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.

Suicide rate in 2008 by World Health Organization (WHO)

Japan … 23.7

Major countries

France	Germany	Canada	USA	Italy	UK
17.6	13.0	11.3	11.0	7.1	6.7

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.

Suicide rate in 2008 by World Health Organization (WHO)

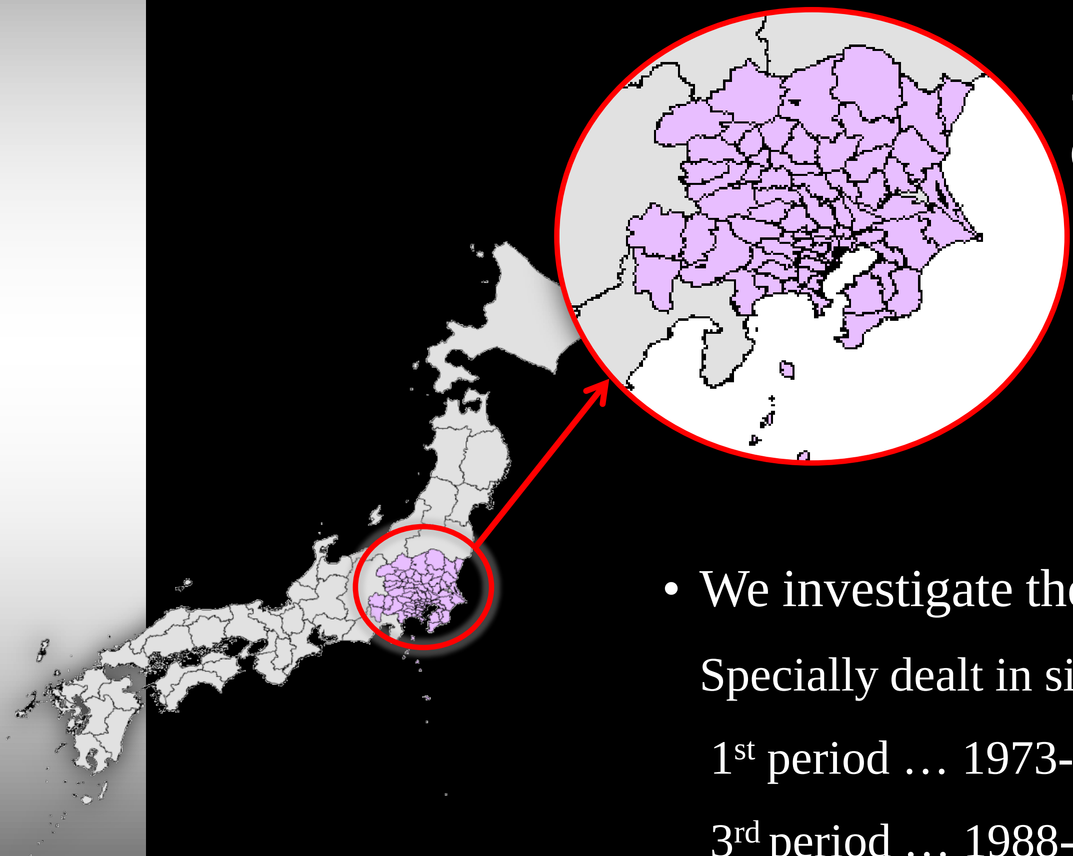
Japan … 23.7

Major countries

France	Germany	Canada	USA	Italy	UK
17.6	13.0	11.3	11.0	7.1	6.7

About data

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



70 regions at Kanto area
(secondary medical care zone)

- We investigate the suicides among men in 1973-2007.

Specially dealt in six time periods;

1st period … 1973-1982 2nd period … 1983-1987

3rd period … 1988-1992 4th period … 1993-1997

5th period … 1998-2002 6th 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.

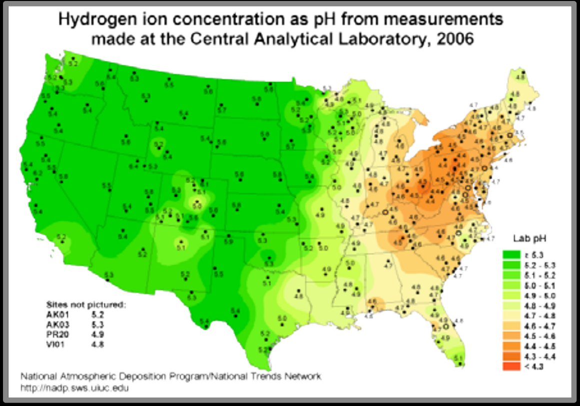
About Spatial Data

$D \subset \mathbf{R}^{d}$ Random field at locations in fixed subset D of d-dimensional Euclidean space $\mathbf{R}^{d}$.

1. Geostatistical data

- Measurements taken at fixed locations.
- The locations are generally spatially continuous.

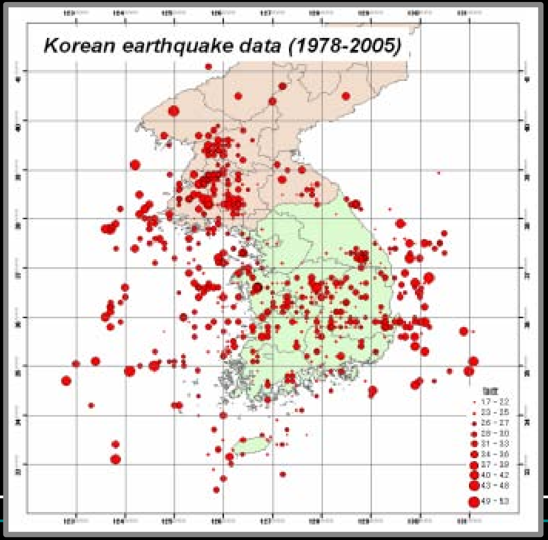
Example: Rainfall recorded at weather stations.



2. Spatial Point Patterns

- Locations themselves are the variable of interest.
- They consist of a finite number of locations.

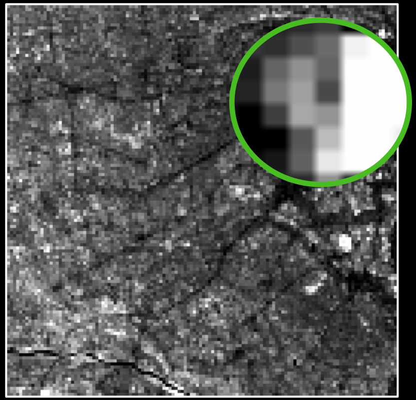
Example: Positions of an earthquake center.



3. Lattice data

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

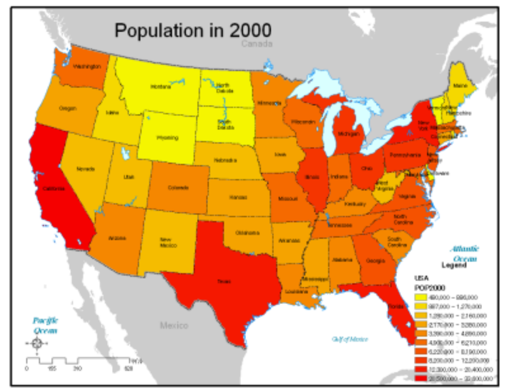
Regularly example: Information obtained by remote sensing from satellites.



Regular

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

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



Irregular

$$D_i, i = 1,2,\ldots,n$$

- 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

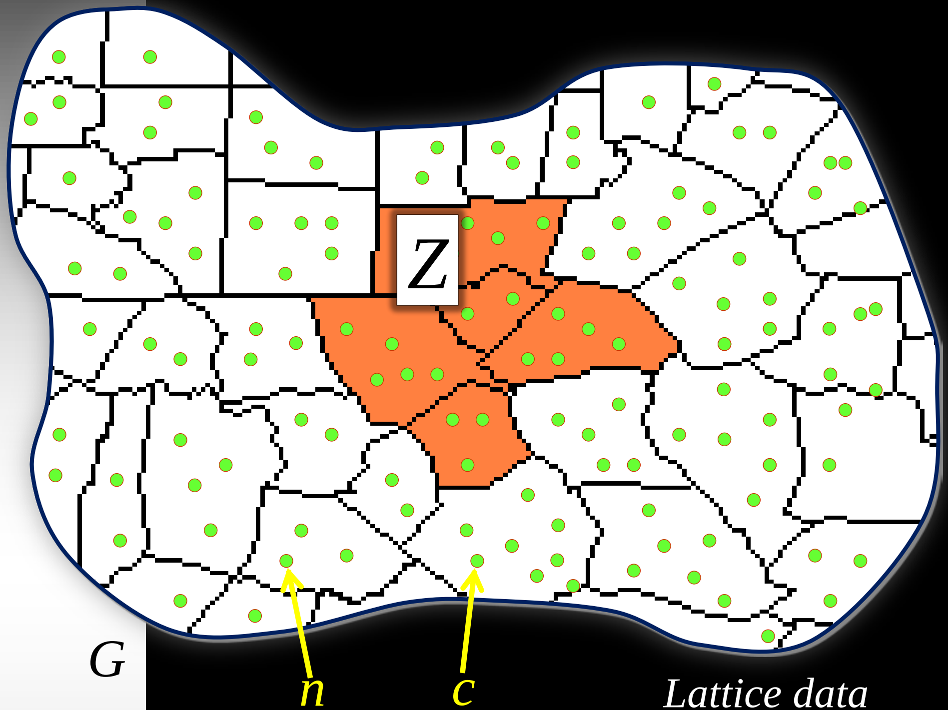
- Spatial scan statistic (Kulldorff, 1997) can detect areas of markedly high rates based on likelihood ratio.

