Detection of Spatial Cluster for Suicide Data using Echelon Analysis

Fumio Ishioka (Okayama University, Japan)

Makoto Tomita (Tokyo Medical and Dental University, Japan)

Toshiharu Fujita (The Institute of Statistical Mathematics, Japan)
Introduction

• The number of suicides in Japan is around 25,000 per year until 1997.

• However, in 1998 it was suddenly more than three million people and it has remained at that level until now.

• For the number of suicides in Japan by the vital statistics of the Ministry of Health, Labour and Welfare, 30,827 people in 2007 is number two after in 2003, which is a major social problem.

<table>
<thead>
<tr>
<th>Suicide rate in 2008 by World Health Organization (WHO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan … 23.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Major countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>17.6</td>
</tr>
</tbody>
</table>
Introduction

- The number of suicides in Japan is around 25,000 per year until 1997.

- However, in 1998 it was suddenly more than three million people and it has remained at that level until now.

- For the number of suicides in Japan by the vital statistics of the Ministry of Health, Labour and Welfare, 30,827 people in 2007 is number two after in 2003, which is a major social problem.

For this serious problem, it is clear that a statistical implication is important.

<table>
<thead>
<tr>
<th>Suicide rate in 2008 by World Health Organization (WHO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan … 23.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Major countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>17.6</td>
</tr>
</tbody>
</table>
About data

• As an analysis area, we use 70 regions at **Kanto area (secondary medical care zone)** in central part of Japan.

• We investigate the suicides among **men** in **1973-2007**.
  
  Specially dealt in six time periods;
  
  1\(^{st}\) period … 1973-1982  \hspace{0.5cm} 2\(^{nd}\) period … 1983-1987
  
  3\(^{rd}\) period … 1988-1992  \hspace{0.5cm} 4\(^{th}\) period … 1993-1997
  
  5\(^{th}\) period … 1998-2002  \hspace{0.5cm} 6\(^{th}\) period … 2003-2007
Spatial Cluster for the Suicide Data
Background

- The importance of statistical analyses for spatial data has increased in various scientific fields.

- A statistical technique for the spatial data has ever been established.

- One interesting aspect of spatial data analysis is detection of cluster areas that have significantly higher values: so-called hotspot.

Objective – Detection of hotspots for spatial data

It is very important to find areas where disease outbreak, abnormal environment, aberration, something unusual, etc.
About Spatial Data

\[ D \subset \mathbb{R}^d \] Random field at locations in fixed subset \( D \) of \( d \)-dimensional Euclidean space \( \mathbb{R}^d \).

1. Geostatistical data
   - Measurements taken at fixed locations.
   - The locations are generally spatially continuous.
   
   Example: Rainfall recorded at weather stations.

2. Spatial Point Patterns
   - Locations themselves are the variable of interest.
   - They consist of a finite number of locations.

   Example: Positions of an earthquake center.
3. **Lattice data**

- Observations associated with spatial regions.
- The regions can be regularly or irregularly spaced.

**Regularly example:** Information obtained by remote sensing from satellites.

\[
D_{ij} = \{(x, y) \mid x_{i-1} < x < x_i, y_{j-1} < y < y_j\}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, m
\]

**Irregularly example:** Population corresponding to each county in a state.

\[
D_i, i = 1, 2, \ldots, n
\]

- A neighborhood information for the spatial regions is available.

In this study, the suicide data is a type of **irregular lattice data**.
**Spatial scan statistic**

- Spatial scan statistic (Kulldorff, 1997) can detect areas of markedly high rates based on likelihood ratio.

- It is currently a very popular and useful method, and it has been mainly used in a field of epidemiology.

- Kulldorff established the spatial scan statistic based on Poisson model.
Spatial scan statistic

✓ “G” is a whole area.
✓ “n”s are population in G.
✓ “c”s are observed cases in G.

