Two-stage spatio-temporal modelling of the health effects of air pollution in England

Alastair Rushworth

University of Glasgow

Spatial Statistics - Emerging Patterns, Avignon

Wednesday June 10^{th}, 2015

Acknowledgements

Work in this talk part of a larger collaboration:

University of Glasgow: Duncan Lee, Richard Mitchell

University of Southampton: Sujit Sahu, Sabyasachi Mukhopadhyay

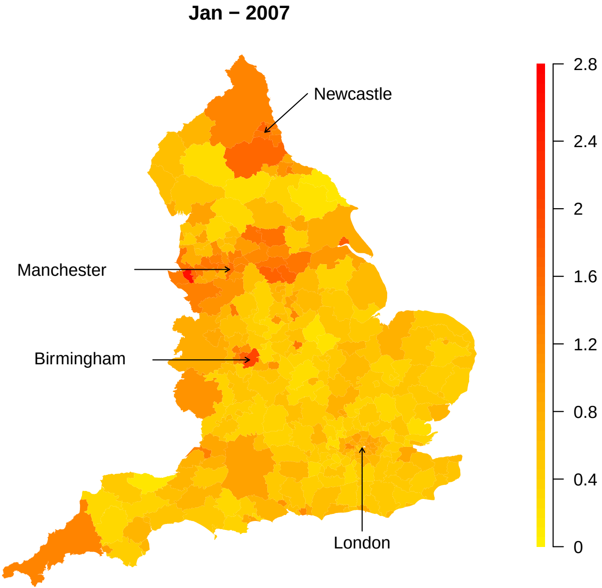
UK Met Office: Paul Agnew, Christophe Sarran

Funded by the EPSRC: EPSRC Engineering and Physical Sciences Research Council

Background

- Air pollution is well known to have a negative impact on human health and is still a major public health issue.
- A recent UK government report estimates that elevated particulate matter reduces life expectancy by 6 months, incurring an annual health cost of £19 billion.
- In London, an additional 4000 deaths attributable to poor air quality annually (Miller, 2010).
- Still difficult to adequately quantify the ill effects of air pollution without first accounting for *confounding variables* and understanding *pollution exposure*.

England LHA respiratory admissions data



- Jan 2007 to Dec 2011 (60 months)
- 323 Local Health Authorities
- N = 19380

Typical model for spatial heath counts

$$\begin{aligned} Y_{kt}|E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}) \\ \ln(R_{kt}) &= \beta_0 + x_{kt}\beta + \mathbf{z}_{kt}^\top \boldsymbol{\alpha} + \phi_{kt}, \\ t &= 1, \ldots, T \quad \text{time points} \\ k &= 1, \ldots, N \quad \text{regions} \end{aligned}$$

Where

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

Statistical considerations

Unmeasured confounding: Air pollution, and the other measured 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?

Misalignment: The air pollution model estimates the true exposure surface $Z(s_{kj}, t)$, by a set of predictive distributions at grid locations, $\{s_{kj}\}$.

Health counts are regional totals. How can we reconcile these quantities? Could we simply average the air pollution?

Uncertainty: The posterior density of $Z(s_{kj}, t)$ 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? What effect does this have on estimation?

Unmeasured confounding: An existing model for ϕ_{kt}

Rushworth et al. (2014) propose the 'global' model:

$$\ln(R_{kt}) = \beta_0 + x_{kt}\beta + \mathbf{z}_{kt}^\top \boldsymbol{\alpha} + \phi_{kt}$$

Letting $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{Nt})$, where $t = 1, \ldots, T$, then:

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

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

Unmeasured confounding: An existing model for ϕ_{kt}

$\mathbf{W}$ is an $N \times N$ matrix that encodes neighbourhood relationships in the study region such that

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

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

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

Unmeasured confounding: a more flexible model for ϕ_{kt}

$\mathbf{Q}(\mathbf{W}, \rho)$ restricts the range of surfaces that can be fitted.

Solution: Treat non-zero elements of $\mathbf{W}$ as random variables $w_{ij}^{+} \in [0, 1]$.

Control model complexity using normal prior on transformed w_{ij}^{+}:

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

μ is chosen to be large and positive reflecting prior preference for spatial smoothness.

