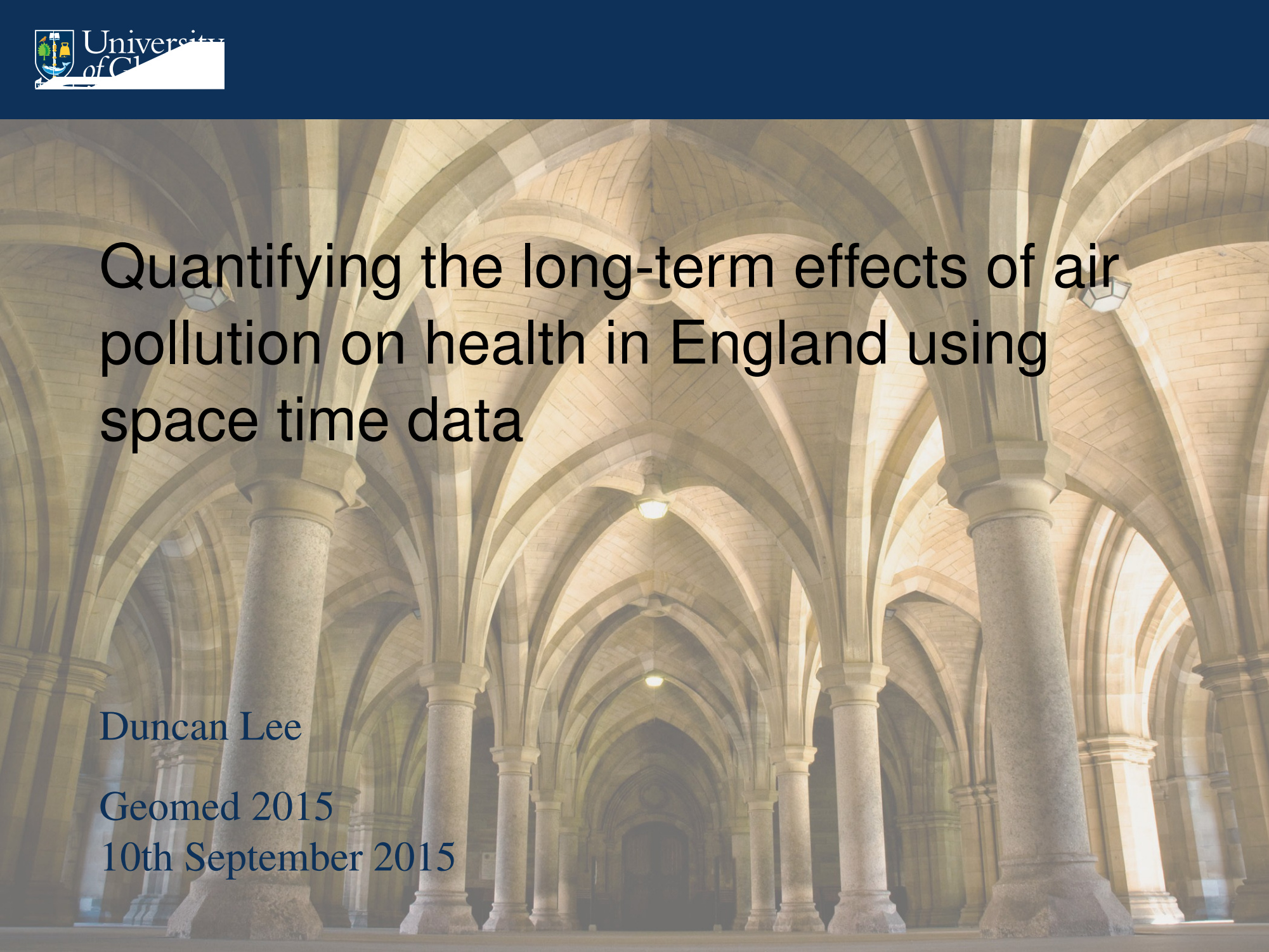


Quantifying the long-term effects of air pollution on health in England using space time data

