Aggregated cancer incidence data: spatial models

$5^{ième}$ Forum du Cancéropôle Grand-est - November 2, 2011

Erik A. Sauleau

Department of Biostatistics - Faculty of Medicine
University of Strasbourg

ea.sauleau@unistra.fr

Haut-Rhin Cancer Registry
Mulhouse

Outline

Outline

Context

- Cancer registries collect exhaustively and actively individual data on new cases
  - Focus on incidence $\Rightarrow$ aggregated outcome
  - Small area data
- Measure of relative risk = Standardized Incidence Ratio (SIR)
  - Ratio of the observed cases in each geographical unit on the expected cases:
  $$\mathrm{SIR}_i = \frac{O_i}{E_i}$$
  where $E_i = \hat{p}_i N_i$
  - What about these $\hat{p}_i$?
    1. Global risk in the study region
    $$\forall i, \hat{p}_i = \hat{p} = \frac{\sum\sum\cdots\sum O_.}{\sum\sum\cdots\sum N_.}$$
    2. Adjusted risk on several categorical variable(s)

Very brief history of mapping

- Spot map: cholera and water street pumps (Snow, 1854)
- Choropleth map: disease mortality in England and Wales (Haviland, 1878)
- Recent developments enhanced by
  - Development of Geographical Information Systems (GIS)
  - Increasing availability of spatially-referenced data
  - Development of statistical methods

Standard practice was (is?) to map risks per small area BUT sparse data need more sophisticated statistical analysis techniques

Outline

1. Why mapping small area incidence rates?

- Mapping geographical variations in health outcomes
  - Sources of heterogeneity and spatial patterns
  - Suggest public health determinants
  - Etiological clues
- Small scale
  - Less susceptible to ecological bias
  - Better able to detect highly localised effects

2. Why smoothing small area incidence rates?

1. Rare events ⇒ imprecision: $\hat{\sigma}_{\text{SIR}} \propto \frac{1}{\text{E}}$
   - SIR very imprecise for rare disease and small population
   - Precision can vary widely between geographical units
2. SIR in each geographical unit is estimated independently
   - Ignores possible spatial correlation (see further)
   - Problem of multiple significance testing

⇒ These problems may be addressed by spatial smoothing of the crude data

3. How smoothing small area incidence rates?

- Idea is to "borrow information" for neighbouring geographical units to produce better estimates of the risk
- Different methods
  - Local smoothing algorithms (spatial moving averages)
  - Trend surface (kriging, spline)
  - Random effects models (empirical Bayes, Bayes)

Outline

Autocorrelation: definition

- Phenomenon ”is much more alike” between two neighbouring geographical units than between two random geographical units
- Neighbourhood $\rightarrow$ sharing a common boundary
- Assessment by a statistic like Moran’s I
- SIRs are spatially correlated because they reflect (?) supra-small area level spatially varying risk factors

$\Rightarrow$ Incorporate spatial correlation in the modelling of SIRs

General spatial model with autocorrelation

Poisson regression for SIRs

- Assume Poisson sampling for count (random variable)

$$O_i \sim \mathcal{P}(\mathrm{E}_i \theta_i)$$

$\Rightarrow$ $\log\left[\mathbb{E}(O_i|\theta_i)\right] = \log\left(\mathrm{E}_i\right) + \log\left(\theta_i\right)$

$\rightarrow$ Generalised linear model (GLM)

1. $\log\left(\mathrm{E}_i\right)$ is an offset
2. Then $\log\left(\theta_i\right)$ is something like $\mu + U_i$

- Spatial structure on U_i
  1. Gaussian Markov Random Field = Intrinsic Conditional AutoRegressive process
  2. Geospline
- Bayesian or frequentist inference?

