数理解析研究所講究録
第1752巻 2011年 63-73

統計的手法を用いた時空間データの集積性について

岡山大学大学院法務研究科 石岡文生 (Fumio Ishioka)
School of Law, Okayama University

1. Introduction

近年、環境リスク解析や環境保全のため、空間データ解析の必要性が高まっている。中でも、ある郡における病気の発生率などのように、領域毎に得られるデータに対して、有意に高いまたは低い値を示す地域（ホットスポット）の検出は、各種の空間データの大きな課題である。ホットスポット領域の検出手法として、これまでに様々な手法が提案されてきた。空間的自己相関の観点からホットスポットを検出する手法（Anselin, 1995）や、全領域の中を一定の規則に基づいた小領域で走査（スキャン）していき、ホットスポットを検出する手法（Openshaw et al., 1987; Besag and Newel, 1991）などが提唱されている。また、疾病の地域集積性を検討するための手法として、Tangoの集積性の検定（Tango, 1995）も提唱されている。そうした中、ホットスポット検出のための優れたツールの一つに、空間スキャン統計量（Kulldorff,1997）がある。しかし、Kulldorffの提案した手法は、データが得られた領域の中心を円の中心とし、円状に領域をスキャンしてホットスポットを見つける手法であるため、円形状のホットスポットしか検出することができない。この問題を克服するため、我々はスキャンの方式として Echelon 解析（Myers et al.,1997; Kurihara, 2004）を利用する。得られたデータに対しEchelon解析を行い、それによって作られた位相的な階層構造に基づいてスキャンを行うことで、円状に限らない領域からなるホットスポットの検出が可能となる（栗原, 2002; Ishioka et al. 2007)。本研究では、病気の発生率のような地域空間データに対してEchelon解析と空間スキャン統計量に基づいたホットスポットを検出する方法について紹介し、さらに他のホットスポット検出法との結果の比較を行う。さらに、対象のデータを時間と空間の広がりの中で観測される時空間データに拡張し、時系列的なホットスポットの推移の表現についても検討する。

2. 空間スキャン統計量

空間スキャン統計量は、ある領域内の地点に起きた現象が偶然によるものか否かを検定し、有意に高い地域群（ホットスポット）を検出するための尤度比検定統計量である。今、対象とするすべての領域をG、その部分集合の領域をZとし、領域Zの内部では個人はある属性を確率p_1、領域Zの外では確率p_2で持つものとする。また、属性を持つ確率は互いに独立とする。このとき、帰無仮説を$H_0: p_1 = p_2$、対立仮説を$H_1: p_1 > p_2$とする。

ここでは、ポアソン分布に基づくモデルを考える。$n(G)$をすべての領域Gでの母集団の数、$n(Z)$を領域Z内での母集団の数、$c(G)$をすべての領域Gで属性を持つものの数、$c(Z)$を領域Z内で属

性を持つものの数としたとき、全領域Gで属性をもつ数が$c(G)$になる確率は以下の式で表される。

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

全ての領域内での地点xでの密度は、

$$\begin{cases} \dfrac{p_1 n(x)}{p_1 n(Z)+p_2 n(Z^c)} & if\ x \in Z \\ \dfrac{p_2 n(x)}{p_1 n(Z)+p_2 n(Z^c)} & if\ x \notin Z \end{cases}$$

そのとき、ポアソンモデルに対する尤度関数は以下のように与えられる。

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

尤度関数を最大にするために、領域Zを与えた下での最大尤度関数を計算する。ここで、最尤推定量は$\hat{p}_1=c(Z)/n(Z)$かつ$\hat{p}_2=c(Z^c)/n(Z^c)$とする。また、尤度比λは、ホットスポットを見つけるために全領域の部分集合の領域Zで最大のものとする。

$$\lambda=\frac{\underset{Z}{Max}\,L(Z)}{L_0}=\frac{(c(Z)/n(Z))^{c(Z)}(c(Z^c)/n(Z^c))^{c(Z^c)}}{(c(G)/n(G))^{c(G)}}$$

