Methods to control for the concurvity in spatial ecological regressions (IneqCities project)

Abstract

Estimates of ecological regression coefficients may be biased when the ecological exposure is correlated with other factors which explain the spatial aggregation and a spatially autocorrelated model is specified. The problem is probably due to a near concurvity between the clustering term and the covariate.

In order to assess the consequences and also possible solutions of the concurvity, we simulated a Poisson response using a neighbourhood matrix based on the Girona Metropolitan Area, Spain, by census tracts (9 municipalities and 85 census tracts). In particular, we considered three scenarios (high, moderate and low concurvity) and three sub-scenarios for each scenario (spatial dependence equal to heterogeneity, much greater and much lower). Then, we fitted four different models for the spatial dependence: intrinsic CAR; proper CAR; space varying regression; and a Gaussian random field representation of Markov (GMRF) constructed explicitly from a stochastic partial differential equation (SPDE) whose solution is a Gaussian field (GF) with covariance function Matérn.

In presence of concurvity, not only the variance of the estimators is inflated, but also the estimates of the parameters are biased. These effects increase as the higher the correlation between the explanatory variable and the spatial random effect (i.e. the concurvity), as well as the more this (spatial) effect in relation to the heterogeneity. However, when we standardized the explanatory variable, however, the estimates of the parameters were not biased and their variances reduced considerably. We show that, with the explanatory variable standardized or not, all four models considered (iCAR, pCAR, Space Varying Regression and SPDE) provided similar results. SPDE models were those with a better goodness-of-fit.

Key words: concurvity, intrinsic and proper CAR, SPDE models.
1 Introduction

The inclusion of the clustering term in ecological analysis has been considered as a way to adjust for unmeasured confounders that vary locally over the study region (Clayton et al., 1993).

The problem is that estimates of ecological regression coefficients may be biased when the ecological exposure is correlated with other factors which explain the spatial aggregation and a spatially autocorrelated model is specified (Breslow and Clayton, 1993).

Catelan et al. (2009) show that the problem is probably due to a near concurvity between the clustering term and the covariate. In fact, they conclude that the parameter $\rho$, that controls the overall strength of the spatial dependence, could be thought of as a neg-penalty: when the penalty is low (i.e. $\rho$ is near to 1), the spatially structured term tends to overfitting, capturing part of the covariate effect and leading to biased estimates of the covariate effect.

2 Methods

2.1 Simulation

We simulated the following response,

$$Y_i \sim \text{Poisson}(\mu_i, \text{Pop}_i)$$

where $\text{Pop}$ was the population of males in the census tract $i$ using a neighbourhood matrix based on the Girona Metropolitan Area by census tracts (9 municipalities and 85 census tracts).

$\mu_i$ was

$$\log(\mu_i) = -5.1096 + 8.7128 \text{Pop65}_i + \log(\text{Pop}_i) + \eta_i + S_i + \beta \text{index}_i$$

where $\text{Pop65}_i$ was the percentage of males aged 65 or older in census tract $i$ over all males in the census tract $i$; $\beta = 1$; $S_i \sim N(0.1522, 0.0648)$; and $\text{index}_i \sim \text{Normal}$.

These figures are based on an estimation of a Besag, York and Mollié (BYM) model (Besag et al., 1991; Mollié, 1996) with real data (incidence of prostate cancer in the Girona Metropolitan Area, 1993-2006).