We say it as a hotspot.

- 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



Lattice data

- ✓ "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.

$$Z \subset G,\ \ G = Z \cup Z^c$$

- ✓ Here, "p_1" and "p_2" are internal and external probability of area Z , respectively.

$$\begin{cases} p_1 = \dfrac{c(Z)}{n(Z)} \\ p_2 = \dfrac{c(Z^c)}{n(Z^c)} = \dfrac{c(G)-c(Z)}{n(G)-n(Z)} \end{cases}$$

Spatial scan statistic

Null hypothesis		Alternative hypothesis
$H_0: p_1 = p_2 = p$	v.s.	$H_1: p_1 > p_2$

✓ The likelihood function for the Poisson model is expressed as

$$\frac{\exp[-p_1 n(Z) - p_2(n(G)-n(Z))][p_1 n(Z) + p_2(n(G)-n(Z))]^{c(G)}}{c(G)\,!} \quad (1)$$

✓ The density function $f(x)$ is

$$\begin{cases} \dfrac{p_1 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} & if \;\; x \in Z \\[2ex] \dfrac{p_2 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} & if \;\; x \notin Z \end{cases} \quad (2)$$

Spatial scan statistic

✓ We can hence, write the likelihood function as

$$L(Z, p_1, p_2) = \frac{\exp[-p_1 n(Z) - p_2(n(G)-n(Z))][p_1 n(Z) + p_2(n(G)-n(Z))]^{c(G)}}{c(G)\,!} \times \prod_{x_i \in Z}^{n \in Z} \frac{p_1 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} \times \prod_{x_i \notin Z}^{n \notin Z} \frac{p_2 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} \qquad (3)$$

$$= \frac{\exp[-p_1 n(Z) - p_2(n(G)-n(Z))]}{c(G)\,!} p_1^{c(Z)} p_2^{c(G)-c(Z)} \prod_{x_i}^{n} n(x_i)$$

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

✓ The maximum likelihood estimator

$$\left.\begin{aligned} \hat{p}_1 &= c(Z)/n(Z) \\ \hat{p}_2 &= (c(G)-c(Z))/(n(G)-n(Z)) \end{aligned}\right\} \text{are substituted in the (3).}$$

Spatial scan statistic

$$L(Z)=\frac{\exp[-c(G)]}{c(G)\,!}\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}\prod_{x_i}^{n} n(x_i) \qquad (4)$$

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

$$\lambda(Z)=\frac{\underset{Z}{Max}\,L(Z)}{L_0}=\frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}} \qquad (5)$$

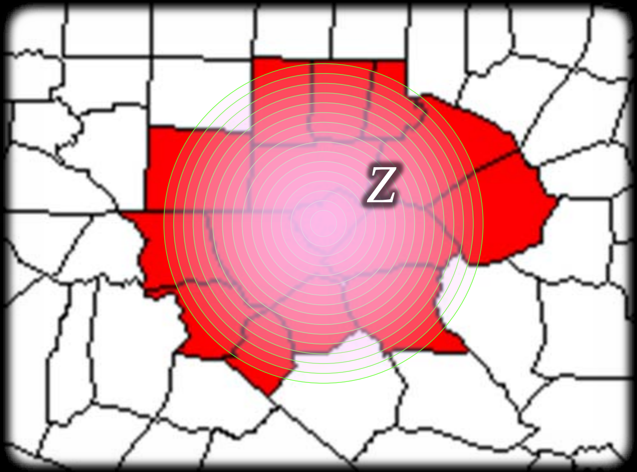
✓ Here, the L_0 means the likelihood function under the null hypothesis.

$$L_0=\sup_p \frac{\exp[-pn(G)]}{c(G)\,!}\,p^{c(G)}\prod_{x_i}^{n} n(x_i)=\frac{\exp[-c(G)]}{c(G)\,!}\left(\frac{c(G)}{n(G)}\right)^{c(G)}\prod_{x_i}^{n} n(x_i) \qquad (6)$$

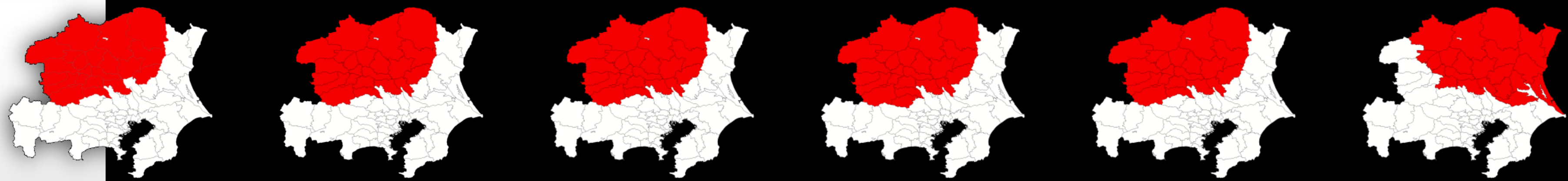
The regions Z that attain the maximum λ is regarded as a hotspot.

Application to suicide data

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



Method of circular window's scan



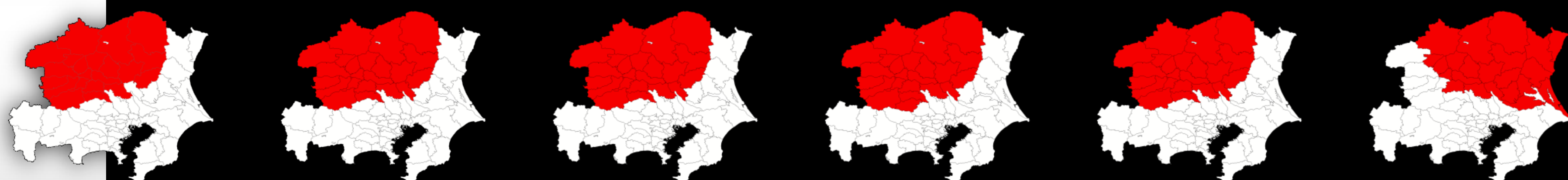
	# regions	# cases	# expected	Incidence rate	Log $\lambda(Z)$	p-value
1st. (1973-1982)	21	5507	4459.52	1.23	134.70	< 0.001
2nd. (1983-1987)	22	3884	3081.51	1.26	114.25	< 0.001
3rd. (1988-1992)	22	3183	2589.65	1.23	74.87	< 0.001
4th. (1993-1997)	23	3822	3298.84	1.16	47.06	< 0.001
5th. (1998-2002)	22	5149	4593.74	1.12	37.95	< 0.001
6th. (2003-2007)	22	6531	5612.04	1.16	87.28	< 0.001

Application to suicide data

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

Method of circular window's scan

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



1st period 2nd period 3rd period 4th period 5th period 6th period

	# regions	# cases	# expected	Incidence rate	Log	p-value
1st. (1973-1982)	21	5507	4459.52	1.23	134.70	< 0.001
2nd. (1983-1987)	22	3884	3081.51	1.26	114.25	< 0.001
3rd. (1988-1992)	22	3183	2589.65	1.23	74.87	< 0.001
4th. (1993-1997)	23	3822	3298.84	1.16	47.06	< 0.001
5th. (1998-2002)	22	5149	4593.74	1.12	37.95	< 0.001
6th. (2003-2007)	22	6531	5612.04	1.16	87.28	< 0.001

discussion

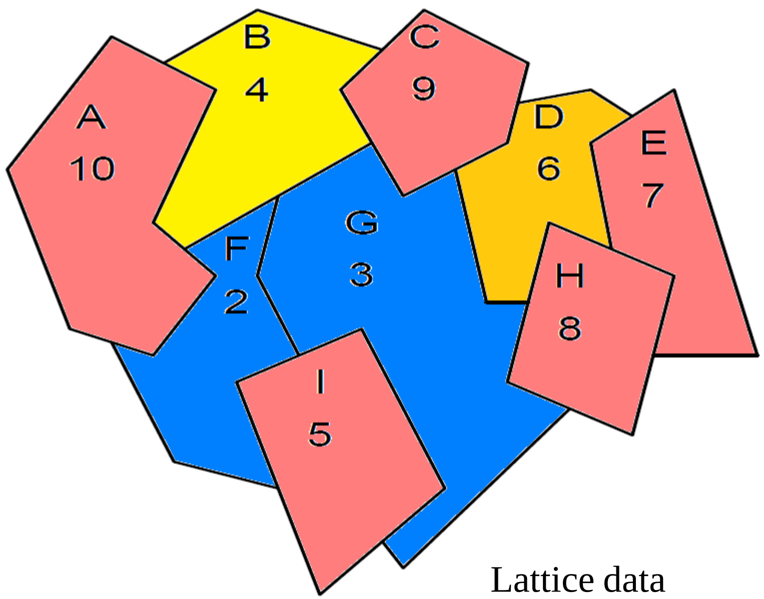
- Kullorff's scan method is useful to find circular-shaped clusters.
- However, it is difficult to detect clusters when they follow the shape of a river or a road.
- To overcome this problem, several non-circular scan techniques were proposed.
  (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Tango 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

