Spatio-temporal Methods in Environmental Epidemiology

Alexandra M. Schmidt

Session 8 - Modelling areal data

Department of Epidemiology, Biostatistics and Occupational Health
McGill University

alexandra.schmidt@mcgill.ca

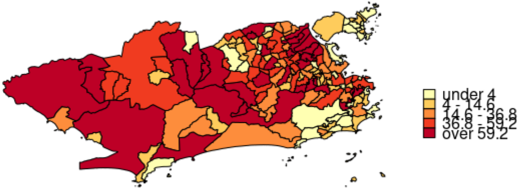
Applied Bayesian Statistics School 2025

University of Genova, Department of Architecture and Design
03-06, June 2025

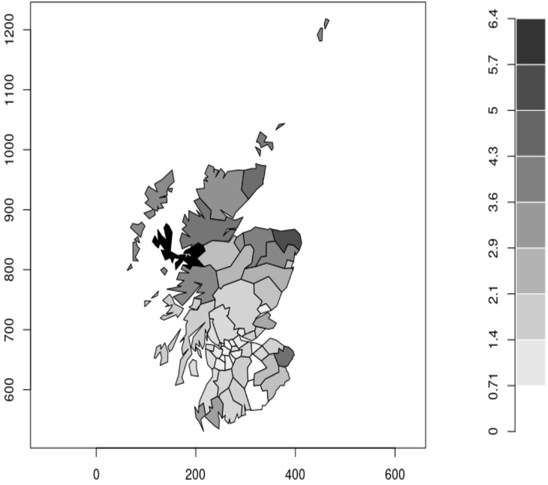
Definition

- Nontrivial observations are taken at a finite number of sites whose whole constitutes the entire study region
- Data locations may be points or regions, but most cases where the data are points can be handled using geostatistics
- We shall focus primarily on situations where data locations are regions
- Examples: Presence, absence of a plant species in square quadrats over a study area; number of deaths due to AIDS in the counties of North Carolina; Pixel values from remote sensing (satellites)

Distribution of Number of Cases of Dengue Fever



Standardized Mortality Ratio Lip Cancer in Scotland



Exploratory Data Analysis

- For regions lying on a regular grid, methods such as plots of row mean or column mean versus row index or column index, median polish, and same-lag scatterplots are applicable
- For irregularly spaced regions, the most useful methods includes:
  - stem-and-leaf plots
  - 3-D scatterplot
  - Variogram cloud and sample variogram
  - Plots of each datum against the average of its nearest neighbors
  - Grayscale maps
  - Plot of response versus area of region

Measures of Spatial Autocorrelation

- Our objective is to measure how strong the trend is for observations from nearby regions to be more (or less) alike than observations from regions farther apart, and then judge whether any apparent trend is sufficiently strong that it is unlikely to be due to chance alone

- We shall focus on spatial autocorrelation measures for two types of response variables:
  - Binary (0-1)
  - Continuous

Measures of Spatial Autocorrelation

The General Cross-Product Statistic

Notation:

- Let Z_i denote the response at the *ith* location ($i = 1, \cdots, n$)
- Let Y_{ij} be a measure of how similar or dissimilar the responses are at locations i and j
- Let W_{ij} be a measure of the spatial proximity of locations i and j
- For future reference, define matrices $\mathbf{Y} = (Y_{ij})$ and $\mathbf{W} = (W_{ij})$

The general cross-product statistic is

$$C = \sum_i \sum_j W_{ij} Y_{ij}$$

Measures of Spatial Autocorrelation

The General Cross-Product Statistic

- Hypothetical example, with binary Z_i's, $Y_{ij} = (Z_i - Z_j)^2$, and

$$W_{ij} = \begin{cases} 1, & \text{if locations } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases}$$

Note:

- C too small ⇒ positive spatial autocorrelation
- C too large ⇒ negative spatial autocorrelation

Measures of Spatial Autocorrelation

How the statistical significance of C should be judged? Comparison

to randomization distribution

- list all possible arrangements of the observed responses over the locations obtained by permutation of responses
- compute C for each arrangement, and rank these
- determine where the data's C-values fits in; P-value for the test is the number of C-values in the randomization distribution as extreme or more extreme than the observed C
- can do one-sided or two-sided tests

Measures of Spatial Autocorrelation

Moran's and Geary's Statistics (Continuous Data)

- Moran (*Biometrika*, 1950) proposed the following statistic which can be used for continuous data:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j W_{ij}(Z_i - \overline{Z})(Z_j - \overline{Z})}{\sum_i (Z_i - \overline{Z})^2}$$

  where $\overline{Z} = \sum_i Z_i / n$, $S_0 = \sum_i \sum_j w_{ij}$.

- Geary proposed a similar statistic:

$$c = \frac{n-1}{S_0} \frac{\sum_i \sum_j W_{ij}(Z_i - Z_j)^2}{\sum_i (Z_i - \overline{Z})^2}$$

- Remarks:
  - Note the superficial resemblance of I to the ordinary correlation coefficient
  - Note that $I = \frac{n}{S_0 \sum_i (Z_i - \overline{Z})^2} . C$ if we take $Y_{ij} = (Z_i - \overline{Z})(Z_j - \overline{Z})$

Measures of Spatial Autocorrelation

- Similarly, c may be related to C by taking $Y_{ij} = (Z_i - Z_j)^2$
- Consequently, I is more sensitive to extreme Z-values whereas c is more sensitive to differences between pairs of Z-values
- However, I is more popular than c
- $E(I) = -\frac{1}{n-1}$ under independence
- $I > -\frac{1}{n-1} \Rightarrow$ positive autocorrelation
- $I < -\frac{1}{n-1} \Rightarrow$ negative autocorrelation

Measures of Spatial Autocorrelation

- Normal approximation to distribution of I under independence ($n > 25$):
  - $E(I)$ as given above
  - $$var(I) = \frac{n[(n^2-3n+3)S_1 - nS_2 + 3S_0^2]}{n(n-1)(n-2)} - \frac{k[n(n-1)S_1 - 2nS_2 + 6S_0^2]}{n(n-1)(n-2)} - \frac{1}{(n-1)^2}$$

  where $k = \frac{n\sum_i (Z_i - \overline{Z})^4}{(\sum_i (Z_i - \overline{Z})^2)^2}$, $S_1 = \frac{1}{2}\sum_i \sum_j \left(w_{ij} + w_{ji}\right)^2$,
  $S_2 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji}\right)^2$

- for smaller samples sizes, can use randomization distribution or Monte Carlo approach to evaluate significance.

