Data mining cancer registries
Retrospective surveillance of small area time trends in cancer incidence using BaySTDetect

Guangquan Li, Sylvia Richardson, Lea Fortunato, Ismail Ahmed, Anna Hansell, Mireille Toledano and Nicky Best

Department of Epidemiology and Biostatistics
Imperial College London

International Workshop on Spatial Spatio-temporal Data Mining, Dec. 11, 2011
Definition of surveillance

“... (disease surveillance is) the continued watchfulness over the distribution and trends of incidence through the systematic collection, consolidation, analysis and dissemination” of data in public health practice.

Outline

Motivation

BaySTDetect: Bayesian model choice for detecting unusual temporal patterns in small area data

Application: Thames Cancer Registry, UK

Conclusions
Outline

Motivation

BaySTDetect: Bayesian model choice for detecting unusual temporal patterns in small area data

Application: Thames Cancer Registry, UK

Conclusions
The Thames Cancer Registry, UK

- The Thames Cancer Registry (TCR) collects data on newly diagnosed cases of cancer in the population of London and South East England.
- It is one of the largest cancer registries in Europe, covering a population of over 12 million, and holds nearly 3 million cancer registration records.
- For illustration purposes, in this talk, we focus on
  - 3 common cancer types: Colon cancer (male and female combined); Breast cancer (female only) and Lung cancer (separately for male and female) for population of age $\geq 30$;
  - Time: 1981-2008 aggregate by 4 year intervals $\rightarrow$ 7 time points;
  - Space: London and South East England at the electoral ward level (1899 wards in total and avg. pop. 5k).
Motivation

- For chronic diseases/cancer, local time trends tend to resemble each other closely, reflecting the national trend.
- Here, we are interested in detecting areas with “unusual” time profiles, which may be due to
  - emergence of localized risk factors (e.g., a new source of pollution);
  - local policy implementation (e.g., health awareness or screening campaigns);
  - changes to health care provision or social structure of the local population
  - local variations in diagnostic or coding practice.
- Important to flag-up areas for further investigations.
Figure: Time profiles of breast cancer incidence for 200 wards

Remark

▶ Difficult to assess departures of the local temporal patterns by “eyeballing”
Problems in small area surveillance

1. Sparse data (small number of cases)
   - BaySTDetect employs the Bayesian multilevel modelling framework to allow appropriate information borrowing.
Problems in small area surveillance

1. Sparse data (small number of cases)
   - BaySTDetect employs the Bayesian multilevel modelling framework to allow appropriate information borrowing.

2. Multiple comparisons are made
   - A Bayesian procedure is used in BaySTDetect to derive decision rules which enable the control of the false discovery rate (FDR).
Outline

Motivation

BaySTDetect: Bayesian model choice for detecting unusual temporal patterns in small area data

Application: Thames Cancer Registry, UK

Conclusions
BaySTDetect: modelling framework

Data level

\[ y_{it} \sim \text{Poisson}(\mu_{it} \cdot E_{it}) \]

Modelling underlying risks

**Model 1:** Time trend pattern is the same for all areas

**Model 2:** Time trends are estimated independently for each area
BaySTDetect: modelling framework

Data level

\[ y_{it} \sim \text{Poisson}(\mu_{it} \cdot E_{it}) \]

Modelling underlying risks

\[ \log(\mu_{it}) = \text{Common time trend} + \text{Common spatial pattern} + \text{Area-specific time trends} \]

Model 1: Time trend pattern is the same for all areas
Model 2: Time trends are estimated independently for each area

Selection

A model indicator \( z_i \) indicates for each area whether Model 1 (\( z_i = 1 \)) or Model 2 (\( z_i = 0 \)) is supported by the data.

\[ \mu_{it} = z_i \cdot \mu^{(M1)}_{it} + (1 - z_i) \cdot \mu^{(M2)}_{it} \]
A detection rule based on FDR

- Define $p_i = P(z_i = 1|\text{data})$ which is the probability that area $i$ belongs to the common trend model (Model 1)
  - A small $p_i$ suggests that area $i$ is unlikely to follow the common trend.

- We need to set a suitable cut-off value, $C$, such that areas with $p_i < C$ are declared to be unusual.

- Put another way, if we declare area $i$ to be unusual, then $p_i$ can be thought of as the probability of false detection for that area.

- We chose $C$ in such a way that we ensure that the average probability of false detection (i.e. the average value of $p_i$) amongst areas declared to be unusual is less than some pre-set level $\alpha$.

- This procedure ensures that, on average, the number of false positives is no more than $(k \times \alpha)$, where $k$ is the number of declared unusual areas.
A detection rule based on FDR

- Define $p_i = P(z_i = 1|\text{data})$ which is the probability that area $i$ belongs to the common trend model (Model 1)
  - A small $p_i$ suggests that area $i$ is unlikely to follow the common trend.
- We need to set a suitable cut-off value, $C$, such that areas with $p_i < C$ are declared to be unusual.
A detection rule based on FDR

- Define $p_i = P(z_i = 1 | \text{data})$ which is the probability that area $i$ belongs to the common trend model (Model 1).
  - A small $p_i$ suggests that area $i$ is unlikely to follow the common trend.

- We need to set a suitable cut-off value, $C$, such that areas with $p_i < C$ are declared to be unusual.

- Put another way, if we declare area $i$ to be unusual, then $p_i$ can be thought of as the probability of false detection for that area.

- We chose $C$ in such a way that we ensure that the average probability of false detection (i.e. the average value of $p_i$) amongst areas declared to be unusual is less than some pre-set level $\alpha$. 
A detection rule based on FDR

- Define $p_i = P(z_i = 1|\text{data})$ which is the probability that area $i$ belongs to the common trend model (Model 1)
  - A small $p_i$ suggests that area $i$ is unlikely to follow the common trend.