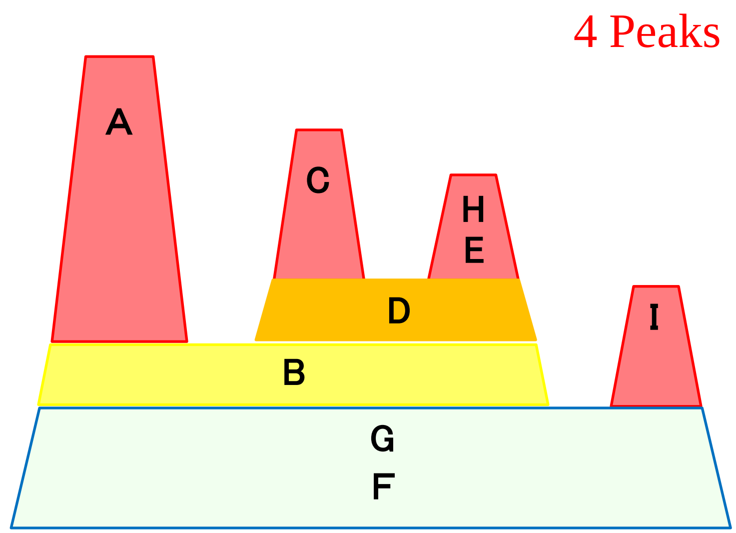
Echelon analysis

- 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.



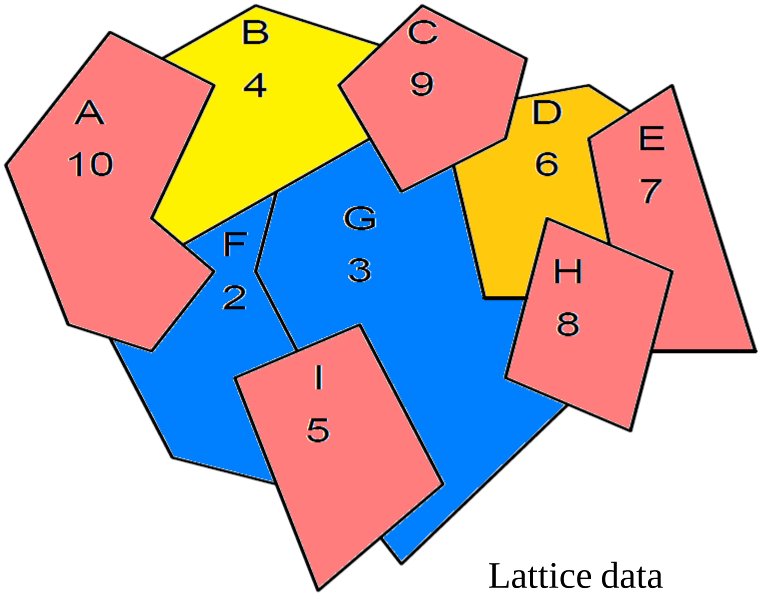
Lattice data



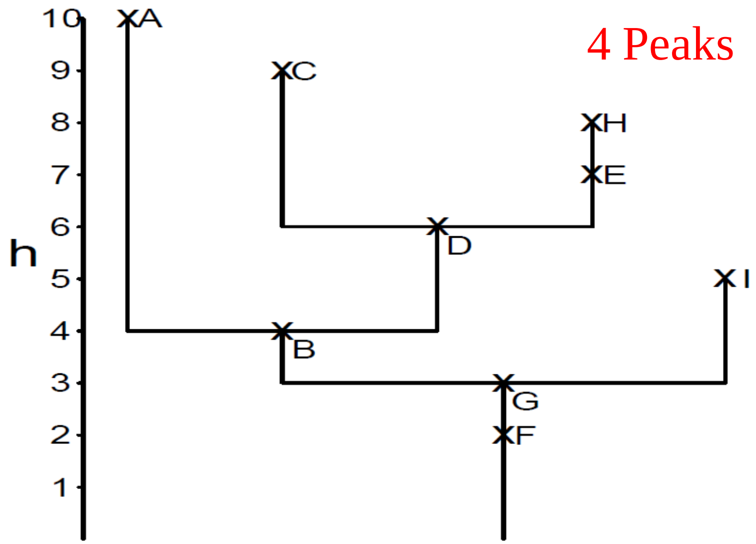
Division to the same phase based on echelon

Echelon dendrogram

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



Lattice data



Division to the same peaks based on echelon

Echelon dendrogram

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).

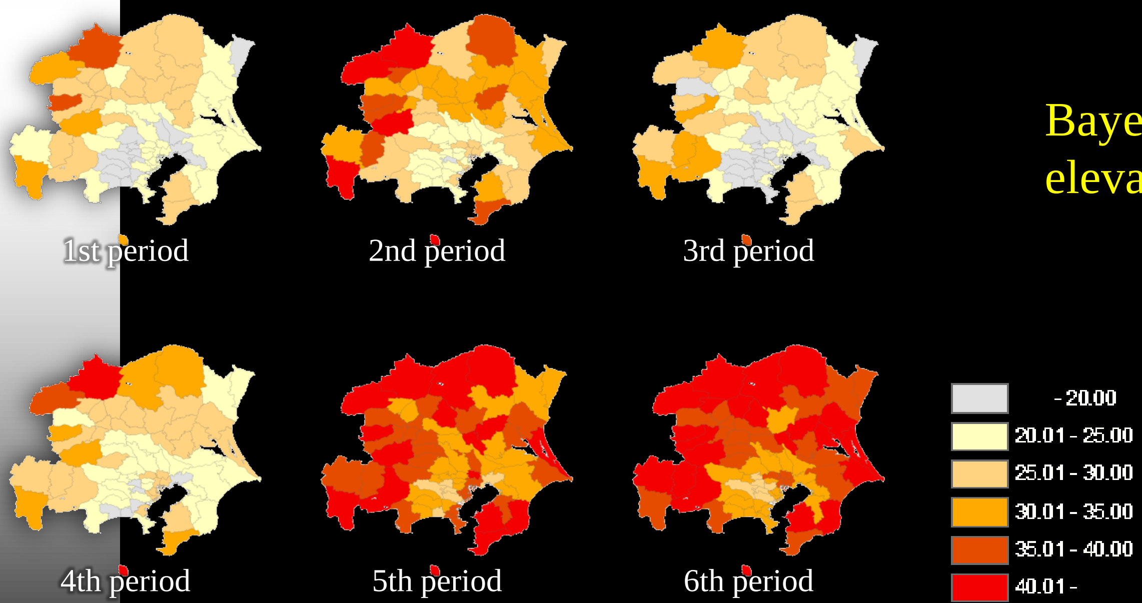
In this study, we use a Bayesian estimates as *h* for echelon analysis.