Measures of Spatial Autocorrelation

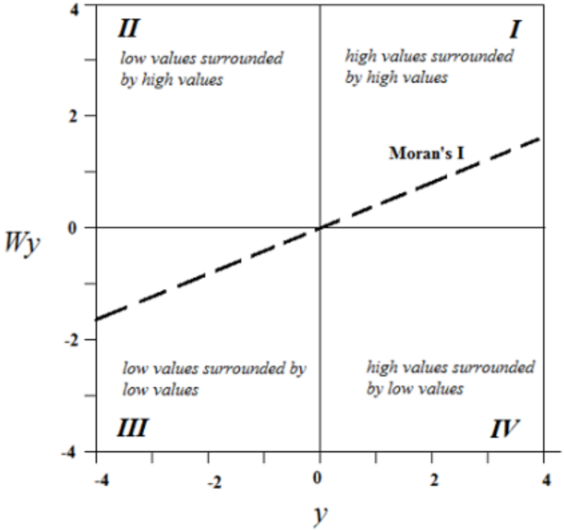
- So far we have assumed that W_{ij}'s are binary. This is rather crude. In many situations we may be able to measure spatial proximity on a more refined scale. Possible refinements:
  1. Use lengths of common boundary; this may more accurately reflect the amount of inter-site interaction
  2. Use actual distance between locations or centroids of locations, e.g.: the inverse Euclidean or city-block distance between locations. This recognizes the fact that interaction between sites not usually terminate sharply beyond places that share a boundary
  3. Incorporate directionality (e.g. upstream vs. downstream) by allowing $W_{ij} \neq W_{ji}$
- Non-binary W_{ij}'s : dist. of the test statistic under indep. is better approx. by the normal

Moran scatter plot

- The Moran scatter plot provides a visual representation of spatial associations in the neighborhood around each observation
- Depending on their position on the plot, the Moran plot data points express the level of spatial association of each observation with its neighboring ones
- You can find the data points on the Moran scatter plot in any of the four quadrants defined by the horizontal line and the vertical line
  - Points in the upper right (or high-high) and lower left (or low-low) quadrants indicate positive spatial association of values that are higher and lower than the sample mean, respectively
  - The lower right (or high-low) and upper left (or low-high) quadrants include observations that exhibit negative spatial association; that is, these observed values carry little similarity to their neighboring ones.

Moran scatter plot

See function `moran.plot` in the `spdep` package


II
low values surrounded by high values
I
high values surrounded by high values
Moran's I
low values surrounded by low values
III
high values surrounded by low values
IV
Wy
4
2
0
-2
-4
y
-4
-2
0
2
4

Disease Mapping

- Disease maps are useful for:
  - Description of geographical distribution of disease
  - Surveillance, to highlight areas at apparently high risk
  - Aiding policy formation and resource allocation
  - Placing point source/disease cluster investigations in context
  - More recently, disease mapping methods have been used to investigate specific hypotheses concerning health and the environment.

Disease Mapping

Disease mapping may be carried out at a variety of scales:

- International comparisons between countries
  - Large international differences in rates of heart disease
  - High rates of liver cancer in Africa and SE Asia, related to Hepatitis B infection
- National comparisons between e.g. counties or states
- Small area studies

Disease Mapping - Some notation

- Map is divided into n contiguous areas labelled $i = 1, \dots, n$
- Let $y = (y_1, \dots, y_n)$ denote $\sharp$ of deaths from the disease of interest
- 'Expected' number of deaths $(e_1, \dots, e_n)$ can be calculated as

$$e_i = \frac{\sum_i y_i}{\sum_i pop_i} \times pop_i.$$

- This e_i corresponds to a kind of "null hypothesis", where we expect constant disease rate in every county
- A better approach might be to make reference to an existing role standard table of age-adjusted rates for the disease. After stratifying the population by age group the e_i emerge as $e_i = \sum_j n_{ij} r_j$
  - n_{ij} is the person-years at risk in area i for age group j (i.e., the number of persons in age group j who live in area i times the number of years in the study), and r_j is the disease rate in age group j (taken from the standard table)

Disease Mapping - Some notation

- When disease is non-contagious and rare, number of deaths in each area are assumed to be mutually independent and follow a Poisson distribution
- y_i has mean $e_i r_i$, $r = (r_1, \ldots, r_n)$ are the unknown area-specific relative risks, and the likelihood of the relative risk r_i is:

$$y_i | r_i = \exp\{e_i r_i\} \frac{(e_i r_i)^{y_i}}{y_i!}$$

Disease Mapping

- MLE of r_i, standardized mortality ratio (SMR), for the i^{th} area, is: $\hat{r}_i = y_i/e_i$
- Naive use of disease mapping can be very misleading
  - Maps often show raw SMRs (y_i/e_i): $SE(SMR_i) \propto 1/e_i$
  - for small areas and/or rare diseases, risk estimates are very unstable
  - Makes no use of data on disease rates in other areas
  - Problems of multiple significance testing
  - Data often exhibit extra-Poisson variation
  - Estimation methods often ignore spatial dependence between disease rates in nearby areas

Disease Mapping - Smoothing of SMRs

Statistical **smoothing** or **shrinkage** methods can be used to overcome these problems:

- Idea is that smoothed estimate for each area ’borrows strength’ (precision) from data in other areas, by an amount depending on the precision of the raw estimate for each area
- Rate in area *i* is estimated by taking a form of weighted average of:
  - observed rate (SMR) in that area
  - mean rate in surrounding areas

