統計数理(2012)
第 60 巻 第 1 号 93–108

特集「時空間統計解析：新たなる分野横断的展開」
［原著論文］

Echelon解析に基づくスキャン法による ホットスポット検出について

石岡 文生[1] ・栗原 考次[2]

(受付 2011 年 7 月 1 日；改訂 9 月 20 日；採択 11 月 7 日)

要　　旨

本論文では，領域ごとに得られるデータ(空間データ)に対してEchelon 解析を適用し，それによって得られる階層構造に基づく尤度比の高いホットスポットの検出手法について述べた．次に，シミュレーションデータを用いて先行研究のホットスポット検出法との比較を行った．また，与えられた空間データにおいて，ホットスポットとなる可能性のある全ての領域の形状のパターンを検出するためのアルゴリズムを提案した．さらに，そのアルゴリズムから得られた全ての形状に対して，対数尤度比と relative risk を計算し，その関係性を検証することで，他の検出法の問題点と Echelon によるホットスポット検出法の妥当性について検討した．

キーワード： ホットスポット，空間データ，空間スキャン統計量，Echelon 解析．

1. はじめに

ある地方における感染症の発生状況の把握や，自然災害におけるハザードマップなどのように，“どの場所で問題が起きているのか”を知ることは，安全対策や環境保全のため最も基本的かつ重要な事であるといえる．近年，そういった問題を解析するため，市区町村別や州別などの領域ごとに得られるデータ(空間データ)を取り扱った研究が盛んに行われている．中でも，ある特定の領域において有意に高い値を示す集積地域(hotspot：ホットスポット)を検出することは，環境状況の把握や，将来の環境や健康への影響を早期に発見するためにも大変重要である．これまで各種の空間データに対して様々な観点からホットスポット検出のための研究が行われてきた．Moran (1948), Anselin (1995)は，空間的自己相関の観点からホットスポットの有無を検定した．また，Openshaw et al. (1987), Besag and Newell (1991)などは，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)することで，ホットスポットを検出する手法を提唱した．疾病の地域集積性を検討するための手法として Tango (1995, 2000)の手法も提唱されている．

そうした中，Kulldorff (1997)は，ホットスポットの存在の有無を検定すると同時にその位置も検出する空間スキャン統計量を提唱した．しかし，Kulldorff の方法は領域内の任意の地点から同心円状に一定の限界まで円を拡大していくことでホットスポットを探索するため，円形状のホットスポットしか検出することができない．それに対し，道路や河川に接するような非円形状のホットスポットを同定するため各種のスキャン法が提唱されてきている(Patil and

[1] 岡山大学大学院 法務研究科：〒700–8530 岡山県岡山市津島中 3–1–1
[2] 岡山大学大学院 環境学研究科：〒700–8530 岡山県岡山市津島中 3–1–1

Taillie, 2004; Duczmal and Assunção, 2004; Tango and Takahashi, 2005)．ところがこれらの先行研究による方法は，ホットスポットの形状が非現実的に大きくなりすぎたり，計算コストの問題から大容量のデータには適用が困難なのが現状である．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．

Echelon 解析は，空間的な位置を表面上のデータの高低に基づき分割し，空間データの位相的な構造を系統的かつ客観的に見つけるために開発された．Echelon 解析の応用として，Kurihara et al. (2000)は，メッシュ型の構造をもつ都心の人口データおよびリモートセンシングデータに対し，その空間的な構造の類似性について分析した．Ishioka et al. (2007)は，廃棄物処理場における地下への汚染水流出を想定したシミュレーションデータに適用し，そこから得られた空間的な構造を基に高汚染濃度地帯を同定した． Kurihara et al. (2006)は，多変量空間データに対し Echelon 解析を適用した．栗原・石岡(2007), Ishioka and Kurihara (2008)は，Echelon によって得られた階層構造を利用する新たな空間クラスタリングの手法を提案した．また，Tomita et al. (2008)は，格子状にデータを得られた遺伝子の連鎖不平衡ブロックの同定問題に対して Echelon 解析を適用し，従来法との比較検討を行っている．

本論文では，Echelon 解析を利用するホットスポット検出法の有効性について，他の検出手法と比較しながら検討を行う．第 2 章，第 3 章では，Echelon 解析ならびに空間スキャン統計量について説明する．第 4 章では，全てのホットスポットの候補を探索するための新たなスキャン法を提唱するとともに，先行研究のホットスポット検出法と Echelon 解析を利用したホットスポット検出のアルゴリズムについて述べる．第 5 章ではシミュレーションデータを用いて実際に解析を行いながら，既存のスキャン法により空間スキャン統計量を求める際の問題点と，Echelon に基づくスキャン法の妥当性について述べる．

2. Echelon 解析に基づく空間データの構造分析

Echelon 解析は，市区町村や州などに分けられた領域上の 1 変量値に対して，空間的な位置を表面上のデータの高低に基づき分割し，空間データの位相的な構造を系統的かつ客観的に見つけるために開発された解析法である．Echelon 解析で使われる Echelon デンドログラムは，それら空間データの構造を的確に表現したグラフである．ここで簡単な例としてリモートセンシングやメッシュデータの様な 2 次元で与えられる空間データに対し，Echelon デンドログラムを作成する方法を紹介する．いま，図 1 (a)の様にデータの高低が 5 × 5 のメッシュ上の位置

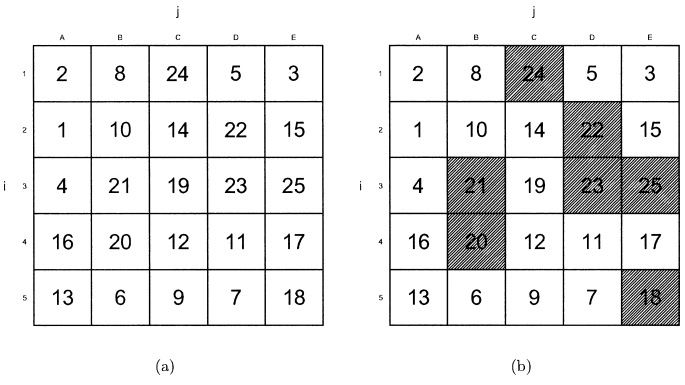


図 1. 5 × 5 のメッシュ上で与えられた空間データ (a) と，そのピーク (b)．

$(i,j), i=1,2,\ldots,5, j=1,2,\ldots,5$ に対して $h_{i,j}$ で与えられているとする．

ここで，ある領域 $x_{i,j}$ における連結情報は，通常上下左右の 4 近傍，または斜め方向も含めた 8 近傍が用いられる(間瀬・武田, 2001)．今回の例では，以下のような縦横の最大 4 方向を連結と定義した．

$$NB(x_{i,j}) = \{\{a,b\}|a=i,\ j-1\leq b\leq j+1\}\cup\{\{a,b\}|i-1\leq a\leq i+1,\ b=j\}$$
$$\cap\{(a,b)|1\leq a\leq 5,\ 1\leq b\leq 5\}-\{(i,j)\}$$

