A spatio-temporal model for cancer incidence data with zero-inflation

Monica Musio $^1$, Erik A. Sauleau $^2$

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

**Spatial 2**
September 1-2 2011, Foggia
Cancer registries: background

- Collect exhaustively individual data on cases of cancer
- Data are aggregated to calculate incidence and mortality summaries
- We analyse data from Haut-Rhin cancer registry:
  - General registry
  - Covering a "departement" of 750,000 inhabitants
  - Website: www.arer68.org
  - Noth-East of France adjacency with Germany and Switzerland.
Dataset: Prostate

- **Covariates:**
  - **Age** into 9 groups: [0-45 years), 5-year intervals and [80 or more]
  - **Time:** year of diagnosis, from 1988 to 2005
  - **Geographical unit of residence**, with centroid coordinates (we have 27 GU)
  - Total number of cases **6878**

- **Population counts**
  - 1999 census for 1998 to 2002
  - 2005 census for 2003 to 2005

- Dataset counts were spread over 4374 cells with 1935 zeros (the **44.2% are zeros**).
Response variable: standardized incidence ratio (SIR)

- Ratio of the observed cases in each geographical unit on the expected cases:
  \[ \text{SIR}_i = \frac{O_i}{E_i} \]

- Expected cases are the result of the exposition of the population at-risk to a certain risk \( E_i = \hat{p}_i N_i \).

- Global risk in the study region
  \[ \forall i, \hat{p}_i = \hat{p} = \frac{\sum \sum \cdots \sum O}{\sum \sum \cdots \sum N}. \]
Review of methodology: ZIP model

Higher incidence of zeros than expected under Poisson distribution

⇒ zero-inflated Poisson distribution (Lambert (1992)):

\[ Pr(O, \lambda, \omega) = \begin{cases} 
\omega + (1 - \omega)e^{-\lambda} & \text{if } O = 0 \\
(1 - \omega)e^{-\lambda} \frac{\lambda^O}{O!} & \text{if } O > 0 
\end{cases} \]
Review of methodology: ZIP model

- Covariates may then enter into the model through the mean $\lambda$ and/or through the probability $\omega$.
- Zero inflation probability constrained to be proportional to the Poisson mean (Lambert (1992));
- Zero-inflated generalized additive model (Chiogna, Gaetan (2007)), where both $\lambda$ and $\omega$ can be modelled as a function of some nonparametric smooth predictors. The two smooth predictors are unconstrained;
- The two processes may be partially constrained ((Liu and Chan (2010))).
Review of methodology

- **ZIGAM:**

\[
g_{\lambda}(\lambda) = \alpha + \sum_{j=1}^{n_1} f_j(x_j) + \sum_{j=1}^{n_2} h_j(y_j)
\]

\[
g_{\omega}(\omega) = \beta + \sum_{j=1}^{m_1} f_j^*(x_j) + \sum_{j=1}^{m_2} h_j^*(y_j)
\]

- **COZIGAM:**

\[
g_{\omega}(\omega) = \beta + \delta_1 f_1(x_1) + \cdots + \delta_{n_1} f_{n_1}(x_{n_1}) + \sum_{j=1}^{m_2} h_j^*(y_j)
\]

\[g_{\lambda}\] and \[g_{\omega}\] link functions.
Notations

- Indices: $a$ for age category (1-9), $t$ for year (1988-2005) and $r$ for GU (1-27)
- Number of cases: $O \rightarrow O_{atr}$
- Expected cases: $E \rightarrow E_{atr}$

The model

$$O_{atr} | \lambda_{atr} \sim f(o_{atr}) = \begin{cases} 0 & \text{with probability } 1 - \omega_{atr} \\ \mathcal{P}(\lambda_{atr}) & \text{with probability } \omega_{atr} \end{cases}$$