Disease Mapping

- Rates can be smoothed either towards:
  - mean of overall map (*global* smoothing) to account for overdispersion
  - mean of nearby areas only (*local* smoothing) to account for spatial dependence in rates

Disease Mapping - Statistical Model

Smoothing is best carried out using Bayesian methods

- Bayesian inference combines:
    - Information from the data (likelihood)
    - Prior distribution for the disease risks to obtain a posterior distribution for the risks, the mean or the median of which is used as the point estimate for each area
- The prior distribution contains information about how the risks in each area are related to one another
- Spatial dependence in disease risk may be modelled via appropriate choice of prior distribution

Disease Mapping - Statistical Model

Prior for the relative risk, r_i

$$\begin{aligned} \log r_i &= \alpha + \sum_q \beta_q X_{qi} + S_i \\ S_i &\sim N(0, v) \end{aligned}$$

Full Bayesian analysis ⇒ need to specify hyperprior distributions for

- v (variance of the log relative risks)
- α (overall mean log relative risk)
- β_q (regression coefficients)

Disease Mapping - Statistical Model

Parameter Interpretation

- S_i are random effects
- e^{S_i}=residual (unexplained) relative risk in area i after adjusting for known covariates (X_q)
- S_i can also be thought of as a latent variable which captures the effects of unknown or unmeasured area level covariates
- If we believe these unmeasured covariates are spatially structured (e.g. environmental effects), then our model for S_i should allow for this

Conditional autoregressive model

- Intrinsic conditional autoregressive (ICAR) prior for **S** (Besag (1974); Clayton and Kaldor (1987)):

$$S_i | S_j = s_j, j \neq i \quad \sim \quad N(m_i, v_i)$$

$$m_i = \frac{\sum_{j \in \delta_i} W_{ij} s_j}{\sum_{j \in \delta_i} W_{ij}} \quad \text{and} \quad v_i = \frac{\tau^2}{\sum_{j \in \delta_i} W_{ij}}$$

- $\delta i =$ set of areas adjacent to i
- Setting $W_{ij} = 1$ for $j \in \delta i$:

$$m_i = \frac{\sum_{j \in \delta i} s_j}{n_i} \text{ and } v_i = \frac{\tau^2}{n_i}$$

where $n_i = \sharp$ neighbours

Interpretation of the variance parameter

- S_i is smoothed towards mean risk in set of neighbouring areas, with variance inversely proportional to the number of neighbours

- The variance parameter τ^2 of the intrinsic CAR random effects is a conditional (spatial) variance and is difficult to interpret

Conditional autoregressive model

Under the conditional specification in the previous slide, the joint density of $S = (S_1, S_2, \cdots, S_n)'$ is proportional to

$$\begin{aligned} p(s) &\propto \tau^{-g} \exp\left\{ -\frac{1}{2\tau^2} s'(D_w - W)s \right\} \\ &\equiv \tau^{-g} \exp\left\{ -\frac{1}{2\tau^2} \sum_{i=1}^{n} \sum_{j \in \partial_i} W_{ij}(s_i - s_j)^2 \right\}, \end{aligned} \tag{1}$$

where D_w is a diagonal matrix with elements $(D_w)_{ii} = W_{i+}$

- If all areas are connected, the ICAR distribution is proper in the $(n-1)$ dimensional space, so one suggestion is to fix g = (n - 1)/2 (Knorr-Held, 2003)
- Hodges et al. (2003) discuss the value of g and show that in the presence of islands (disconnected groups of areas) g should be fixed at (n-I)/2, where I is the number of islands

Conditional autoregressive model

- The distribution in equation (1) is *improper*
- We can add any constant to the elements of S and the density does not change.
- Similarly, the precision matrix $(D_w - W)$ is not of full rank
- For this reason, the ICAR model is not used to define the distribution of observed data
- Although the joint distribution of S is improper, all the conditional distributions $S_i \mid s_{-i}$ are proper
- One way to make this distribution proper is to impose a sum-zero constraint, that is, $\sum_i s_i = 0$
- Typically, ICAR models are used to assign the **prior distribution** for spatially varying **random effects** in hierarchical models.

Conditional autoregressive model

Comments on the intrinsic CAR prior in `WinBugs/OpenBugs`

- CAR prior for $S_i|S_j$ is expressed as a multivariate distribution for the *vector* of spatial random effects, $S = \{S_i\}_{i=1,\ldots,I}$
- Intrinsic CAR prior is improper (overall mean is undefined)
- `car.normal` prior in WinBugs imposes **sum-to-zero** constraint, i.e. $\sum_i S_i = 0 \Rightarrow$ can include separate intercept term α
  - Hyperprior on α should be improper:
    $\alpha \sim \mathit{dflat}()$.

Convolution prior specification

- To allow greater flexibility, Besag, York, and Mollié (1991) recommend combining the intrinsic CAR prior and an exchangeable normal prior

$$\begin{aligned} \log \mu_i &= \log e_i + \alpha + \beta X_i + S_i + U_i \qquad (2) \\ U_i &\sim N(0, \sigma^2) \\ S_i \mid S_j, j \neq i &\sim N(m_i, v_s) \text{ and } S_i \perp U_i \end{aligned}$$

- This model is called a convolution prior
- Total variation risk reflects a combination of spatial dependence and heterogeneity

Convolution prior specification

- Note that if $\tau^2 \to 0$ then the S_i's are constants, whereas large values of τ^2 imply large but spatially structured variation
- On the other hand, if $\sigma^2 \to 0$ then $V_i = 0$, whereas large σ^2 implies large unstructured variability (Besag et al., 1991)
- Note that τ^2 represents a conditional variance, whereas σ^2 a marginal variance $\Rightarrow$ they are not comparable as they are in different scales
- More importantly, the **likelihood** involves the sum $S_i + U_i$, so it is challenging to identify each of these components from a single realization of the process under study
- See Schmidt and Nobre (2018) for a recent review of conditional autoregressive models

Useful posterior Summaries

- Area-specific relative risks:
  `R[i]<-exp(alpha+beta*X[i]+S[i])`

- Area-specific residual relative risks (what’s left over after adjusting for known covariates etc.):
  `residR[i]<-exp(S[i])`