Duncan Lee

Geomed 2015
10th September 2015

Acknowledgements

- This is joint work with Alastair Rushworth from the University of Glasgow, Sujit Sahu and Sabyasachi Mukhopadhyay from the University of Southampton, and Paul Agnew from the UK 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 and data

- Air pollution has long been known to adversely affect public health, in both the developed and developing world.
- Recent reports by the UK government and the World Health Organisation estimate that:
  - particulate matter reduces life expectancy by 6 months, with a health cost of £19 billion per year.
  - there were estimated to be over 23,000 premature deaths from air pollution in 2010.
- Air pollution will remain a key health problem for some time, as nitrogen dioxide emissions are predicted to exceed European Union limits until after 2030 in the urban areas of Greater London, West Midlands and West Yorkshire.

Chronic studies

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

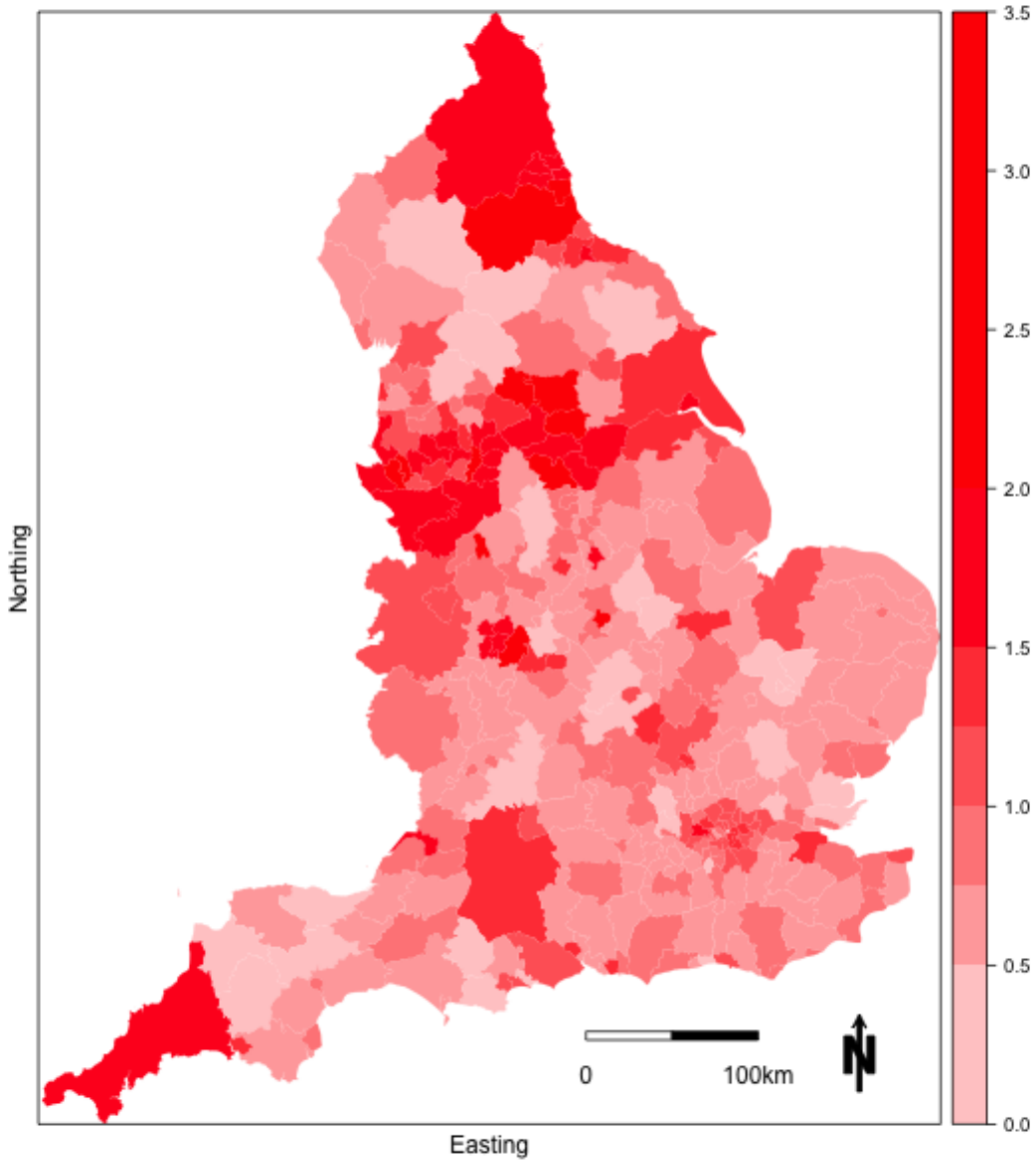
Cohort studies e.g. The Six Cities study by Dockery *et al* (1993) and the Escape study by Beelen *et al* (2014), which relate average air pollution concentrations to the health status of a large pre-defined cohort of people.