ただし、L_0は帰無仮説上での尤度関数の値である。

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

最も尤度の高いホットスポットを検出するためには、領域Gに含まれる全ての部分集合の領域をスキャンし、対数尤度比統計量$\log\lambda$が最大になる領域Zを求める必要がある。しかし、領域内でスキャンする領域Zの取り方は無数である。Kulldorff (1997) は、あらかじめ決められたいくつかの点を中心とし、ある大きさまでを円状にスキャンする方式 (Circular scan 法) を提唱するとともに、ホットスポットのためのソフトウェア SaTScan™ を開発している。また、提唱したスキャン統計量の分布を解析的に求めるのは難しいので、モンテカルロ法 (Dwass, 1957) により分布を求めるとともにp値を計算している。しかし、円状に領域をスキャンすることにより、円状のホットスポットの検出には優れているが、線状や他の形状をしたホットスポット検出には適しない。近年、この問題を解決するため、Upper level set scan 法 (Patil and Taillie, 2004)、Simulated annealing scan 法 (Duczmal and Assunção, 2004)、Flexible scan 法 (Tango and Takahashi, 2005)などの新たなスキャン法が提唱されている。

3. エシェロン解析

3.1 1次元空間データのエシェロン解析

地形図の断面図のような一次元空間データの場合、データは水平位置 x とデータの高度 $h(x)$を用いて$(x, h(x))$として与えられる。いま、データが k 個の区間 $I(i)=(i-1,i]$, i=1, 2, ..., k に分けられた lattice (interval) データを考える。表 1 は、A から Y と名前が付けられた区間とその区間での高度を示している。

表 3.1　1 次元 lattice データ

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
$h(i)$	1	2	3	4	3	4	5	4	3	2	3	4	5	6	5	6	7	6	5	4	3	2	1	2	1

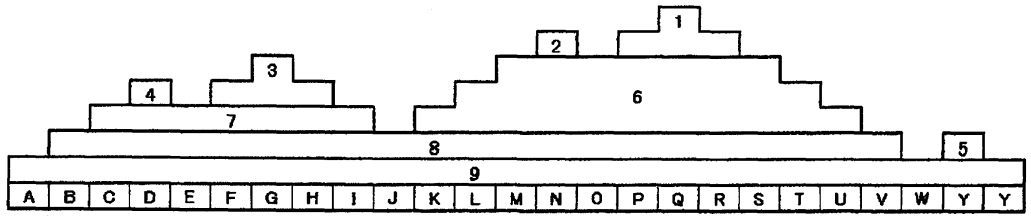

図 3.1　エシェロン解析における同じ位相領域への分割

図 3.1 は、表 3.1 の空間データの断面図を表している。このような断面図が与えられた場合、位相的に同じ領域（エシェロン）へ分けることができる。図で与えられている番号がエシェロン番号であり、1 から 5 までのピークと 6 から 9 までのファウンデーションから構成される。エシェロン番号の 1 から 5 はピークであり、エシェロン番号の 6 と 7 は 2 つ以上のピークのファウンデーションである。エシェロン番号 8 は 2 つ以上のファウンデーションのファウンデーションであり、エシェロン番号 9 は、ルートである。これらの関係はエシェロン番号を利用して 9(8 (7 (4 3) 6(2 1)) 5)と表すことができる。エシェロンデンドログラムは、エシェロン解析で使われるエシェロン地図や構造を階層的に表しており、空間データの構造を的確に表現することのできるグラフである。図 3.1 で示されるデータの構造は図 3.2 のようなエシェロンデンドログラムで与えられる。

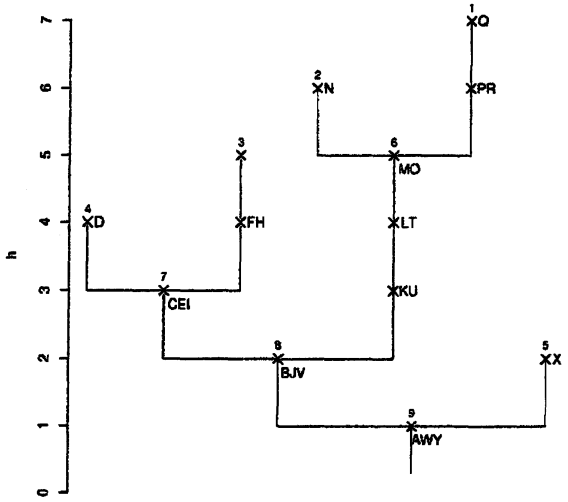

図 3.2　次元空間データのエシェロンデンドログラム

3.2 2次元空間データのエシェロン解析

リモートセンシングやメッシュデータなどの 2 次元で与えられる空間データは、$D_1 \times D_2$ 上の値 h_{ij} で与えられる。

$$l_2(i,j) = \{(x,y) \mid x_{i-1} \le x \le x_i, y_{j-1} \le y \le y_j\}, i = 1,2,\ldots,D_1, j = 1,2,\ldots,D_2$$