’Statistical Significance’ for Bayesian estimates

It is straightforward to decide if there is a ’statistically significant’ excess risk in area *i*:

- Calculate the posterior probability of
  $R_i > 1$
  =Area under posterior distribution curve to the right of 1.0
  =Proportion of values in the posterior sample of `R[i]` which are > 1.0

- This can be interpreted directly as a marginal probability

Prior choice for Poisson-Lognormal models[1]

- We need to specify priors for:
  - The intercept β_0 and regression coefficient β_1
  - The variance of the normal random effects σ^2
- An improper prior

  $$p(\beta_0, \beta_1) \propto 1$$

  may often be used, but in some circumstances such a choice may lead to an improper posterior
- If there are a large numbers of covariates, or high dependence amongst multiple covariates then more informative priors will be beneficial
- If an informative prior is required, then a multivariate normal distribution is the natural choice
- This is equivalent to a multivariate lognormal distribution for the relative risks

[1] Based on Wakefield (2013)

Prior choice for Poisson-Lognormal models

- It is convenient to specify lognormal priors for a positive parameter $\exp(\beta)$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the lognormal
- Denote by $LogNormal(\mu,\sigma)$ the lognormal distribution for a generic parameter θ with

  $$E[\log(\theta)] = \mu,\ var(\log(\theta)) = \sigma^2,$$

  and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior
  In our example, $\theta = \exp(\beta)$
- Then it is straightforward to show that

$$\begin{aligned}\mu &= \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2}-z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_1}-z_{q_2}}\right),\\ \sigma &= \frac{\log(\theta_1)-\log(\theta_2)}{z_{q_1}-z_{q_2}}\end{aligned}$$

Prior choice for Poisson-Lognormal models

As an example, suppose that for the ecological relative risk

$$\theta = \exp(\beta)$$

we believe there is a 50% chance that the relative risk is less than 1 and a 95% chance that it is less than 5

This gives $q_1 = 0.5$, $\theta_1 = 1.0$, $q_2 = 0.95$, $\theta_2 = 5.0$,

we obtain lognormal parameters

$$\mu = 0,\ \sigma = \log \frac{5}{1.645} = 0.98.$$

The density is shown in the next slide

Prior choice for Poisson-Lognormal models

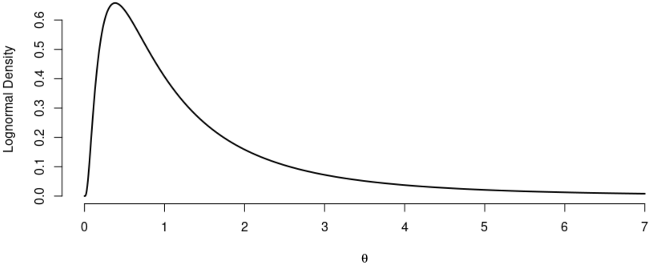


Figure: Lognormal density with 50% point 1 and 95% point 5

Prior choice for Poisson-Lognormal models

The priors $\tau_\varepsilon = \sigma_\varepsilon^{-2} \sim Ga(1, 0.0260)$ or $\tau_\varepsilon = \sigma_\varepsilon^{-2} \sim Ga(0.5, 0.0005)$ will often be suitable in a mapping context.

τ_ε is the precision, i.e., the reciprocal variance

For the $Ga(1, 0.026)$ prior the 2.5%, 50% (median) and 97.5% quantiles for σ^2 are:

(0.014, 0.047, 1.01)

For the $Ga(0.5, 0.0005)$ prior the 2.5%, 50% (median) and 97.5% quantiles for σ_ε are:

(0.084, 0.194, 1.01)

So the Ga(1,0.026) prior favors smaller values, i.e. more shrinkage is anticipated

Prior choice for Poisson-Lognormal models

Interpretation is helped by approximation of the residual relative risk

$$\exp(\varepsilon) \approx 1 + \varepsilon$$

for small ε and so

$$s.d(e^{\varepsilon}) = \sigma_{\varepsilon}$$

is approximately the standard deviation of the residual relative risks.

Sensitivity of the results to the specification should be carried out, particularly if the number of areas is not large.

Bayesian disease mapping with Stan

For examples of implementations of Bayesian disease mapping using `Stan` see

CARLipCancerScotland.pdf and

Denguefit.R

Proper CAR specification

- The ICAR specification leads to an improper joint distribution for S as $(D_w - W)^{-1}$ is singular
- One way to turn it into a proper distribution is to assume a scaled version of the known neighboring matrix, ρW, such that $(D_w - \rho W)$ is nonsingular, and (1) becomes the kernel of a zero mean **multivariate normal** distribution with covariance matrix $\Sigma_u = (D_w - \rho W)^{-1}$
- In this case, ρ must be chosen in a way that Σ_u^{-1} is nonsingular
- A common choice is to assume $\rho \in (1/\gamma_{(1)}, 1/\gamma_{(n)})$, where $\gamma_{(1)}$ and $\gamma_{(n)}$ are, respectively, the minimum and maximum eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$

Proper CAR specification

- In this case, the full conditional is given by

$$S_i \mid s_{-i} \sim N\left(\rho \sum_{j \in \partial_i} \frac{W_{ij}\, s_j}{W_{i+}}, \frac{\tau^2}{W_{i+}}\right), \tag{3}$$

  and the joint distribution of S is $N(0, \tau^2(D_w - \rho W)^{-1})$
- As $\Sigma_s^{-1} = \tau^{-2}(D_w - \rho W)$ is the precision matrix of the distribution of S, it follows that $Var(S_i \mid s_j, j \neq i) = 1/(\Sigma_{s_{ii}}^{-1})$. Therefore, if $W_{ij} = 0$ then S_i and S_j are conditionally independent given s_k, $k \neq i, j$
- Usually, CAR models induce sparse precision matrices making the computation of these matrices very fast in the inference procedure. The ICAR model is the limiting case when $\rho = 1$
- Also, it can be shown that 0 belongs to $(1/\gamma_{(1)}, 1/\gamma_{(n)})$, leading to an independent model for the S_i's if $\rho = 0$, and in this case $S_i \sim N(0, \tau^2/w_{i+})$