Bayesian estimates for suicide data

Age - adjusted death rate (Bayesian estimation)

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

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

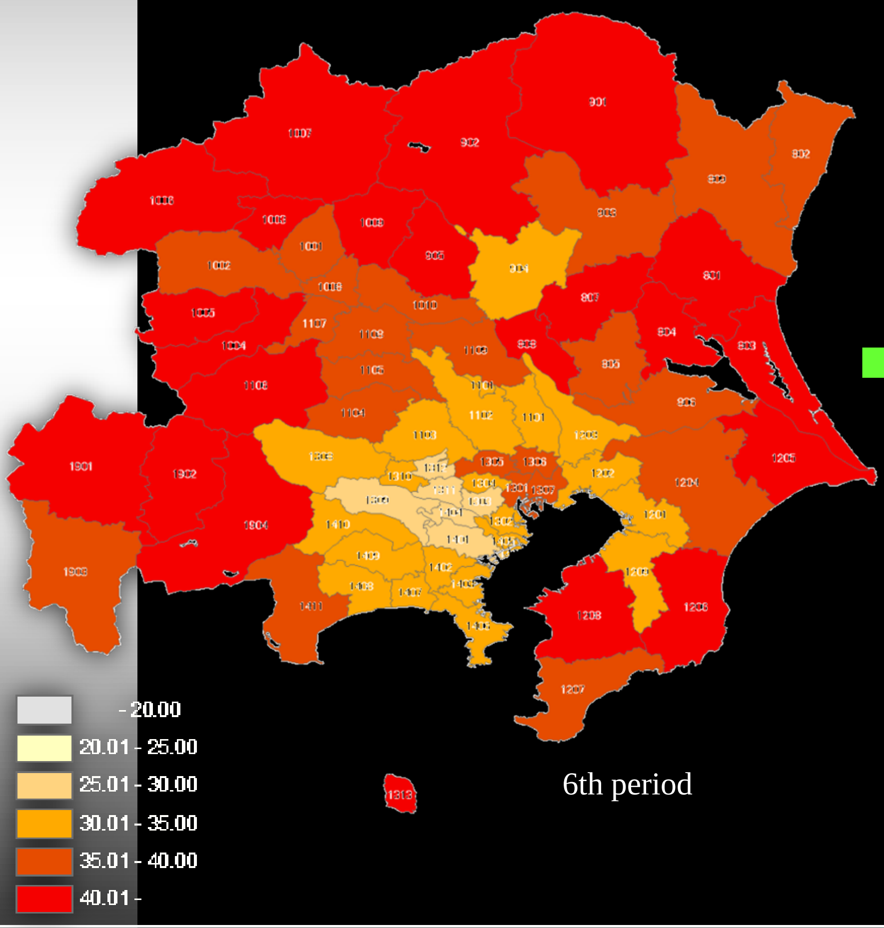


Bayesian estimates are markedly-elevated from 5th (1998-) period.

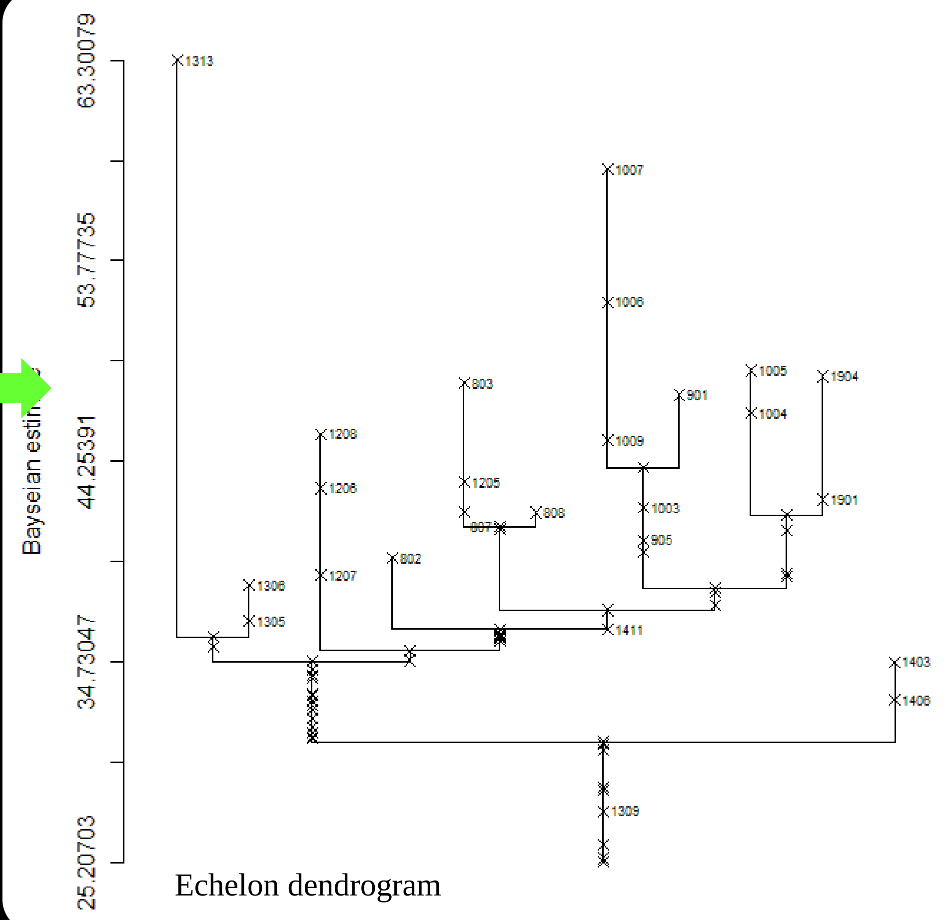
Choropleth map of Bayesian estimation

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.