この時、セル $l_2(i,j)$ の隣接情報は次のように与えらえる。

$$NB(l_2(i,j)) = \{(a,b) \mid i-1 \le a \le i+1, j-1 \le b \le j+1\} \cap \{(a,b) \mid 1 \le a \le D_1, 1 \le b \le D_2\} - \{(i,j)\}$$

$$where\ A - B = A \cap B^c.$$

図 3.3 のような 5×5 で与えられる 2 次元空間データの場合、エシェロンデンドログラムは次の手順に従って作成される。

	A	B	C	D	E
1	2	24	8	15	3
2	10	1	14	22	5
3	4	13	19	23	25
4	20	21	12	11	17
5	16	6	9	18	7

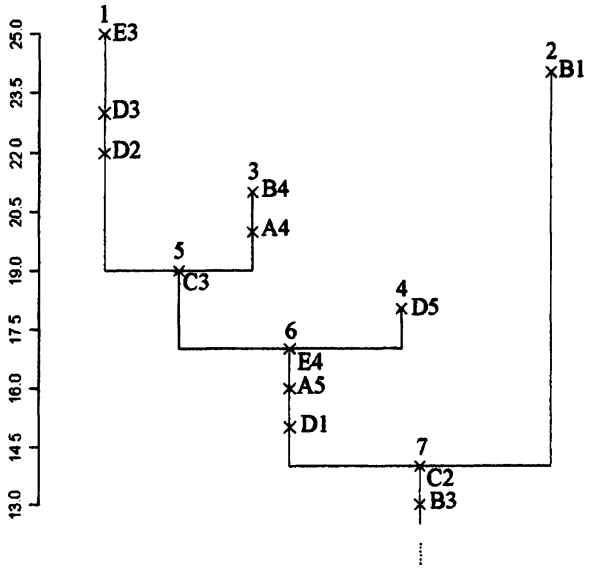

図 3.3 5×5 の空間データとそのエシェロンデンドログラム

Step 1) ピークの検出

ピークに属するデータ値は、同じピークに属するデータ以外の隣接するデータ値より大きい。図 3.3 の 5×5 の空間データにおいて、最大値は 25 である。従って、セル{E3}は第 1 ピークに属する。{E3}に隣接するセルの中で最大値となるのは{D3}の 23 で、そのセルは{E3, D3}に隣接するデータより大きいので、{D3}も第 1 ピークに属する。{E3, D3}に隣接するデータ値の最大は{D2}の 22 で、そのセルは{E3, D3, D2}に隣接するデータ値よりも大きいので{D2}も第 1 ピークに属する。{E3, D3, D2}に隣接するデータ値の最大は{C3}の 19 である。しかし、19 は{E3, D3, D2, C3}に隣接する{B4}の 21 より小さいので第 1 ピークに属さない。よって第 1 ピークはデータ値 25, 23, 22 の{E3, D3, D2}から構成され、エシェロン番号は 1 である。

第 1 ピークを除いたデータ値の最大は{B1}の 24 である。まず、{B1}は第 2 ピークに属する。{B1}に隣接するデータ値の最大は{C2}の 14 であるが、隣接する{D3}の 23 より小さいので第 2 ピークに属さない。よって第 2 ピークは（エシェロン番号 2）は{C2}

からのみ構成される。同様な手順により、第3ピーク（エシェロン番号3）は{B4, A4}、第4ピーク（エシェロン番号4）は{D5}から構成される。

Step 2) ファウンデーションの検出

4つのピークに属するセルを除いた最大値は{C3}の19である。{C3}は第1ピークと第2ピークのファウンデーションであり、エシェロン番号は5となる。エシェロン番号{1, 3, 5}に隣接するデータ値の最大は{E4}の17である。しかし、{E4}はそれに隣接する{D5}の18よりも小さいので{E4}はエシェロン番号5には属さない。以後、ファウンデーションを見つける際、エシェロン番号1とエシェロン番号3は使用されず、代表してエシェロン番号5を用いる。

