each areal unit with an associated measure of uncertainty.

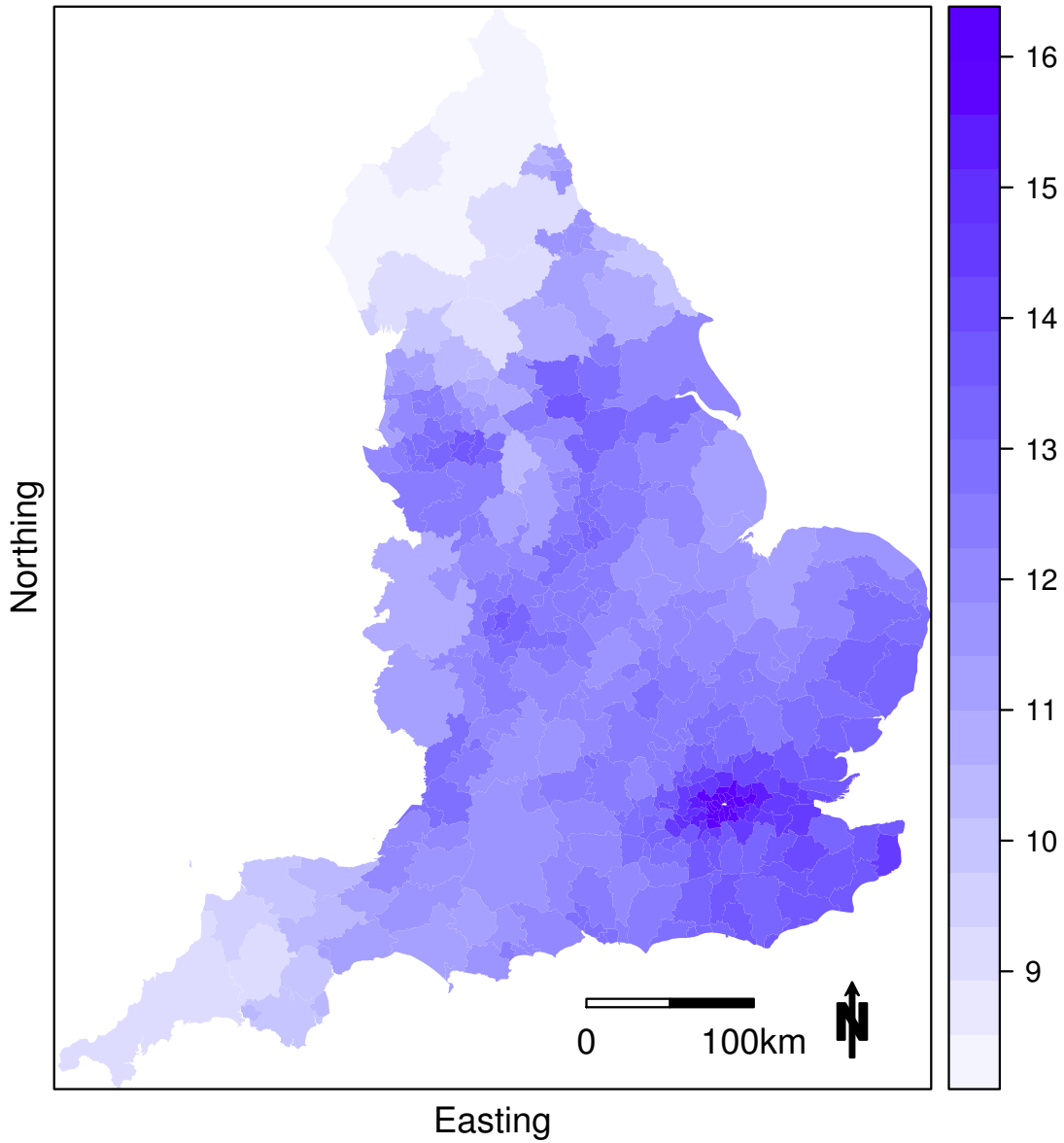
2. Extending the health model so that it allows for the uncertainty in x_k.

3. Extending the health model so that it models the spatial autocorrelation (a prior for $(\phi_1, \ldots, \phi_n)$) more flexibly than the global CAR model, allowing for local rather than global patterns of spatial structure.