このとき，次のステップで Echelon 解析が進められる．

Step1. ピークの検出

1) 空間データ上で，連結している周辺領域の値よりも高い値からなる領域の集団をピークという．図 1 において，最大値は $h_{3,5}=25$ (位置ラベル；E3)である．従って 25 は第 1 ピークに含まれる．25 に連結する最大値は $h_{3,4}=23$ (D3)で，23 は {25, 23} に連結しているどの領域の値よりも大きいので 23 も第 1 ピークに含める．{25, 23} に連結する最大値は $h_{2,4}=22$ (D2)で，22 は {25, 23, 22} に連結している値よりも大きいので 22 も第 1 ピークに含める．{25, 23, 22} に連結する中で最大の値は $h_{3,3}=19$ (C3)である．しかし 19 は {25, 23, 22, 19} に連結する $h_{3,2}=21$ (B3)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 {25, 23, 22} から構成され，その階層集団を **En**(1)={25, 23, 22} と表すことができる．これらの値は同じピーク以外の連結するどの値より大きい．

2) 第 1 ピークを除いた最大値は $h_{1,3}=24$ (C1)である．まず，24 は第 2 ピークに含まれる．24 に連結する最大値は $h_{2,3}=14$ (C2)であるが，それに連結している $h_{2,4}=22$ (D2)よりも小さいので第 2 ピークに属さない．よって，第 2 ピークは 24 からのみ構成され，**En**(2)={24} となる．同様な手順により，第 3 ピーク **En**(3)={21, 20}，第 4 ピーク **En**(4)={18} が得られる(図 1 (b))．

Step2. ファウンデーションの検出

1) ピークを形成する集団に属さず，2 つ以上の階層集団の根を連結するための土台となる下位階層集団をファウンデーションという．4 つのピークに含まれる領域を除いた最大値は $h_{3,3}=19$ (C3)である．19 は **En**(1)と **En**(3)に属する領域と連結しているため，これらのファウンデーションとなる．{**En**(1), **En**(3), 19} に隣接する最大値は $h_{4,5}=17$ (E4)であるが，17 は隣接する **En**(4)の 18 より小さいので 17 はこのファウンデーションに属さない．従って，**En**(5)={19} となる．**En**(5)は **En**(1)と **En**(3)のペアレントであり，この関係は **En**(5(1 3))と表すことができる．以後，ファウンデーションを見つける際，**En**(1)と **En**(3)は使用されず，代表して **En**(5)を用いる．

2) **En**(1)から **En**(5)に含まれる領域を除いた最大値は $h_{4,5}=17$(E4)である．17 は **En**(5)と **En**(4)に連結することから，これらのファウンデーションとなり，**En**(6)={17} と表される．**En**(6)は **En**(5)と **En**(4)のペアレントとなり，**En**(6(5(1 3)4))である．以後，**En**(1)，**En**(3)，**En**(4)，**En**(5)は代表して **En**(6)を用いる．以後，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 のメッシュデータの構造は図 2 のような階層構造(Echelon デンドログラム)として表すことができる．また，これらの関係は **En**(7(6(5(1 3)4)2))と表すことができる．

3. 空間スキャン統計量

空間スキャン統計量(Kulldorff, 1997)は，全領域内でホットスポットの候補となる領域群を評価

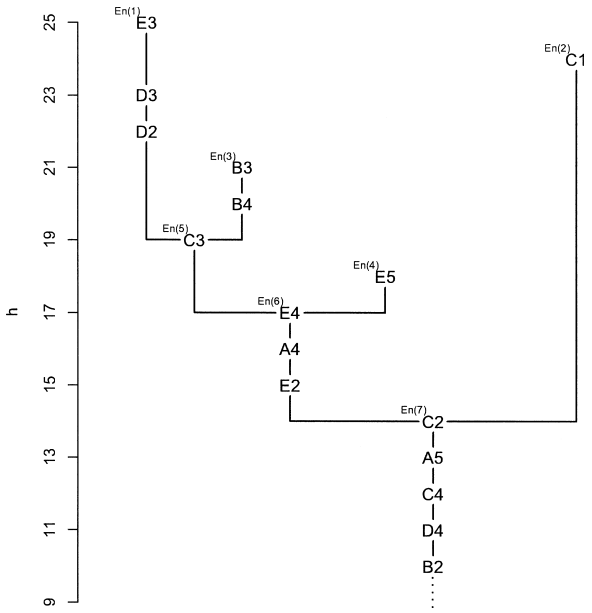


図 2. 5×5 の空間データの Echelon デンドログラム.

する指標である．いま，解析を行う対象の全ての領域 G が市区町村，州などいくつかの領域に分割されているものとする．それら各領域の母集団の数を n，属性を持つものの数を c で表すと，全領域 G での母集団の数，属性を持つものの数はそれぞれ $n(G)$，$c(G)$ で表され，G 内のある連結した領域の群 Z 内ではそれぞれ $n(Z)$, $c(Z)$ と表すことができる．このとき，Z における属性確率 p_z は $c(Z)/n(Z)$，Z の外部 Z^c における属性確率 p_{z^c} は $c(Z^c)/n(Z^c) = (c(G) - c(Z))/(n(G) - n(Z))$ と表すことができる．このとき，Z がホットスポットとなるか否かを検定する仮説は以下の通りである．

$$\text{帰無仮説} \quad H_0 : p_z = p_{z^c} \qquad \text{v.s.} \qquad \text{対立仮説} \quad H_1 : p_z > p_{z^c}$$

