A spatial model for hospital recruitment via INLA

Monica Musio ¹, Erik A. Sauleau ²

¹ Department of Mathematics, University of Cagliari, Italy
² Medicine Faculty, University of Strasbourg, France

SCO 2011
September 19-21 2011, Padova
Aims

- Explain spatial recruitment in Haute Alsace a region in the north-east part of France;
- Consider an explicative model which account for continuous covariates with possible non linear effects;
- Select a "best" explicative model; identify a set of covariates which explain differences in the observed recruitment;
- Use this selected model
  - For health service planning;
  - To study the effect on recruitment of the modification of some of these variables → what happens if a new road decreases the access time to a certain hospital from some cities of a region?
Data

- Full time hospitalized patients from the public hospital of Mulhouse in 2009;
- Total number of cases 33572;
- **Response variable: recruitment ratio (SRR)**
  - Ratio between the number of patients living in each geographical unit $y_i$ and the population $N_i$ in the same unit:
    \[
    \text{SRR}_i = \frac{y_i}{N_i}
    \]
  - $N_i$ can be considered as the number of persons at risk to visit an healthcare provider.
Data

- **Covariates:**
  - Geographical unit of residence, with centroid coordinates (377 GU):
    - between 0 and 11.029 cases per GU;
    - The mean is 89 while the median is 19.
  - Age, gender;
    - Age is categorized in 18 categories from [0,5) up to [80,85) and more than 85 years.
  - The distance and the access time between the healthcare provider and the GU of residence;
  - The distance and the access time between each GU and a second healthcare provider;
  - The density of practitioners in each GU (for 1000 inhabitants);
  - Proximity zone: each region is divided into several zones, centered by an healthcare provider to which its patients are recruited.
Observed recruitment ratio per geographical unit
Observed recruitment ratio per age category and gender

<table>
<thead>
<tr>
<th>Age category</th>
<th>Recruitment ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,5)</td>
<td>5%</td>
</tr>
<tr>
<td>[25,30)</td>
<td>10%</td>
</tr>
<tr>
<td>[50,55)</td>
<td>15%</td>
</tr>
<tr>
<td>More than 85</td>
<td>20%</td>
</tr>
</tbody>
</table>

- **Female**
- **Male**
Our model

Notations

- Indices: $a$ for age category (1-18), $s$ for gender (1,2) and $r$ for GU (1-377)
- Number of observed cases: $Y \rightarrow Y_{ras}$

The model

$$Y_{ras} \sim \mathcal{B}(N_{ras}, p_{ras})$$

where:

- $N_{ras}$ population;
- $p_{ras}$ the risk;

the linear predictor $\eta_{ras}$ is modelled as a semiparametric additive predictor:
Our model

\[ \logit(p_{ras}) = \eta_{ras} = \mu + \sum_{j=1}^{m} f^{(j)}(u_{jras}) + \sum_{k=1}^{K} \beta_{k} z_{kras} + f^{(s)} + \varepsilon_{r}. \]

Structured Additive Regression (StAR) model (see Fahrmeir and Tutz (2001)).

- \( f^{(j)}(\cdot) \)s are unknown smoothing functions of the \( m \) covariates in \( u \);
- \( \beta_{k} \)s represent the vector parameters for the linear effect of covariates in \( z \);
- \( f^{(s)} \) is a spatially structured component;
- \( \varepsilon \) is a spatially unstructured component.
Our model

We consider the following prior distributions:

- Smoothing functions $f^{(j)}, j = 1, \cdots, m \rightarrow$ first or second-order random walk models with precisions $\tau^{(j)}$:
  
  $u(1) < \cdots < u(i) < \cdots < u(m)$
  
  $f = (f_1, \cdots, f_m)$ where $f(i) = f(u(i))$

  then

  $f(i) = f(i - 1) + w(i)$ with $w(i) \sim \mathcal{N}(0, \tau^{(i)})$

  $\pi(w(1)) \propto \text{const}$

- Spatial structured component $f^{(s)} \rightarrow$ ICAR process with precision $\tau^{(s)}$, (Besag et al. (1991)):