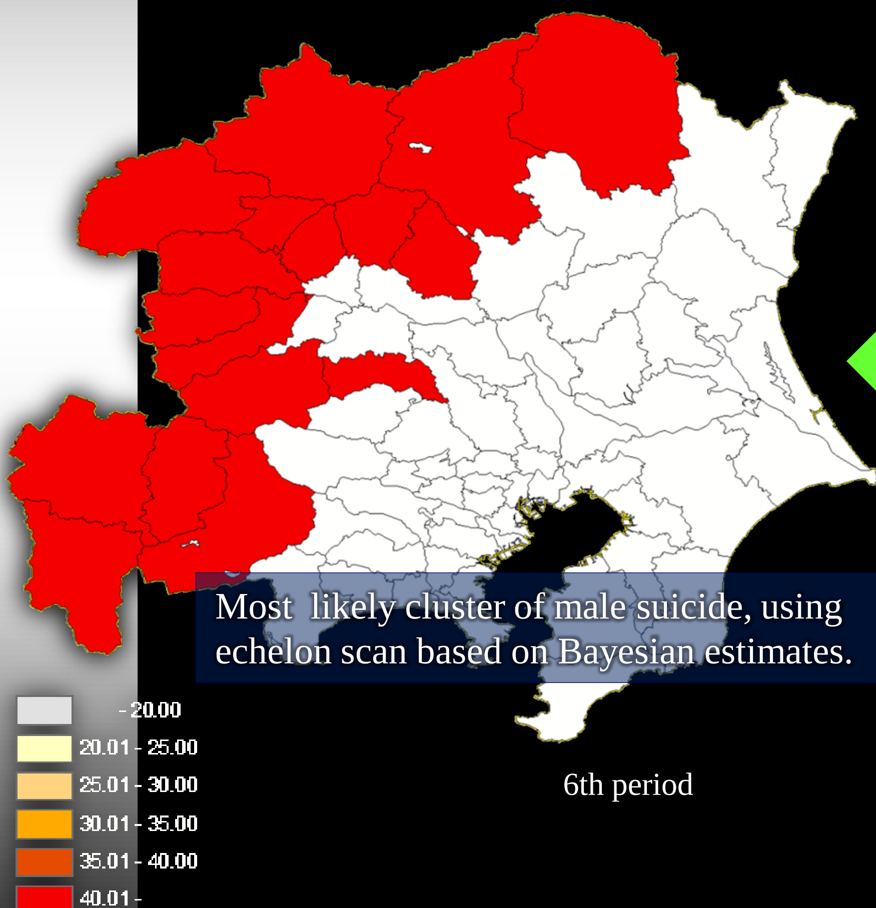
6th period

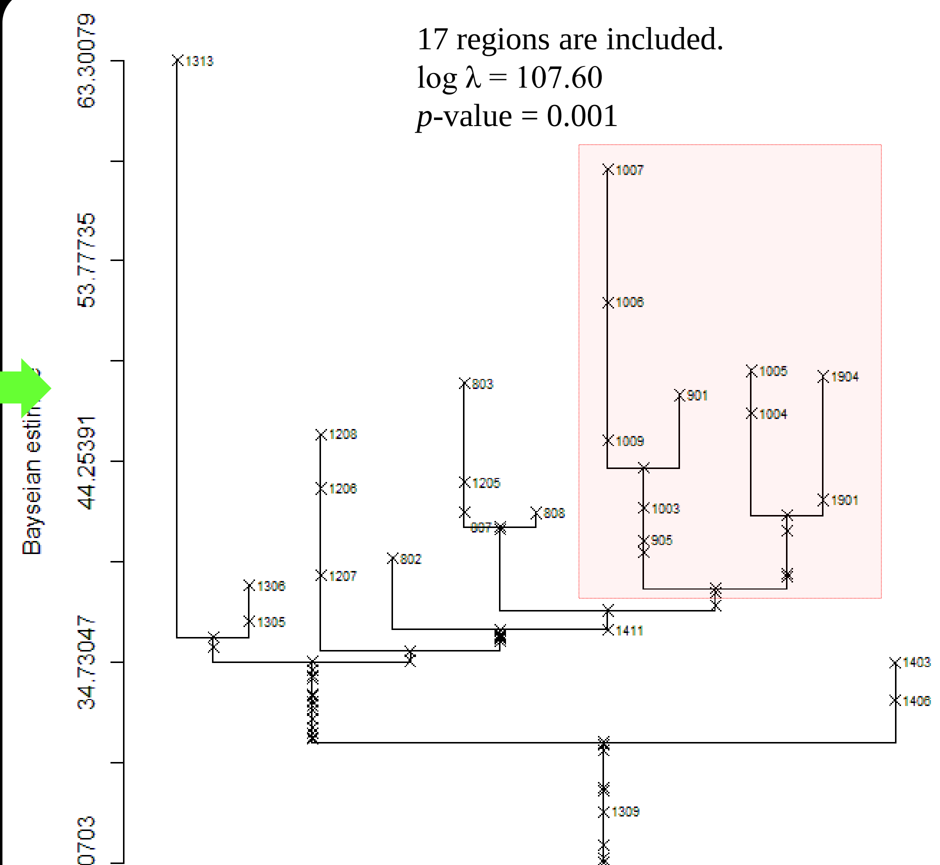


Echelon dendrogram

Hotspot detectoin

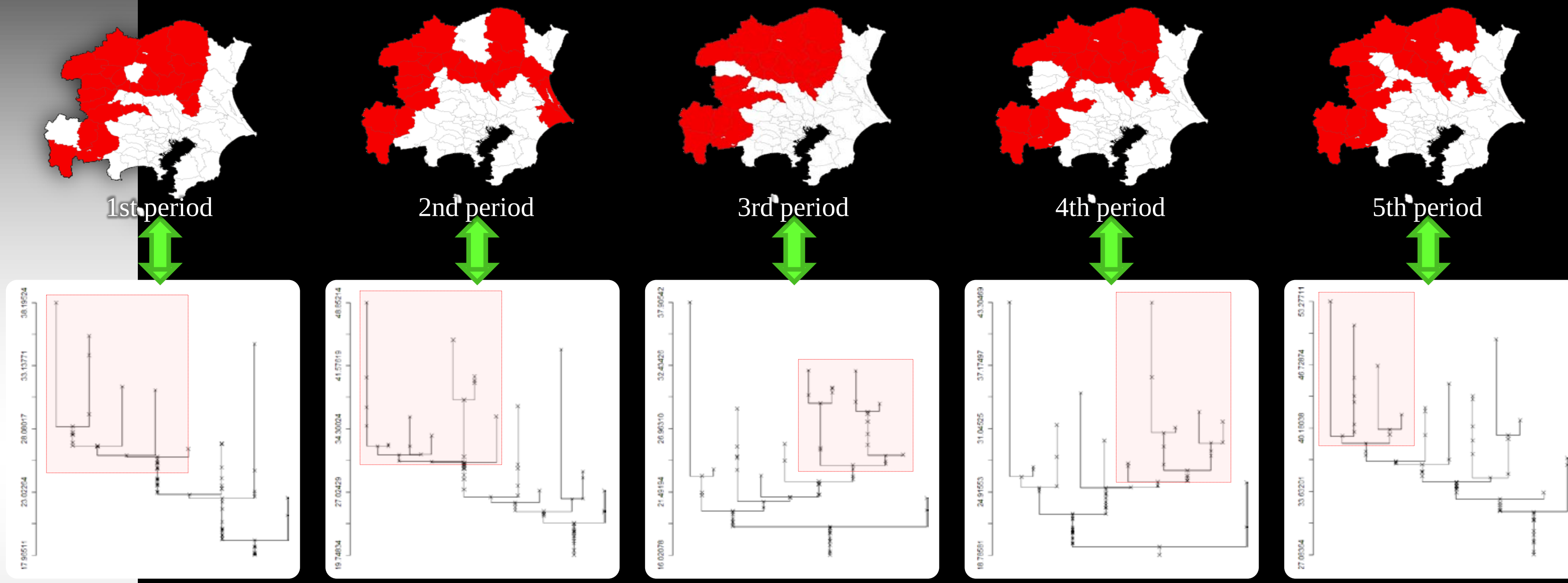
- We find most likely cluster by scanning from the regions included in upper echelon to the regions included in bottom echelon.





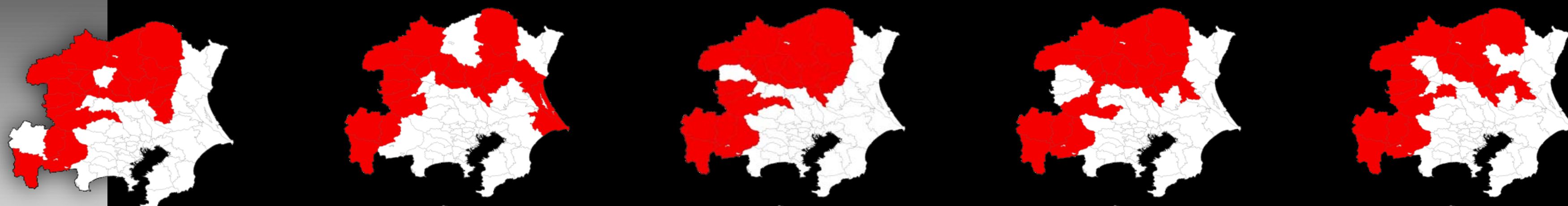
Echelon dendrogram

Hotspot detection



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

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.

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

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.

Kulldorff's circular scan

	Log $\lambda(Z)$	p-value
1st. (1973-1982)	134.70	< 0.001
2nd. (1983-1987)	114.25	< 0.001
3rd. (1988-1992)	74.87	< 0.001
4th. (1993-1997)	47.06	< 0.001
5th. (1998-2002)	37.95	< 0.001
6th. (2003-2007)	87.28	< 0.001

Echelon scan

	Log $\lambda(Z)$	p-value
1st. (1973-1982)	151.45	< 0.001
2nd. (1983-1987)	152.88	< 0.001
3rd. (1988-1992)	118.00	< 0.001
4th. (1993-1997)	69.49	< 0.001
5th. (1998-2002)	72.80	< 0.001
6th. (2003-2007)	107.60	< 0.001

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

Because…

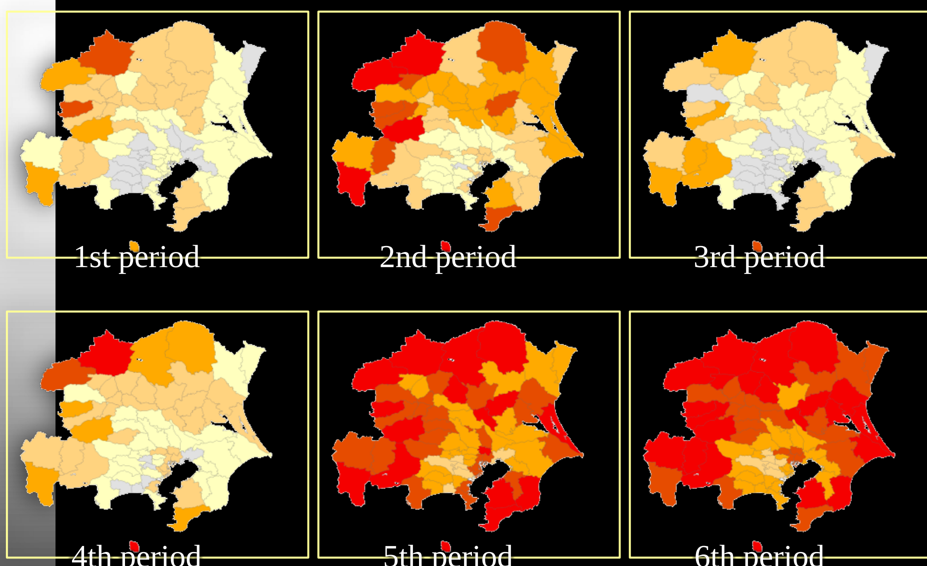
It is not limited to the shape of circularly.

It scans from regions which create the peak structure having a high value.

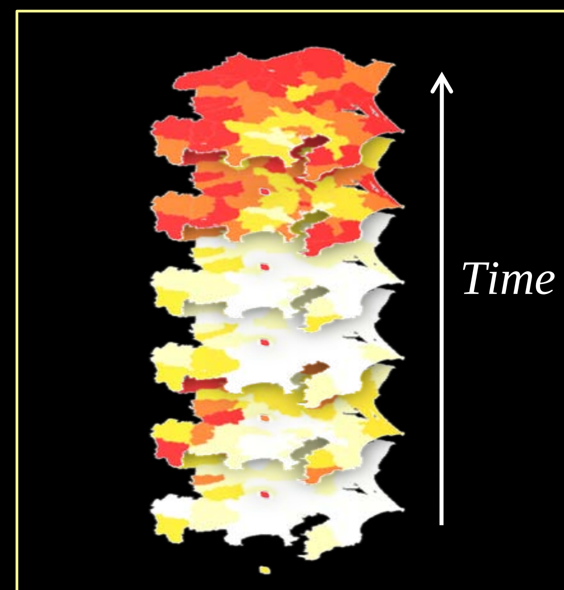
Space-time Hotspot for the Suicide Data

Spatial-temporal data

- In many case, 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.