同様な手順により、ファウンデーションを求めると、最終的にこの5×5の2次元空間データは図3.3のようなエシェロンデンドログラムによって与えられる。

3.3 地域空間データのエシェロン解析

病気の発生率のような地域空間データは、対象とする地域が市や郡などいくつかの区画 D_i, $i=1, 2, \ldots, k$ に分割され、データは $h(D_i)$で与えられる。例として、アメリカ合衆国ノースカロライナ州の乳幼児突然死症候群（Sudden Infant Death Syndrome; SIDS）データ（Cressie and Chan, 1989）を用いる。データは、ノースカロライナ州の100郡において1974年7月から1978年6月の期間に観測されたデータである。こうした郡別に得られた空間データの構造を可視化するツールとして主に図3.4のようなコロプレスマップなどの統計地図が多く利用されている。統計地図を利用することにより、色の濃淡と位置情報を基に、SIDSの高い地域や低い地域を把握できる。

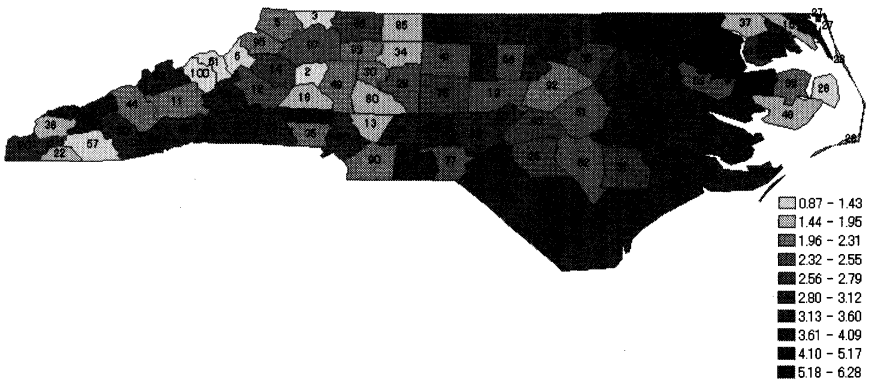

図3.4 ノースカロライナ州100郡のFreeman-Tukey変換後のSIDSデータのコロプレスマップ

しかし、この分析は郡別に与えられたデータに対して、単純にその値に応じて色の濃さを変えて白地図上に描いたに過ぎず、SIDSに関する構造に関する情報がない。この種の地域空間データのような場合も、領域間の近隣情報 $NB(D_i)$を与えることにより、位相的な構造を階層構造で表

す事ができる。ここでは、各郡における生誕数と SIDS 死亡数との分散を Freeman-Tukey 変換式 $Y_i = \sqrt{1000(SID_i)/BIR_i} + \sqrt{1000(SID_i+1)/BIR_i}$ (Cressie and Chan, 1989)によって調整した値 Y_i を基に Echelon デンドログラムを作成する。ここで、$SID_i : i = 1,2,...,100$, $BIR_i : i = 1,2,...,100$ は、それぞれ各郡における SIDS 死亡数と生誕数を意味している。各郡の隣接情報と Y_i から作成される Echelon デンドログラムを図 3.5 に示す。

