University of Glasgow

Using small-area spatial data for statistical and epidemiological research

Duncan Lee Duncan.Lee@glasgow.ac.uk

Scottish Government 16th December 2015

Acknowledgements

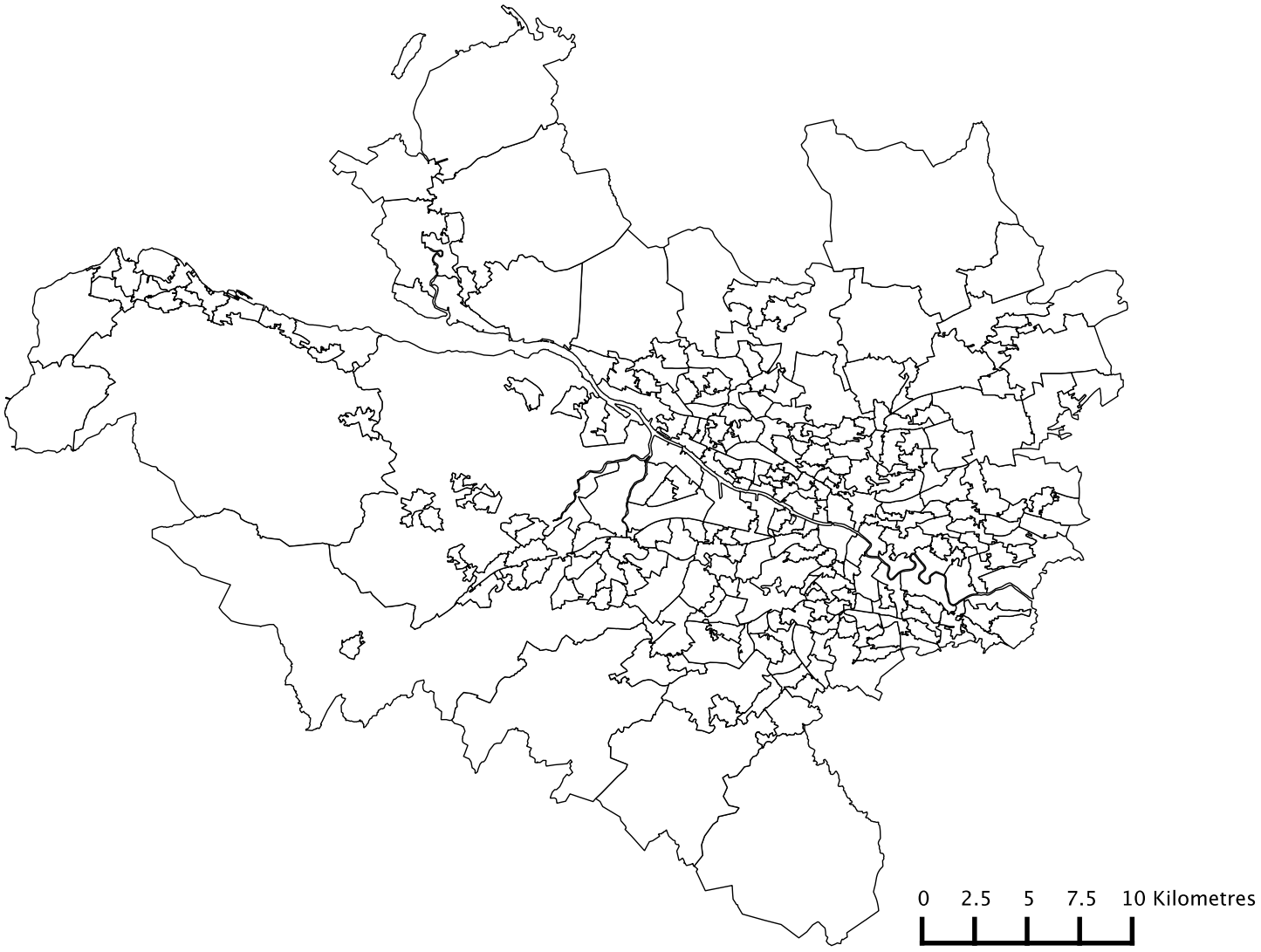
This is joint work with colleagues from two main projects:

- Andrew Lawson, Gary Napier, Kevin Pollock and Chris Robertson on an MRC funded project focusing on modelling spatio-temporal patterns in disease risk.
- Guowen Huang and Marian Scott on an EPSRC funded project focusing on the effects of air pollution on human health.

Introduction

- The availability of small-area spatial data has dramatically increased in the last decade or so, including:
  - Scottish Neighbourhood Statistics (SNS).
  - Health and Social Care Information Centre (HSCIC).
- This increase has been accompanied by the widespread development of statistical methodology and software for mapping and modelling these data.
- The latter include Geographical Information Systems software and statistical modelling software, such as `ArcGIS / QGIS`, and packages in the statistical software `R` such as `CARBayes` and `RINLA`.
- The analysis of small-area data is performed in many fields, such as econometrics, epidemiology, and social science.

Example - Intermediate Geographies



Modelling spatial data

The key feature when modelling spatial data is that of **spatial autocorrelation**, which is summarised by Tobler's first law of geography which says

> Everything is related to everything else, but near things are more related than distant things.

Which means that standard regression models that assume independence in the residuals are likely to be inappropriate and potentially result in misleading conclusions.

How to model spatial aucorrelation

There are a number of commonly used models for capturing spatial autocorrelation in data / residuals from a regression model, including:

- Conditional AutoRegressive (CAR) models.
- Simultaneous AutoRegressive (SAR) models.
- Point level (Geostatistical) models based on each areas centroid.

CAR and SAR models are most commonly used for small-area data.

