Statistical modelling of health-environment relationships: *handling ecological bias*

Nicky Best

Department of Epidemiology and Public Health

In collaboration with Sylvia Richardson, Sara Morris and Chris Jackson
Introduction

Recent trends in environmental epidemiology:

- ↑ use of Bayesian hierarchical models to smooth risk estimates
- ↑ availability of high-resolution geo-referenced health and exposure data
- ↑ use of ecological regression models at small-area scale
Standard ecological regression model:

\[ Y_i \sim \text{Poisson}(E_i \theta_i) \]

\[ \log \theta_i = \alpha + \beta X_i \]

- \( \exp(\beta) = \text{RR associated with exposure } X \)
Interpretation of $\beta$

- $\beta$ measures the **ecological or group-level association** between exposure and disease risk.

- Not necessarily equal to the individual-level association $\rightarrow$ **ecological bias**

- Can arise for various reasons, including:
  - non-linear exposure-response relationship, combined with within-area variability of exposure
  - within- and between-area confounding
  - area-level effect modification
  - spatial dependence in the residuals
Spatially dependent residuals

- Observed risk factor(s) $X$ are unlikely to explain all of the between-area variation in risk
- Residual variation (in excess of Poisson) can be captured using hierarchical model
  \[
  Y_i \sim \text{Poisson}(E_i \theta_i)
  \]
  \[
  \log \theta_i = s_i + \beta X_i
  \]
  \[
  s_i \sim \text{spatial prior}
  \]
- $s_i$ acts as proxy for unidentified area risk factors
  - captures *residual variation* in risk not explained by $X$
  - necessary for correct estimation of *precision* of exposure effects, $\beta$ (see Wakefield 2003)
Area-level effect modification

• Effect modification occurs when dose-response relationship depends on level of a third variable

• Common examples of (individual) effect modifiers:
  – age, time, genetic predisposition

• In ecological regression, may suspect geographical heterogeneity in effects of various risk factors:
  – subgroups of people particularly susceptible
  – interaction with location due to contextual effects, effectiveness of health system, socio-economic / cultural / ethnic factors, …..
Spatially varying coefficient models

• Instead of letting spatial structure only influence residual effects in ecological regression model
  → introduce spatial structure on covariate effects, $\beta$
  → spatially varying coefficient models

• Widely used in econometrics and geography (Assunçao, 2003)

• Links with generalised additive models (gam) where regression coefficients vary as smooth function of other variables (effect modifiers)
Statistical model:

\[ Y_i \sim \text{Poisson}(E_i \theta_i) \]
\[ \log \theta_i = s_i + \beta_i X_i \]
\[ s_i \sim \text{spatial prior} \]
\[ \beta_i \sim \text{spatial prior} \]

- \( \exp(\beta_i) = \text{RR associated with exposure } X \text{ in area } i \)
  - assumed to vary across areas
  - location is acting as proxy for effect modifier
Questions

• How well do models capture different patterns of covariate effects when they are present?

• Can the spatial structure of the coefficients be distinguished from that of the spatial residuals?

• Do these models “invent” spatial structure in the coefficients, even when it is not there?

• How much is lost by over-fitting: constant vs varying coefficient model?
# Simulation study

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression coefficient</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constant</td>
<td>Spatial</td>
</tr>
<tr>
<td>2</td>
<td>Random</td>
<td>Spatial</td>
</tr>
<tr>
<td>3</td>
<td>Spatial (Patchy)</td>
<td>Spatial</td>
</tr>
<tr>
<td>4</td>
<td>Spatial (Smooth)</td>
<td>Spatial</td>
</tr>
</tbody>
</table>

- **Geographical scenario**: 525 wards in NW London
- **Expected counts**: based on real population and incidence of rare cancer (median $E_i = 2.8$)
- **50 datasets** simulated for each model using $E_i \times 1$ and $E_i \times 5$
- Analysed using (i) **spatially varying** coefficient model and (ii) **constant** coefficient model
Regression coefficients

- True RR (constant)
- Average estimated RR (E x 1)
- Average estimated RR (E x 5)

Color scale:
- 0.4 to 0.8
- 0.8 to 3.3
- 1.134 to 1.146
- 1.146 to 1.17
True RR (Random)

Average estimated RR (E x 1)

Corr=0.54

Average estimated RR (E x 5)

Corr=0.69
True RR (Patchy A)

Average estimated RR (E x 1)
Corr=0.6

Average estimated RR (E x 5)
Corr=0.76
True RR (Patchy B)

Average estimated RR (E x 1)

True RR (Patchy C)

Average estimated RR (E x 1)

Corr=0.75

Corr=0.79
True RR (Smooth)

Average estimated RR (E x 1)
Corr=0.99

Average estimated RR (E x 5)
Corr=0.99

Legend:
- 0.4
- 0.8
- 3.3
Spatial residuals

True Residual

Posterior mean

True Residual

Posterior mean

Legend:

0.2 0.7 4.5
Model comparison

- Compare constant and varying coefficient models using **Deviance Information Criterion (DIC)**
  - similar to AIC but for hierarchical models
  - model with **smaller DIC** is preferred

<table>
<thead>
<tr>
<th>Model</th>
<th>True coefficient</th>
<th>DIC\textsubscript{var} – DIC\textsubscript{const} (10 datasets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constant</td>
<td>-0.9 to 1.3 across datasets</td>
</tr>
<tr>
<td>2</td>
<td>Random</td>
<td>-23 to -55 across datasets</td>
</tr>
<tr>
<td>3</td>
<td>Spatial (Patchy)</td>
<td>-19 to -42 across datasets</td>
</tr>
<tr>
<td>4</td>
<td>Spatial (Smooth)</td>
<td>-32 to -90 across datasets</td>
</tr>
</tbody>
</table>
Non-linear exposure-response

- Most epidemiological models assume multiplicative relationship between exposure and risk
  - Individual and group (area)-level relationships have different functional form, e.g.

  $x_{ik}$ is a binary exposure for person $k$ in area $i$
  $p_{ik} =$ risk of person $k$ developing (rare) disease
  $\log p_{ik} = \alpha + \beta x_{ik} \Rightarrow p_{ik} = e^\alpha$ if unexposed; $p_{ik} = e^{\alpha + \beta}$ if exposed
  $X_i =$ proportion of people exposed to $x$ in area $i$ (mean of $x_{ik}$)
  $\theta_i =$ average risk of disease in area $i$

  $\theta_i = e^\alpha (1-X_i) + e^{\alpha + \beta} X_i = e^\alpha (1 + (e^\beta - 1) X_i)$
  $\Rightarrow \log \theta_i \neq \alpha + \beta X_i$ (unless $X_i = 0$ or 1)
• Similar result holds for continuous exposures
  – For example, if the $x_{ik}$ are approximately Normally distributed with mean $X_i$ and variance $V_i$ in area $i$, and
  \[ \log p_{ik} = \alpha + \beta x_{ik} \] as before, then
  \[ \log \theta_i = \alpha + \beta X_i + \frac{\beta^2 V_i}{2} \neq \alpha + \beta X_i \] (unless $V_i = 0$)

• If exposure varies within areas, and multiplicative risk model holds at individual level
  – appropriate integrated (aggregated) functional form should be used for the ecological regression model

• Even if correct model used, ecological data often contain little information about some of the risks
Simulation study to investigate bias of \( \beta \) coefficient in ecological regression model with binary exposure.

<table>
<thead>
<tr>
<th>Exposure range</th>
<th>(Areas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 25%</td>
<td>(100 areas)</td>
</tr>
<tr>
<td>0 - 50%</td>
<td>(100 areas)</td>
</tr>
<tr>
<td>0 - 100%</td>
<td>(100 areas)</td>
</tr>
<tr>
<td>0 - 25%</td>
<td>(25 areas)</td>
</tr>
</tbody>
</table>
Simulation study to investigate bias of $\beta$ coefficient in ecological regression model with continuous exposure

- True log RR
- Moderate between-area exposure variance (CV=22%) (100 areas)
- Small between-area exposure variance (CV=6%) (100 areas)
- Moderate between-area exposure variance (CV=22%) (25 areas)

log RR (95%CI) of disease per unit increase in x
• Substantial improvements can be achieved by including individual-level data on a small sub-sample of people in each area

• Simultaneously estimate individual-level and ecological regressions (easily implemented in Bayesian paradigm)
Effect of including sample of 10 individuals per area on estimates of binary exposure effect

- Log RR (95%CI) of disease for exposed individuals
- Interaction (modelled)
- Interaction (ignored)
- Ve correlation (modelled)
- Ve correlation (ignored)
- Exposure 0 - 25%
- Exposure 0 - 50%
- Exposure 0 - 100%

25 areas

Log RR (95%CI) of disease for exposed individuals
Conclusions

• Bayesian hierarchical models allow “borrowing of information” about disease risk across areas.

• This property allows estimation of varying coefficient models:
  – Models have reasonable power to detect true spatial variation in covariate effects, even with sparse data.
  – Over-fitting does not appear to be a problem when no effect modification is present.
  – Able to separate spatial pattern of effect modification from that of the residuals.
Conclusions continued

• When exposures vary within areas, care needed to fit appropriate aggregated risk model
  – Need large exposure contrasts between areas
  – Inclusion of even small sub-samples of individual level data can reduce bias and improve precision
  – More work needed on optimal study design

• Combining varying coefficient models and individual sub-samples should further improve our ability to handle ecological bias
Thank you