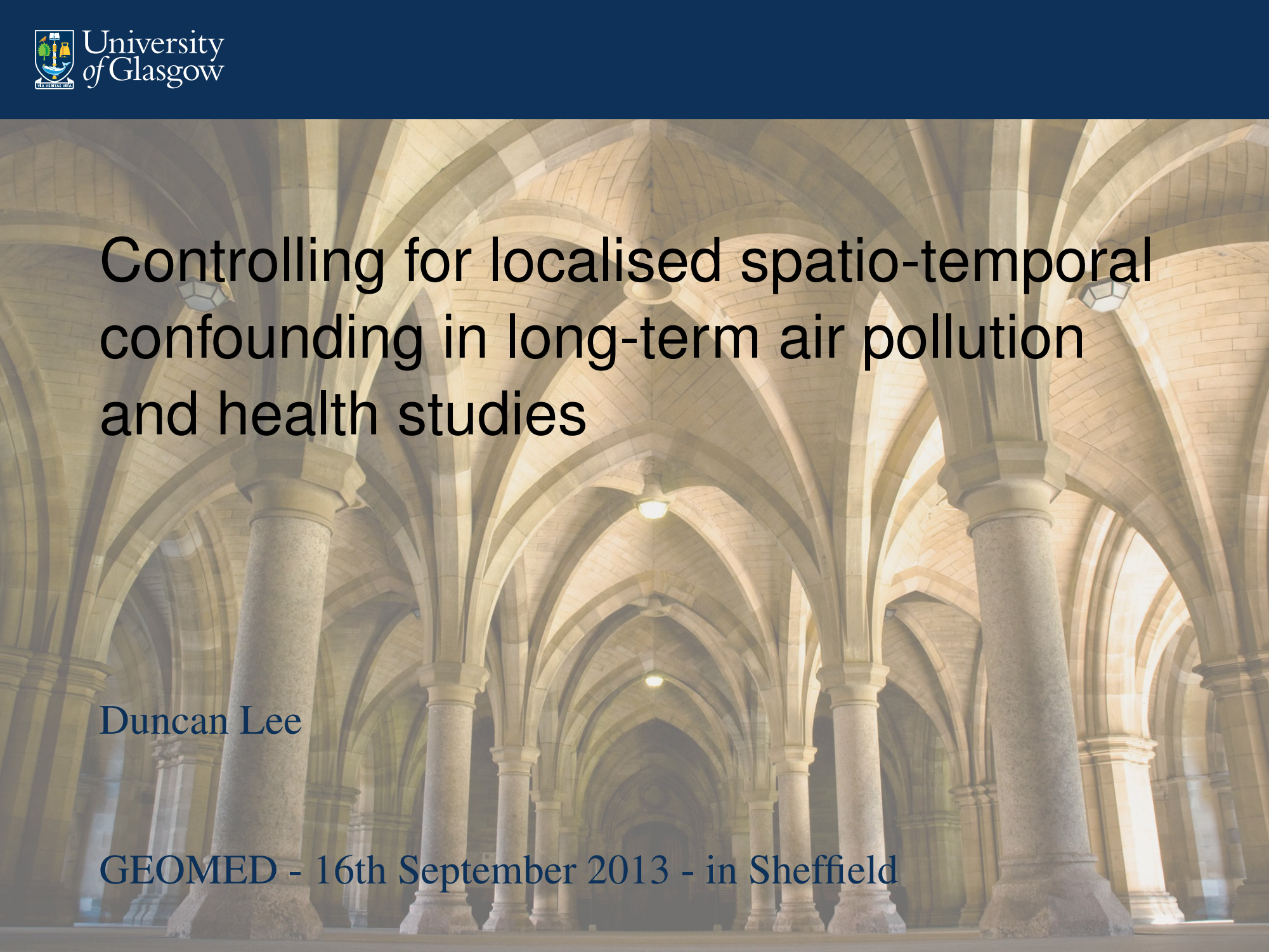


University of Glasgow

Controlling for localised spatio-temporal confounding in long-term air pollution and health studies

Duncan Lee

GEOMED - 16th September 2013 - in Sheffield

University of Glasgow

Acknowledgements

- Thank you to Andrew Lawson for inviting me to give this talk.
- This is joint work with Richard Mitchell from the University of Glasgow.
- The work is funded by the EPSRC grants EP/J017442/1 and EP/J017485/1.