Capturing spatial similarity

All models require the spatial closeness between each pair of areal units to be defined, and CAR and SAR models use an $K \times K$ *neighbourhood or adjacency* matrix $\mathbf{W}$, where K is the number of areal units in the data set. Then, the kjth element of $\mathbf{W}$ is typically defined as:

$$w_{kj} = \begin{cases} 1 & \text{Areas (k,j) share a border} \\ 0 & \text{Otherwise} \end{cases}$$

so that if $w_{kj} = 1$ data in areal units (k, j) are modelled as spatially autocorrelated, while if $w_{kj} = 0$ they are assumed to be conditionally independent.

My research

Most of my research is interested in developing new statistical methodology for epidemiological applications:

- Identifying trends in disease risk over time and how those trends vary in space.
- Identifying the locations of clusters of small-areas that exhibit substantially higher disease risk than their neighbours.
- Estimating the effects of air pollution on human health.

However, I am also involved in a social science led project AQMeN (*http://aqmen.ac.uk/*) looking at changes in urban segregation over time.

Motivations

My choice of research topics is motivated by two main goals:

1. Finding an important public health problem for which existing statistical models are inadequate, and hence providing an innovative modelling solution.
2. Developing free to use software in `R` for others to be able to apply both standard and novel statistical models for small-area data to their own problems.

Examples of the latter include the `CARBayes` and `CARBayesST` packages in `R` for spatial and spatio-temporal modelling of areal unit data.

This talk

In this talk I give 2 examples of statistical research you can do focusing on addressing key questions in epidemiology, and using small-area data such as that from the Scottish Neighbourhood Statistics database. The examples are:

1. Identifying temporal trends in measles susceptibility over time.
2. Estimating the long-term effects of air pollution on health in Scotland.

3. Susceptibility to measles

- An article published in 1998 by Andrew Wakefield *Wakefield et al. (1998)* linked the measles, mumps and rubella (MMR) vaccine with an increased risk of autism.
- This scare led to a reduction in vaccination rates, which by 2003 reached a low of around 80% in some parts of the UK.
- Repercussions from these decreased vaccination rates were felt in 2013, with large outbreaks of measles occurring in England and Wales (*Pollock et al. 2014*).
- The Wakefield article was partially retracted in 2004 before being fully discredited in 2010 after multiple epidemiological studies failed to find any association.

Data

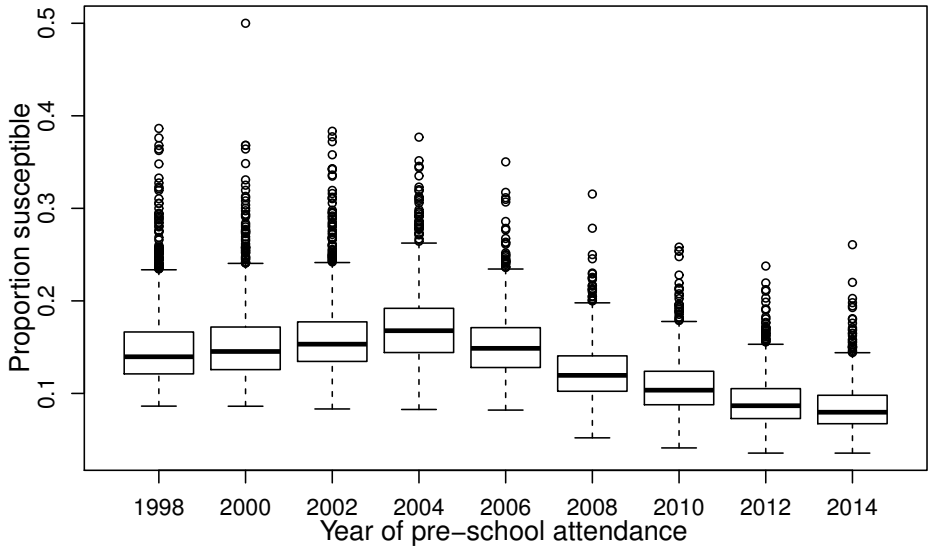
- Since 1988 individual vaccination records of all children in Scotland are kept in the Scottish Immunisation & Recall System (SIRS) .
- The data analysed here are based on the estimated numbers of children eligible to attend pre-school (aged 2-4) from non-overlapping two-year birth cohorts.
- We have the estimated number susceptible to measles (Y_{kt}), and the total number of pre-school children (N_{kt}) in the 1235 (indexed by k) intermediate geographies in Scotland between 1998 and 2014 (indexed by t).
- Thus $\hat{\theta}_{kt} = Y_{kt}/N_{kt}$ is the estimated proportion of children who are susceptible to measles, where measles susceptibility is based upon the receipt of one or two vaccinations that each have a failure rate of 10%.

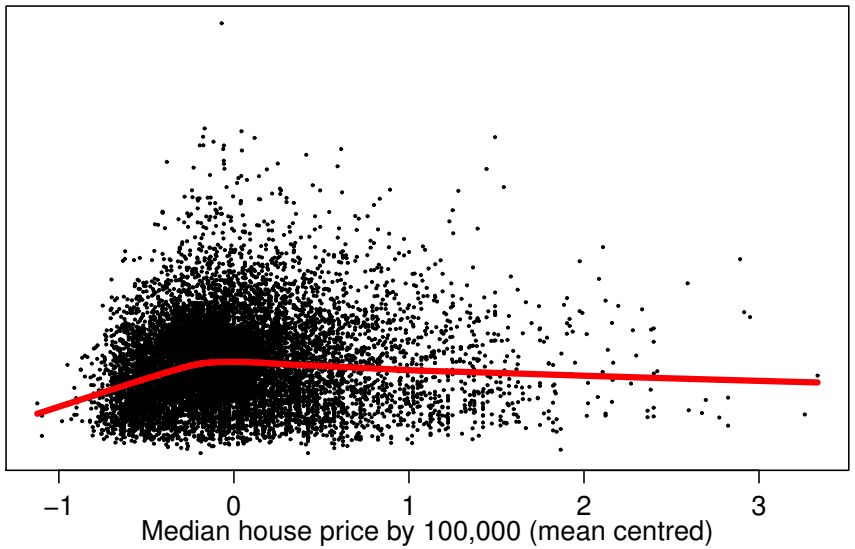
Covariate factors

We use three covariate factors here to capture the potential impacts of socio-economic deprivation and ethnicity:

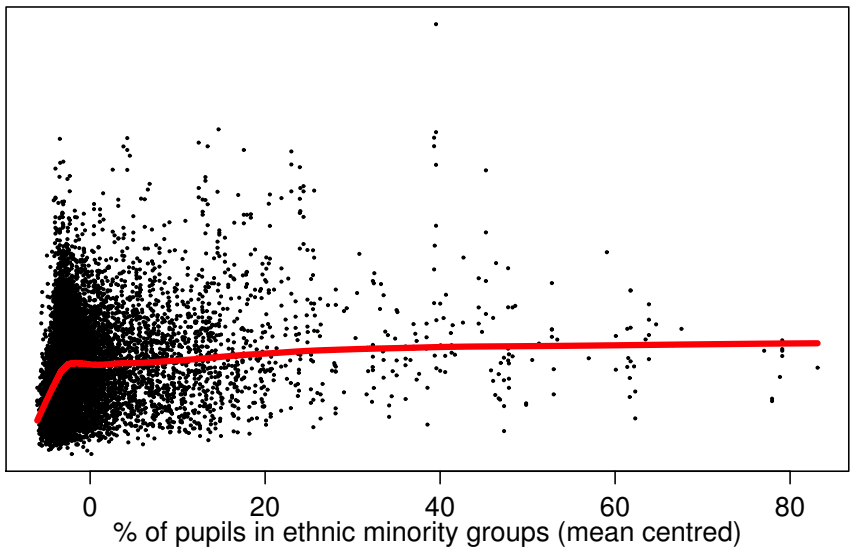
- Median House Price (MHP) in each IG and year.
- Percentage of working age people in receipt of Job Seekers Allowance (JSA) in each IG and year.
- Percentage of school children from ethnic minorities (EM).

Data


Proportion susceptible
0.1
0.2
0.3
0.4
0.5
1998
2000
2002
2004
2006
2008
2010
2012
2014
Year of pre−school attendance


−1
0
1
2
3
Median house price by 100,000 (mean centred)


Proportion susceptible
0.1
0.2
0.3
0.4
0.5
0
5
10
% of working age population claiming JSA (mean centred)


0
20
40
60
80
% of pupils in ethnic minority groups (mean centred)

Exploratory analysis

Initially a simple binomial logistic regression model was fitted to these data:

$$\begin{aligned} Y_{kt} &\sim \text{Binomial}(N_{kt}, \theta_{kt}), \\ \ln\left(\frac{\theta_{kt}}{1-\theta_{kt}}\right) &= \beta_0 + S_1(\text{MHP}_{kt}) + S_2(\text{JSA}_{kt}) + S_3(\text{EM}_{kt}), \end{aligned}$$

where $S_i(\cdot) = \sum_{j=1}^{3} B_i(\cdot)\beta_{ij}$, $i = 1, 2, 3$ is a natural cubic spline of each covariate with 3 degrees of freedom to allow for the non-linear relationships observed above.

Residual autocorrelation

Spatial autocorrelation in the residuals was assessed by computing Moran's I statistics and performing permutation tests on the residuals from the above model separately for each year.

The values of the Moran's I statistics obtained ranged between 0.12 and 0.38, with the statistics generated from 10 000 random permutations of the data yielding p-values of less than 0.0001 in all cases. Thus spatial autocorrelation has to be modelled.

Goals of the analysis

The goals of the analysis are as follows:

1. Estimate the magnitude of the increased measles susceptibility associated with the MMR vaccination scare linked to the Wakefield article and assess whether susceptibility has decreased in recent years.

2. Assess whether the magnitude of the inequality, as measured by the spatial variability, in measles susceptibility in Scotland increased at the same time and whether spatial variation has now decreased.

3. Determine whether any area based covariates, such as deprivation, have any impact on measles susceptibility in Scotland.

Modelling approach

We propose a Bayesian hierarchical model to answer these questions, which is given by:

$$\begin{aligned} Y_{kt}|N_{kt},\theta_{kt} &\sim \text{Binomial}\left(N_{kt},\theta_{kt}\right), \\ \ln\left(\frac{\theta_{kt}}{1-\theta_{kt}}\right) &= \beta_0 + S_1(\text{MHP}_{kt}) + S_2(\text{JSA}_{kt}) + S_3(\text{EM}_{kt}) + \phi_{kt} + \delta_t. \end{aligned}$$

where δ_t is the Scotland-wide temporal trend at time t while ϕ_{kt} is a spatio-temporal effect for area k and time t.