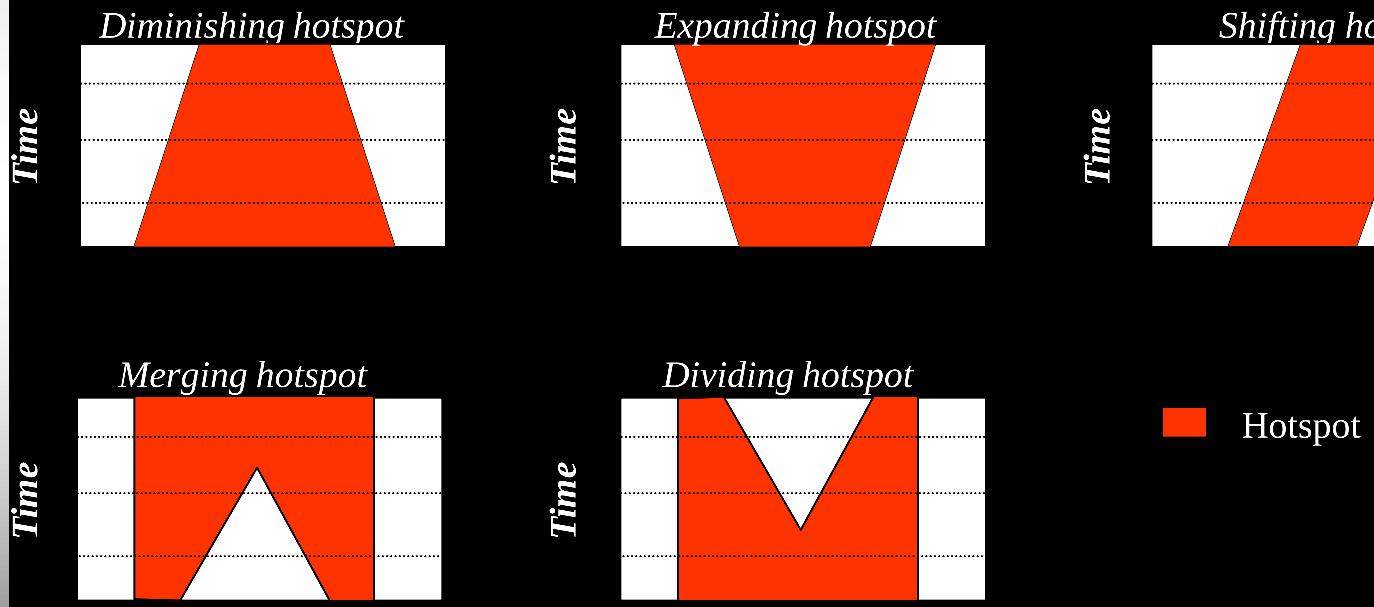
Spatial data of each period (fixed time series)



Spatial-temporal data

Space-time Hotspots

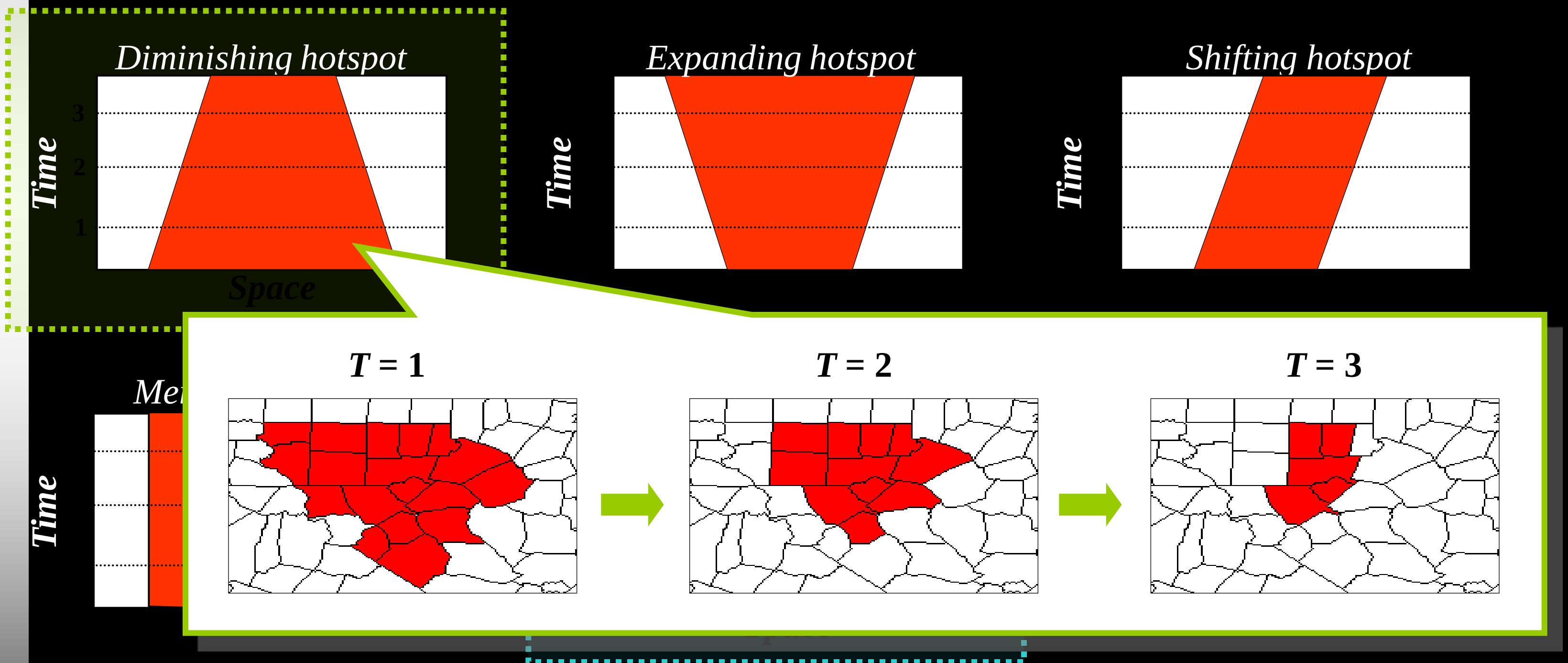
- Space-time hotspots mean the hotspots where the regions change into time-series.
- The space-time hotspots vary variously with time.



- 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.

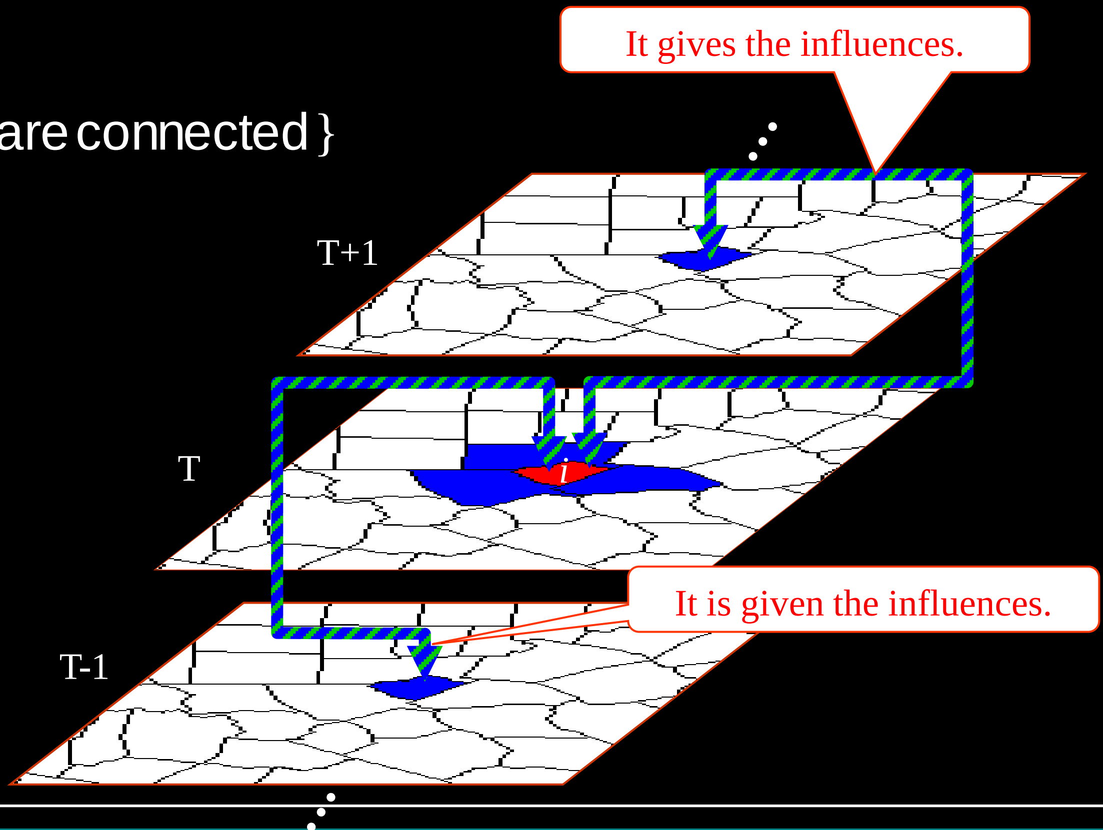


- The spatial regions are represented schematically on the horizontal axis.
- The time is represented on the vertical axis.