  $f_{r}^{(s)} | f_{-r}^{(s)}, \tau^{(s)} \sim \mathcal{N}
  \left( \frac{1}{n_{r}} \sum_{j \in \partial r} f_{j}^{(s)}, n_{r} \tau^{(s)} \right)$.
Our model

- The unstructured spatial effect \( \varepsilon \rightarrow \) random effect
  \[ \varepsilon = (\varepsilon_1, \ldots, \varepsilon_{377}) \sim \mathcal{N}(0, \tau^{(\varepsilon)} I) \]

- We assign independent \( \Gamma(0.001, 0.001) \) priors to the hyperparameters \((\tau^{(1)}, \ldots, \tau^{(m)}, \tau^{(s)}, \tau^{(\varepsilon)})^t\) and a \( \mathcal{N}(0, 0.01) \) prior to \( \mu \) and \( \beta_k \)s.

→ Our model is a **Latent Gaussian model**

- \( x \) vector of all parameters;
  \( \theta \) the vector of all hyperparameters.

- We are interested in estimating the posterior distribution
  \[
  \pi(x_i | y) = \int \pi(x_i | \theta, y) \pi(\theta | y) d\theta.
  \]

- To compute \( \pi(x_i | y) \) we use **INLA method** (see Rue et al. (2009)).
Review of methodology: INLA method

The posterior marginals $\pi(x_i|y)$ are approximated by INLA using the finite sum

$$\tilde{\pi}(x_i|y) = \sum_k \tilde{\pi}(x_i|\theta_k, y) \tilde{\pi}(\theta_k|y) \Delta_k \quad (1)$$

1. Utilise of latent Gaussian random field $\rightarrow$ Laplace approximations;
2. Utilise the conditional independence properties of the latent Gaussian fields $\rightarrow$ Numerical algorithms for sparse matrices;
   - In contrast with McMC, the INLA method does not sample from the posterior. It approximates the posterior with a closed form expression;
   - Very fast when the number of hyperparameters does not exceed 6;
   - For comparisons between McMC and INLA methods see Rue et al. (2009), Held et al. (2010).
Model selection

1. We consider a reference model including an age-gender effect and a structured spatial effect;
   ▶ the age-gender effect is modelled using a random walk prior on age for each sex;
   ▶ the spatial effect is estimated using an ICAR prior;

2. Starting from the reference model we build as many models as there are explanatory variables. Models are retained if their DIC is lower than that of the reference model (Spiegelhalter et al. (2002));

3. A multivariate "best" model is then estimated using previous "significant" effects;

4. We add to the selected model an unstructured spatial effect to account for possible remaining spatial structure;

5. We check the fit of the best model by testing alternatives to random walk priors (linear or quadratic trends).
Results: Model comparisons via the DIC criterion

<table>
<thead>
<tr>
<th>Model</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference: age, gender and ICAR</td>
<td>256.48</td>
<td>28,041.10</td>
</tr>
<tr>
<td>+ distance to provider</td>
<td>248.97</td>
<td>28,036.25</td>
</tr>
<tr>
<td>+ access time to provider</td>
<td>250.49</td>
<td>28,028.58</td>
</tr>
<tr>
<td>+ distance to the second provider</td>
<td>224.73</td>
<td>28,030.89</td>
</tr>
<tr>
<td>+ access time to the second provider</td>
<td>252.91</td>
<td>28,040.96</td>
</tr>
<tr>
<td>+ proximity zone (as factor)</td>
<td>214.51</td>
<td>28,019.32</td>
</tr>
<tr>
<td>+ medical density</td>
<td>258.29</td>
<td>28,041.54</td>
</tr>
<tr>
<td>+ access time to provider + dist to 2nd provider</td>
<td>201.63</td>
<td>28,010.59</td>
</tr>
<tr>
<td>+ access time to provider + prox zone</td>
<td>212.60</td>
<td>28,018.32</td>
</tr>
<tr>
<td>+ dist to 2nd provider + prox zone</td>
<td>198.97</td>
<td>28,016.04</td>
</tr>
<tr>
<td>+ dist to 2nd provider + access time to prov. + prox zone</td>
<td>197.69</td>
<td>28,007.01</td>
</tr>
<tr>
<td>+ dist to 2nd provider + access time to prov. + prox zone + unstructured spatial effect</td>
<td>198.59</td>
<td>28,007.21</td>
</tr>
</tbody>
</table>

Analysis carried out using the package INLA of R.
Sensitivity analysis

We check the robustness of the best selected model.