Spatio-temporal trend models

- The overall temporal trend is modelled by a first order random walk process:

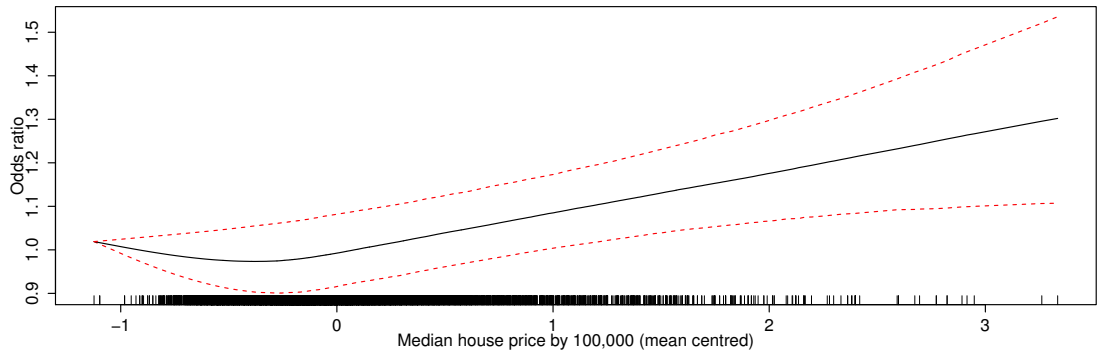
$$\delta_t \sim \mathrm{N}\left(\delta_{t-1}, \sigma^2\right).$$

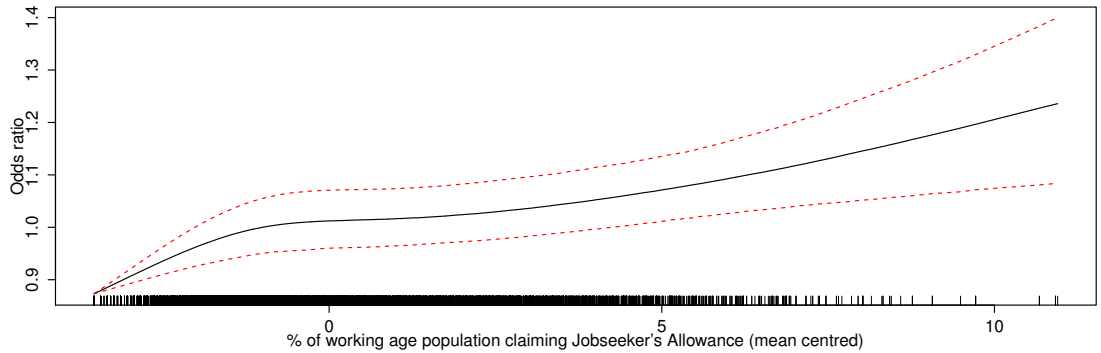
- The spatial trend in year t, $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{Kt})$ is modelled by a conditional autoregressive (CAR) model:

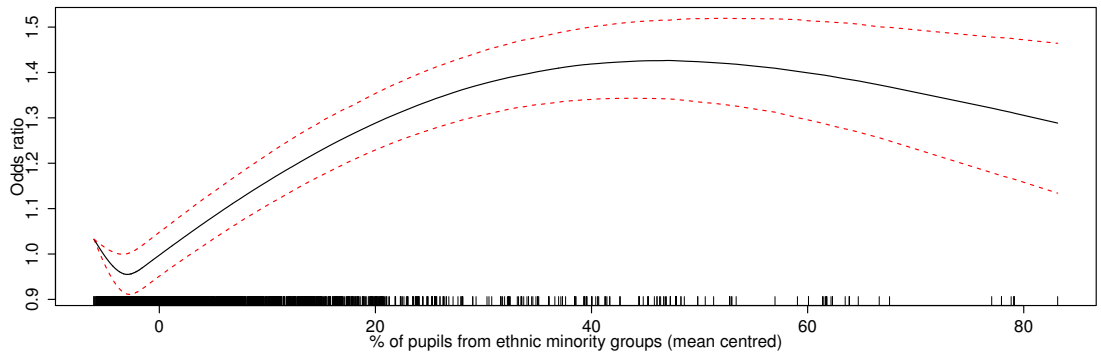
$$\phi_{kt} | \boldsymbol{\phi}_{-kt}, \mathbf{W}, \rho, \tau_t^2 \sim \mathrm{N}\left(\frac{\rho \sum_{j=1}^{K} w_{kj}\phi_{jt}}{\rho \sum_{j=1}^{K} w_{kj} + 1 - \rho}, \frac{\tau_t^2}{\rho \sum_{j=1}^{K} w_{kj} + 1 - \rho}\right),$$

where each year has its own variance τ_t^2 to allow for temporally varying spatial variability.