EPSRC
Engineering and Physical Sciences Research Council

1. Background and motivation

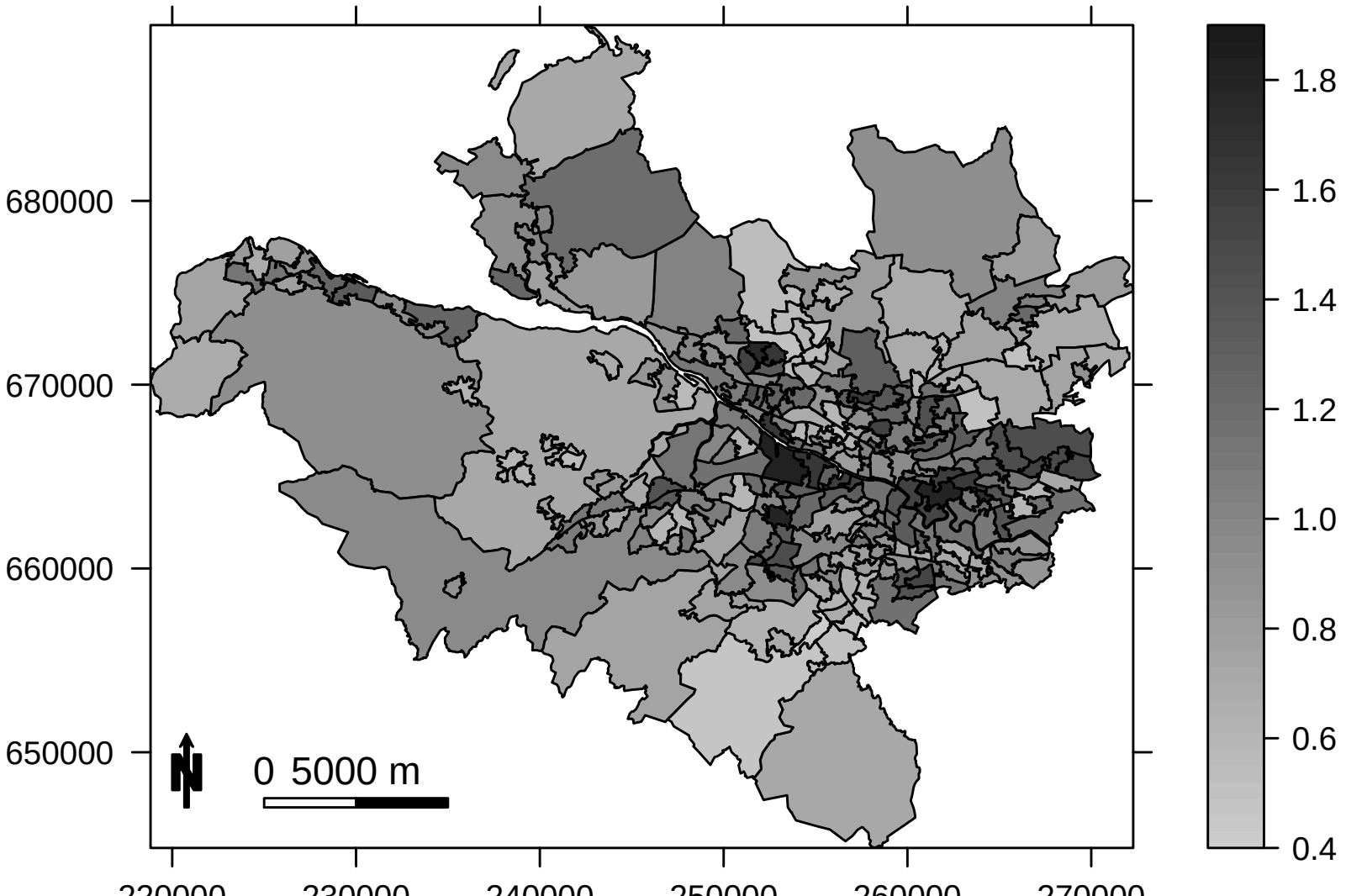
- Air pollution has long been known to adversely affect public health, in both the developed and developing world.
- A recent report by the UK government estimates that particulate matter alone reduces life expectancy by 6 months, with a health cost of £19 billion per year.
- Epidemiological studies into the effects of air pollution have been conducted since the 1990s, focusing on both short-term and long-term exposure.
- The long-term effects can be estimated by cohort or small-area ecological studies, and in this talk we focus on the latter.