Reparametrizations of the BYM model

- Because of the identifiability issue with the BYM model, some authors have proposed different reparametrizations of the BYM model.
- Leroux *et al.* (1999) remove U_i from equation (2) and include a component λ in the CAR prior specification, such that

$$E(S_i \mid s_{-i}) = \left(\frac{\lambda}{1-\lambda q_{ii}}\right) \sum_{j \in \partial_i} s_j, \text{ and } V(S_i \mid s_{-i}) = \frac{\tau^2}{1-\lambda+\lambda q_{ii}}, \tag{4}$$

  where $0 \leq \lambda \leq 1$ denotes a spatial dependence parameter, such that $\lambda = 0$ defines a nonspatial model, and the level of spatial dependence increases with λ

Reparametrizations of the BYM model

- In this case, the joint prior distribution of the vector S is a zero mean normal distribution, with covariance matrix $\Sigma_s = \tau^2 \left[\lambda Q + (1-\lambda) I_n\right]^{-1}$, where the diagonal elements of Q, q_{ii}, contain the number of neighbors of region i, and $q_{ij} = -1$ if $i \sim j$, and 0 otherwise

- When $\lambda = 1$ the distribution of S is improper, as we have the ICAR specification. Lee (2011) performs a simulation study comparing the mostly used CAR prior specifications in **disease mapping**, and conclude that the model proposed by Leroux *et al.* (1999) is the best among the fitted ones

- Lee (2011) claims that this is because, different from the ICAR and the BYM prior specifications, the prior proposed by Leroux *et al.* (1999) can represent a range of strong and weak spatial correlation structures with a single set of random effects

Reparametrizations of the BYM model

- More recently, Simpson *et al.* (2017) claim that the components S and U in equation (2) should not be assumed to be independent, suggesting that the priors on τ^2 and σ^2 should be dependent

- Riebler *et al.* (2016) propose a reparametrisation of the model in equation (2), wherein $g(\mu_i) = \mu + x_i^T\beta + \frac{1}{\tau}\left(\sqrt{1-\phi}\,U_i + \sqrt{\phi}\,S_i^*\right)$, with $0 \leq \phi \leq 1$ being a mixing parameter

- The component S^* is a scaled spatially structured component where the generalised variance, computed as the geometric mean of the marginal variances, is equal to one. In this case, $1/\tau$ represents the marginal precision contribution from S^* and U, with ϕ representing the fraction of this variance explained by S^*, and $(1-\phi)$ the fraction explained by U

Reparametrizations of the BYM model

- Riebler *et al.* (2016) perform a simulation study comparing the performance of the parametrisation proposed by Simpson *et al.* (2017) in contrast with the BYM, and the models proposed by Leroux *et al.* (1999) and Dean (2001), who proposed another reparametrisation of the BYM model

- Their results show that in terms of model comparison criteria, their proposal performs at least equally well as the compared parametrisations. But they claim that their proposal is the only one that allows for interpretation of the parameters and of the hyperpriors Riebler *et al.* (2016).

Part I

Zero-Inflated Poisson Model

Zero-inflated Poisson Model

- In some applications the specification of Poisson or Binomial models for counts data might be inappropriate due to the excess of zeros in the data compared with what expected from the model
- To overcome this issue the so-called zero-inflated models can be specified
- These are a mixture of two components: a point mass at zero and a count distribution
- Such models distinguish between structural zeros, for units where zero is the only observable value, and sample zeros, for units on which we observe a zero, but others values might also have been recorded
- Following Blangiardo and Cameletti (2015), we will analyze brain cancer incidence in Navarra (Spain) (Gómez-Rubio and Lopez-Quilez, 2010)
- These models might be useful when there are many zeros in the data, or you might want to account for the probability of detecting the disease

Zero-inflated Poisson Model

- The PMF is given by

$$f(y_i \mid \pi_0, \lambda_i) = \pi_0 I(y_i = 0) + (1 - \pi_0) \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{5}$$

  where $I(y_i = 0)$ is the indicator variable
- The probability of observing a zero in the $i - th$ area is $\pi_0 + (1 - \pi_0)\exp(-\lambda_i)$
- Mean and variance are given by

$$\begin{aligned} E(y_i|lambda_i) &= (1 - \pi_0)\lambda_i \\ Var(y_i) &= (1 - \pi_0)\lambda_i + \frac{\pi_0}{1 - \pi_0}((1 - \pi_0)\lambda_i) \end{aligned}$$

- The spatial structure can be included in model for λ_i

Navarre Example

- The number of brain cancer cases reported in each of the 40 health districts in the Navarra region is modeled following the ZIP model described in the previous slide
- Then conditional on y_i not being a structural zero, the logarithmic transformation of λ_i is modeled as

  $$\log(\lambda_i) = b_0 + v_i + u_i + \log(E_i),$$

  allowing for a global intercept and a BYM specification $(u_i + v_i)$ as done previously with the simpler Poisson model
- See NavarreFit.r

Zero-inflated Poisson Model

- When only structural zeros are present in the data, then equation (5) becomes

$$f(y_i \mid \pi_0, \lambda_i) = \quad \pi_0 I(y_i = 0) + (1 - \pi_0) I(y_i > 0) \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

- In R-INLA this alternative model can be accessed typing `zeroinflatedpoisson0` in the likelihood specification

Part II

DENGUE FEVER IN RIO DE JANEIRO WITH DIFFERENT NEIGHBORHOOD STRUCTURES

Spatial Modelling of the Relative Risk of Dengue Fever in Rio de Janeiro for the epidemic period between 2001 and 2002

Gustavo S. Ferreira and Alexandra M. Schmidt
Brazilian Journal of Probability and Statistics, vol. 20, 1, 29–47, 2006

- Observations were obtained from the Health Secretary of Rio de Janeiro;
- Observed data correspond to 125.368 notifications of the number of cases of Dengue Fever in the city of Rio de Janeiro, from December 2001 until May, 2002 (aggregated over time);
- Rio has 159 districts;

Main aims

- search for social, geographic and economic factors which might be related to the relative risk of dengue fever at a certain district;
- study if there is any spatial trend of the relative risk of dengue across the districts of the city;
- evaluate the influence of different specifications of the neighborhood structure of the random effects (mountains);

Model

The number of cases o dengue fever at district i, y_i, is assumed to follow a Poisson distribution, that is

$$y_i \mid r_i, e_i \sim Poisson(e_i r_i) \quad i = 1, \ldots, n = 159, \text{ where}$$

The second level of hierarchy is defined as

$$\log(r_i) = \mathbf{X}_i' \beta + S_i, \text{ where} \tag{6}$$

- **X** represents the vector with q covariates. Available covariates: strategic points, quota 100m, average income, life expectancy at birth, adult literacy rate, index of human development, garbage not collected, public pipeline system, area of the district covered by slums.