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.

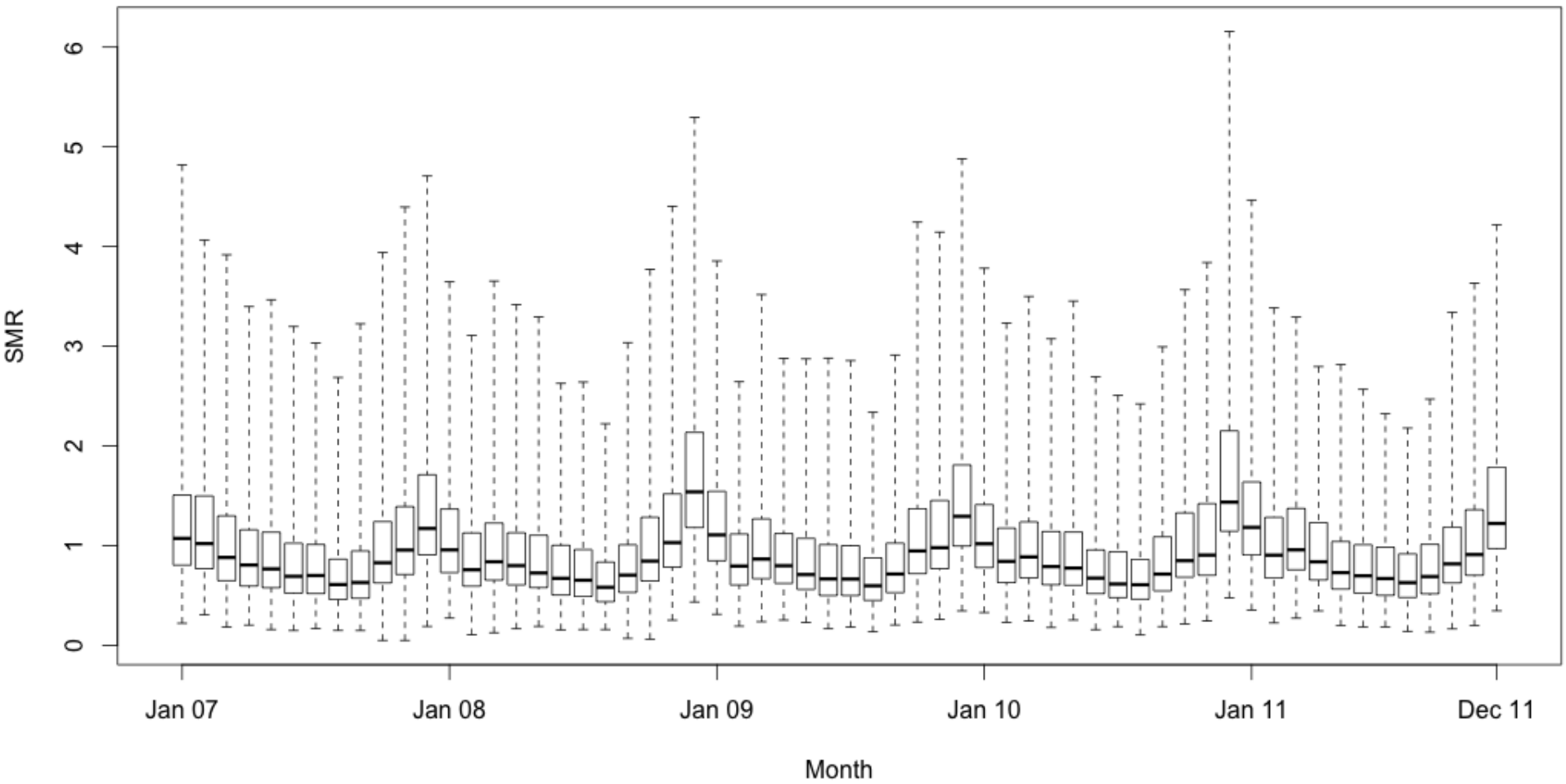
Ecological study design

- In ecological studies the data relate to populations living in a set of $k = 1, \ldots, K$ non-overlapping areal units for $t = 1, \ldots, T$ time periods, rather than to individuals.
- In this study we have $K = 323$ local and unitary authorities (LUA) that make up mainland England, and data are collected for $T = 60$ months between 2007 and 2011.
- For LUA k and month t the observed number of hospital admissions due to respiratory disease is denoted by Y_{kt}, while the expected number of admissions based on population demographics is denoted by E_{kt}.
- The standardised mortibidy ratio is given by $\text{SMR}_{kt} = Y_{kt}/E_{kt}$, an exploratory measure of disease risk.

Spatial pattern in SMR



Temporal pattern in SMR



