Comparison of Bayesian methods for flexible modeling of spatial risk surfaces in disease mapping.

S. Sturtz, K. Ickstadt, Biometrical Journal 56 (2014)

Bernd Taschler, 07 May 2015
covariates: excess or relative risk factors
  ▶ main covariate of interest: benzene emission levels

model comparison
  ▶ Poisson/Gamma random field model\(^1\) (PGRF)
  ▶ Markov random field based ecologic regression model\(^2\) (MRF)
  ▶ Bayesian Detection of Clusters and Discontinuities model\(^3\) (BDCD)

simulation study

real data example - Leukemia

\(^1\) Wolpert, Ickstadt (1998)
\(^2\) Besag et al (1991); Clayton, Bernardinelli (1992)
\(^3\) Knorr-Held, Rasser (2000)
Figure 1: Number of observed (a) and expected cases (b) in the period of 1985–1996.
295 cases of leukemia in children under 15 years old in the area of Inner London (Thames Cancer Registry, 1985-1996)

310 electoral wards

national leukemia rate $r_{st}$ (for age-sex stratum $s$ and year $t$)

expected number of cases per region $i = 1, \ldots, n$:

$$E_i = \sum_s \sum_t r_{st} N_{ist}$$

standardised mortality ratio:

$$\text{SMR}_i = \frac{O_i}{E_i}$$

($O_i$ ... number of observed cases)
Figure 2: Calculated SMRs for the leukemia data set (a). In (b) the increased SMR of 135.1 in the central region of (a) is set to be NA (white) for a better presentation of the risk surface.
Benzene is known to be associated with increased risk of various types of leukemia.

Effects of low level (environmental) Benzene concentrations is unknown.

**Data:** Benzene is monitored indirectly on 1km x 1km grid cells (atmospheric emissions inventory for London).
Estimated Benzene emission rate

Figure 3: Benzene exposure data on 1 km $\times$ 1 km grid cells.
Model description - Poisson / Gamma RF model

- take points $y \in \mathcal{Y}$

- Poisson process $N(dy)$ on $\mathcal{Y}$ with mean $\Lambda(y)w(dy)$, where $w(dy)$ is the population reference measure

- mean risk surface or 'intensity' $\Lambda(y)$ depends on a set of risk factors

- excess risk factors $a_j \in J_A$ and relative risk factors $a_j \in J_M$ (all with corresponding regression coefficients $\theta_j$)
Poisson / Gamma RF model

- accounting for non-measured covariates: introduce latent covariate $X_*$ with regression coefficient $\theta_*$

- model $X_*$ as a kernel mixture of a random measure $\Gamma(ds)$ on an auxiliary space $S$

- $\Gamma(ds)$ consists of latent sources located at $\mu_s$ with magnitudes $\gamma_s$

- bivariate Gaussian kernels $k(y, s) = \exp \left\{ -\frac{1}{2} \left( \frac{(y_1 - \mu_1)^2}{\sigma_1^2} + \frac{(y_2 - \mu_2)^2}{\sigma_2^2} \right) \right\}$

Thus, influence of the latent covariate at location $y$ is given by:

$$\sum_s k(y, s) \gamma_s$$
Poisson / Gamma RF model

Level 1: Points

Intensity

\[ N(dy) \sim \text{Pois}(\Lambda(y)w(dy)), \ y \in \mathcal{Y} \]

\[ \Lambda(y) = \left( \theta_0 + \sum_{j \in J_A} a_j \theta_j + \sum_{s \in S} k(y, s) \gamma_s \theta_* \right) \]

\[ \times \exp \left( \sum_{j \in J_M} a_j \theta_j \right) \]

Level 2: Latent sources

\[ \Gamma(ds) \sim \text{Gamma}(\alpha(ds), \beta(ds)) \]

Level 3: Parameter

\[ \theta \sim \pi(\theta)d\theta. \]
Alternative models

MRF-based ecologic regression model\textsuperscript{4}

- observed cases are assumed to follow a Poisson distribution
- Poisson rate depends on number of expected cases

\[ N_i \sim \text{Poi}(\mu_i) \]

\[
\log(\mu_i) = \log(E_i) + \alpha + V_i + U_i + \beta_{\text{benz}}(B_i)
\]

\textsuperscript{4} details in dissertation by Sturtz (2007)
Alternative models

**BDCD model**\(^5\)

- spatial partition model
- also assumes a Poisson distribution for the observed cases
- partition of regions into clusters is estimated by Reversible Jump MCMC
- Poisson rate depends on the expected cases as well as on a relative risk that is assumed to be constant across clusters
- does not allow for continuous covariates

\(^5\)details in dissertation by Sturtz (2007)
Simulation studies

**Simulation goals:**
- identification and modelling of clustered structures
- estimation of latent field (risk surfaces)
- different covariate interpretations: benzene as excess (additive) or relative (multiplicative) risk factor

- low vs high risk scenarios for benzene

- different spatial patterns (e.g. benzene only, linear trend, increased risk in 3 random clusters of different sizes, etc.)

- model comparison: DIC, MSE

- data generation:

  \[ O_i \sim \text{Poi}(\Lambda_i E_i) \]
Table 1  Structures used for data generation: × marks structures not described in detail in this paper, capital letters correspond to structures involving benzene additively (A) or multiplicatively (M).