✓ Suppose a geographical cluster candidate area “Z” within the G.
\[ Z \subset G, \quad G = Z \cup Z^c \]

✓ Here, “p_1” and “p_2” are internal and external probability of area Z, respectively.
\[
\begin{align*}
  p_1 &= \frac{c(Z)}{n(Z)} \\
  p_2 &= \frac{c(Z^c)}{n(Z^c)} = \frac{c(G) - c(Z)}{n(G) - n(Z)}
\end{align*}
\]
The likelihood function for the Poisson model is expressed as

\[
\exp[-p_1n(Z) - p_2(n(G) - n(Z))] \left[ \frac{p_1n(Z) + p_2(n(G) - n(Z))}{c(G)} \right]^{c(G)}
\]

(1)

The density function \( f(x) \) is

\[
\begin{cases} 
\frac{p_1n(x)}{p_1n(Z) + p_2(n(G) - n(Z))} & \text{if } x \in Z \\
\frac{p_2n(x)}{p_1n(Z) + p_2(n(G) - n(Z))} & \text{if } x \notin Z 
\end{cases}
\]

(2)
Spatial scan statistic

✓ We can hence, write the likelihood function as

\[
L(Z, p_1, p_2) = \frac{\exp[-p_1 n(Z) - p_2 (n(G) - n(Z))] [p_1 n(Z) + p_2 (n(G) - n(Z))]^{c(G)}}{c(G)!} \\
\times \prod_{x_i \in Z} \frac{p_1 n(x)}{p_1 n(Z) + p_2 (n(G) - n(Z))} \\
\times \prod_{x_i \notin Z} \frac{p_2 n(x)}{p_1 n(Z) + p_2 (n(G) - n(Z))}
\]

(3)

✓ In order to maximize the likelihood function, we calculate the maximum likelihood function conditioned to the area \(Z\).

✓ The maximum likelihood estimator

\[
\hat{p}_1 = \frac{c(Z)}{n(Z)}
\]

\[
\hat{p}_2 = \frac{(c(G) - c(Z))/(n(G) - n(Z))}{(c(G) - c(Z))/(n(G) - n(Z))}
\]

are substituted in the (3).
Spatial scan statistic

\[ L(Z) = \frac{\exp[-c(G)]}{c(G)!} \left( \frac{c(Z)}{n(Z)} \right)^{c(Z)} \left( \frac{c(G) - c(Z)}{n(G) - n(Z)} \right)^{c(G) - c(Z)} \prod_{x_i} n(x_i) \]  

(4)

✓ The likelihood ratio \( \lambda(Z) \) is maximized over all the subset area to detect the hotspot.

\[ \lambda(Z) = \frac{\text{Max } L(Z)}{L_0} = \frac{\left( \frac{c(Z)}{n(Z)} \right)^{c(Z)} \left( \frac{c(G) - c(Z)}{n(G) - n(Z)} \right)^{c(G) - c(Z)}}{\left( \frac{c(G)}{n(G)} \right)^{c(G)}} \]  

(5)

✓ Here, the \( L_0 \) means the likelihood function under the null hypothesis.

\[ L_0 = \sup_p \frac{\exp[-pn(G)]}{c(G)!} p^{c(G)} \prod_{x_i} n(x_i) = \frac{\exp[-c(G)]}{c(G)!} \left( \frac{c(G)}{n(G)} \right)^{c(G)} \prod_{x_i} n(x_i) \]  

(6)

The regions \( Z \) that attain the maximum \( \lambda \) is regarded as a hotspot.
Application to suicide data

• Kulldorff proposed using a circular window to detect regions $Z$ consisting of high $\lambda(Z)$.

<table>
<thead>
<tr>
<th>Period</th>
<th># regions</th>
<th># cases</th>
<th># expected</th>
<th>Incidence rate</th>
<th>$\log\lambda(Z)$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>21</td>
<td>5507</td>
<td>4459.52</td>
<td>1.23</td>
<td>134.70</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>22</td>
<td>3884</td>
<td>3081.51</td>
<td>1.26</td>
<td>114.25</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>22</td>
<td>3183</td>
<td>2589.65</td>
<td>1.23</td>
<td>74.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>23</td>
<td>3822</td>
<td>3298.84</td>
<td>1.16</td>
<td>47.06</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5th. (1998-2002)</td>
<td>22</td>
<td>5149</td>
<td>4593.74</td>
<td>1.12</td>
<td>37.95</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6th. (2003-2007)</td>
<td>22</td>
<td>6531</td>
<td>5612.04</td>
<td>1.16</td>
<td>87.28</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
**Application to suicide data**

- Kulldorff proposed using a circular window to detect regions $Z$ consisting of high $\lambda(Z)$.

We can see that the most likely cluster is located on a little outside from big cities such as Tokyo.