このとき，ひとつひとつの Z に対して検定を繰り返すと検定の多重性の問題が発生してしまう（丹後 他, 2007）．そこで Kulldorff は次のような尤度比に基づく統計量を考案した．

ある癌による死亡数など，属性をもつものの数が Poisson 分布に従う場合を想定するとき，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．

$$\frac{\exp[-p_z n(Z) - p_{z^c} n(Z^c)][p_z n(Z) + p_{z^c} n(Z^c)]^{c(G)}}{c(G)!} \tag{3.1}$$

全ての領域内での地点 x での密度は，

$$\begin{cases} \dfrac{p_z n(x)}{p_z n(Z) + p_{z^c} n(Z^c)} & \text{if } x \in Z \\ \dfrac{p_{z^c} n(x)}{p_z n(Z) + p_{z^c} n(Z^c)} & \text{if } x \in Z^c \end{cases}$$

そのとき，Poisson model に対する尤度関数は次のように与えられる．

$$L(Z,p_z,p_{z^c}) = \frac{\exp[-p_z n(Z) - p_{z^c} n(Z^c)][p_z n(Z) + p_{z^c} n(Z^c)]^{c(G)}}{c(G)!} \times \prod_{x_i \in Z}^{c(Z)} \frac{p_z n(x_i)}{p_z n(Z) + p_{z^c} n(Z^c)} \prod_{x_i \in Z^c}^{c(Z^c)} \frac{p_{z^c} n(x_i)}{p_z n(Z) + p_{z^c} n(Z^c)} = \frac{\exp[-p_z n(Z) - p_{z^c} n(Z^c)]}{c(G)!} {p_z}^{c(Z)} {p_{z^c}}^{c(Z^c)} \prod_{x_i} n(x_i) \tag{3.2}$$

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．最尤推定量 $\hat{p_z} = c(Z)/n(Z)$，$\hat{p_{z^c}} = c(Z^c)/n(Z^c)$ を式(3.2)に代入すると次式が得られる．

$$L(Z) = \frac{\exp[-c(G)]}{c(G)!} \left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{c(Z^c)}{n(Z^c)}\right)^{c(Z^c)} \prod_{x_i} n(x_i) \tag{3.3}$$

尤度比 λ は，ホットスポットを見つけるために全領域内の連結した部分集合の領域群 Z で最大のものとする．

$$\lambda(Z) = \max_z L(Z)/L_0 \tag{3.4}$$

ただし，L_0 は帰無仮説上 $p_z = p_{z^c} = p$ での尤度関数の値である．

$$L_0 \stackrel{\text{def}}{=} \sup_p \frac{\exp[-pn(G)]}{c(G)!} p^{c(G)} \prod_{x_i} n(x_i) = \frac{\exp[-c(G)]}{c(G)!} \left(\frac{c(G)}{n(G)}\right)^{c(G)} \prod_{x_i} n(x_i) \tag{3.5}$$

したがって，尤度比検定統計量 $\lambda(Z)$ は

$$\lambda(Z) = \begin{cases} \dfrac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{c(Z^c)}{n(Z^c)}\right)^{c(Z^c)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}} & \text{if } \dfrac{c(Z)}{n(Z)} > \dfrac{c(Z^c)}{n(Z^c)} \\ 1 \quad \text{otherwise} \end{cases} \tag{3.6}$$

と表される．このとき，尤度比 $\lambda(Z)$ を最大にするような領域群 Z をホットスポット候補と考える．

4. ホットスポット検出のためのスキャン手法

4.1 All possible scan 法

与えられた空間データに対して，真に尤度比を最大にする領域群 Z を検出するには，互いに連続した領域群全てをスキャンする必要があるが，通常その数は膨大になりすぎて現実的に不可能である．しかし，全体の領域数が極端に少ない場合は全ての Z をスキャンし，その内容を検証する必要があるだろう．本節では全ての Z を求めるための次のようなアルゴリズム(All possible scan 法)を提案する．

いま，全領域が M 個の領域に分けられた空間データを考える．続いて，全領域内で m 個の連結した領域から形成される Z の集合体を $\boldsymbol{Z}_m\,(m=1,2,\ldots,M)$ と表す．また，各 $\boldsymbol{Z}_m$ に含まれる Z の総数は，それぞれ K_m 個あるとする．このとき，1 個の領域からなる Z は必ず M 個存在するので，$K_1 = M$ と表すことができる．次に，ある領域 $i_k \in \boldsymbol{Z}_1\,(k=1,2,\ldots,K_1)$ に対し，それに連結している領域 $j \in NB(i_k)$ を求める．i と j を併合させ，ホットスポット候補 $Z=\{i,j\}$ とし，$\boldsymbol{Z}_2$ に格納する．このとき $\boldsymbol{Z}_2$ に含まれる Z の全体集合は，$\{(i_k,j)|1 \le k \le K_1, j \in NB(i_k)\}$

として得られる．最後に，$\boldsymbol{Z}_2$ 内で重複する形状のものは一つを除いて全て削除する．これにより，連結した 2 個の領域からなる全ての Z を求めることができる．続いて，3 個の領域から形成される Z を求めるには，ある 2 個の連結した領域 $i_k \in \boldsymbol{Z}_2\,(k=1,2,\ldots,K_2)$ に連結する領域 $j \in NB(i_k)$ を求め，先ほどと同様の手順により $\boldsymbol{Z}_3$ を求める．このように，m 個の領域からなる形状の集団 $\boldsymbol{Z}_m$ を，$\boldsymbol{Z}_{m-1}$ を利用して探索していく．それを $m=M$ となるまで行うことで，重複する形状を除いた全ての Z を求めることが可能となる．得られた $\boldsymbol{Z_m}\,(m=1,2,\ldots,M)$ において，$\max_{Z\in\boldsymbol{Z_m}}\lambda(Z)$ となる Z をホットスポットと同定する．得られたホットスポットの有意性の評価については，スキャン統計量の分布を解析的に求めるのは困難であるので，モンテカルロ法により分布を推定し p 値を計算する方法(Dwass, 1957)が広く使われている．それに伴い，本論文における各種のスキャン手法で同定されるホットスポットの有意性の評価についてもこれに従った．