English respiratory data: random effects

We will compare the random effects models

Model type	Random effects	Adjacency model
GLM	NA	—
Non-adaptive	ϕ_{kt}	$w^+_{kt} = 1$
Adaptive	ϕ_{kt}	$\mathrm{logit}(w^+_{kt}) \sim N(\mu, \tau^2)$

Under the risk specification

$$\ln(R_{kt}) = \beta_0 + x_{kt}\beta + \texttt{jobseekers}_{kt}\alpha_1 + \texttt{houseprice}_{kt}\alpha_2 + \phi_{kt}$$

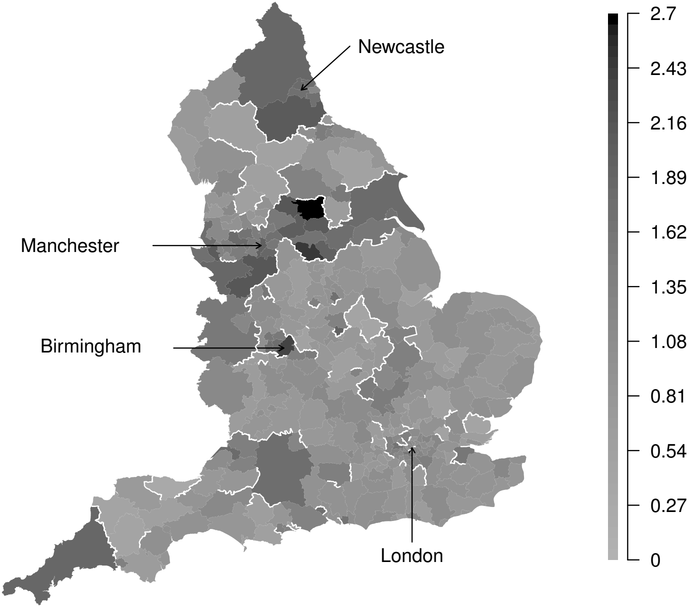
English respiratory data: random effects

Pollutant	No random effects (GLM)	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)
O_3	0.981 (0.974, 0.987)	0.983 (0.972, 0.995)	0.980 (0.965, 0.993)

Table : Risks and 95% CIs for 1-standard deviation increases in pollutant

Simpler models have a tendency to overestimate air-pollution effects.

ϕ_{kt} estimates and adjacency model



Air pollution - uncertainty

1^{st} stage model yields predictive distributions for air pollution in space and time.

This uncertainty should be passed through the 2^{nd} stage health model so that resulting health estimates represent all available information.

Some 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)

Exploring the English respiratory data: uncertainty

Compare approaches to incorporating pollution uncertainty:

$$\begin{array}{lcll} \textbf{(1)} & x_{kt} & = & \bar{x}_{kt} \\ \textbf{(2)} & x_{kt} & \sim & DU \text{ over posterior air pollution samples} \\ \textbf{(3)} & x_{kt} & \sim & MVN \text{ estimated from posterior samples} \end{array}$$

Again, under the risk specification

$$\ln(R_{kt}) = \beta_0 + x_{kt}\beta + \texttt{jobseekers}_{kt}\alpha_1 + \texttt{houseprice}_{kt}\alpha_2 + \phi_{kt}$$

Results – uncertainty

Pollutant	**(1)** $x_{kt} = \bar{x}_{kt}$	**(2)** $x_{kt} \sim DU$	**(3)** $x_{kt} \sim MVN$
$\mathrm{NO_2}$	1.048 (1.036, 1.060)	1.001 (0.999, 1.003)	1.035 (1.030, 1.041)
$\mathrm{PM_{10}}$	1.006 (0.995, 1.015)	1.000 (0.998, 1.003)	1.025 (0.999, 1.043)
$\mathrm{PM_{2.5}}$	1.006 (0.997, 1.016)	1.001 (0.997, 1.004)	1.008 (0.995, 1.062)
$\mathrm{O_3}$	0.980 (0.965, 0.993)	1.000 (0.999, 1.001)	0.996 (0.967, 1.000)

Table : Risks and 95% CIs for 1-standard deviation increases in pollutant

Conclusions

- Choices for handling spatio-temporal autocorrelation have important consequences for the estimating the 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.

Future work:

- Simulate study to determine bias an coverage properties for β
- Improve on current Gaussian approximation to air pollution posterior
- Multivariate pollution model

Thank you very much for listening!