Study design

- Small area studies have an ecological design, because the data are population level summaries relating to a set of N non-overlapping areal units for T consecutive years, rather than to individuals.
- Examples 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.
- The covariates, including air pollution concentrations, are contained in an $n \times p$ matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Example - Greater Glasgow, Scotland

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



Statistical model

$$
\begin{aligned}
Y_k &\sim \text{Poisson}(E_k R_k), \\
\log(R_k) &= \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta} + \phi_k, \\
\phi_k | \boldsymbol{\phi}_{-k}, \tau^2, W &\sim \text{N}\left(\frac{\sum_{i=1}^{n} w_{ki} \phi_i}{\sum_{i=1}^{n} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{n} w_{ki}} \right),
\end{aligned}
$$

where

- R_k quantifies disease risk in area k.
- $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are random effects to model residual spatial correlation, and are assigned an intrinsic Conditional autoregressive (ICAR) prior where $W = (w_{ki})$ is a binary $n \times n$ neighbourhood matrix.

Limitations and motivation

- The ICAR prior forces the random effects ϕ to be spatially smooth, which leads to problems of collinearity with covariates that are also smooth such as air pollution.
- Existing work in this area include Reich *et al.* (2006) and Hughes and Haran (2013).
- Furthermore, the random effects are unlikely to be spatially smooth anyway, because the disease data (e.g. the SIR) are not spatially smooth so the residuals after removing covariate effects are also unlikely to be.
- We propose an extension to ICAR priors that can capture localised spatial smoothness in the random effects surface, i.e. subregion of spatial smoothness and step changes.

2. Methodology

- Our general approach follows Lu *et al.* (2007) and Lee and Mitchell (2012, 2013), and models $\mathcal{W} = \{w_{kj} | k \sim j, k > j\}$ as binary random variables, rather than w_{kj} being fixed at 1.
- This is because if $w_{kj} = w_{jk} = 1$ then (ϕ_k, ϕ_j) are correlated and are smoothed over, where as if $w_{kj} = w_{jk} = 0$ they are conditionally independent and are not smoothed over.
- We follow the terminology of graphical models and refer to $w_{kj} \in \mathcal{W}$ as *edges*, and define any edge w_{kj} that equals zero as having been removed.

Our methodological innovation is a *Localised Conditional AutoRegressive (LCAR)* prior, which decomposes the joint distribution for an extended set of random effects $\tilde{\phi}$ and the set of edges $\mathcal{W}$ as

$$f(\tilde{\phi}, \mathcal{W}) = f(\tilde{\phi}|\mathcal{W})f(\mathcal{W})$$

Standard ICAR models consist of the first of these distributions, the latter is fixed.

Prior distribution - $f(\tilde{\phi}|\mathcal{W})$

- The ICAR prior is inappropriate if $\mathcal{W}$ is random, because one could get $\sum_{i=1}^{n} w_{ki} = 0$ leading to an infinite variance.
- Therefore we introduce $\tilde{\phi} = (\phi, \phi_*)$, where ϕ_* is a global random effect that prevents any unit from having no edges.
- The corresponding $(n+1) \times (n+1)$ neighbourhood matrix is given by

$$\tilde{W} = \begin{bmatrix} W & \mathbf{w}_* \\ \mathbf{w}_*^{\mathrm{T}} & 0 \end{bmatrix},$$

where $\mathbf{w}_* = (w_{1*}, \ldots, w_{n*})$, $w_{k*} = \mathrm{I}[\sum_{i \sim k}(1 - w_{ki}) > 0]$.