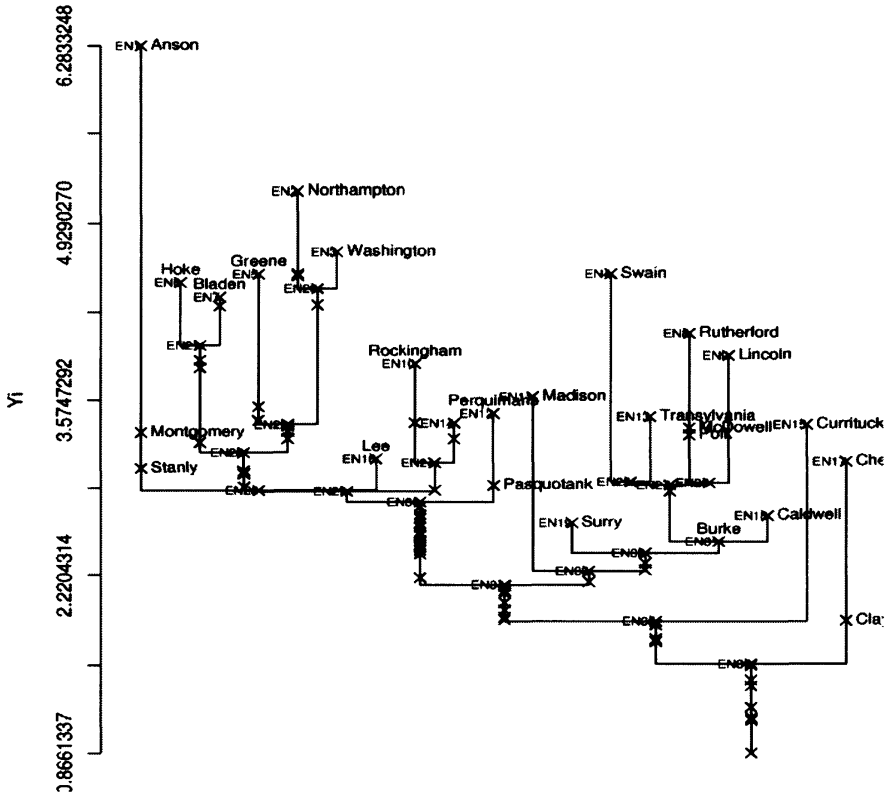

図 3.5　SIDS データの Echelon デンドログラム

4. SIDS データのホットスポット検出

4.1　先行研究のスキャン法によるホットスポット検出

Kulldorff (1997) は，データが得られた地点を中心に円状に領域をスキャンし、有意に尤度の高い比率を示す領域を見つける手法を提案した (Circular スキャン法, 図 4.1)。

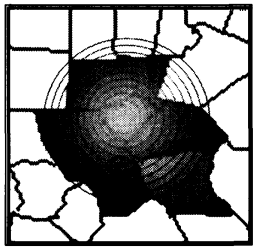

図 4.1 Circular スキャン

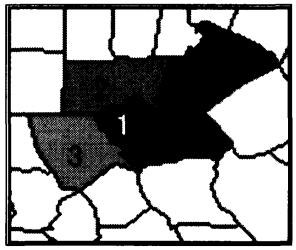

図 4.2 Flexible スキャン

しかし，円状に領域をスキャンするため、円形状のホットスポットの検出には優れているが、線状や他の形状をしたホットスポットの検出は不向きである事が指摘されている。この問題を解決するために、Tango, Takahashi（2005）らは Flexible スキャン法を提唱した。これは、あらかじめスキャンされる領域数を決めておき、その中で総当り的に領域のパターンをスキャンすることで、尤度の高い領域を検出するというものである（図 4.2)。この手法により、円形状によらないホットスポット領域の検出は可能となったが、計算時間の問題から大規模なホットスポット領域を検出するような場合には向かない。先ほどの SIDS データに対し、ホットスポットの最大領域数を 15 とした時の Circular scan 法、Flexible scan 法によるホットスポット検出結果をそれぞれ図 4.3.、図 4.4 に示す。