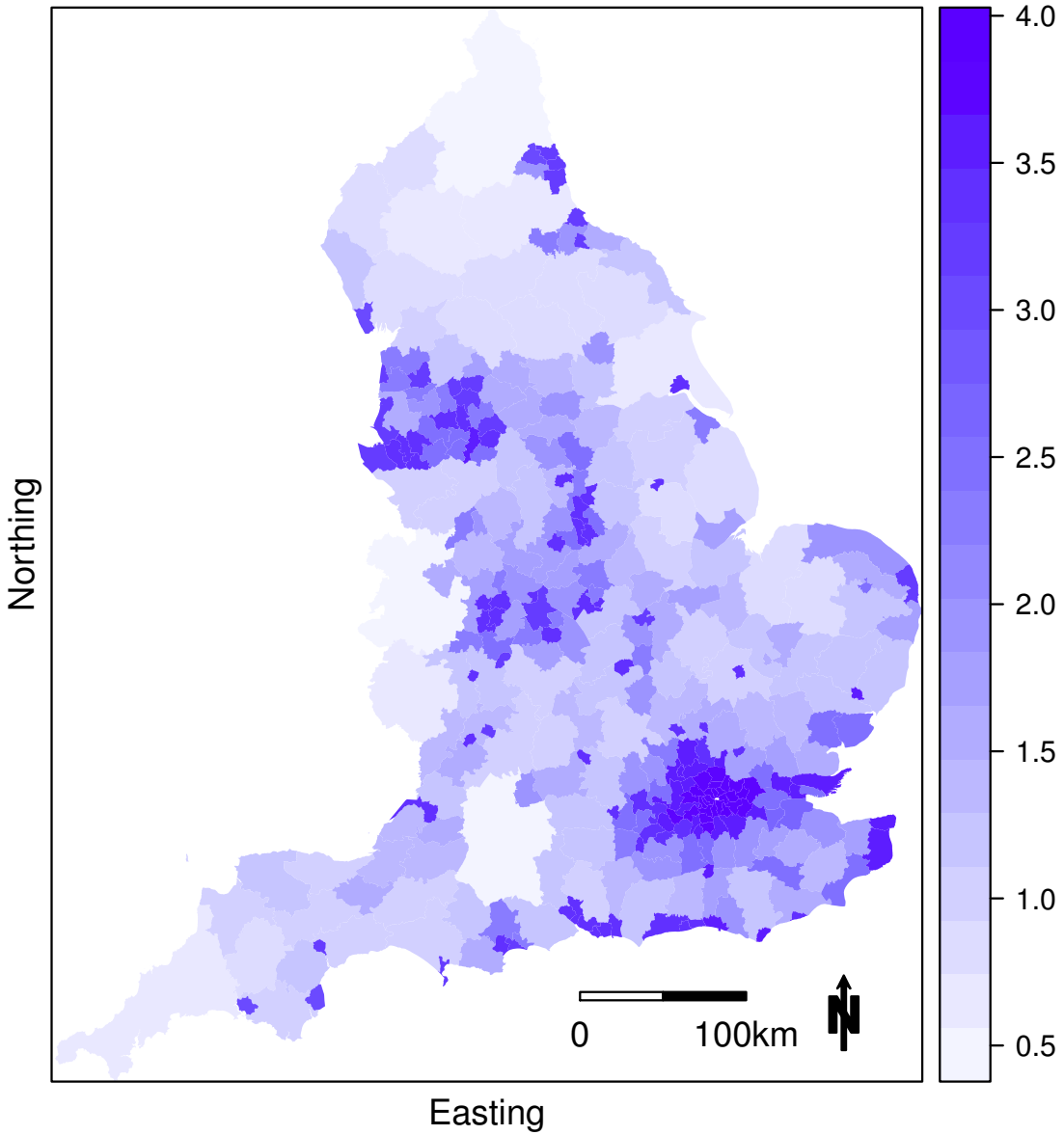
3.1. A model for the pollution data

- A spatial regression model was developed, which has the monitoring data as the response and the modelled data as a covariate.
- This model allows for spatial autocorrelation in the pollution data, which is used to predict the monitoring values at unmeasured locations.
- Predictions were made at each 12 kilometre grid square (to align with the modelled data), and the predictions were averaged over each local authority to give the predicted average concentration of pollution in each area.
- A Bayesian modelling approach was taken, allowing a set of N predictions to be made for each local authority average concentration x_k. That is, x_k is predicted by the set $(x_k^{(1)}, \ldots, x_k^{(N)})$, which quantifies its uncertainty.

Average $PM_{2.5}$ concentrations in 2010



Standard deviation in $PM_{2.5}$ concentrations



3.2. Allowing for uncertainty in x_k

The standard regression model assumes x_k is a known value measured without error, which is unrealistic in this setting. Two main statistical approaches have been proposed for correcting this.

1. **Measurement error models** - This approach assumes the predicted values $(x_k^{(1)}, \ldots, x_k^{(N)})$ are error prone measurements of the single true but unknown concentration x_k.

2. **Ecological bias models** - This approach assumes the predicted values $(x_k^{(1)}, \ldots, x_k^{(N)})$ represent the range of concentrations in area k, allowing for within area variation in the concentrations.