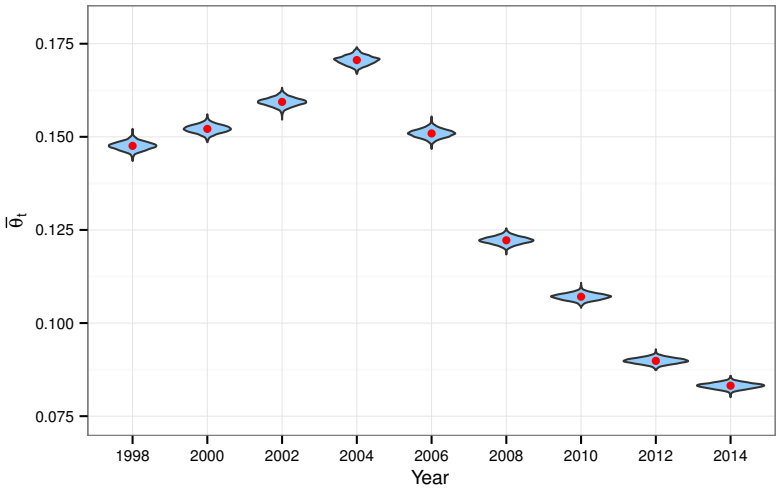
Results - covariate effects

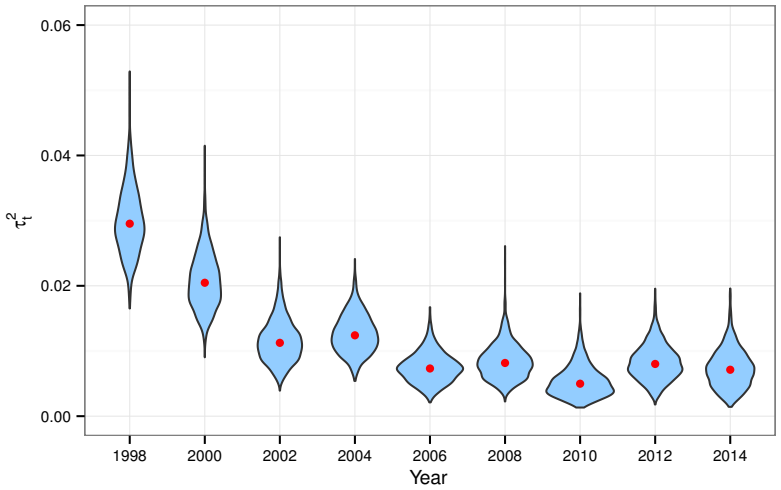


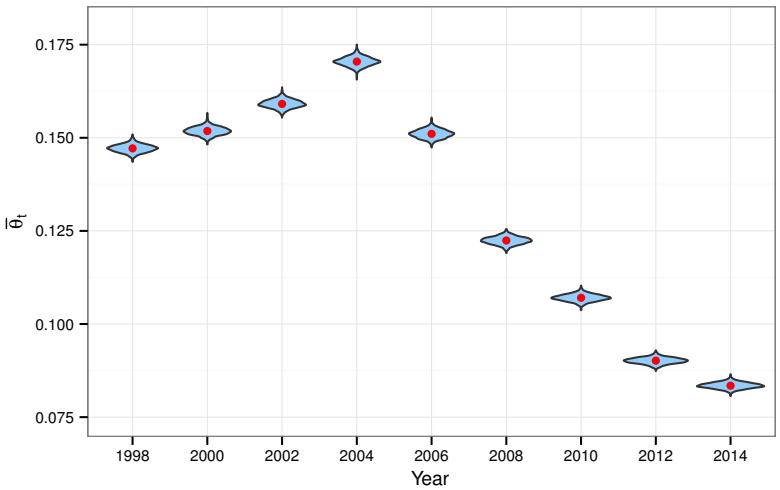


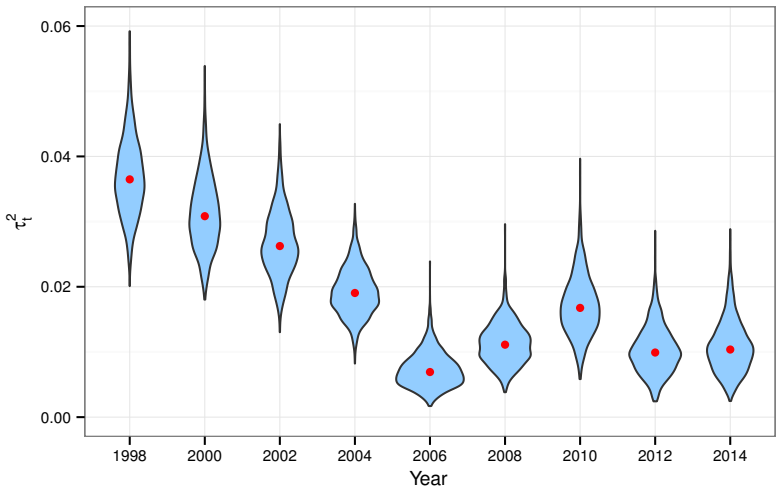


Results - Time trends and spatial variability



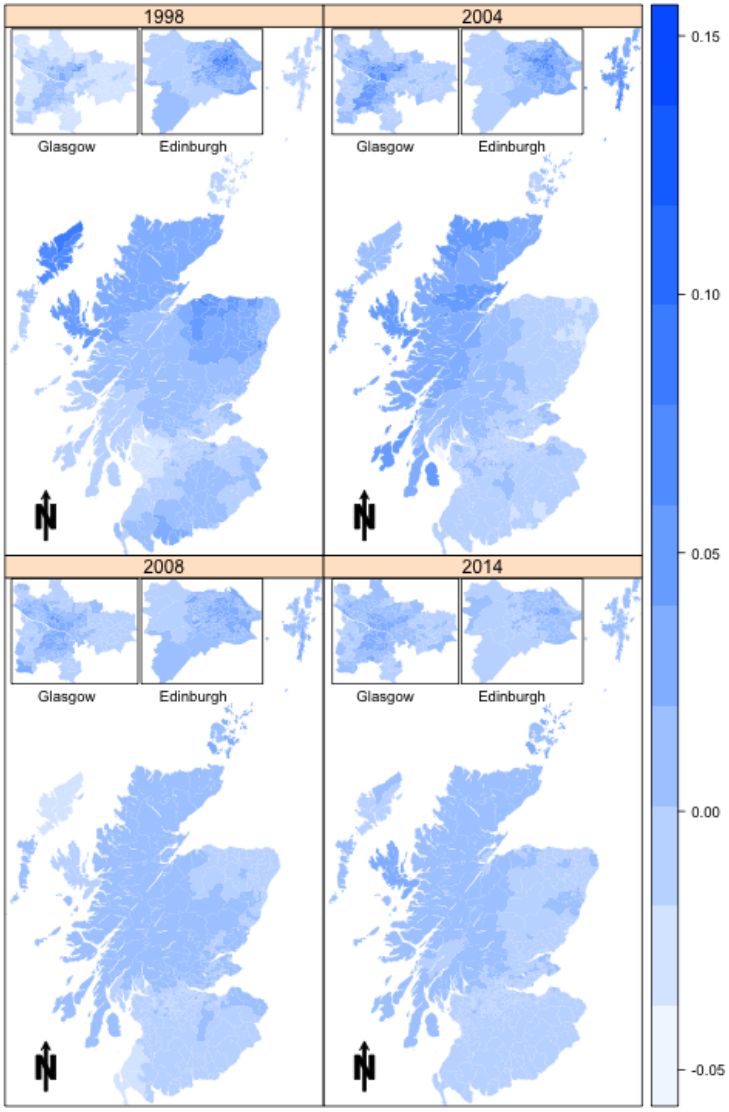






The top row is with covariates and the bottom panel is without.

Results - Spatial trends



Conclusions

- Measles susceptibility increased to a peak in 2004 coinciding with the media coverage surrounding the Wakefield article, before dropping dramatically until the present day.
- Spatial variation in measles susceptibility decreased until 2006, whereafter it has stayed relatively constant over time.
- Socio-economic deprivation appears to have a *U-shaped* relationship with measles susceptibility, with increasing susceptibility for very poor and very affluent communities.