We propose the multivarite prior $\tilde{\phi} \sim \mathrm{N}(\mathbf{0}, \tau^2 Q(\tilde{W}, \epsilon)^{-1})$, where

$$Q(\tilde{W}, \epsilon) \;=\; \mathrm{diag}(\tilde{W}\mathbf{1}) - \tilde{W} + \epsilon I.$$

This is an ICAR prior for $\tilde{\phi}$, except for the addition of ϵI, for small ϵ, which ensures the matrix is invertible. The full conditional distributions are given by:

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

Prior distribution - $f(\mathcal{W})$

- The dimensionality of $\mathcal{W}$ is $N_{\mathcal{W}} = \mathbf{1}^T W \mathbf{1}/2$, and as each edge is binary the sample space has size $2^{N_{\mathcal{W}}}$.
- Previous studies have shown that modelling each element in $\mathcal{W}$ separately results in weakly identifiable parameters.
- Therefore we treat $\mathcal{W}$ as a single random quantity, and propose the following prior for $\tilde{W}$;

$$\tilde{W} \sim \text{Discrete Uniform}(\tilde{W}^{(0)}, \tilde{W}^{(1)}, \ldots, \tilde{W}^{(N_{\mathcal{W}})}).$$

- Candidate $\tilde{W}^{(j)}$ has j edges retained in the model (i.e. j elements in $\mathcal{W}$ equal 1), so $(\tilde{W}^{(0)}, \tilde{W}^{(N_{\mathcal{W}})})$ correspond to independence and IAR priors respectively.

Eliciting $(\tilde{W}^{(0)}, \tilde{W}^{(1)}, \ldots, \tilde{W}^{(N_W)})$

- We propose eliciting $(\tilde{W}^{(0)}, \tilde{W}^{(1)}, \ldots, \tilde{W}^{(N_W)})$ from disease data prior to the study period, because it should have a similar spatial structure to the response.
- Let $((\mathbf{Y}_1^p, \mathbf{E}_1^p), \ldots, (\mathbf{Y}_r^p, \mathbf{E}_r^p))$ denote disease data for the r years prior to the study.
- The study data have expectation $\mathbb{E}[\mathbf{Y}] = \mathbf{E}\exp(X\boldsymbol{\beta} + \boldsymbol{\phi})$, which is equivalent to $\ln(\mathbb{E}[\mathbf{Y}]/\mathbf{E}) = X\boldsymbol{\beta} + \boldsymbol{\phi}$. Thus we make the approximation:

$$\phi_j^p = \ln\left[\frac{\mathbf{Y}_j^p}{\mathbf{E}_j^p}\right] \approx \ln\left[\frac{\mathbf{Y}}{\mathbf{E}}\right] \sim_{approx} \mathrm{N}(X\boldsymbol{\beta}, \tau^2 Q(\tilde{W}, \epsilon)_{1:n}^{-1}).$$

Elicitation algorithm

1. Start at $\tilde{W}^{(N_{\mathcal{W}})}$ which has all edges retained in the model ($w_{kj} = 1$) and corresponds to the IAR prior for strong spatial smoothing.