Ecological bias models

- As the health data Y_k summarises disease burden over a population, the standard model naively assumes the pollution concentration x_k is the same for each individual within area k.
- Work by *Wakefield and Salway (2001)* and others has shown that when there is within area variation in pollution concentrations, the estimated β_x from the ecological level model does not equal the individual level association β_I.
- They show this by taking a hypothesised individual level models and aggregating it to the population level, and showing that is has a different mathematical form to the standard ecological model.

$$\mathbb{E}[\exp(x_k^{(i)}\beta_I)] \neq \exp(\mathbb{E}[x_k^{(i)}]\beta_x)$$

A model for ecological bias

Richardson et al (1987) show that one solution to overcome this problem is to change the model for R_k from

$$R_k = \exp(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta}_z + \mu_k\beta_x + \phi_k),$$

to

$$R_k = \exp(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta}_z + \mu_k\beta_x + \sigma_k^2\beta_x^2/2 + \phi_k),$$

thus explicitly incorporating the variation in x_k. Here (μ_k, σ_k^2) respectively denote the mean and variance of $(x_k^{(1)}, \ldots, x_k^{(N)})$.

This change is based on assuming the within area exposure distribution is Gaussian, and comes direct from the moment generating function $\mathbb{E}[\exp(x_k^{(i)}\beta_I)]$.

The impact of ecological bias

The difference between the naive ecological model and the corrected model is the term $\sigma_k^2 \beta_x^2 / 2$, which is likely to be small if:

- β_x is small.
- σ_k^2 is small.

The former is likely to be true and the latter may or may not be, and preliminary analyses show that the impact of ecological bias may be small for these studies.

We are currently investigating the exact extent of this problem.

3.3. The problem with CAR models for ϕ

Recall that in the CAR spatial autocorrelation model for $(\phi_1, \ldots, \phi_n)$, spatial autocorrelation is induced by the binary neighbourhood matrix W, where if areas (k, i) share a common border (are spatially close) then $w_{ki} = 1$. This induces spatial autocorrelation between (ϕ_k, ϕ_i) as can be seen from their partial correlations:

$$\text{Corr}[\phi_k, \phi_i | \boldsymbol{\phi}_{-ki}] = \frac{w_{ki}}{\sqrt{(\sum_{j=1}^{n} w_{kj})(\sum_{l=1}^{n} w_{il})}}.$$

Hence all pairs of areas sharing a common border ($w_{ki} = 1$) will be correlated. This assumes the same level of spatial smoothness across $(\phi_1, \ldots, \phi_n)$ and is not realistic.

Alternatives?

- Ignore the spatial autocorrelation entirely and let $\phi_k = 0$ for all areas k.
- Replace the random effects $(\phi_1, \ldots, \phi_n)$ with an alternative spatial smoothing component that is forced to be orthogonal (unrelated) to the air pollution covariate, so that spatial confounding cannot occur. Such an approach was proposed by Hughes and Haran in (2013).
- Extend the CAR model to make it more flexible and allow for localised smoothness in the random effects, so that geographically adjacent values can be modelled as similar or very different.

Locally smooth CAR models

A number of approaches have been proposed to extend the standard CAR model to allow for localised spatial smoothness. They can be broken into two main approaches:

- Treat each w_{ki} relating to neighbouring areas as binary random quantities, so that if w_{ki} is estimated as one then (ϕ_k, ϕ_i) are spatially smoothed, while if w_{ki} is estimated as zero they are not. Examples include Lee and Mitchell (2013) and Lee, Rushworth and Sahu (2014).

- Augment the spatially smooth random effects with a piecewise constant jump component with different mean levels, so that if two areas close together have different mean levels their residuals will not be similar. Examples include Lawson and Clark (2002), Lee et al (2014).

Results

The choice of residual spatial autocorrelation model can make a large difference on the estimated pollution-health relationship. The estimated relative risks and 95% uncertainty intervals for a $1\mu gm^{-3}$ increase in $PM_{2.5}$ were:

Ignore correlation - 1.032 (1.005, 1.060).

CAR model - 1.065 (1.043, 1.091).

Orthogonal model - 1042 (1.040, 1.043).

Localised CAR model - 1.046 (1.033, 1.060).

Both the estimates and uncertainty intervals can differ substantially.

5. Conclusions

1. Developing a statistically valid approach for estimating the health effects of air pollution using spatial ecological data is a challenging task, and requires complex models for both the pollution and health data.

2. Using an inappropriate statistical model results in estimated health effects that are likely to be biased.

3. Future work will extend this model into the spatio-temporal domain, and the replication of the spatial data over time will enable more precise estimation of the air pollution effects.

4. The effects of air pollution will be investigated across Scotland, to see if the England results are replicated here.