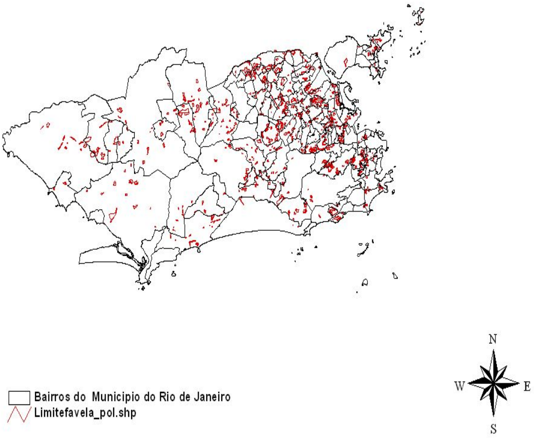


Figure: Distribution of the slums across the districts of Rio de Janeiro.

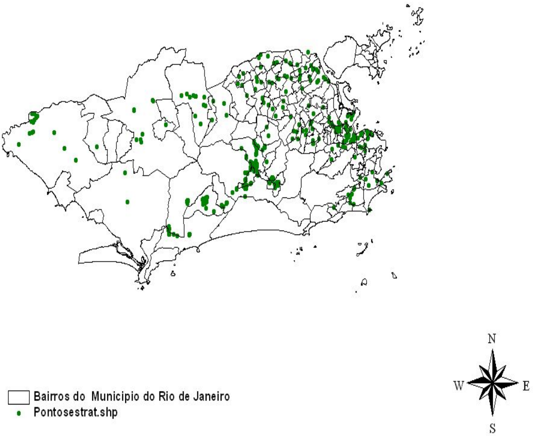


Figure: Distribution of the strategic points across the districts of Rio de Janeiro.

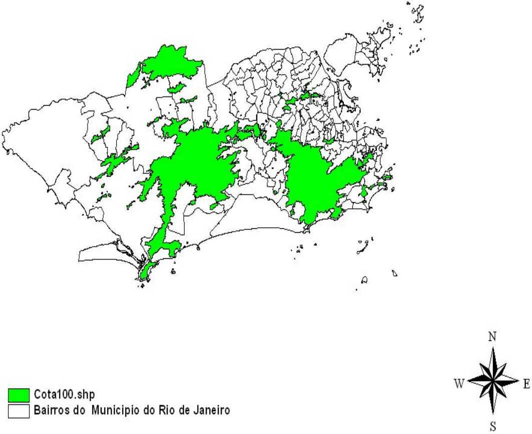


Figure: Distribution of the mountains with quota at least 100m across the districts of Rio de Janeiro.

Fitted models

Prior distributions considered for the spatial random effect:

1. CAR 0-1 neighborhood
2. CAR with neighborhood weighted by the length of the borders of the districts
3. CAR with neighborhood weighted by the length of the borders of the districts and the proportion covered by mountains.

Table: Posterior Summary of the Coefficients for Each Covariate Under Each of the Priors.

One Covariate	PRIORS		
X_1	I	II	III
STR	+	+	+
SLU	-	-	-
C10	+	+	0
AVI	-	-	-
AVS	-	-	-
GAR	+	+	0
PPS	+	+	+
ALR	(+)	(+)	+
IHD	0	0	0
LEB	0	0	0

		PRIORS					
Two Covariates		I		II		III	
X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
STR	SLU	+	-	+	-	+	-
STR	C10	+	+	+	(+)	+	0
STR	AVI	+	0	+	-	+	-
STR	GAR	+	(-)	+	-	+	-
STR	PPS	+	+	+	+	+	0
SLU	AVI	-	-	-	-	-	-
SLU	GAR	-	(+)	-	(+)	-	0
SLU	PPS	-	+	-	+	-	+
SLU	C10	-	+	-	+	-	0

For table elements, + or - are positive or negative coefficients with 95% credible intervals that do not overlap 0, (+) or (-) are positive or negative coefficients with 90% credible intervals that do not overlap 0, and 0 are coefficients which were not significant.