2. For $j = N_{\mathcal{W}}, \ldots, 1$ move from $\tilde{W}^{(j)}$ to $\tilde{W}^{(j-1)}$ by removing a single edge from $\mathcal{W}$ (i.e. by setting an element in $\mathcal{W}$ equal to zero). This corresponds to localised spatial smoothing

3. When $j = 0$ $\tilde{W}^{(0)}$ contains no edges ($w_{kj} = 0$), and corresponds to non-spatial smoothing.

Moving from $\tilde{W}^{(j)}$ to $\tilde{W}^{(j-1)}$

At step j compute the joint approximate Gaussian log-likelihood for $(\phi_1^p, \ldots, \phi_r^p)$ given by

$$\begin{aligned}
\ln[f(\phi_1^p, \ldots, \phi_r^p | \tilde{W}^{(*)})] &= \sum_{j=1}^{r} \ln[\mathrm{N}(\phi_j^p | X\hat{\beta}, \hat{\tau}^2 Q(\tilde{W}^*, \epsilon)_{1:n}^{-1})], \\
&\propto \frac{r}{2} \ln(|Q(\tilde{W}^*, \epsilon)_{1:n}|) - \frac{nr}{2} \ln(\hat{\tau}^2) \\
&\quad - \frac{1}{2\hat{\tau}^2} \sum_{j=1}^{r} (\phi_j^p - X\hat{\beta})^{\mathrm{T}} Q(\tilde{W}^*, \epsilon)_{1:n} (\phi_j^p - X\hat{\beta}),
\end{aligned}$$

for all matrices $\tilde{W}^{(*)}$ that differ from $\tilde{W}^{(j)}$ by having one additional edge removed. Then set $\tilde{W}^{(j-1)}$ equal to the value of $\tilde{W}^{(*)}$ that maximises the log-likelihood.

Overall model

The overall model is given by

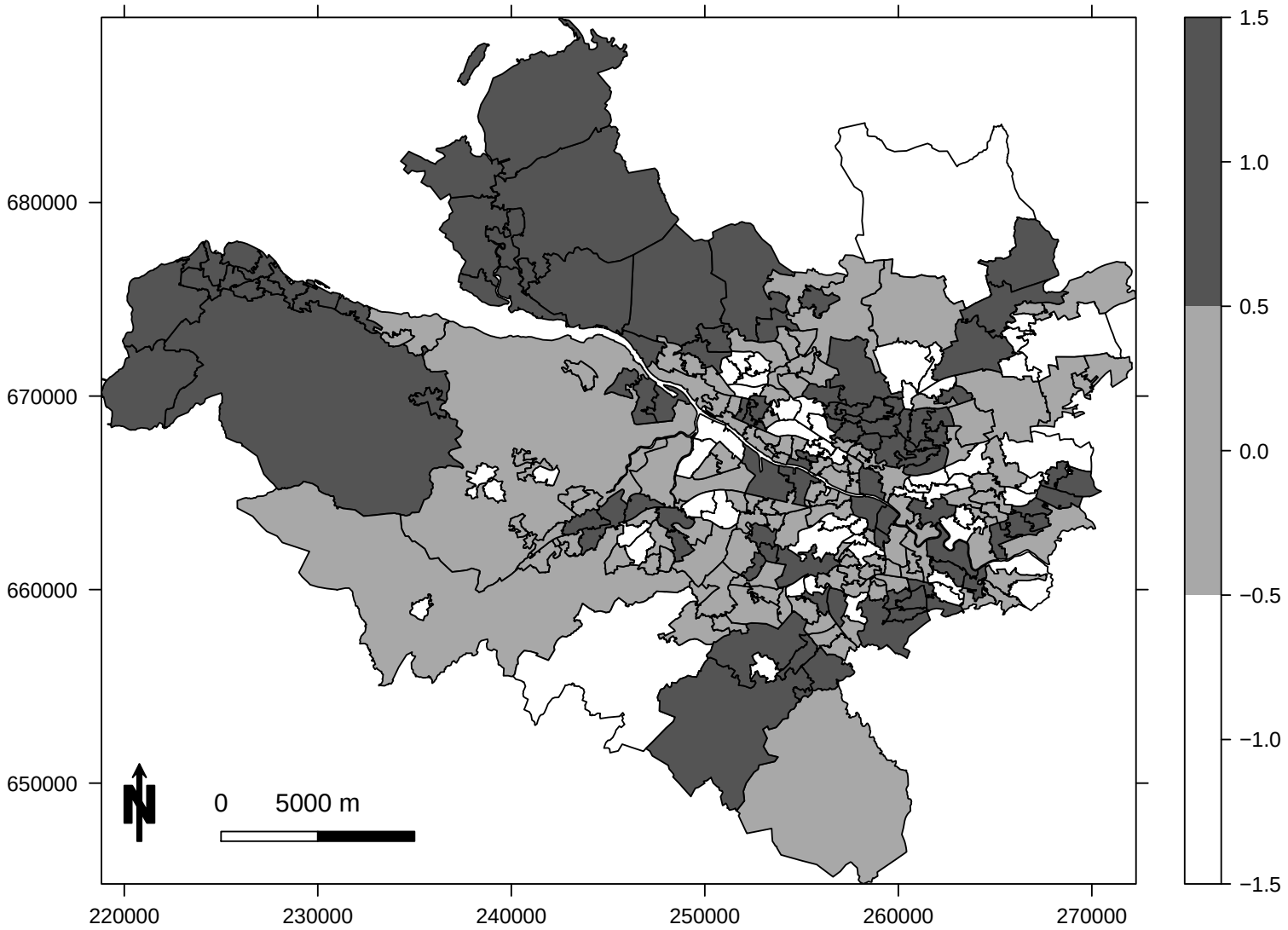
$$
\begin{aligned}
Y_k | E_k, R_k &\sim \text{Poisson}(E_k R_k) \quad \text{for } k = 1, \ldots, n, \\
\ln(R_k) &= \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta} + \phi_k, \\
\tilde{\boldsymbol{\phi}} &\sim \text{N}(\mathbf{0}, \tau^2 Q(\tilde{W}, \epsilon)^{-1}), \\
\tilde{W} &\sim \text{Discrete Uniform}(\tilde{W}^{(0)}, \tilde{W}^{(1)}, \ldots, \tilde{W}^{(N_{\mathcal{W}})}), \\
\beta_j &\sim \text{N}(0, 1000) \quad \text{for } j = 1, \ldots, p, \\
\tau^2 &\sim \text{Uniform}(0, 1000).
\end{aligned}
$$