- Sensitivity analysis is focused on the precision parameters: crucial problem in bayesian generalized mixed models is the specification of distributions for the random effect precision (see Roos and Held (2011)) parameters;

- We compute a mean root square of the squared percentage error: for a given covariate, we define

\[
Err = \frac{100}{N} \sqrt{\sum_{i=1}^{N} \left( \frac{\hat{f}(i) - \hat{f}(i)^{(B)}}{\hat{f}(i)^{(B)}} \right)^2}
\]

- \( N \) is the dimension of the dataset \( N = 13572 \);
- \( \hat{f}(i)^{(B)} \) is the posterior mean estimate of the point \( u_i \) by the model with default prior;
- \( \hat{f}(i) \) is the corresponding estimate by the model with alternative prior.
Sensitivity analysis on the best multivariate model

<table>
<thead>
<tr>
<th>Prior</th>
<th>Covariate</th>
<th>Time</th>
<th>Dist.</th>
<th>Age</th>
<th>ICAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HNorm $\tau^{-1}$</td>
<td>$\mathcal{N}_+ (0.01)$</td>
<td>0.0020</td>
<td>0.0023</td>
<td>0.0065</td>
<td>0.0030</td>
</tr>
<tr>
<td>Gamma $\Gamma(1,0.01)$</td>
<td>0.0017</td>
<td>0.0022</td>
<td>0.0060</td>
<td>0.0031</td>
<td></td>
</tr>
<tr>
<td>Gamma $\Gamma(10,0.001)$</td>
<td>0.0018</td>
<td>0.0019</td>
<td>0.0057</td>
<td>0.0029</td>
<td></td>
</tr>
<tr>
<td>Gamma $\Gamma(100,0.0001)$</td>
<td>0.0003</td>
<td>0.00004</td>
<td>0.00005</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>Gamma $\Gamma(1,0.001)$</td>
<td>Default prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values of the mean root square of the squared percentage error between estimates using default priors and alternative priors.
Results: Access time effect (exponential scale), best multivariate model
Results: distance to second healthcare provider effect (exponential scale), best multivariate model
Results: Spatial effect (exponential scale), best multivariate model
Results: fixed proximity zone effect (exponential scale), best multivariate model

<table>
<thead>
<tr>
<th>Zone</th>
<th>Posterior mean</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1: Altkirch</td>
<td>0.87</td>
<td>0.69 - 1.11</td>
</tr>
<tr>
<td>Z2: Colmar</td>
<td>0.33</td>
<td>0.19 - 0.58</td>
</tr>
<tr>
<td>Z3: Guebwiller</td>
<td>0.74</td>
<td>0.50 - 1.07</td>
</tr>
<tr>
<td>Z4: Saint-Louis</td>
<td>0.92</td>
<td>0.70 - 1.20</td>
</tr>
<tr>
<td>Z5: Sélestat</td>
<td>0.08</td>
<td>0.02 - 0.31</td>
</tr>
<tr>
<td>Z6: Thann</td>
<td>1.10</td>
<td>0.87 - 1.39</td>
</tr>
<tr>
<td>Z7: Mulhouse</td>
<td></td>
<td>Reference zone</td>
</tr>
</tbody>
</table>

Reference zone
Conclusions

- Comprehensive analysis, results can be easily interpreted;
- The "best" model explains (some) of the processes going on;
- The "best" model can be used for cost-effective care planning;
- INLA is a useful and flexible tool for fitting spatial models with even a complex dependence structure;
- Running time is very fast.
Bibliography


Bibliography