ICAR and convolution prior

- Intrinsic conditional autoregressive process

$$U_i|\mathbf{U}_{-i} \sim \mathcal{N}\left(\frac{\sum_{j\in\partial} U_i}{n_i}, \frac{\sigma^2_{\mathbf{U}}}{n_i}\right)$$

  - n_i: number of neighbours around i
  - Mean: average risk in neighbouring
  - Variance inversely proportional to number of neighbours
- Proxy for unobserved covariates which, if observed, would display a spatial autocorrelation
- What about proxy for unobserved covariates which, if observed, would not display a spatial autocorrelation?

$\Rightarrow$ Add a second term for "heterogeneity": $V_i \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{V}}\right)$

Aggregated spatial data as continuous

- Geographical unit $\Leftrightarrow$ coordinates of its centroid
- Spatial trend $U_i = \alpha \cdot \mathrm{lon}_i + \beta \cdot \mathrm{lat}_i$
- Bi-dimensional smoothing is much more powerful (if necessary)

$\Rightarrow$ Geospline and generalised additive mixed models (`GAMM`)

- Thin plate spline (isotropic)
- Tensor product of cubic P-splines

$$U_i = \sum^{m_1} \sum^{m_2} \omega_{jk} \mathrm{a}_j(\mathrm{lon}_i) \mathrm{b}_k(\mathrm{lat}_i)$$

- Idem in Bayesian inference but on a regular grid with random walk priors on the ωs

Autocorrelation: summary

- Aggregated spatial data (adjacency matrix)
  - Bayesian ICAR or convolution prior
- ”Continuous” spatial data (centroids)
  - Geospline (Bayesian or frequentist)
  - Distance model $\approx$ geostatistics (Bayesian or frequentist)

Outline

General principle

- Assume a continuous explanation covariate (even latent) known by geographical unit, say `Townsend`
- So simple to incorporate covariates in previous GLM, GLMM or GAMM

$$\log\left[\mathbb{E}(O_i|\boldsymbol{\theta}_i)\right] = \begin{cases} \log\left(\mathrm{E}_i\right) + \mu + U_i & \text{(previously)} \\ \log\left(\mathrm{E}_i\right) + \mu + U_i + \beta\cdot\texttt{Townsend}_i & \text{(from now)} \end{cases}$$

- More general: structured additive regression model (**StAR**)

$$\log\left[\mathbb{E}(O_{ij}|\boldsymbol{\theta}_{ij})\right] = \log\left(\mathrm{E}_{ij}\right) + \mu + U_i + \boldsymbol{\beta X} + \sum^{K} f_k(\tilde{x}_{ij})$$

where

j Stands for combination of the strata of covariates we are interest in (even if $x_{ij} = x_i, \ \forall j$)

$f_k(\cdot)$ May be multidimensional (or not) smoothing function

Modelling a covariate x

1. Main effect
   - Categorical: dummy-variables and "fixed" effect $\sum^{C-1} \beta_c I(x = c)$
   - Ordinal: recoding with contrast or as discrete with scoring
   - Discrete: "fixed" effect βx
   - Continuous: "fixed" effect βx or smoothing $f(x)$ (for example spline)
2. Interaction with y
   - "Fixed" effect: $\beta \cdot x \cdot y$
   - Varying coefficient model: $x \cdot f(y)$ or $y \cdot f(x)$
   - Multidimensional smoothing: $f(x, y)$ (even if y is geospline)

Strongly depends on how the covariate is and on the aim of modelling

Some more issues

- Adjusted relative risk (multiplicative assumption)
  - $\log(\mathrm{E}_i) + \mu + U_i + V_i + \beta \cdot \texttt{Townsend}_i$
  - $\exp(\beta)$ is the spatially adjusted relative risk of `Townsend`
  - $\exp(U_i + V_i)$ is the global adjusted spatial relative risk
- Misalignment: different scales for variables
- Spatial autocorrelation of O and of `Townsend` $\Rightarrow$ spatial confounding
  - Introduce or remove bias in estimating β
  - "Restricted spatial regression"