4.2 Circular scan 法

All possible scan 法は，必ず対数尤度比が最大となるホットスポットを同定する事ができるため，ある意味理想的ではあるが，実際のデータへ適用するのは困難な場合が多い．そこで Kulldorff (1997)は，ホットスポット候補 Z の決め方として同心円状にスキャンしていく Circular scan 法を提唱した．この方法は，ある領域 i の代表点 1 点(市区町村役場の所在地や人口重心など)からその周りに半径 r の同心円を描いていく．その際，領域 j の代表点がその同心円に含まれると i と j を併合させ，このときホットスポット候補 $Z=\{i,j\}$ とする．半径 r は 0 から Z の値がある臨界値(最大距離，人口，領域数など)に達するまで拡大させる．スキャンされた Z の全体集合 $\boldsymbol{Z}$ において，$\max_{Z\in\boldsymbol{Z}}\lambda(Z)$ となる Z をホットスポットと同定する．この方法は，円状に領域をスキャンすることにより，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検出には適しないことが指摘されている．そこで近年，非円形状の $\boldsymbol{Z}$ を生成するため Upper level set scan (Patil and Taillie, 2004), Simulated annealing scan (Duczmal and Assunção, 2004), Flexible scan 法(Tango and Takahashi, 2005)などの新たなスキャン法が提案されている．

4.3 Flexible scan 法

非円形状のホットスポットを検出するためのスキャン法として，ここでは Tango and Takahashi (2005)の Flexible scan 法について触れる．この手法は，まずある領域 i を中心として，そこから近い順に K 個の領域からなる集合を求める．その集合内で i 自身を含み，互いに連結している部分集合を Z としてスキャンする．Z の全体集合 $\boldsymbol{Z}$ に対し，$\max_{Z\in\boldsymbol{Z}}\lambda(Z)$ となる Z をホットスポットと同定する．Flexible scan 法を利用するためのソフトウェアとして，FlexScan (Takahashi et al., 2010)が開発されている．

4.4 Echelon scan 法

我々は Echelon デンドログラムによって得られる空間データの構造に基づき領域をスキャンしていく方法(Echelon scan 法)を提案している．そのアルゴリズムは次の通りである．いま，N 個の階層から形成される Echelon デンドログラムにおいて，各階層の集合 $\mathbf{En}(k)\,(k=1,2,\ldots,N)$ は n_k 個 の領域から構成されているものとする．このとき各階層内における領域を，上位から $e(k,1)\geq e(k,2)\geq\cdots\geq e(k,n_k)$ とする．まず，$k=1$ (第 1 ピーク)の最上位の領域 $e(1,1)$ を Z としてスキャンする．続いて，$e(1,2)$ を $e(1,1)$ に併合させ，ホットスポット候補 $Z=\{e(1,1),e(1,2)\}$ としてスキャンする．以下同様に，Echelon を構成する上位の領域から順に，Echelon を構成する領域を Z に加えながらスキャンする．これをあらかじめ定めておいた Z の値がある臨界値(最大距離，人口，領域数など)に達するまでスキャンするものとする．このとき Echelon scan

表 1. 各スキャン法の特徴.

	必要な空間情報	ホットスポットの形状	尤度の高いホットスポット	計算コスト
All possible scan 法	連結情報	任意	◎	×
Circular scan 法	距離情報	円形	×	○
Flexible scan 法	連結情報かつ距離情報	任意	◎ ($K \leq 20^{*}$)	$\times$ **
Echelon scan 法	連結情報	任意	○	◎

*Flex scan ソフトウェアの推奨するホットスポット領域制限値.
**Tango(2008) の制約付き尤度比統計量を用いれば計算コストを抑えられる.

法では Z の全体集合として $\boldsymbol{Z}=\{e(k,l)|1\leq k\leq N, 1\leq l\leq n_k\}$ を得る．なお，ファウンデーションとなっている階層をスキャンする場合には，その上位階層に含まれる領域も全て含めてスキャンする．こうして得られた $\boldsymbol{Z}$ において，$\max_{Z\in\boldsymbol{Z}}\lambda(Z)$ となる Z をホットスポットと同定する．なお，この方法でスキャンされる Z は，連結情報に基づいて作成される Echelon デンドログラムを利用して求めているため，必ず互いに連結する領域群から成り立っている．

4.5 各手法の特徴

各種のスキャン法の特徴についてまとめたものを表 1 に示す．All possible scan 法は必ず尤度比が最大となるホットスポットを検出することができるが，大量データに対してはスキャンされる Z が多くなりすぎるため，適用は難しい．また，最大尤度比を求めるため，複雑な形状の大きなホットスポットを同定してしまう傾向がある．Circular scan 法は簡便な反面，スキャンの方式上，形状が円でないホットスポットの同定には検出力が低いことが報告されており (Tango and Takahashi, 2005)，たとえ有意なホットスポットを得られた場合でも，それが真のホットスポットを同定できているのかどうかは疑問が残る．それに対し，Flexible scan 法は非円形上のホットスポットを検出でき，かつ尤度比の高いホットスポットの同定が出来るよう工夫されている．しかしある種の総当たり的な要素を含んでいるため，K の値が大きくなると，非現実的な形状をした大きなホットスポットを検出してしまったり，また，大容量のデータへの適用には計算コストの面で問題がある．この問題に対し，Tango (2008) は，同定するホットスポット領域が広範囲になり過ぎない制約付き尤度比統計量を用いた Flexible scan 法を提唱している．一方，Echelon scan 法は非円形状のホットスポットが同定でき，かつデータの本来もつ階層構造のピークから優先的にスキャンしていくため，計算コストが抑えられ大量データにも適用が可能である．

5. シミュレーションデータを用いた性質の評価

5.1 データ適用例

ここでは，スキャン法の違いによるホットスポット同定の様子を検証する．本論文では，All possible scan 法でも解析できるよう，6×4 程度の領域の少ないメッシュデータを用いた．ここで，各領域の中心間の距離は互いに等しく，各領域は縦横の最大 4 方向に隣接しているものとする．いま，各領域は等しい母集団となるように $n(G)=24000$ と設定し，領域群 {C1, B2, C2} と {A6, B6, C6, D6} の 2 つの群だけ値が 3 倍高くなるような条件のもとで 1 組の Poisson 乱数を発生させた (図 3)．

まず，このデータに All possible scan 法によりホットスポットを検出する．この 6×4 のデータに All possible scan 法を適用するイメージを図 4 に示す．この図は，ある $Z\in\boldsymbol{Z}_1$ の領域の連結情報を基に $Z\in\boldsymbol{Z}_2$ を探索していき，さらには得られた $\boldsymbol{Z}_2$ 内で重複する形状の物は一つを除いて全て削除する様子を示している．例えば A1 は B1 と A2 に連結していることから，

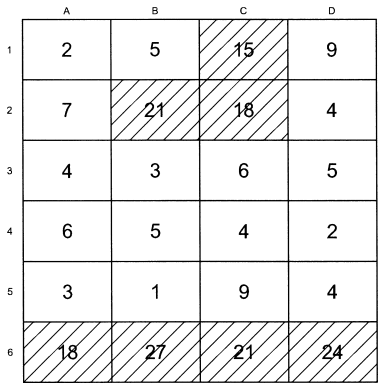


図 3. 6×4 のメッシュ上で与えられた空間データ.