Confounding factors

- The SMR exhibit a strong seasonal pattern, and this was accounted for by seasonally adjusting the expected disease counts E_{kt} by a monthly correction factor.

- The main confounding factor is socio-economic deprivation, and this is accounted for using two proxy variables, the percentage of people in each LUA and month who are in receipt of Job Seekers Allowance (JSA), and the average property price in each LUA and month.

Pollution data

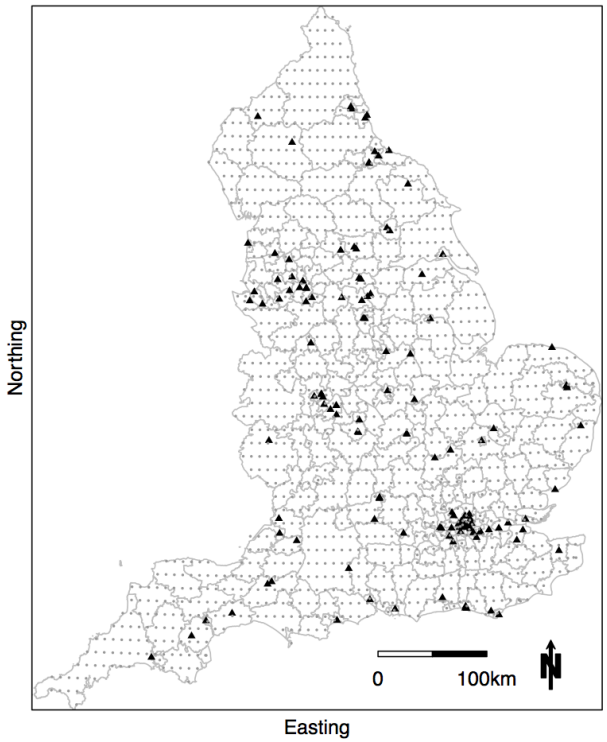
Recall from the previous talk that we consider concentrations of the following pollutants:

- Nitrogen dioxide (NO_2).
- Particulate matter measured as $PM_{2.5}$ and PM_{10}.
- Ozone (O_3).

For simplicity we only consider the first 3 here. Data on these pollutants come from the AURN monitoring network and the AQUM modelled concentrations provided by the Met Office.