		PRIORS								
Covariate(s)		**PRIOR I**			**PRIOR II**			**PRIOR III**		
X_1	X_2	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC
Const		1398.56	190.42	1588.98	1390.09	182.29	1572.39	1443.81	235.03	1678.84
STR		1397.24	188.92	1586.17	1388.48	180.50	1568.98	1440.07	231.11	1671.18
SLU		1398.41	190.35	1588.76	1389.35	181.68	1571.02	1440.53	231.89	1672.42
C10		1398.60	190.43	1589.03	1389.73	181.94	1571.68	1444.76	236.10	1680.86
AVI		1398.9	190.90	1589.80	1389.95	182.22	1572.18	1444.28	235.55	1679.83
AVS		1397.58	189.12	1586.70	1389.06	180.95	1570.01	1442.10	233.09	1675.18
GAR		1398.56	190.40	1588.95	1390.03	182.23	1572.26	1444.71	236.04	1680.75
PPS		1396.55	188.12	1584.67	1388.11	180.08	1568.19	1443.13	234.38	1677.52
ALR		1398.54	190.44	1588.98	1441.07	235.58	1676.65	1614.04	408.26	2022.30
IHD		1398.10	190.03	1588.13	1389.90	182.07	1571.97	1442.98	234.22	1677.20
LEB		1398.13	190.00	1588.13	1474.92	269.74	1744.66	1736.03	530.73	2266.77
STR	SLU	1396.10	187.83	1583.93	1387.85	180.05	1567.90	1437.49	228.64	1666.12
STR	C10	1396.96	188.26	1585.59	1388.43	180.49	1568.92	1439.75	230.83	1670.58
STR	AVI	1396.85	188.53	1585.38	1388.48	180.50	1568.97	1439.00	230.11	1669.11
STR	GAR	1396.74	188.45	1585.19	1388.06	180.08	1568.14	1439.00	230.13	1669.14
STR	PPS	1396.63	188.17	1584.80	1387.66	179.59	1567.25	1440.83	231.91	1672.73
SLU	AVI	1397.98	189.91	1587.89	1389.08	181.42	1570.50	1439.88	231.23	1671.11
SLU	GAR	1398.25	190.25	1588.50	1389.42	181.69	1571.11	1439.74	231.12	1670.87
SLU	PPS	1396.42	188.06	1584.48	1387.46	179.57	1567.03	1440.38	231.74	1672.13
SLU	C10	1397.26	189.16	1586.42	1388.71	181.05	1569.77	1442.01	233.44	1675.45

Table: DIC's for all fitted models according to the covariate used and the prior specification of the random effects.

		PRIORS								
Covariate(s)		PRIOR I			PRIOR II			PRIOR III		
X_1	X_2	P	G	EPD	P	G	EPD	P	G	EPD
Const		341.67	4.41	346.09	335.85	3.87	339.72	387.27	5.08	392.35
STR		341.42	4.56	345.98	335.02	4.30	339.32	382.00	5.34	387.34
SLU		346.25	4.53	350.78	338.46	3.53	341.99	387.31	5.06	392.37
C10		342.24	4.64	346.88	337.64	3.92	341.56	394.20	4.49	398.69
AVI		343.86	4.50	348.36	335.89	3.85	339.74	387.73	4.96	392.69
AVS		342.77	4.86	347.63	336.39	3.95	340.34	388.26	5.66	393.92
GAR		343.89	4.47	348.35	335.34	3.78	339.12	390.68	5.14	395.83
PPS		342.15	4.55	346.71	333.40	4.03	337.43	387.47	5.09	392.56
ALR		345.70	4.48	350.18	390.22	1.08	391.31	559.72	1.84	561.56
IHD		343.22	4.34	347.55	335.75	3.84	339.59	384.48	4.93	389.40
LEB		342.17	4.46	346.64	424.73	0.55	425.27	682.16	0.77	682.93
STR	SLU	340.11	4.73	344.84	333.17	3.88	337.05	384.77	5.43	390.20
STR	C10	342.27	4.87	347.14	332.45	4.21	336.65	382.56	5.31	387.88
STR	AVI	344.19	4.55	348.73	335.26	4.13	339.39	385.42	5.23	390.65
STR	GAR	339.80	4.75	344.55	333.92	4.09	338.01	383.84	5.19	389.03
STR	PPS	342.58	4.82	347.41	332.03	3.05	335.08	386.42	5.33	391.76
SLU	AVI	341.84	4.66	346.50	335.1	3.76	338.87	384.05	5.05	389.11
SLU	GAR	339.67	4.22	343.89	335.43	3.97	339.40	385.54	4.77	390.30
SLU	PPS	342.29	4.74	347.02	334.49	3.86	338.35	386.50	5.23	391.72
SLU	C10	341.23	4.49	345.72	336.60	3.80	340.40	383.28	4.90	388.18

Table: EPD’s for all fitted models according to the covariate used and the prior specification of the random effects.

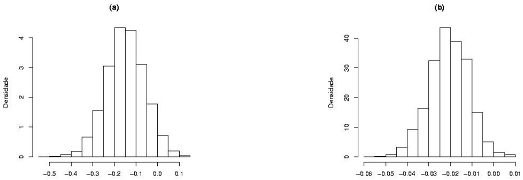
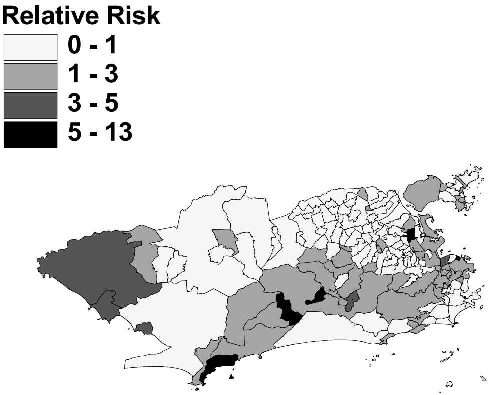
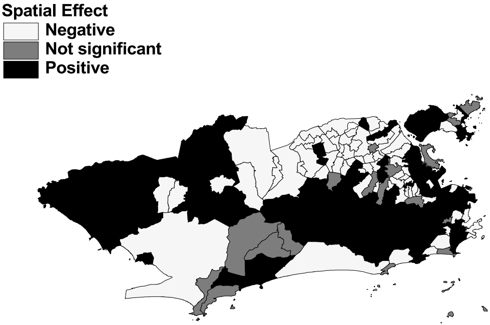


Figure: Posterior distribution of (a) α_0, (b) α_1 for the best model according to DIC.