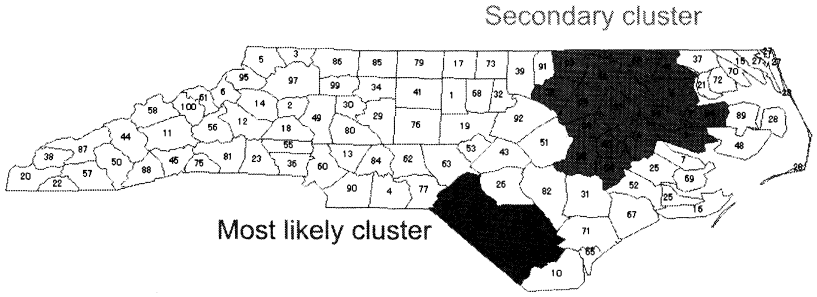

図 4.3　Circular scan 法に基づく SIDS データのホットスポット

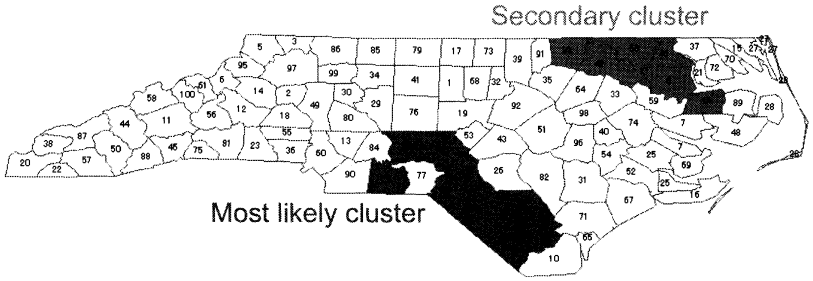

図 4.4　Flexible scan 法に基づく SIDS データのホットスポット

4.2　エシェロン解析に基づくホットスポットの検出

Echelon に基づくホットスポットの検出は、以下の手順に従って行う。

Step1) Echelon 解析によって、位相的な階層構造を明らかにする。

Step2) 求められた階層構造から、上位の Echelon を構成する領域を Z に加えながらスキャンする。

Step3) あらかじめ最大ホットスポット領域数 K を決めておき、K 以下で最も統計量が高くなった時の領域 Z をホットスポット候補とする。

Step4) ホットスポット候補 Z に対して、モンテカルロ検定により p 値を計算する。

空間スキャン統計量を利用することにより、尤度の高いホットスポットを見つけることができる。ここでは、K=15 として SIDS データに対するホットスポット検出を行った。図 3.5 で得られた Echelon デンドログラムに基づいてスキャンを行なった結果、1 番目のホットスポット (Most likely cluster) は、Echelon 番号 22(5 20(2 3))に含まれる Beaufort(7)、Bertie(8)、Edgecombe(33)、Greene(40)、Halifax(42)、Hertford(46)、Lenoir(54)、Northampton(66)、Pitt(74)、Warren(93)、Washington(94)、Wayne(96)、Wilson(98)の 13 領域となり、その時の統計量は 16.506、p 値は 0.001 となった。また、2 番目のホットスポット (Secondary cluster) として、Echelon 番号 21(6 7)に含まれる Bladen(9)、Columbus(24)、Hoke(47)、Pender(71)、Robeson(78)、Scotland(83)の 6 領域が得られ、その時の統計量は 15.303、p 値は 0.001 となった。これらのホットスポット領域を図 4.5 に示す。

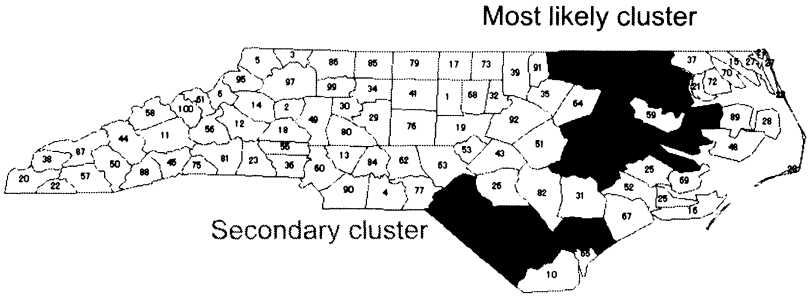

図 4.5　Echelon scan 法に基づく SIDS データのホットスポット

表 4.1 に先行研究の手法と、我々の Echelon による手法の SIDS データのホットスポット検出結果を示す。いずれの手法の結果も、北部と南部にホットスポットが存在することが示唆された。南部においては、Flexible scan による手法が統計量の高いホットスポットを検出したが、北部のホットスポットは、Echelon に基づく手法が最も統計量の高いホットスポットを検出した。

表 4.1　各スキャン法における SIDS データの北部と南部のホットスポット検出結果

北部	領域数	生誕数	SIDS 数	統計量	p 値
Echelon scan	13	36005	123	16.506	0.001
Circular scan	15	42006	131	12.585	0.001
Flexible scan	6	9763	49	15.968	0.001

南部	領域数	生誕数	SIDS 数	統計量	p 値
Echelon scan	6	17998	73	15.303	0.001
Circular scan	5	16770	69	14.930	0.001
Flexible scan	8	22246	92	20.649	0.001