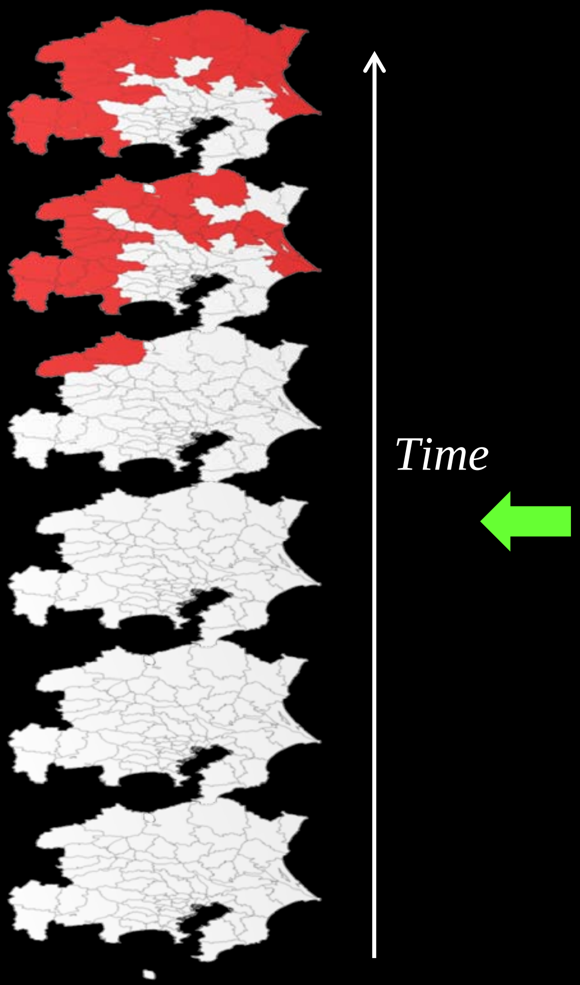
Echelon analysis for Spatial-temporal data

- 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.

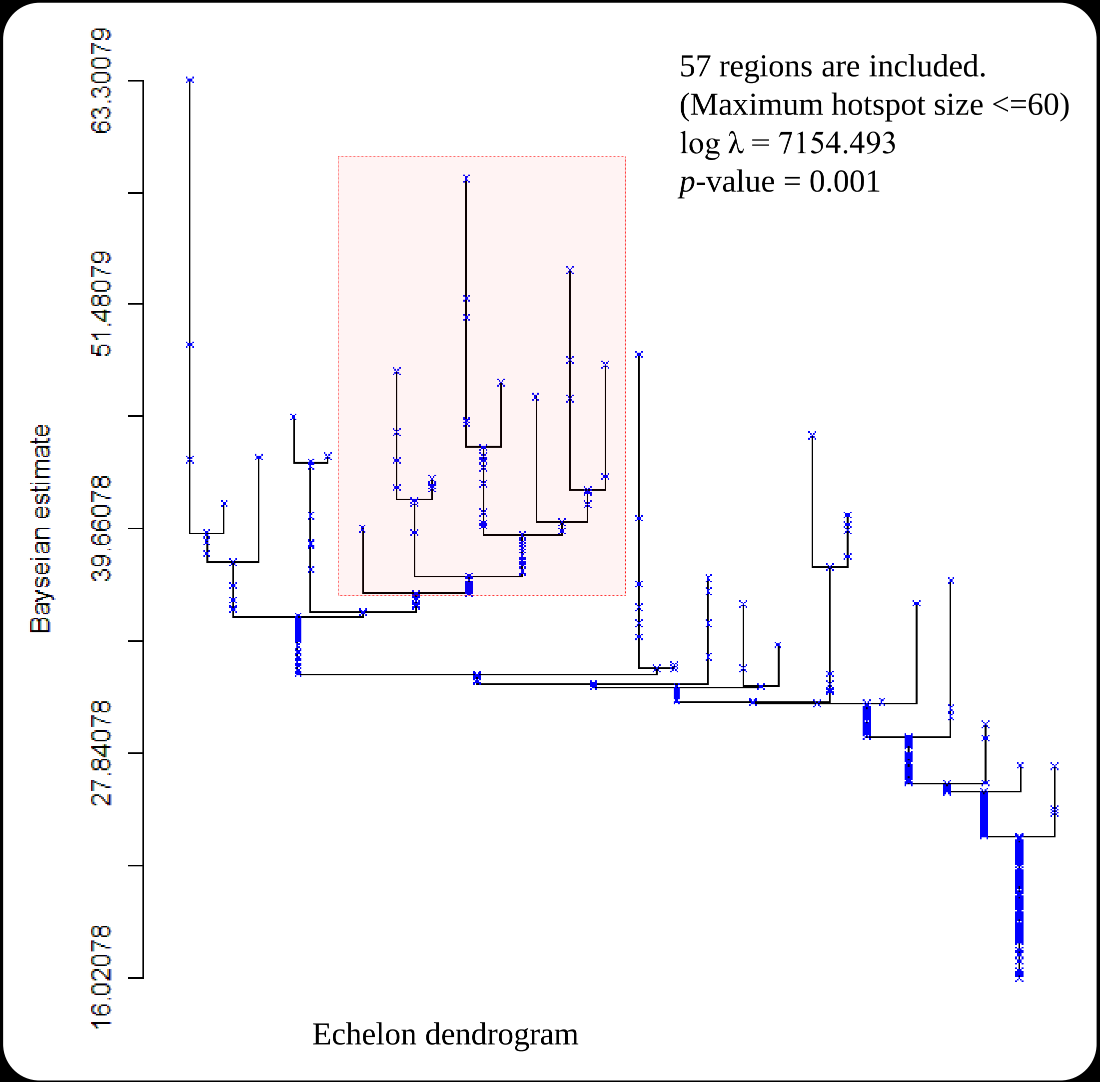
$$
\begin{aligned}
& NB(X(T,i)) \\
&= \{(T,k) \mid \text{regions } i \text{ and } k \text{ are connected}\} \\
&\cap X(T+1,i) \\
&\cap X(T-1,i)
\end{aligned}
$$