Relative Risk
0 - 1
1 - 3
3 - 5
5 - 13


Spatial Effect
Negative
Not significant
Positive

Part III

MODEL COMPARISON

Model selection

- We now have many potential models in our arsenal
- For a given dataset, how do determine whether a simple model is sufficient or if we need to bring out the ‘big guns”?
- Is there a “right” model? Probably not
- A statistical model is a mathematical representation of the system that includes errors and biases in the observation process
- All models are simplifications of reality
- We want a model that is as simple as possible yet seems to fit the data reasonably well

Model Comparison

- For model comparisons there are a finite number of candidate models and we want to select one
  - Bayes factors
  - Stochastic search variable selection
  - Cross validation
  - Deviance information criteria (DIC)
  - Watanabe-Akaike information criteria (WAIC)
- In cases where multiple models fit well, we will discuss
  - Bayesian model averaging (BMA)
- After selecting a model, we want to test whether it fits the data well
  - Posterior predictive checks

Information criteria

- Several information criteria have been proposed that do not require fitting the model several times

- Many are functions of the deviance, i.e., twice the negative log likelihood

$$D(\mathbf{y} \mid \theta) = -2\log[f(\mathbf{y} \mid \theta)]$$

- Ideally, models will have small deviance

- However, if a model is too complex it will have small deviance between be unstable (over-fitting)

- The Akaike information criteria has a complexity penalty

$$AIC = D(\mathbf{y} \mid \widehat{\theta}) + 2p$$

where $\widehat{\theta}$ is the MLE

- Model with smaller AIC are preferred

Bayesian information criteria (BIC)

- The Bayesian information criteria is similar

$$BIC = D(\mathbf{y} \mid \widehat{\theta}) + \log(n)\, p$$

- This is motivated as an approximation to the log Bayes factor of the model compared to the null model

- However, this is only an asymptotic (large *n*) approximation

- With large n the prior is irrelevant, and so this is not satisfying to a subjective Bayesian

Deviance information criteria (DIC)

- DIC is a popular Bayesian analog of AIC or BIC

- Unlike CV, DIC requires only one model fit

- Unlike BF, it can be applied to complex models

- However, proceed with caution

- DIC really only applies when the posterior is approximately normal, and will give misleading results when the posterior far from normality, e.g., bimodal

- DIC is also criticized for selecting overly-complex models

Deviance information criteria (DIC)

- Let $\overline{D} = E[D(\mathbf{Y} \mid \theta) \mid \mathbf{Y}]$ be the posterior mean of the deviance

- Denote $\widehat{\theta}$ as the posterior mean of θ

- The effective number of parameters is

$$p_D = \overline{D} - D(\mathbf{Y} \mid \theta)$$

- DIC can be written like AIC,

$$DIC = \overline{D} + p_D = D(\mathbf{Y}|\widehat{\theta}) + 2p_D$$

- Models with small $\overline{D}$ fit the data well

- Models with small p_D are simple

- We prefer models that are simple and fit well, so we select the model with smallest DIC

Deviance information criteria (DIC)

- The effective number of parameters is a useful measure of model complexity
- Intuitively, if there are p parameters and we have uninformative priors then $p_D \approx p$
- However, $p_D << p$ if there are strong priors
- For example, how many free degrees of freedom do we have with $\theta \sim Beta(1,1)$ versus $\theta \sim Beta(1000,1000)$?
- In some cases p_D has a nice closed form

Deviance information criteria (DIC)

- As with AIC or BIC, we compute DIC for all models under consideration and select the one with smallest DIC
- Rule of thumb: a difference of DIC of less than 5 is not definitive and a difference greater than 10 is substantial
- As with AIC or BIC, the actual value is meaningless, only differences are relevant
- DIC can only be used to compare models with the same likelihood

Watanabe-Akaike information criteria (WAIC)

- WAIC is an alternative to DIC
- It is motivated as an approximation to leave-one-out CV
- In the end WAIC has model-fit and model-complexity components
- It is used the same as DIC with smaller WAIC being preferred
- In practice the two often give similar results, but WAIC is arguably more theoretically justified

Watanabe-Akaike information criteria (WAIC)

- WAIC is written in terms of the posterior of the likelihood rather than parameters
- Let m_i and v_i be the posterior mean and variance of

$$\log[f(Y_i|\theta)]$$

- The effective model size is $p_W = \sum_{i=1}^{n} v_i$
- The criteria is

$$WAIC = -2\sum_{i=1}^{n} m_i + 2p_W$$

For examples of DIC and WAIC calculations see
https://www4.stat.ncsu.edu/~reich/BSMdata/DIC_WAIC.html
and
https://www4.stat.ncsu.edu/~reich/BSMdata/Election.html

Posterior predictive checks

- After comparing a few models, we settle on the one that seems to fit the best
- Given this model, we then verify it is adequate
- The usual residual checks are appropriate here: qq-plots; added variable plots; etc.
- A uniquely Bayesian diagnostic is the posterior predictive check
- This leads to the Bayesian p-value

Posterior predictive distributions

- Before discussing posterior predictive checks, let's review Bayesian prediction in general

- The plug-in approach would fix the parameters θ at the posterior mean $\hat{\theta}$ and then predict $Y_{new} \sim f(y|\hat{\theta})$

- This suppresses uncertainty in θ

- We would like to propagate this uncertainty through to the predictions

Posterior predictive distributions (PPD)

- We really want the PPD

$$f(Y_{new}|\mathbf{Y}) = \int f(Y_{new}, \theta \mid \mathbf{y}) d\theta = f(Y_{new} \mid \theta) f(\theta \mid \mathbf{y}) d\theta$$

- MCMC easily produces draws from this distribution

- To make S draws from the PPD, for each of the S MCMC draws of θ we draw a Y_{new}

- This gives draws from the PPD and clearly accounts for uncertainty in θ

Posterior predictive checks

- Posterior predictive checks sample many datasets from the PPD with the identical design (same n, same $\mathbf{X}$) as the original data set
- We then define a statistic describing the dataset, e.g.,

$$d(\mathbf{Y}) = max\{Y_1, \ldots, Y_n\}$$

- Denote the statistic for the original data set as d_0 and the statistic from simulated data set number s as d_s
- If the model is correct, then d_0 should fall in the middle of the $d_1, \ldots, d_S$

Posterior predictive checks

- A measure of how extreme the observed data is relative to this sampling distribution is the Bayesian $p - value$

$$p = \frac{1}{S}\sum_{s=1}^{S} I(d_S > d_0)$$

- If p is near zero or one the model doesn't fit

- This is repeated for several d to give a comprehensive evaluation of model fit

See
`https://www4.stat.ncsu.edu/~reich/BSMdata/checks_guns.html`
for an example