- Spatially, the rural northwest part of Scotland appears to have the highest rates of susceptibility, as it has stayed consistently higher than other parts of the country for all time periods.

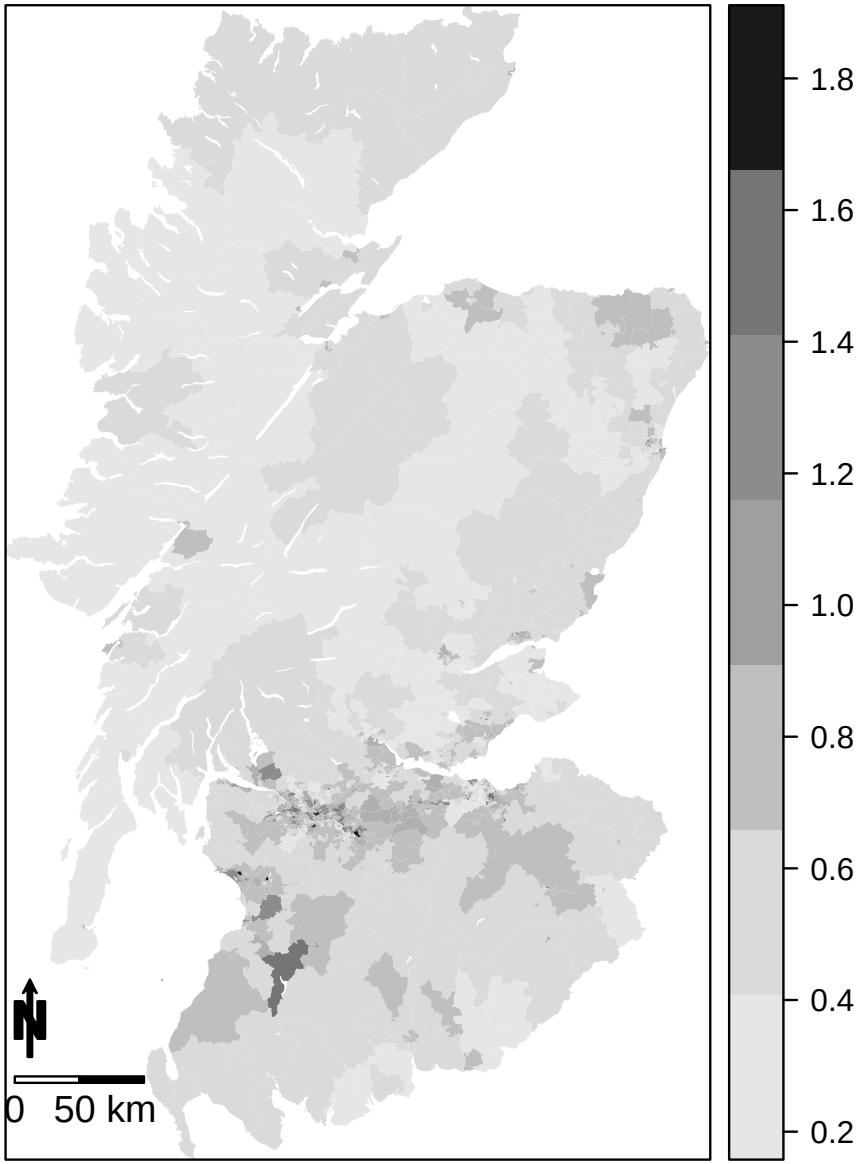
2. Air pollution and health

- 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 2020 in key parts of the UK, including in Glasgow.

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 $K = 1207$ Intermediate Geographies that make up mainland Scotland, and data are collected for $T = 5$ years between 2007 and 2011.
- For IG k and year 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 morbidity ratio is given by $\text{SMR}_{kt} = Y_{kt}/E_{kt}$, an exploratory measure of disease risk.