Application to the suicide data

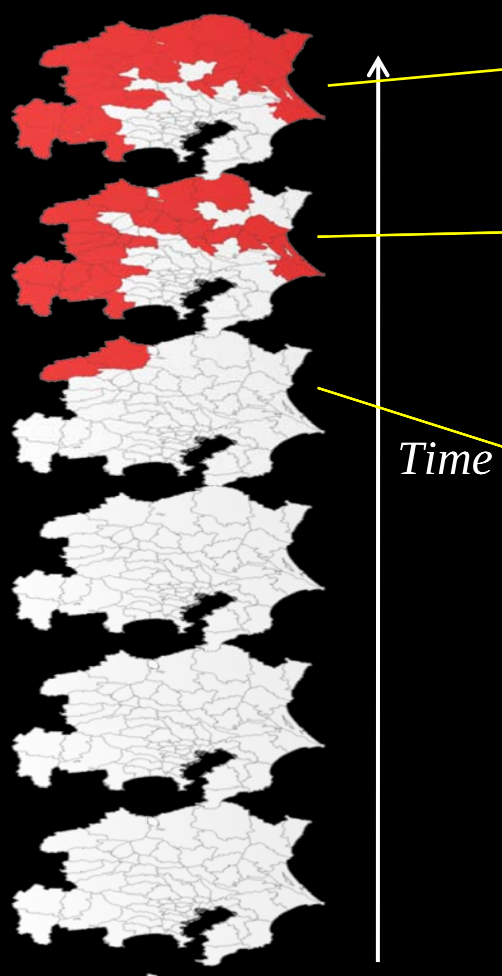


Spatial-temporal data

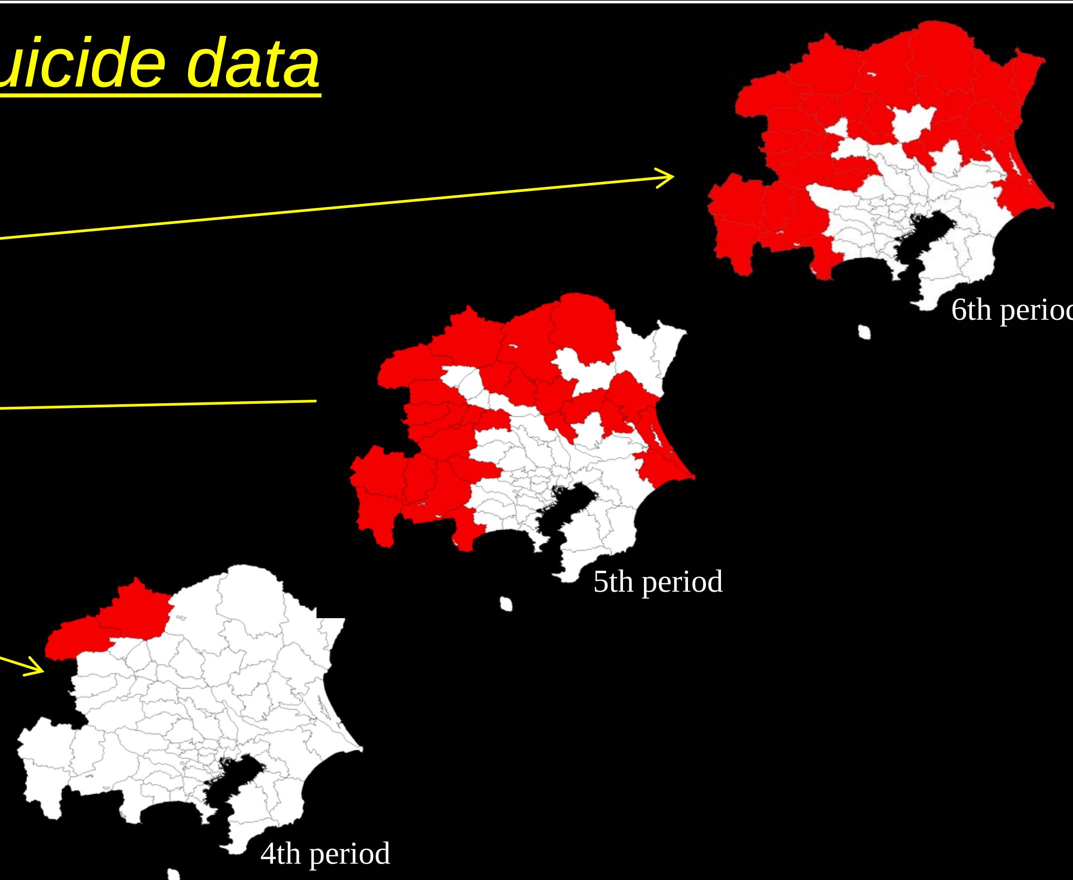


Echelon dendrogram

Application to the suicide data

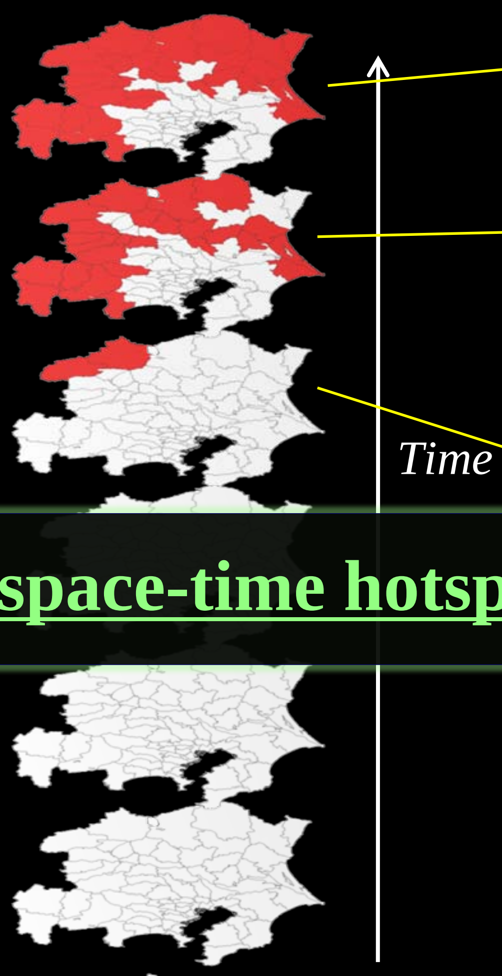


Spatial-temporal data

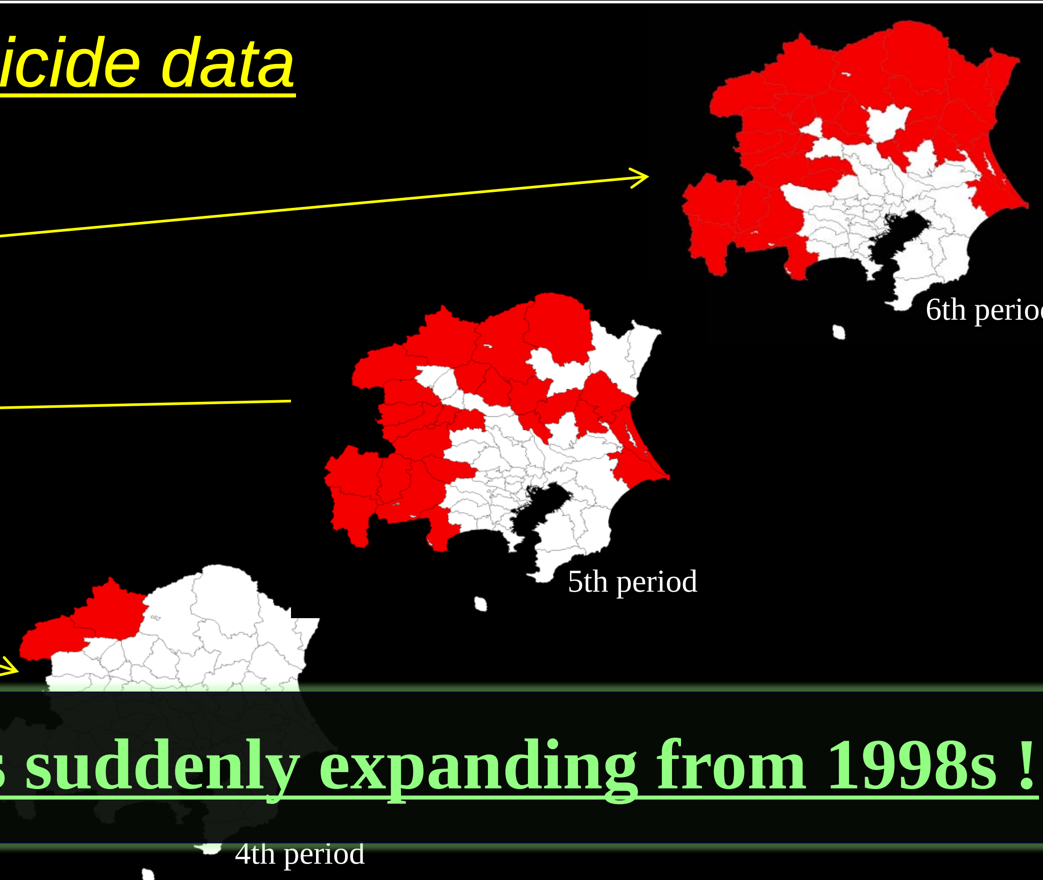


	# regions	# cases	Log	p-value
1st. (1973-1982)	0	13252	7154.493	< 0.001
2nd. (1983-1987)	0			
3rd. (1988-1992)	0			
4th. (1993-1997)	2			
5th. (1998-2002)	25			
6th. (2003-2007)	30			

Application to the suicide data



Spatial-temporal data



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

	# regions	# cases	Log	p-value
1st. (1973-1982)	0	13252	7154.493	< 0.001
2nd. (1983-1987)	0			
3rd. (1988-1992)	0			
4th. (1993-1997)	2			
5th. (1998-2002)	25			
6th. (2003-2007)	30			

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.

Conclusion

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

- Cabinet Office. (2008): White Book for Strategy to Prevent Suicide. Saiki Printing Co.
- Duczmal, L. and Assunção, R.A. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, 45, 269-286.
- Fujita, T., Tanihara, T. and Miura Y. (2003). Geographical Features of the Increasing Number of Suicide After 1998 in Japan. *Journal of Health and Welfare Statistics*, 50(10), 27-34.
- Kulldorff, M. (1997). A spatial scan statistics. *Communications in Statistics*, *Theory and Methods*, 26, 1481-1496.
- Kulldorff, M. (2006): Information Management Services Inc: SaTScan v7.0: Software for the spatial and space time scan statistics, http://www.satscan.org/.
- Myers, W.L., Patil, G.P. and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, 4, 131-152.
- Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11, 183-197.
- Tango, T. and Takahashi, K. (2005). A flexible spatial scan statistic for detecting clusters, International Journal of Health Geographics, 4,11.