{A1}$\in \boldsymbol{Z}_1$ を基にして {{A1, B1}, {A1, A2}} $\in \boldsymbol{Z}_2$ が生成されている．また，{{A1, B1}, {B1, A1}} $\in \boldsymbol{Z}_2$ は互いに同じ形状であるので，{B1, A1} は削除される．この結果，全ての連結する領域群 $Z \in \boldsymbol{Z}_m\,(m = 1, 2, \ldots, 24)$ の総数は $\sum_m^{24} K_m = 1168587$ 個存在した．Kulldorff (1997) は，領域群 Z に含まれる母集団の数が全母集団の半分になるまでスキャンすることを推奨していることから，$\boldsymbol{Z}_m\,(m = 1, 2, \ldots, 24)$ の内，その条件にあう 198806 個の Z をスキャン対象とした．そしてそこから尤度比が最大となる Z^* を検出すると，11 個からなる領域群 Z^* ={C1, D1, B2, C2, C3, C4, C5, A6, B6, C6, D6} となり (図 5 (a))，その対数尤度比は $\log\lambda(Z^*) = 45.55$，モンテカルロ推定に基づく p 値は 0.001 となった．さらに，帰無仮説の下での相対的な比率 (相対リスク比：relative risk) は 2.18 であった．

続いて，Kulldorff (1997) の提唱した Circular scan 法で解析を行った．Z を全母集団の半分になるまでスキャンした結果，尤度比を最大にする Z^* は，4 個の連結した領域群 Z^* ={C5, B6, C6, D6} となり (図 5 (b))，そのときの対数尤度比 $\log\lambda(Z^*) = 24.90$，relative risk は 2.18，モンテカルロ推定に基づく p 値は 0.001 となった．また，このときスキャンされた Z の総数は 288 個となった．

次に，Flexible scan 法について，スキャンする領域の制限を $K = 15, K = 20$ の場合でホットスポットの検出を行った．検出には FlexScan ソフトウェア (v3.1) を使用した．$K = 15$ のとき，領域群 $Z^*_{(15)}$ ={B6, D6, C6, A6} が最大対数尤度比となり (図 5 (c))，あらかじめ想定していたホットスポットが正しく同定される結果となった ($\log\lambda(Z^*_{(15)}) = 35.11$，relative risk=2.18，$p = 0.001$)．ところが $K = 20$ のときは 9 個からなる領域群 $Z^*_{(20)}$ ={B2, C2, C3, C4, C5, A6, B6, C6, D6} がホットスポットとして同定され (図 5 (d))，そのとき $\log\lambda(Z^*_{(20)}) = 38.01$，relative risk=1.77，$p = 0.001$ であった．K を大きくすることで，対数尤度比こそ高くなったものの，本来ホットスポットと同定されては不自然な領域までもが取り込まれたため，その relative risk は低くなったと考えられる．なお，Tango の制限付き尤度比統計量による Flexible scan 法を用いると，$K = 20$ の制限の場合であっても $K = 15$ のときと同様の結果を得ることが出来る．

最後に，Echelon scan 法によりホットスポットの検出を行う．この 24 個の各領域の連結情報と relative risk を基に作成された Echelon デンドログラムを図 6 に示す．大きなピーク集団が 2 つあり，それぞれ $\mathbf{En}(5\ (1\ 2))$ ={B6, D6, C6, A6, C5}，$\mathbf{En}(3)$ ={B2, C2, C1, D1, A2, C3, B1, D3} となっている．他と同様，Z を全母集団の半分になるまでスキャンした結果を表 2 に示す．$\mathbf{En}(5\ (1\ 2))$ における Z^* ={B6, D6, C6, A6} までスキャンしたとき対数尤度比が最も高

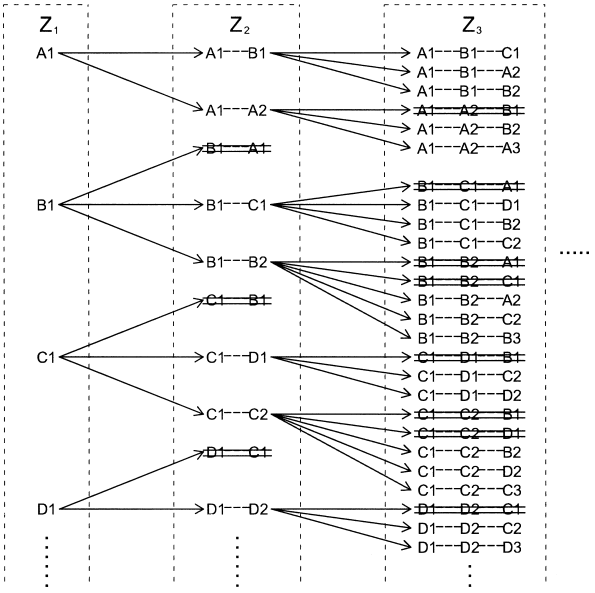
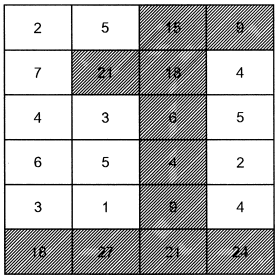
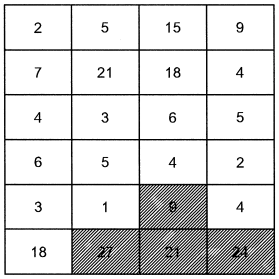


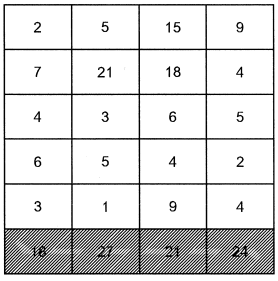
図 4. 6×4 の空間データに対する All possible scan の様子.