Spatial pattern in SMR



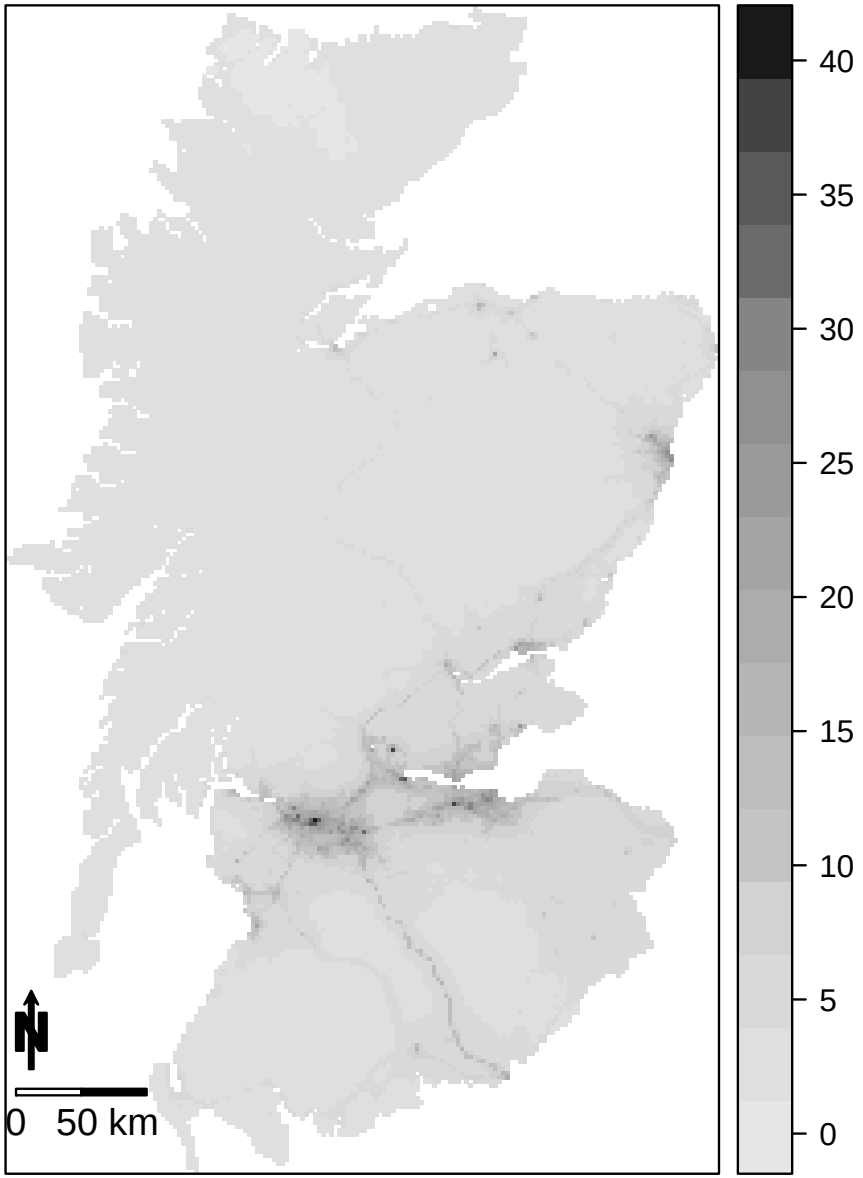
Air pollution

Air pollution concentration data are available from two sources:

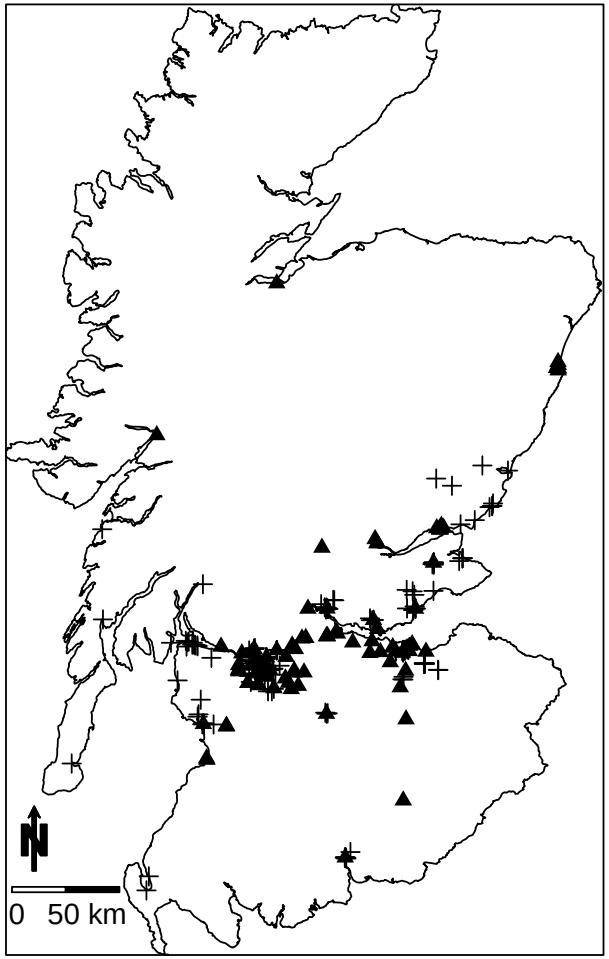
- Measured data from a small number of monitoring sites such as the Automatic Urban and Rural Network (AURN).
- Modelled concentrations from a dispersion model on a 1kilometre regular grid, such as the data provided by DEFRA.

Neither data source is ideal, as the measured data are spatially sparse and the modelled data are only estimates and not real measurements.

Modelled nitrogen dioxide (NO_2) data



Locations of measured NO_2 data



Triangles are monitors and + are diffusion tubes.

Goal 1 - Estimating NO_2 concentrations

The first goal in this study is to estimate representative concentrations of nitrogen dioxide for each IG from these two sets of data. However, there are two natural questions one has to overcome.

- How should the different spatial scales of the three data sets be resolved (point, grid square, IG)?
- How can we use both the modelled and monitored data to get the *best* estimate of NO_2 concentrations.