- We need to set a suitable cut-off value, $C$, such that areas with $p_i < C$ are declared to be unusual.

- Put another way, if we declare area $i$ to be unusual, then $p_i$ can be thought of as the probability of false detection for that area.

- We chose $C$ in such a way that we ensure that the average probability of false detection (i.e. the average value of $p_i$) amongst areas declared to be unusual is less than some pre-set level $\alpha$.

- This procedure ensures that, on average, the number of false positives is no more than ($k \times \alpha$), where $k$ is the number of declared unusual areas.
Outline

Motivation

BaySTDetect: Bayesian model choice for detecting unusual temporal patterns in small area data

Application: Thames Cancer Registry, UK

Conclusions
TCR results: number of detected wards

Table: Summary of detected wards at various FDR pre-set levels.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>FDR=0.1</th>
<th>FDR=0.2</th>
<th>FDR=0.3</th>
<th>FDR=0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>19</td>
<td>54</td>
<td>101</td>
<td>177</td>
</tr>
<tr>
<td>Colon</td>
<td>3</td>
<td>8</td>
<td>18</td>
<td>40</td>
</tr>
<tr>
<td>Male Lung</td>
<td>14</td>
<td>39</td>
<td>84</td>
<td>147</td>
</tr>
<tr>
<td>Female Lung</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>18</td>
</tr>
</tbody>
</table>

Remarks:

- The FDR controlling level reflects the confidence in the detection results:
  - For instance, wards detected as unusual at the 0.1 FDR level on average have a lower probability of being false discoveries than those detected FDR=0.4.
- However, one may miss out some truly unusual wards at a lower FDR level.
Detected wards: Colon cancer

- FDR=0.1 (3 wards detected)
- FDR=0.2 (8 wards detected)
- FDR=0.3 (18 wards detected)
- FDR=0.4 (40 wards detected)
Detected wards: Lung cancer

Male

Female

FDR=0.1 (14 wards detected)
FDR=0.2 (39 wards detected)
FDR=0.3 (84 wards detected)
FDR=0.4 (147 wards detected)
Interactive map on Google Earth
Selected unusual trends: colon cancer (FDR=0.2)
Post-processing the detected trends

- With a relatively large number of detected areas (e.g., breast and male lung cancer), examination of the individual trends becomes difficult;

- For the detected areas, the estimated RR trends from the local trend model are fed into a standard hierarchical clustering method (hclust in R);

- The cluster-specific trends are then compared to the overall RR trend.
Post-processing the detected trends

Cluster Dendrogram

1 cluster

2 clusters

3 clusters

4 clusters

5 clusters

Breast cancer
FDR=0.2

Black line = common trend
Coloured lines = average local trend in each cluster
Outline

Motivation

BaySTDetect: Bayesian model choice for detecting unusual temporal patterns in small area data

Application: Thames Cancer Registry, UK

Conclusions
Conclusions

- We have demonstrated the data mining feature of BaySTDetect using the TCR data;
- We have explored various ways to visualise the detection results;
- Implemented in R and WinBUGS, BaySTDetect enables real-time analysis of routinely collected data;
- In the conference paper, we have presented a simulation study on increasing detection power in BaySTDetect for dealing with rare cancer types.
- Papers and WinBUGS codes for this model are available on www.bias-project.org.uk.
Acknowledgement

▶ This project is funded by the ESRC National Center for Research Methods through the BIAS II project.
▶ Thanks to the Thames Cancer Registry and the Small Area Health Statistics Unit (SAHSU) for providing the cancer incidence data.

Thank you!!
References


2. Li G, Best N, Hansell A, Ahmed I, and Richardson S. BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice (under revision);

BaySTDetect: Model specification

Let $y_{it}$ and $E_{it}$ denote the observed and expected cases, respectively, in area $i$ at time $t$, we have

$$y_{it} \sim \text{Poisson}(E_{it} \cdot \mu_{it})$$

$$\log(\mu_{it}) = \begin{cases} \alpha_0 + \eta_i + \gamma_t & \text{Model 1 for all } i, t \\ u_i + \xi_{it} & \text{Model 2 for all } i, t. \end{cases}$$

Model 1

$$\eta_i \sim \text{spatial BYM model}$$

$$\gamma_t \sim \text{random walk [RW}(\sigma_\gamma)^2)]$$

Model 2

$$u_i \sim N(0, 1000)$$

$$\xi_{i,t} \sim \text{random walk [RW}(\sigma_{\xi,i})^2]$$

Common spatial pattern

Common temporal pattern

Area-specific temporal pattern
Simulation: Setup

- Simulated data were based on the observed COPD mortality data (see Li et al. 2012).
- Three departure patterns were considered.
- When simulating the data, either the original set of expected counts from the COPD data or a reduced set (multiplying the original by 1/5) were used.
- 15 areas (approx. 4%) were chosen to have the unusual trend patterns.
  - areas were chosen to cover a wide range expected count values and overall spatial risks.
- Results were compared to those from the popular SaTScan space-time scan statistic.
Simulation: Unusual patterns

**Figure:** Illustration of the three departure patterns (red), compared to the common trend pattern in black. The departure magnitude in this plot is 1.5.

**Pattern 1**

**Pattern 2**

**Pattern 3**
Simulation: Sensitivity

Figure: Sensitivity of detecting the 15 truly unusual areas with departure pattern 2 (departure magnitude = 1.5 (top row) and 2 (bottom row)).

BaySTDetect (FDR = 0.1) SaTScan (p = 0.05)