Air pollution model

Recall from the previous talk that pollution concentrations were predicted on a 12km by 12km regular grid across England using a Bayesian hierarchical model, resulting in LUA k and month t having matrix of $l = 1, \ldots, 5000$ pollution predictions $\{z^{(l)}(\mathbf{v}_{kj}, t)\}$ at prediction locations $(\mathbf{v}_{k1}, \ldots, \mathbf{v}_{kn_k})$.



Summarising pollution concentrations

For LUA k and month t we have $5000 \times n_k$ predictions from the pollution model, and the simplest approach is to average them, that is

$$\hat{z}_{kt} = \frac{1}{Ln_k} \sum_{j=1}^{n_k} \sum_{l=1}^{L} z^{(l)}(\mathbf{v}_{kj}, t)$$

This ignores two different sources of uncertainty in the pollution concentrations:

- Spatial variation in pollution within each LUA.
- Posterior uncertainty in the pollution predictions at each prediction location.

2. Methodology and results

A typical Bayesian hierarchical model for the disease data is

$$\begin{aligned} Y_{kt}|E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}) \\ \ln(R_{kt}) &= \beta_0 + \hat{z}_{kt}\beta + \mathbf{x}_{kt}^{\top}\boldsymbol{\beta}_x + \phi_{kt}, \end{aligned}$$

where

Y_{kt}	health counts	E_{kt}	expected cases
R_{kt}	health risk	ϕ_{kt}	random effect
$\mathbf{x}_{kt}^{\top}\boldsymbol{\beta}_x$	other covariate effects	β_0	intercept
$\hat{z}_{kt}$	air pollution	β	pollution effect

Statistical considerations

1. Unmeasured confounding: Air pollution and the other covariates do not account for all variation. Adding a set of spatio-temporal random effects, ϕ_{kt} can offer a solution.

How should ϕ_{kt} be structured in space and time?

2. Spatial Misalignment: The air pollution model estimates the true exposure surface $Z(v_{kj}, t)$ at grid locations, $\{v_{kj}\}$ not as LUA regional averages.

How can we reconcile these quantities?

3. Uncertainty: The posterior distribution of $Z(v_{kj}, t)$ is available via MCMC samples, and therefore uncertainty in air pollution is quantified.

How should this source of uncertainty be incorporated into the health model?

Issue 1 - unmeasured confounding

Rushworth et al. (2014) propose the 'global' autoregressive conditional autoregressive (CAR) model for the random effects $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{Kt})$ at time t:

$$\begin{aligned} \boldsymbol{\phi}_1 &\sim \text{N}\left(\mathbf{0}, \sigma^2 \mathbf{Q}(\mathbf{W}, \rho)^{-1}\right) \\ \boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1} &\sim \text{N}\left(\alpha\boldsymbol{\phi}_{t-1}, \sigma^2 \mathbf{Q}(\mathbf{W}, \rho)^{-1}\right) \quad \text{for } t \geq 2 \end{aligned}$$

where

$$\begin{aligned} \mathbf{Q}(\mathbf{W}, \rho) &= \rho\left[diag(\mathbf{W1}) - \mathbf{W}\right] + (1-\rho)\mathbf{I} \\ \mathbf{W} &= \text{spatial (binary) neighbours matrix.} \end{aligned}$$

W is a $K \times K$ matrix that encodes neighbourhood relationships in the study region such that

$$\begin{aligned} w_{kj} = 1 \quad &\Longleftrightarrow \quad \text{units } k \text{ and } j \text{ share a common border} \\ w_{kj} = 0 \quad &\qquad\quad \text{otherwise, or if } \; k = j \end{aligned}$$

If $T = 1$ then the conditional distribution for ϕ_{k1} is

$$\phi_{k1} | \boldsymbol{\phi}_{-k1} \quad \sim \quad \mathrm{N}\left(\frac{\rho \sum_{j=1}^{n} w_{kj} \phi_{j1}}{1 - \rho + \rho \sum_{j=1}^{n} w_{kj}}, \frac{\tau^2}{1 - \rho + \rho \sum_{j=1}^{n} w_{kj}} \right)$$

A more flexible model for ϕ_{kt}

$\mathbf{Q}(\mathbf{W}, \rho)$ restricts the range of surfaces that can be fitted, as ρ controls the level of spatial autocorrelation globally across the entire region.