3. Simulation study

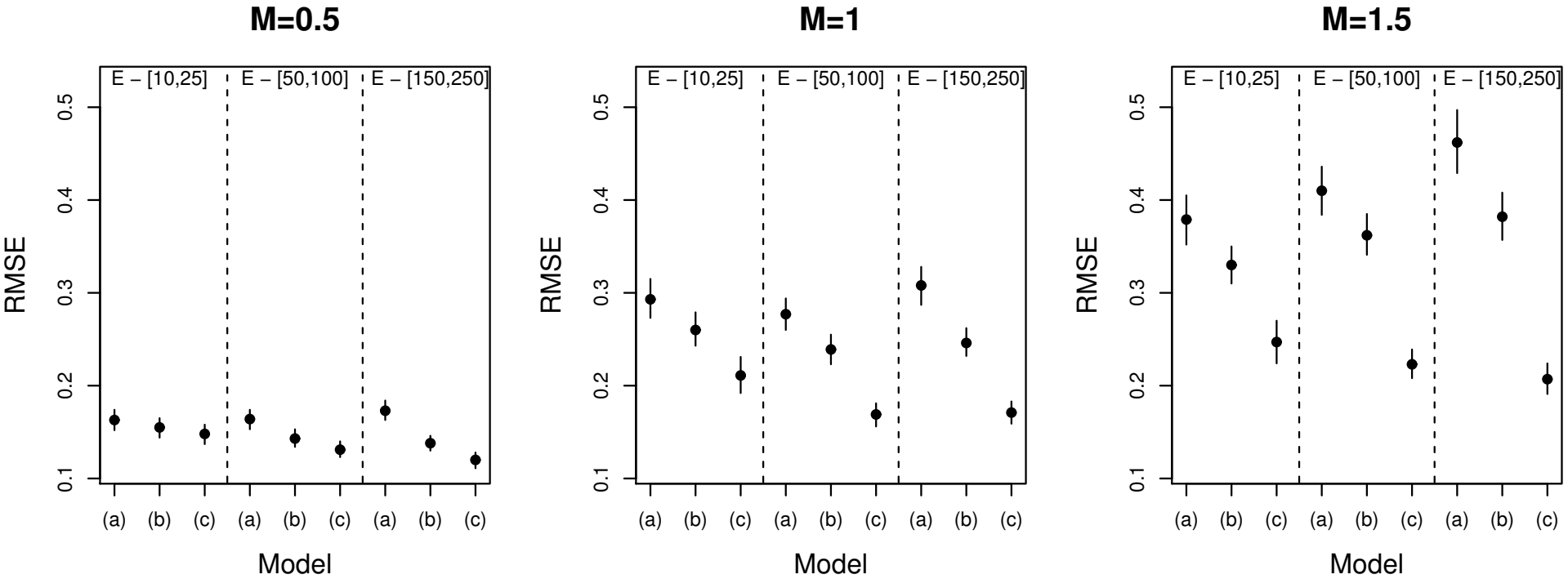
- Five hundred data sets were generated for Greater Glasgow under a number of different scenarios.
- Each data set consisted of study data and three years of prior data, and the LCAR model was compared with the commonly used ICAR and BYM models.
- For each data set the log-risk surface was generated as a linear combination of a spatially smooth covariate (representing air pollution) and localised residual spatial structure (to be modelled by the random effects).
- The localised residual spatial structure was generated from a multivariate Gaussian distribution with a piecewise constant mean, the template for which is shown on the next slide.