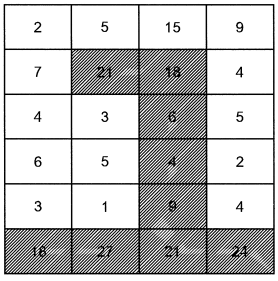
(a) All possible scan 法



(b) Circular scan 法



(c) Flexible scan 法 ($K = 15$), Echelon scan 法



(d) Flexible scan 法 ($K = 20$)

図 5. 各スキャン法によるホットスポットの同定の結果.

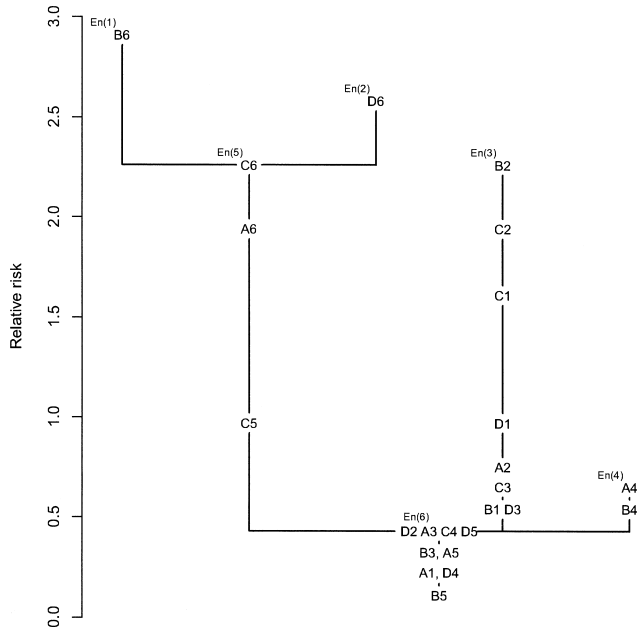


図 6. 6×4 の空間データの Echelon デンドログラム.

表 2. 6×4 の空間データへの Echelon scan 法の適用結果.

Z	属性値	期待値	Relative risk	$\log\lambda(Z)$	p
B6	27	9.29	2.91	11.85	0.001
D6	24	9.29	2.58	8.58	0.005
B6, D6, C6	72	27.88	2.58	29.61	0.001
B6, D6, C6, A6	90	37.17	2.42	35.11	0.001
B6, D6, C6, A6, C5	99	46.46	2.13	31.09	0.001
B2	21	9.29	2.26	5.74	0.042
B2, C2	39	18.58	2.10	9.55	0.001
B2, C2, C1	54	27.88	1.94	11.42	0.001
B2, C2, C1, D1	63	37.17	1.70	9.30	0.001
B2, C2, C1, D1, A2	70	46.46	1.51	6.80	0.026
B2, C2, C1, D1, A2, C3	76	55.75	1.36	4.58	0.112
B2, C2, C1, D1, A2, C3, B1, D3	86	74.33	1.16	1.34	0.928
A4	6	9.29	0.65	0	1
A4, B4	11	18.58	0.59	0	1

い値 $\log\lambda(Z^*) = 35.11$ となった(図 5 (c)). また, そのとき relative risk は 2.42, モンテカルロ推定に基づく p 値は 0.001 となった. また, **En**(3) の集団へのスキャンでは $Z^* = \{\text{B2, C2, C1}\}$ のとき対数尤度比が最も高い値 $\log\lambda(Z^*) = 11.42$ となり, こちらも有意なホットスポットが正しく同定できている.

5.2 考察

スキャン法の違いによるホットスポット検出結果を比較した結果を表 3 に示す．Echelon scan 法は，総スキャン数がわずか 14 個だったにもかかわらず，対数尤度比，relative risk ともに Circular scan 法よりも高い値を得た．これは，Echelon デンドログラムを利用することにより，relative risk のピークを形成する領域から優先的にスキャンするため，高尤度比となりやすい Z を効率よく探索できたことによるものである．また，互いに連結している領域を取り込みながらスキャンしていくので，今回の例のような非円形状のホットスポットの同定も可能となっている．一方，Circular scan 法は円形状に領域をスキャンするため，あらかじめ想定されていた線形状のホットスポットは同定できなかった．Kulldorff et al. (2006) は，この問題を解決するため，楕円形状にスキャンする Elliptic scan 法を提案しているが，大きな改善には至っていない．

また，All possible scan 法，Flexible scan 法 ($K = 20$) では，尤度比こそ高い値となったが，relative risk は Echelon scan 法に比べてかなり低い値となった．これは，最大尤度比をとる様な Z を求めるとき，今回の例の様に 2 つの別々のホットスポットが存在しているにもかかわらず，それらを 1 つのホットスポットとして同定してしまった事により，ホットスポットと同定されては不自然な値の小さな領域までもが Z に取り込まれたためと考えられる．一方，Echelon scan 法では，値の小さい領域は，階層構造的に下位の方に位置されるので，これらがスキャンされる優先度はかなり低くなる．そのため今回の様に 2 つの別々のホットスポットを正しく同定できたと考えられる．図 7 は，All possible scan 法によってスキャンされた 198806 個の Z に対する対数尤度比を横軸，そのときの relative risk の値を縦軸にプロットしたものである．All possible scan 法はあらゆるスキャン法でスキャンされる Z を包括的にスキャンしているため，表 2 に示した Echelon scan 法でスキャンされた 14 個の Z も同様に図 7 上にプロットした．Echelon scan 法は，データの持つ階層構造のピークから順にスキャンしていくため，極力 relative risk が低くならない範囲で，高い対数尤度比をもつ Z がスキャンできている．

5.3 Echelon scan 法の検出力の評価

ここでは，Tango and Takahashi (2005) の提唱した，シミュレーションによってホットスポッ

表 3. 各スキャン法によるスキャン結果.