Therefore we treat the non-zero elements of $\mathbf{W}$ as random variables $w_{kj}^{+} \in [0, 1]$. We control model complexity using a normal shrinkage prior on the logit transformed w_{kj}^{+}:

$$\ln\left(\frac{w_{kj}^{+}}{1 - w_{kj}^{+}}\right) \sim \mathrm{N}\left(\mu,\ \tau^2\right)$$

If μ is positive then w_{kj}^{+} are *a-priori* close to one reflecting prior preference for spatial smoothness.

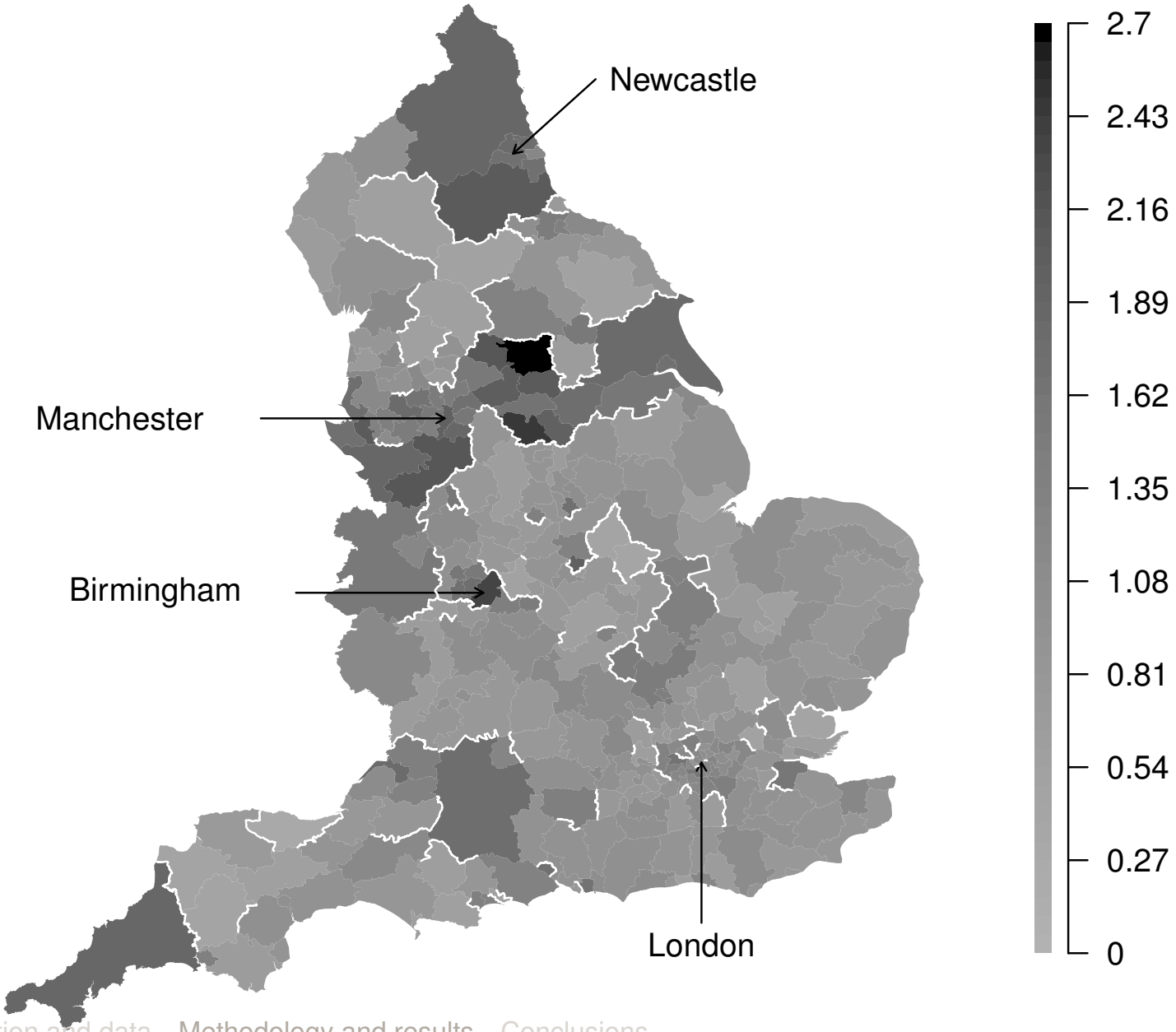
English respiratory data results

Table: Posterior median relative risks and 95% credible intervals for a 1-standard deviation increase in each pollutant, which is 16.07, 4.90, and 4.11 μgm^{-3} respectively.

Pollutant	No random effects	Non-adaptive ϕ_{kt}	Adaptive ϕ_{kt}
NO_2	1.151 (1.144, 1.158)	1.057 (1.045, 1.069)	1.048 (1.036, 1.060)
PM_{10}	1.013 (1.007, 1.020)	1.007 (0.998, 1.015)	1.006 (0.995, 1.015)
$PM_{2.5}$	1.013 (1.007, 1.019)	1.006 (0.997, 1.014)	1.006 (0.997, 1.016)

Simpler models have a tendency to have larger estimated air-pollution effects.

ϕ_{kt} estimates and adjacency model



Issue 2 - Spatial misalignment

- The spatial misalignment between the pollution and disease data means there is within LUA variation in pollution which can result in ecological bias.
- Wakefield and Shaddick (2006) among others derive an appropriate aggregate model that overcomes this ecological bias.
- Under simplifying assumptions the bias term can be derived to be of the order of β^2, thus as β is small here the effect of this bias is likely to be negligible.
- This negligible effect was confirmed empirically by Lee and Sarran (2015) in this air pollution and health context, and the naive ecological model and aggregate model similar to Wakefield and Shaddick (2006) give identical results.

Issue 3 - Posterior uncertainty

The pollution model yields predictive distributions, based on 5,000 MCMC samples, for average air pollution in each LUA and month.

This uncertainty should be fed into the health model so that the resulting health estimates account for this variation.