<table>
<thead>
<tr>
<th>⊕ benzene</th>
<th>⊗ benzene</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>A₂</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>330</td>
<td>770</td>
</tr>
</tbody>
</table>
Table 2  Summary of the DICs achieved for different structures; capital letters correspond to the underlying simulated structures, lower case letters to the applied Poisson/gamma random field model (benzene included as excess risk factor: model a; benzene included as relative risk factor: model m; benzene not included: model o), the MRF-based ecologic regression model (MRF) and the BDCD model (BDCD), bold numbers indicate the best-fitting model.

<table>
<thead>
<tr>
<th>Structure</th>
<th>model a</th>
<th>model m</th>
<th>model o</th>
<th>MRF</th>
<th>BDCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>(+1 cluster)</td>
<td>346.2</td>
<td>+28.5</td>
<td>+38.2</td>
<td>+41.4</td>
</tr>
<tr>
<td>M₁</td>
<td>(benz only)</td>
<td>321.9</td>
<td>+0.9</td>
<td>+4.0</td>
<td>+0.5</td>
</tr>
<tr>
<td>M₂</td>
<td>(+1 cluster)</td>
<td>373.2</td>
<td>+0.3</td>
<td>+2.0</td>
<td>+13.7</td>
</tr>
<tr>
<td>M₂^{high}</td>
<td>(+1 cluster)</td>
<td>+83.9</td>
<td>341.1</td>
<td>+137.3</td>
<td>+59.8</td>
</tr>
<tr>
<td>M₃</td>
<td>(+smooth cov)</td>
<td>+1.4</td>
<td>338.3</td>
<td>+4.4</td>
<td>+14.7</td>
</tr>
<tr>
<td>M₄</td>
<td>(+step)</td>
<td>+25.2</td>
<td>+25.5</td>
<td>+23.1</td>
<td>+28.7</td>
</tr>
<tr>
<td>M₅</td>
<td>(+3 clusters)</td>
<td>+34.2</td>
<td>+34.2</td>
<td>+58.7</td>
<td>+68.1</td>
</tr>
</tbody>
</table>
Simulation Results

- MRF models assume spatial dependencies similar to multiplicative PGRF models, but results are inferior.

- MRF models don’t reliably reproduce the size of the true covariate effect and in all simulation scenarios never represent the best fitting model.

- BDCD model performs best on clustered structures.

- PGRF outperforms the other two models in all scenarios except for clustered structures that are not dominated by excess risk factors.

- Number of kernels used in PGRF: inclusion of 5-10 latent risk sources is enough (models not including benzen require more kernels).
Figure 4: Scatterplot matrix of generated $\Lambda_i E_i$ for structure $M_5$ and estimated $\hat{\Lambda}_i E_i$ of models $M_5 m15$, $M_5 a15$, $M_5 o7$, $M_5 z$ (MRF model), and $M_5 y$ (BDCD model), and plot of rates $\hat{\Lambda}_i$ for model $M_5 o7$. 
Figure 1: Number of observed (a) and expected cases (b) in the period of 1985–1996.
Figure 6: Scatterplot matrix of estimated $\hat{\Lambda}_i E_i$ of models applied to the leukemia data set: BDCD model with the lowest overall DIC, Poisson/gamma random field model a1, model m10, model o11 and MRF model including benzene, and plot of rates $\hat{\Lambda}_i$ for model o11.
Figure 7: Spatial risk surfaces of the rates $\hat{\Lambda}_i$ estimated by the MRF model including benzene (a) and BDCD (b).
Figure 8: Poisson/gamma random field model including benzene as a relative risk factor and 10 Gaussian kernels: (a) spatial risk surface of the rates $\hat{\Lambda}_i$, (b) latent random random field.
**Table 4**  DIC values for Poisson/gamma random field models with different number of kernels as well as for the BDCD and MRF model applied to the leukemia data set.

<table>
<thead>
<tr>
<th># latent factors</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>model a</td>
<td>417.0</td>
<td>390.8</td>
<td>394.6</td>
<td>417.1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>model m</td>
<td>415.5</td>
<td>391.0</td>
<td>397.3</td>
<td>391.8</td>
<td>388.9</td>
<td>386.9</td>
<td>385.2</td>
<td>384.1</td>
<td>383.4</td>
<td>382.8</td>
<td>382.5</td>
<td>382.5</td>
</tr>
<tr>
<td>model o</td>
<td>—</td>
<td>390.8</td>
<td>396.6</td>
<td>393.3</td>
<td>390.5</td>
<td>388.2</td>
<td>386.4</td>
<td>384.8</td>
<td>383.8</td>
<td>383.2</td>
<td>383.0</td>
<td>382.9</td>
</tr>
<tr>
<td>BDCD model</td>
<td>375.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRF model</td>
<td>390.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with benzene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRF model</td>
<td>391.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>without benzene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- PGRF is much more flexible than other models
- PGRF can model excess risk factors and continuous covariates
- Drawbacks
  - Difficulties with abruptly changing risk (i.e., clusters)
  - Computational cost
Thank you!