Then:

$$\log(\lambda_{atr}) = \log(E_{atr}) + \mu_{atr}$$

and the linear predictor $\mu_{atr}$ is modelled as a semiparametric additive model.
\[ \mu_{atr} = \alpha + f_a(\text{age}_a) + f_{a,t}(\text{age}_a, \text{year}_t) + f_{e,n}(\text{east}_r, \text{north}_r) + f_t(\text{year}_t) \]

- \( f_a(\text{age}_a), f_t(\text{year}_t) \) cubic regression spline;
- \( f_{e,n}(\text{east}_r, \text{north}_r) \) thin plate regression spline basis - isotropic;
- \( f_{a,t}(\text{age}_a, \text{year}_t) \) 2-d tensor product smooth
  - Tensor products are \textbf{scale invariant}: allow to model interactions between two or more variables which have different scale of measures.
Concerning \( \omega \)...

- COZIGAM: We assume
  
  \[
  \text{logit}(\omega_{\text{atr}}) = \beta + \delta_1 f_a(\text{age}_a) + \delta_2 f_{e,n}(\text{east}_r, \text{north}_r),
  \]

- ZIGAM: In this case \( \text{logit}(\omega_{\text{atr}}) = \beta + \sum_{k=1}^{K} f_k(x_k) \).
  Different sets of covariates can be used for modelling \( \mu \) and \( \omega \).
Model selection

- In a bayesian framework the posterior probability of model \( M_i \) equals

\[
P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)},
\]

under a flat prior \( P(M_i) \) the posterior model probability is proportional to the marginal likelihood

\[
P(D|M_i) = \int P(D|\theta, M_i)P(\theta|M_i)d\theta;
\]

- Following (Liu and Chain (2010)) we use marginal likelihood as the model selection criterion which maximises the posterior model probability;

- Preference will be given to model with larger marginal likelihood;

- Use Laplace methods to approximate the marginal likelihood for GAM, COZIGAM and ZIGAM.
Compare different models via their marginal likelihood $L$:

<table>
<thead>
<tr>
<th>Model</th>
<th>Effects in $\mu$</th>
<th>Effects in $\omega$</th>
<th>$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod0</td>
<td>a+s+t+(a,t)</td>
<td></td>
<td>-5 765.933</td>
</tr>
<tr>
<td>mod1</td>
<td>a+s+t+(a,t) a</td>
<td></td>
<td>-5 701.080</td>
</tr>
<tr>
<td>mod2</td>
<td>a+s+t+(a,t) s</td>
<td></td>
<td>-5 733.567</td>
</tr>
<tr>
<td>COZIGAM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analysis carried out using packages COZIGAM and mgcv of R.
Results: age effect in log scale
Results: age-time effect in log scale
Results: Spatial effect in log scale

Spatial effect estimated by model 3 (on the natural scale)
Results: time effect in log scale
Results: parameter estimations for $\omega$ mod1

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>standard error</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>5.429</td>
<td>1.302</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>$\delta_{age}$</td>
<td>1.572</td>
<td>0.359</td>
<td>&lt;0.00001</td>
</tr>
</tbody>
</table>
Results: residuals

- $E(O) = \omega \mu$
- $\text{Var}(O) = \omega \mu (1 + \mu - \omega \mu)$

Standardized residuals are given by

$$\hat{e} = \frac{O - \hat{\omega} \hat{\mu}}{\sqrt{\hat{\omega} \hat{\mu} (1 + \hat{\mu} - \hat{\omega} \hat{\mu})}}.$$

Q-Q plot (García Ben and Yohai (2004)).
Results: Model check

Proportions of observed values (red dots) and boxplot of simulated values from mod0 (left side) and mod1 (right side).
Conclusions

- First comprehensive analysis;
- COZIGAM provide framework for modelling incidence data with zero inflation
  - Concerning the Poisson process: There is a strong age-time effect;
  - Concerning the zero inflation process: the main effect is this one of the age;
- It is possible to include more complex covariance structures.
Bibliography