同定された領域	属性値	期待値	Relative risk	$\log\lambda(Z)$	p	スキャンされた Z の数
All possible scan 法						
C1, D1, B2, C2, C3, C4, C5, A6, B6, C6, D6	172	102.21	1.68	45.55	0.001	198806
Circular scan 法						
C5, B6, C6, D6	81	37.17	2.18	24.90	0.001	288
Flexible scan 法 ($K = 15$)						
A6, B6, C6, D6	90	37.17	2.42	35.11	0.001	*
Flexible scan 法 ($K = 20$)						
B2, C2, C3, C4, C5, A6, B6, C6, D6	148	83.63	1.77	38.01	0.001	*
Echelon scan 法						
A6, B6, C6, D6	90	37.17	2.42	35.11	0.001	14

*Circular scan 法以上，All possible scan 法以下の数になる．

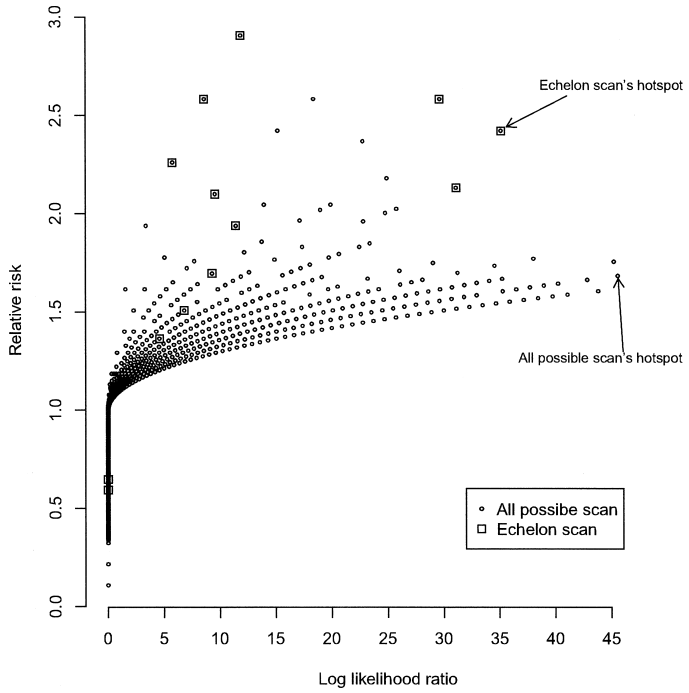


図 7. 6×4 の空間データの対数尤度比と relative risk.

トの検出力を評価する 2 変量の検出力指標を基に，Echelon scan 法における真のホットスポットの検出力評価を試みる．これらの指標を用いて，Tango and Takahashi (2005)，Tango (2010) は，Circular scan 法と Flexible scan 法の検出力に関する分析を行っている．

いま，ホットスポットとして同定された領域の数を l，その中に含まれる真のホットスポット領域の数を s とし，l に対する s の数を計測することを考える．このとき，s^* を真のホットスポットの領域の数とすると，$l = s = s^*$ の周囲の割合が高ければ，真のホットスポットを同定し，かつ，大きめな領域群をホットスポットと同定していないことになり，よい性能といえる．ここでは，先ほどの例と同様，6×4 のメッシュデータに対し，母集団を一定の下，パラメータに幅を持たせて 1000 回の Poisson 乱数を発生させた．そこから，$Z^* = \{\text{C2, B3, C3, D3, C4}\}$ (円状) を真のホットスポットと仮定した場合と，$Z^* = \{\text{C2, C3, C4, C5}\}$ (線状) を真のホットスポットと仮定した場合 (ともに Z^* 内の値が 3 倍高くなるよう設定) を想定し，Circular scan 法と Echelon scan 法の性能を比較する．ここで，それぞれ母集団が半分の値になるまでスキャンを行った．その結果をそれぞれ表 4，表 5 に示す．

ここでは，一つの目安として $l = s = s^*$ とその周辺 4 方向までの合計の割合を用いて真のホットスポット検出力を推し量る．円状を想定した場合，それぞれ表 4 の $l = s = s^* = 5$ とその周辺の合計の割合 $P(l,s) = \sum_{l=4}^{6} \sum_{s=4}^{5} \{(l,s)\}/1000$ は，Circular scan 法では 0.981，Echelon scan 法では 0.935 となった．これより，どちらの手法も高い割合で真のホットスポットを同定できていることがわかる．

一方，線状を想定した場合には，それぞれ表 5 の $l = s = s^* = 4$ とその周辺の合計の割合 $P(l,s) = \sum_{l=3}^{5} \sum_{s=3}^{4} \{(l,s)\}/1000$ を求めると，Circular scan 法ではわずか 0.114 であったのに

表 4. 円状のホットスポットを仮定した場合 ($s^* = 5$) の真のホットスポットの検出力.

Circular scan 法

	s				
l	1	2	3	4	5
1	4				
2	0	4			
3	0	2	6		
4	0	0	0	45	
5	0	0	1	0	925
6	0	0	0	1	10
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	1
10	0	0	0	0	0
11	0	0	0	0	1
12	0	0	0	0	0

Echelon scan 法

	s				
l	1	2	3	4	5
1	5				
2	0	4			
3	0	0	25		
4	0	0	1	160	
5	0	0	0	7	723
6	0	0	0	2	43
7	0	0	0	2	8
8	0	0	0	1	7
9	0	0	0	0	3
10	0	0	0	0	3
11	0	0	0	2	1
12	0	0	0	0	3

表 5. 線状のホットスポットを仮定した場合 ($s^* = 4$) の真のホットスポットの検出力.