<table>
<thead>
<tr>
<th>Period</th>
<th># Regions</th>
<th># Cases</th>
<th># Expected</th>
<th>Incidence Rate</th>
<th>Log</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>21</td>
<td>5507</td>
<td>4459.52</td>
<td>1.23</td>
<td>134.70</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>22</td>
<td>3884</td>
<td>3081.51</td>
<td>1.26</td>
<td>114.25</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>22</td>
<td>3183</td>
<td>2589.65</td>
<td>1.23</td>
<td>74.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>23</td>
<td>3822</td>
<td>3298.84</td>
<td>1.16</td>
<td>47.06</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5th. (1998-2002)</td>
<td>22</td>
<td>5149</td>
<td>4593.74</td>
<td>1.12</td>
<td>37.95</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6th. (2003-2007)</td>
<td>22</td>
<td>6531</td>
<td>5612.04</td>
<td>1.16</td>
<td>87.28</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
discussion

• Kullorff’s scan method is useful to find circular-shaped clusters.

• However, it is difficult to detect clusters when they follow the shape of a river or a road.

• To overcome this problem, several non-circular scan techniques were proposed. (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Tango and Takahashi, 2005)

• In addition to these methods, we have proposed a non-circular hotspot detection, using Echelon analysis.
Echelon Approach for the Suicide Data
Echelon analysis

- Echelon analysis (Myers et al., 1997; Kurihara, 2004) is a useful technique to study the topological structure of a surface in the systematic and objective manner.

- Echelons are derived from the changes in topological connectivity with decreasing surface level.
**Echelon dendrogram**

- **Echelon dendrogram** is the graph which express exactly the structure of the spatial data.
Bayesian estimates

• When a group observed small population size, mortality rates vary greatly with a slight decrease in the number of suicide.

• In other words, the numbers become unstable because the effect of chance variation, small population size of population for suicide be used to calculate the comparison is often not suitable.

• Above mortality data, therefore, following age-adjusted death rate applied empirical Bayes estimates (Bayesian estimates) are used. (Fujita et al., 2003).

In this study, we use a Bayesian estimates as $h$ for echelon analysis.
Bayesian estimates for suicide data

Age-adjusted death rate (Bayesian estimation)

\[
\text{Death rate} = \sum \left( \frac{\# \text{ death by age class for observation } n + \hat{\beta}_i}{\text{Population by age class for observation } n + \hat{\alpha}_i} \times \frac{\text{Population by age class for base population}}{\text{Population for base population}} \right)
\]

where, \( \hat{\alpha}_i \) and \( \hat{\beta}_i \) are the prior distribution of the suicide situation in the country. (Gamma distribution selection)

Bayesian estimates are markedly elevated from 5th (1998- ) period.
Echelon analysis for suicide data

- A spatial structure of male suicide based on Bayesian estimates (for example, at 6th time period) is given by an echelon dendrogram.
**Hotspot detection**

- We find most likely cluster by scanning from the regions included in upper echelon to the regions included in bottom echelon.

Most likely cluster of male suicide, using echelon scan based on Bayesian estimates.

17 regions are included.
\[ \log \lambda = 107.60 \]
\[ p\text{-value} = 0.001 \]
### Hotspot detection

<table>
<thead>
<tr>
<th>Period</th>
<th># regions</th>
<th># cases</th>
<th># expected</th>
<th>Incidence rate</th>
<th>Log(\lambda(Z))</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>21</td>
<td>5123</td>
<td>3081.51</td>
<td>1.66</td>
<td>151.45</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>22</td>
<td>4078</td>
<td>3138.59</td>
<td>1.30</td>
<td>152.88</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>22</td>
<td>3042</td>
<td>2323.06</td>
<td>1.31</td>
<td>118.00</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>22</td>
<td>3576</td>
<td>2963.94</td>
<td>1.21</td>
<td>69.49</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5th. (1998-2002)</td>
<td>19</td>
<td>4089</td>
<td>3404.84</td>
<td>1.20</td>
<td>72.80</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6th. (2003-2007)</td>
<td>17</td>
<td>3386</td>
<td>2634.98</td>
<td>1.29</td>
<td>107.60</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Hotspot detection

- The most likely cluster exists northwest in all periods.
- However, there are little changes by periods.
- We can see that the most likely cluster is located on a little outside from big cities such as Tokyo, as well as using the circular scan.