3 possible strategies:

(1) Treat posterior mean pollution concentrations as true values (no uncertainty).

(2) Directly feed samples from the posterior air pollution density through the health model.

(3) Treat the posterior pollution densities as prior distributions in the health model (e.g. using a Gaussian approximation).

Results – uncertainty

Table: Posterior median relative risks and 95% credible intervals for a 1-standard deviation increase in each pollutant, which is 16.07, 4.90, and 4.11 μgm^{-3} respectively.

Pollutant	No uncertainty	Posterior	Prior
NO_2	1.048 (1.036, 1.060)	1.001 (0.999, 1.003)	1.035 (1.030, 1.041)
PM_{10}	1.006 (0.995, 1.015)	1.000 (0.998, 1.003)	1.025 (0.999, 1.043)
$PM_{2.5}$	1.006 (0.997, 1.016)	1.001 (0.997, 1.004)	1.008 (0.995, 1.062)

3. Conclusions

- Choices for handling spatio-temporal autocorrelation have important consequences for the estimated effects of air pollution.
- It is important to treat air pollution exposure as uncertain, as it is rarely realistic to assume exposure is observed (or predicted) without error.
- From our study it appears that NO_2 poses the greatest ongoing health risks, and a 3.5% increased risk corresponds to around 21,500 more admissions per year.
- Given that large parts of England are expected to exceed EU emissions targets for the next 15 years, NO_2 continues to be a major health problem.

Future work

Air pollution effect estimates are small and can be volatile, so we will conduct a large sensitivity analysis of our study results looking at how sensitive the estimated pollution effects are to:

- the pollution model described in the previous talk.
- the use of monthly mean concentrations rather than monthly extremes or exceedences of a threshold level.
- the choice of variables used to control for the confounding effects of socio-economic deprivation.
- The choice of random effects model to account for spatio-temporal autocorrelation.