Generating localised spatial structure

The piecewise constant mean below is multiplied by M.



RMSE values for β for different $(M, \mathbf{E})$



Here (a)-ICAR, (b)-BYM and (c)-LCAR. The dot is the root mean square error (RMSE) of the regression parameter β and the line is a bootstrapped 95% confidence interval.

4. Case study

- The methodology was motivated by a study estimating the effects of air pollution on hospitalisation due to respiratory disease in Greater Glasgow, Scotland in 2010.
- The prior distribution for the spatial structure was elicited using three years of respiratory hospitalisation data between 2007 and 2009.
- Modelled concentrations of nitrogen dioxide (NO_2), and particulate matter ($PM_{2.5}$ and PM_{10}) were available for 2009, along with a measure of income deprivation, a major confounder in spatial ecological studies.
- The pollutants were included in separate models to avoid issues of collinearity, and in all cases inference was based on 150,000 samples obtained from 3 Markov chains.

Results 1 - Parameter estimates

	Model		
	ICAR	**BYM**	**LCAR**
DIC (p.d)	2113.3 (164.1)	2096.5 (164.6)	2090.8 (162.1)
NO_2	1.017 (0.975, 1.061)	1.032 (0.994, 1.067)	1.034 (1.002, 1.068)
$PM_{2.5}$	1.033 (0.990, 1.078)	1.042 (1.009, 1.078)	1.043 (1.010, 1.074)
PM_{10}	1.037 (0.997, 1.081)	1.043 (1.007, 1.079)	1.048 (1.017, 1.080)

The pollution effects are relative risks for a one standard deviation increase in the yearly average concentrations, which are: NO_2 - 5.0μgm^{-3}, $PM_{2.5}$ - 1.1μgm^{-3}, PM_{10} - 1.5μgm^{-3}.

5. Conclusions

1. The LCAR prior proposed here has the flexibility to capture both sub-regions of spatial correlation and step changes in the random effects surface, which reduces the effects of collinearity with spatially smooth covariates compared with commonly used CAR models.

2. The improvements in the estimation of the fixed effects can be substantial, as the percentage reductions in RMSE between the BYM and LCAR models ranged between 4.5% and 45.8% in the simulation study presented here.

3. Future work will extend this model into the spatio-temporal domain, as well as applying it to the related fields of disease mapping and Wombling.

Contact details

Dr Duncan Lee

School of Mathematics and Statistics
University of Glasgow
Glasgow
G12 8QQ

Duncan.Lee@glasgow.ac.uk