5. 時空間ホットスポットの検出

5.1 時空間ホットスポット

これまで、ホットスポット検出のために、ある 1 時点における観測結果から得られた空間データのみを取り扱ってきた。しかし、空間データは時系列的に観測された場合が多く、そのためホットスポットの時系列な変化を解析することは大変重要になる。そのような時空間ホットスポットの推移模様の例を図 5.1 に示す。図 5.1 の左側の 3 つの図は、それぞれ横軸に空間、縦軸に時間をとることで、連続する時間の中におけるホットスポット空間の推移を表現しており、上から順に、時間推移とともに縮小するホットスポット、移動するホットスポット、分裂するホットスポットとなる。さらに、右側の 3×3 枚の図は、それぞれの時空間ホットスポットの軌道を、3 時点で取り出し、2 次元空間上に示している。

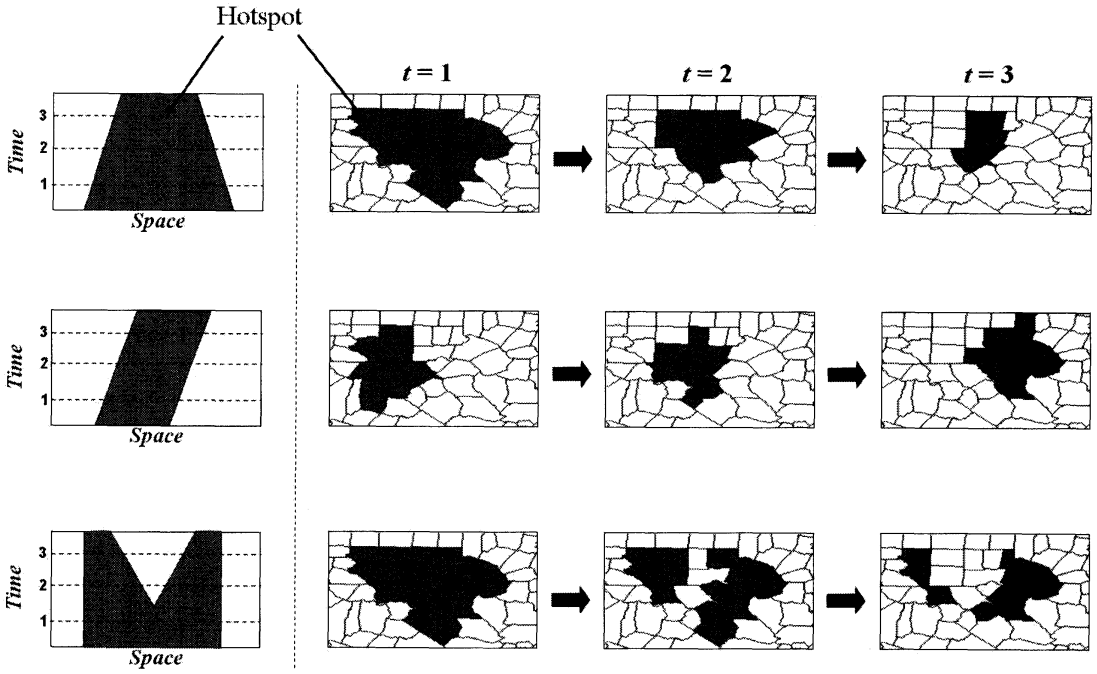

図 5.1　時空間ホットスポットの例

5.2 Echelon 解析に基づいた時空間ホットスポットの検出

地域型の時空間データは、時点 t, t=1, 2, ..., T における、ある区画 D_i, i=1, 2, ..., k として得られ、データは $h(D_{t,i})$で与えられる。この時、各領域間の隣接情報 $NB(D_{t,i})$を以下のように定義する。

$$NB(D_{t,i}) = \{D_{t,j} \mid \text{regions } i \text{ and } j \text{ are connected}\}$$
$$\cap D_{t+1,i}$$
$$\cap D_{t-1,i}$$

データ値 $h(D_{t,i})$と各領域の隣接情報 $NB(D_{t,i})$から、時空間データにおける Echelon デンドログラムの作成が可能になる。

適用例として、アメリカ合衆国ニューメキシコ州 32 郡における 1973 年、1982 年、1991 年の 3 年

間分の悪性脳腫瘍死亡データを用いる。32 郡×3 年間=96 領域を対象にして Echelon 解析を行い、Echelon scan 法に基づいてホットスポットを検出した結果、統計量は 10.02 で p 値は 0.001 となった。これらのホットスポットを図 5.2 に示す。時間の推移とともに、最初北西部に存在したホットスポットが一度分裂し、南東部において再び統合される様子が見てとれる。