<table>
<thead>
<tr>
<th></th>
<th># regions</th>
<th># cases</th>
<th># expected</th>
<th>Incidence rate</th>
<th>Log</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>21</td>
<td>5123</td>
<td>3081.51</td>
<td>1.66</td>
<td>151.45</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>22</td>
<td>4078</td>
<td>3138.59</td>
<td>1.30</td>
<td>152.88</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>22</td>
<td>3042</td>
<td>2323.06</td>
<td>1.31</td>
<td>118.00</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>22</td>
<td>3576</td>
<td>2963.94</td>
<td>1.21</td>
<td>69.49</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5th. (1998-2002)</td>
<td>19</td>
<td>4089</td>
<td>3404.84</td>
<td>1.20</td>
<td>72.80</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6th. (2003-2007)</td>
<td>17</td>
<td>3386</td>
<td>2634.98</td>
<td>1.29</td>
<td>107.60</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Comparison of two methods

- The echelon analysis based on Bayesian estimates provides the clusters with the high likelihood ratio than the circular scan in every period.

<table>
<thead>
<tr>
<th>Period</th>
<th>Log $\lambda(Z)$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>134.70</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>114.25</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>74.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>47.06</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5th. (1998-2002)</td>
<td>37.95</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6th. (2003-2007)</td>
<td>87.28</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

- The echelon scan could detect a high-grade hotspot area, in comparison with the circular scan.

Because...

- It is not limited to the shape of circularly.
- It scans from regions which create the peak structure having a high value.
Space-time Hotspot for the Suicide Data
**Spatial-temporal data**

- In many cases, spatial data is gotten by periodic observation such as year, month, day and so on.
- It is important to detect **hotspots based on spatial-temporal scale** as well as hotspots which obtained under the fixed time series.
- Spatial-temporal data is given by the overlapping same geographical areas for each time.

![Images of spatial-temporal data](image-url)
Space-time Hotspots

- Space-time hotspots mean the hotspots where the regions change into time-series.

- The space-time hotspots vary variously with time.

  - The spatial regions are represented schematically on the horizontal axis.
  - The time is represented on the vertical axis.
**Space-time Hotspots**

- Space-time hotspots mean the hotspots where the regions change into time-series.
- The space-time hotspots vary variously with time.

- The spatial regions are represented schematically on the horizontal axis.
- The time is represented on the vertical axis.
**Echelon analysis for Spatial-temporal data**

- We apply the echelon analysis to the spatial-temporal data. (e.g., polluted air or a contagious disease)

- By defining neighbor information for region $X(T, i)$ as follows, we simultaneously treat a time and a space.

\[
NB(X(T,i)) = \{(T,k) \mid \text{regions } i \text{ and } k \text{ are connected} \}
\]

\[
\cap X(T+1,i) \
\cap X(T-1,i)
\]
Application to the suicide data

57 regions are included. (Maximum hotspot size <=60)
log $\lambda = 7154.493$
$p$-value = 0.001
Application to the suicide data

Spatial-temporal data

<table>
<thead>
<tr>
<th>Period</th>
<th># regions</th>
<th># cases</th>
<th>Log</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>2</td>
<td>13252</td>
<td>7154.493</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Application to the suicide data

The space-time hotspot is suddenly expanding from 1998s!

<table>
<thead>
<tr>
<th>Time Period</th>
<th># Regions</th>
<th># Cases</th>
<th>Log</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. (1973-1982)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd. (1983-1987)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd. (1988-1992)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th. (1993-1997)</td>
<td>2</td>
<td>13252</td>
<td>7154.493</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Conclusion

• In this paper,

1) We investigated the spatial cluster of **male suicide** in Kanto area, by using the circular scan and the echelon scan.

2) Additionally, we investigated the transition and the tendency for six time periods by detecting space-time hotspot based on echelon analysis.
Conclusion

1) We investigated the spatial cluster of male suicide in Kanto area, by using the circular scan and the echelon scan.

• We can see that the most likely cluster is located on a little outside from big cities such as Tokyo.

• The result of echelon scan based on Bayesian estimates is shown to obtain higher likelihood clusters than the result of circular scan.

• The echelon scan is useful tool to detect spatial cluster because…
  1) it is not limited to the shape of circularly.
  2) it is efficient because of scanning from regions which create the peak structure having a high value.
  3) thus it helps a reduction of computation time.
2) Additionally, we investigated the transition and the tendency for six time periods by detecting space-time hotspot based on echelon analysis.

- We can simultaneously treat a time and a space by echelon analysis.
- The echelon analysis can express a time series change in hotspots.
- We could substantiate rapid increase in male suicide at Kanto from 1998s.
References