We considered three scenarios (high, moderate and low concurvity) and three sub-scenarios for each scenario (spatial dependence equal to heterogeneity, spatial dependence much greater than heterogeneity and spatial dependence much lower than heterogeneity):

<table>
<thead>
<tr>
<th>Linear effect</th>
<th>Scenarios</th>
<th>Sub-scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_\eta = \sigma_S$</td>
<td>Cor(S,index) = 0.9</td>
<td>Cor(S,index) = 0.9</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.5</td>
<td>Cor(S,index) = 0.5</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.3</td>
<td>Cor(S,index) = 0.3</td>
</tr>
<tr>
<td>$\sigma_\eta = \frac{1}{5} \sigma_S$</td>
<td>Cor(S,index) = 0.9</td>
<td>Cor(S,index) = 0.9</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.5</td>
<td>Cor(S,index) = 0.5</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.3</td>
<td>Cor(S,index) = 0.3</td>
</tr>
<tr>
<td>$\sigma_\eta = 5 \sigma_S$</td>
<td>Cor(S,index) = 0.9</td>
<td>Cor(S,index) = 0.9</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.5</td>
<td>Cor(S,index) = 0.5</td>
</tr>
<tr>
<td></td>
<td>Cor(S,index) = 0.3</td>
<td>Cor(S,index) = 0.3</td>
</tr>
</tbody>
</table>
2.2 Models

Then, we fitted four different models for the spatial dependence:

**Intrinsic CAR (iCAR)**

\[ Y_i \sim \text{Poisson}(\mu_i, \text{Pop}_i) \]

\[ \log(\mu_i) = \alpha + \gamma_i \text{Pop}4564_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_j) + \eta_i + S_i + \beta \text{ index}_i \]

\[ S_i | S_j \sim \mathcal{N}\left( \frac{1}{n_i} \sum_{i \neq j} S_j, \frac{1}{n_i} \right) \quad \forall i \neq j \]

**Proper CAR (pCAR)**

\[ Y_i \sim \text{Poisson}(\mu_i, \text{Pop}_i) \]

\[ \log(\mu_i) = \alpha + \gamma_i \text{Pop}4564_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_j) + \eta_i + S_i + \beta \text{ index}_i \]

\[ S_i | S_j \sim \mathcal{N}\left( \frac{1}{d + n_i} \sum_{i \neq j} S_j, \frac{1}{d + n_i} \tau \right) \quad \tau = \frac{1}{\rho} \quad \forall i \neq j \]

**Space Varying Regression Model**

\[ Y_i \sim \text{Poisson}(\mu_i, \text{Pop}_i) \]

\[ \log(\mu_i) = \alpha + \gamma_i \text{Pop}4564_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_j) + \eta_i + S_i + \beta \text{ index}_i \]

\[ S_i | S_j \sim \mathcal{N}\left( \frac{1}{n_i} \sum_{i \neq j} S_j, \frac{1}{n_i} \right) \quad \forall i \neq j \]

**SPDE**

\[ Y_i \sim \text{Poisson}(\mu_i, \text{Pop}_i) \]

\[ \log(\mu_i) = \alpha + \gamma_i \text{Pop}4564_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_j) + \eta_i + s(x_{\text{coord}}, y_{\text{coord}}) + \beta \text{ index}_i \]

Following the recent work of Lindgren et al. (2011), we specify a Matérn structure (Stein, 1999) for the spatial dependence.

\[ \text{corr} (d_{ii'}) = M\left( |i - i'|, \sigma_i^2, \rho_i, \nu_i \right) + (1 - \rho_i^2) \sigma_i^2 I(i = i') \]

In short, we use a Gaussian random field representation of Markov (GMRF) constructed explicitly from a stochastic partial differential equation (SPDE) whose solution is a Gaussian field (GF) with covariance function Matérn (Lindgren et al., 2011). Instead of using a regular lattice, as was the standard practice, which would imply an estimate with a high computational cost and also very little efficiency (Lindgren et al., 2011), we specify a Matérn spatial covariance structure in a triangulation (triangulation of Delaunay - Hjelle and Daehlen, 2006 -) of the Girona Metropolitan Area, with very little computational cost and, most importantly in our context, much more efficient.

3 Results

In presence of concurrency, not only the variance of the estimators is inflated, but also the estimates of the parameters are biased. These effects increase as the higher the correlation between the explanatory variable and the spatial random effect (i.e. the concurrency), as well as the more this (spatial) effect in relation to the heterogeneity (Tables 1a and 2a).