Circular scan 法

	s			
l	1	2	3	4
1	48			
2	1	792		
3	0	3	0	
4	0	0	0	0
5	0	0	114	0
6	0	0	13	0
7	0	0	1	0
8	0	0	0	0
9	0	0	1	0
10	0	0	0	22
11	0	0	0	1
12	0	0	0	2

Echelon scan 法

	s			
l	1	2	3	4
1	12			
2	0	23		
3	0	2	87	
4	0	0	9	771
5	0	0	1	51
6	0	0	0	20
7	0	0	0	9
8	0	0	0	2
9	0	0	0	2
10	0	0	0	1
11	0	0	1	2
12	0	0	2	3

対し，Echelon scan 法では 0.919 と高い割合を示した．Echelon scan 法は，形状に依存する事なく $l = s = s^*$ 周辺に多く分布しており，真のホットスポットを検出する力が高いことを示している．

6. まとめ

本論文では空間データに対して，空間スキャン統計量によるホットスポットの検出のためのツールとして Echelon 解析を利用する手法を紹介するとともに，シミュレーションデータに対して All possible scan 法，Circular scan 法ならびに Flexible scan 法を適用することで，Echelon scan 法の妥当性について検討した．また，シミュレーションによって Echelon scan 法の検出力の評価を行った．空間スキャン統計量は尤度比を最大化するというモデル化のため，真のホットスポットのサイズよりかなり大きめの領域群をホットスポットとして同定してしまう (丹後他, 2007)．その結果 relative risk が低くなってしまったり，不自然に値の小さい領域をホットスポットとして含めてしまうという問題点がある．この問題は，任意の連結した Z をある条件の下でスキャンしていく各種の先行研究のスキャン法に共通する問題点である．この問題を

数値解析的に解決する新たな空間スキャン統計量が Tango (2008) によって提案されているが，Echelon scan 法の様にデータが本来のもつ階層構造のピークからスキャンすることは，記述統計の見地からも客観的であり，ホットスポットの意味づけや解釈について受け入れやすいだろう．加えて，尤度比のみならず relative risk の面からも有意義なホットスポットを検出することが可能である．課題として，現状の Echelon scan 法ではデンドログラムのピークとファウンデーションの境い目においてスキャンされない Z が存在してしまう．例えば今回 5.1 節で用いたデータの場合，{B6, C6} や {D6, C6} は比較的高い relative risk (それぞれ 2.58，2.42) をもつが，現状の Echelon scan 法ではスキャンされない．これら 2 つの領域群における対数尤度比の値はそれぞれ 18.36，15.17 であり，今回の例ではホットスポットとはならないが，今後はこれら Echelon の上位に位置する領域をスキャンする際には何らかの改善が必要だろう．しかし，デンドログラムの構造に基づいたスキャンは，これまでの方法で行われていた不必要なスキャンが大幅に省かれるため，各種の先行研究に比べ格段にスキャンされる Z の数が抑えられる．前述した Echelon scan 法でスキャンしきれない Z の存在の問題を差し引いても，これは大きな利点であると言えるだろう．これにより，これまでは計算コストの面から適用が困難であった数千から数万に及ぶ領域からなる様な大容量の空間データに対するホットスポットの検出が可能となる．参考までに，母集団一定の下，Poisson 乱数を発生させた 50×50 のメッシュデータに対し，Echelon デンドログラムを求め母集団の半分までスキャンする Echelon scan 法の一連の解析を行ったところ，その計算時間は 165.26 ± 4.06 (Mean $\pm$ SD) 秒であった (Platform：R2.10，64bit 3GHz Intel Core 系 PC による 30 回の計測)．これより，Echelon scan 法は広範囲にわたって測定された環境データ，リモートセンシングデータ，ハザードマップ等，広い応用分野への適用が期待される．

謝　辞

本研究は，科研費・若手研究 (B) (21700305) の助成を受けたものである．

参 考 文 献

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.

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. and Kurihara, K. (2008). A new approach to spatial clustering based on hierarchical structure, *COMPSTAT2008 Proceedings in Computational Statistics* (ed. P. Brito), 193–200.

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., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistics, *Statistic in Medicine*, **25**, 3929–3943.

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical representation, *Clas-*

sification, Clustering and Data Mining Applications (ed. D. Banks et al.), 251–258, Springer, Berlin, Tokyo.

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.

Kurihara, K., Myers, W. L. and Patil, G. P. (2000). Echelon analysis of the relationship between population and land cover patter based on remote sensing data, *Community ecology*, **1**, 103–122.

Kurihara, K., Ishioka, F. and Moon, S. (2006). Detection of hotspots on spatial data using principal component analysis, *Journal of Korean Data Analysis Society*, **8**(2), 447–458.

間瀬　茂，武田　純(2001)．『空間データモデリング—空間統計学の応用—』，データサイエンスシリーズ 7，共立出版，東京.

Moran, P. (1948). The interpretation of statistical maps, *Journal of the Royal Statistical Society B*, **10**, 243–251.

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.

Takahashi, K., Yokoyama, T. and Tango, T. (2010). *FleXScan v3.1*: *Software for the Flexible Scan Statistic*, National Institute of Public Health, Japan.

Tango, T. (1995). A class of tests for detecting 'general' and 'focuses' clustering of rate diseases, *Statistics in Medicine*, **14**, 2323–2334.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing, *Statistics in Medicine*, **19**, 191–204.

Tango, T. (2008). A spatial scan statistic with a restricted likelihood ratio, *Japanese Journal of Biometrics*, **29**(2), 75–95.

Tango, T. (2010). Statistical methods for disease clustering, *Statistics for Biology and Health*, Springer, New York.

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

丹後俊郎，横山徹爾，高橋邦彦(2007)．『空間疫学への招待』，医学統計学シリーズ 7，朝倉書店，東京.

Tomita, M., Hatsumichi, M. and Kurihara, K. (2008). Identify LD blocks based on hierarchical spatial data, *Computational Statistics & Data Analysis*, **52**(4), 1806–1820.

108 Proceedings of the Institute of Statistical Mathematics Vol. 60, No. 1, 93–108 (2012)

Hotspot Detection Using Scan Method Based on Echelon Analysis

Fumio Ishioka[1] and Koji Kurihara[2]

[1]School of Law, Okayama University
[2]Graduate School of Environmental Science, Okayama University

There are several approaches to detecting hotspots from different kinds of spatial data. Recently, a spatial scan statistical method for finding hotspot areas based on a likelihood ratio has been a very common and useful method. However, this method tends to detect hotspots much larger than the true hotspot. Therefore it does not always detect hotspots with high relative risk. A problem is how to scan regions that have a high likelihood ratio and relative risk. Echelon analysis is a useful technique for systematically and objectively investigating the phase-structure of spatial lattice data. In this study, we use an echelon scan method to explore hotspot regions based on spatial structure, and compare them with those detected by a previous study's method. In addition, we newly propose a method for scanning all hotspot candidate regions. Finally, we evaluate the validity of the echelon scan by comparison with all possible scans for simulated data.

Key words: Hotspot, spatial data, spatial scan statistic, Echelon analysis.