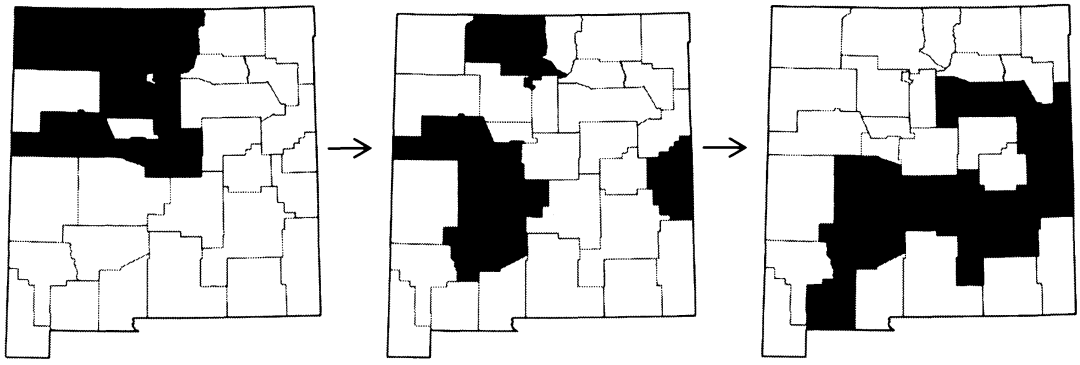

図 5.2 Echelon scan 法に基づく 3 年間分の悪性脳腫瘍死亡データの時空間ホットスポット

6. 最後に

本研究では、空間データに対して Echelon 解析に基づくホットスポット検出法について述べた。空間データのホットスポットを検出する際、Echelon 解析より得られた階層構造に基づき領域をスキャンする方式を適用した。この方式により、円形状に限らない任意の形状のホットスポット検出が可能となる。さらに、この手法は空間データのもつピークからスキャンするので、効率的なスキャンすることができる。そのため、従来の手法では困難であった大量データからなる空間データに対するホットスポット検出が可能になる。また、Echelon 解析の応用として、時空間データに適用し、それによって時空間ホットスポットの検出を可能にした。今後は、様々な大規模な空間データ、時空間データなどへの適用が期待される。

References

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographic Analysis*, 27, 93-115.

Besag, J. and Newell, J. (1991). The detection of clusters in rate diseases. *Journal of the Royal Statistical Society, Series A*, 154, 143-155.

Cressie, N. and Chan, N.H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, 84, 393-401.

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.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.

Ishioka, F., Kurihara, K., Suito, H., Horikawa, Y., and Ono, Y. (2007). Detection of Hotspots for 3-dimensional Spatial Data and Its Application to Environmental Pollution Data. *Journal of Environmental Science for Sustainable Society*, 1, 15-24.

Kulldorff, M. (1997). A spatial scan statistics. *Communications in Statistics, Theory and Methods*, 26, 1481-1496.

Kulldorff M, Athas WF, Feuer EJ, Miller BA and Key CR, (1998). Evaluating luster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health* 88, 1377-1380.

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, 164, 61-72.

Kulldorff, M. (2004). SaTScan™ User Guide for version 5.0.

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical Representation. *Classification, Clustering, and Data Mining Applications (Edited by D. Banks et al.)*, Springer, 251-258.

栗原考次. (2002). 階層的空間構造を利用したホットスポット検出. 計算機統計学, 15(2), 171-183.

栗原考次, 石岡文生. (2007). 空間データの階層構造による分類とその応用. 日本統計学会誌, 37(1), 113-132.

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.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A.W. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1, 335-358.

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. (1995). A class of tests for detecting 'general' and 'focuses' clustering of rate diseases. *Statistics in Medicine*, 14, 2323-2334.

Takahashi, K., Yokoyama, T. and Tango, T. (2005). FleXScan v1.1:Software for the Flexible Scan Statistic. National Institute of Public Health, Japan.

Tango, T. and Takahashi, K. (2005). A flexible spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, 4,11.