Outline

Summary

- Small area estimation
  - Less susceptible to ecological bias
  - Better able to detect highly localised effects
  - Supra-small area risk factors
  - Need for spatial smoothing
- Freeware: `R` (`mgcv`, `INLA`), `WinBUGS`, `BayesX`

A super-short example

Ear-Nose-Throat cancer in Haut-Rhin

- New cases between 01/01/1988 and 31/12/2005
- 12,580,392 people at risk
- Small area: commune of residence
- 3,304 male and 516 female
- A best model includes:
  - Different age-time smoothing surfaces for male and for female
  - Geospline
- Adjusted relative risk

	Minimum	Maximum	Median
Male	0.002	3.07	1.01
Female	0.005	36.9	6.10
Commune	0.651	1.53	1.02

A super-short example

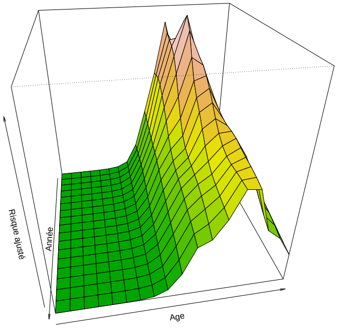
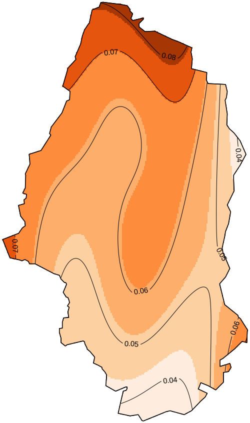


Figure: Age-time smoothing surfaces: male (left) and female (right)

A super-short example



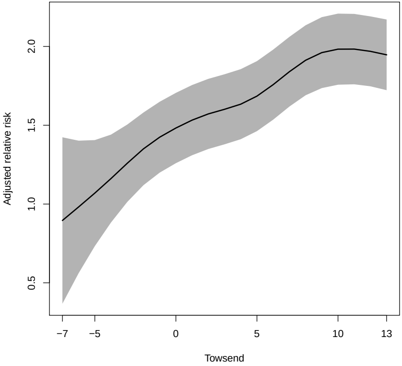


Figure: Substituting spatial effect (left) with deprivation index (right)

Outline

- Books

- Banerjee S, Carlin BP, Gelfan AE. *Hierarchical modeling and analysis for spatial data*. Monographs on Statistics and Applied Probability 101, Chapman and Hall/CRC, Boca Raton, 2004.
- Cressie NAG. *Statistics for spatial data*, Wiley, New York, 1993.
- Rue H, Held L. *Gaussian Markov random fields: theory and applications*. Monographs on Statistics and Applied Probability 104, Chapman and Hall/CRC, Boca Raton, 2005.
- Rupper D, Wand MP, Carrol RJ. *Semiparametric regression*. Cambridge University Press, Cambridge, 2003.
- Waller LA, Gotway CA. *Applied spatial statistics for public health data*. John Wiley and Sons, Hoboken, New Jersey, 2004.
- Wood S. *Generalized additive models: an introduction with R*. Chapman and Hall / CRC, Boca Raton, 2006.

- Papers

Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991:**43**,1-59.

Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987:(**43**),671-81.

Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data : a bayesian perspective. *Statistica Sinica* 2004:(**14**),715-45.

Lang S, Brezger A. Bayesian P-splines. *Journal of Computational and Graphical Statistics* 2004:(**13**),183-212.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B* 2009:(**71**),319-392.

Wood S. Thin plate regression splines. *J. R. Statist. Soc. B* 2003:**65**,95-114.

Wood S. Low-rank scale invariant tensor product smooths for generalized additive mixed models. *Biometrics* 2006:**62**,1025-36.

Thank you for your patient attention

It seems to be a law of science that no discovery or invention is named after its first discoverer.
Stigler's Law of Eponymy, Stigler 1980.

$\Rightarrow$ Who was the first to discover Bayes's Theorem?