Fusion modelling

- The accepted approach in the literature is a statistical fusion model, which essentially regresses the measured pollution data against the modelled pollution data.
- Let $\boldsymbol{X}_t = (X_t(\boldsymbol{s}_1), ..., X_t(\boldsymbol{s}_{n_t}))$ denote the vector of n_t measured NO_2 concentrations (on the natural log scale) at sites $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_t})$ in year t
- These measured concentrations are related to an $n_t \times p$ design matrix of covariates $\boldsymbol{Z}_t$, such as the modelled concentrations, site type, etc.

Pollution model

Then the regression component of the model is given by:

$$\boldsymbol{X}_t \sim \mathrm{N}(\boldsymbol{Z}_t\boldsymbol{\beta}_t, \sigma_t^2\boldsymbol{I}_t) \quad t = 1, ..., T,$$

The regression parameters $\boldsymbol{\beta}_t$ and the variance parameter σ_t^2 vary over time via autoregressive processes:

$$\begin{aligned} \boldsymbol{\beta}_t &\sim \mathrm{N}\left(\boldsymbol{\beta} + \kappa(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}), \tau^2\boldsymbol{V}\right) \quad t = 2, ..., T, \\ \ln\left(\sigma_t^2\right) &\sim \mathrm{N}(\ln\left(\sigma_{t-1}^2\right), \sigma^2) \quad t = 2, ..., T, \end{aligned}$$

Note, no spatial autocorrelation is allowed for here because the modelled concentrations are spatially autocorrelated and account for it.

Prediction

- The model is fitted in a Bayesian setting, using Markov Chain Monte Carlo (MCMC) simulation methods.
- It predicts NO_2 concentrations at each 1km grid square in mainland Scotland.
- We aggregate to the IG level by computing the mean and maximum of the set of 1km grid squares in each IG.
- The mean is commonly used in these studies, while the maximum may better represented densely populated parts of each IG.

Goal 2 - Estimating the health effect of NO_2

Recall that (Y_{kt}, E_{kt}) denote the observed and expected numbers of hospital admissions in the kth IG in year t. Then a Poisson generalised linear mixed model is typically used:

$$\begin{aligned} Y_{kt} \mid E_{kt}, R_{kt} &\sim \text{Poisson}(E_{kt}R_{kt}), \\ \ln(R_{kt}) &= \boldsymbol{b}_{kt}^T\boldsymbol{\alpha} + \tilde{X}_{kt}\lambda + \phi_{kt}, \end{aligned}$$

where $\tilde{X}_{kt}$ is NO_2 (mean or max) and $\boldsymbol{b}_{kt}^T$ are other covariates such as socio-economic deprivation. Finally, ϕ_{kt} is a random effect that accounts for any unmeasured spatio-temporal autocorrelation.

Modelling spatial autocorrelation

Our basic model for spatial autocorrelation is a combination of an autoregressive time series process of order 1 and a conditional autoregressive spatial process, and is given by

$$\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1} \sim \mathrm{N}\left(\gamma\boldsymbol{\phi}_{t-1}, \nu^2\boldsymbol{Q}(\rho, \boldsymbol{W})^{-1}\right),$$

where $\boldsymbol{W}$ is a spatial neighbourhood matrix and (γ, ρ) are temporal and spatial autocorrelation parameters respectively. Again, the model is fitted in a Bayesian framework using MCMC simulation, and the `R` package `CARBayesST` can be used to fit this model.

Results

The results are presented as relative risks for a standard deviation increase in each covariates value, which is NO_2 6.84 μgm^{-3}, Logprice 0.38, JSA 2.35.

Parameter	**Spatial mean NO_2**	**Spatial max NO_2**
NO_2	0.993	1.021
	(0.980,1.008)	(1.004,1.037)
Logprice	0.920	0.921
	(0.909,0.929)	(0.911,0.930)
JSA	1.200	1.196
	(1.185,1.215)	(1.180,1.214)
ρ	0.926	0.911
	(0.891,0.956)	(0.866,0.946)
γ	0.832	0.830
	(0.797,0.867)	(0.792,0.865)

Conclusions

- In this study the choice of spatially representative measure of NO_2 had a large impact on the results, with no effect being seen for the spatial mean but a substantial effect being observed for the spatial max.
- Estimating the effect of air pollution on health is a difficult task because of factors such as:
  - Spatial misalignment between the pollution and disease data.
  - The difficult task of controlling for residual spatio-temporal autocorrelation.
  - The lack of real *exposure* data.
- Future work in this field should look at personal exposures and not outdoor measured data at fixed locations.

Talk summary

- In summary, databases of small-area statistics such as SNS provide a valuable resource for addressing key questions of public and policy interest.
- Although this talk has focused on epidemiology, the SNS has data on topics as diverse as property prices, educational attainment, benefit claimants, access to services and demography.
- A key statistical issue is that of spatial autocorrelation, in that simple models ignoring this are likely to produced biased results, particularly in terms of confidence intervals.
- A second issue is that of the modifiable areal unit problem (MAUP), which essentially means do your inferences remain valid if you change spatial scale, e.g. from IG to DZ?

References

- *An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland*, Guowen Huang, et al., Spatial and Spatio-temporal Epidemiology, 2015, 14-15, 63-74.
- *A model to estimate the impact of changes in MMR vaccine uptake on inequalities in measles susceptibility in Scotland*, Gary, Napier, et al, under revision.
- `CARBayes` and `CARBayesST` R packages are available from *https://cran.r-project.org*.
- Further details about my research can be found at *http://www.gla.ac.uk/schools/mathematicsstatistics/staff/duncanlee/*.