統計数理(2025)
第 73 巻 第 1 号 77–99

特集「空間統計モデリング：理論と応用」
［原著論文］

連続値を対象とした位相的階層構造に基づく空間集積性の検出について

石岡 文生†

（受付 2024 年 6 月 24 日；改訂 10 月 29 日；採択 10 月 31 日）

要　　旨

「興味のある対象が，ある特定の地域に集中しているか(空間的な集積性は存在するか)」を知ることは，空間データ解析における関心事の一つである．近年，空間データの各領域を特定のルールに基づいて走査(スキャン)し，尤度に基づいて空間集積性の有無を評価する空間スキャン検定が，様々な分野において広く利用されている．しかし，空間スキャン検定に関連する研究成果の多くは，主にその対象をカウントデータ(離散値)としており，また，その集積領域群の形状にも制約があった．そこで本研究では，「離散値をとらない空間データに対し，任意の形状をした集積領域群を評価できるか」という問いに対し，エシェロン解析法をベースとしたアプローチによって解決を試みる方法を提案する．エシェロン解析法とは，空間データの各領域が持つ 1 変量値と領域間の近傍情報に基づいて，データを同位相の領域群(エシェロン)に分類し，それらを階層構造のグラフで表現する手法である．提案法により，多くの場合で連続値として得られる「ある予測モデルに基づいた推定値」などのデータに対して，柔軟な空間集積性の議論が可能となる．本稿では，数値実験を通じて提案法の有効性を検証するとともに，従来法との違いについて考察する．また，クリギング予測値やベイズ推定値といった連続値を取るデータへの応用例も紹介する．

キーワード：空間集積性，空間スキャン検定，エシェロン解析法，エシェロンスキャン法．

1. はじめに

「興味のある対象が，ある特定の地域に集中しているか(空間的な集積性は存在するか)」を知ることは，データ解析における関心事の一つである．例えば，「新型肺炎などの感染症がどの地域で集中して発生しているのか？」や，「ある町の生活道路において，交通事故がどの場所で頻発しているのか？」などのように，「どこかに集積性は存在するのか？それとも全体的にばらついているのか？」「集積性が存在しているとしたら，その範囲はどこまでなのか？」を統計的根拠(データ)に基づいて決定できれば，「発生源の特定による原因の解明」や，「将来の環境や健康に対する影響の早期発見」に寄与することが期待される．

そんな中，空間データの各領域を特定のルールに基づいて走査(スキャン)し，尤度に基づいて空間集積性の有無を評価する空間スキャン検定 (Kulldorff, 1997)が，疫学の分野などで広く利用されている．また，この空間スキャン検定を行うためのフリーソフトウェア SaTScan™ と

† 岡山大学 学術研究院環境生命科学学域：〒 700–8530 岡山県岡山市北区津島中 3–1–1

相まって，関連する研究成果が多数提供されている．空間スキャン検定は，1. 尤度算出に利用する統計モデル，および 2. スキャン方式，の 2 つの要素から成り立ち，それぞれの要素を組み合わせることで実行される．ところが，実際の空間スキャン検定の適用において，1 の側面では「ある疾病に罹患した患者数」や「交通事故の発生件数」などのカウントデータ（離散値）を対象にした Poisson モデルを用いたもの，2 の側面では領域を同心円状にスキャンする circular scan 法を用いたもの，がその大半を占める．そのため，「離散値をとらないデータ」を対象とした「任意形状をした空間集積性」を評価するための空間スキャン検定は，十分に確立されていないのが現状である．

離散値をとらないデータに関する空間集積性の先行研究では，連続値を対象とした Normal モデル（Kulldorff et al., 2009; Huang et al., 2009）や，生存時間を対象にした Exponential モデル（Huang et al., 2007）およびそれを拡張した Weibull モデル（Bhatt and Tiwari, 2014）などが提案されている．これらのモデルの一部は，ソフトウェア SaTScan™ に実装されているが，このソフトは同心円状にスキャンを行う circular scan 法（Kulldorff, 1997）を採用しているため，円の形状をした集積領域群しか検出できず，例えば川や道路に沿うような非円状の集積性は評価できない．SaTScan™ には楕円状にスキャンするオプションも存在するが，これはスキャン形状の課題を本質的に解決するものにはなっていない．一方で，検出される集積形状の制約を緩和するスキャン方式が複数提案されており（Duczmal and Assunção, 2004; Patil and Taillie, 2004; Tango and Takahashi, 2005; Costa et al., 2012），そのいくつかは統計ソフト R のパッケージで公開されている．しかしながら，それらの多くは Poisson モデルや Bernoulli モデルなどの離散値を想定したモデルでデザインされている．

このような状況の中で，筆者らはエシェロンスキャン法（栗原, 2003; 石岡・栗原, 2012）と呼ばれるスキャン方式を提案している．筆者らは，この方式によって「集積形状の制約」や「計算コスト」が改善されることを実証してきた（石岡・栗原, 2012; Ishioka et al., 2019; 竹村 他, 2021）．一方で，これら先行研究では主に「領域のスキャンの効率化」に焦点を当てており，他のスキャン方式との比較検証や，実データへの適用の際には，離散値を用いてきた．連続値を対象にエシェロンスキャン法を適用した先行研究としては，掃部 他（2023）があるものの，その有効性についての詳細な検証や応用についてはまだ検討の余地が残されている．そこで本稿では，空間スキャン検定における重要な課題の一つ，すなわち「離散値をとらない空間データに対し，任意の形状をした集積領域群を評価できるか」について，エシェロンスキャン法のアプローチから解決を試みる．

本稿の構成は以下の通りである．第 2 節おいて本稿で扱う統計量のモデルについて述べる．第 3 節ではエシェロンスキャン法について説明する．第 4 節では提案法の有効性について数値実験により検証する．第 5 節では，提案法の応用例としてクリギング予測値やベイズ推定値への適用を試みる．第 6 節でまとめと今後の課題について触れる．

2. 空間スキャン統計量

2.1 連続値を対象とした Normal モデル

空間スキャン統計量は，空間集積性を評価するのに用いられる統計指標であり，最初は Kulldorff and Nagarwalla（1995）によって提案された．その後，データの性質に応じた様々な統計量のモデルが提案されているが，本稿では，連続的なデータに適応するために開発された Normal モデル（Kulldorff et al., 2009; Huang et al., 2009）を扱う．これは，対象データが個々のカウント値として取得されるのではなく，地域ごとに集計された感染率や汚染濃度などのように，個別の観測値を基に推定される集約レベルの観測値に対して，未知の真値の地理的分布

に焦点を当てたモデルとなっている．

いま，解析対象地域全体 $\boldsymbol{G}$ が全部で m 個の領域から構成されているとする．ここで，$\boldsymbol{G}$ 内のある 2 つ以上の領域が連結して形成される部分集合を集積領域群の候補と考え，この部分集合をウィンドウと呼び，$\boldsymbol{Z}$ で表す．このとき，任意の $\boldsymbol{Z}$ において，$i \in \boldsymbol{Z}$ を満たす領域 i での観測値 y_i は，$y_i|w_i \sim N(\mu_{\boldsymbol{z}}, \sigma^2/w_i)$ に従うものと仮定する．また，$i \in \boldsymbol{Z}^c (= \boldsymbol{G} - \boldsymbol{Z})$ においては，$y_i|w_i \sim N(\mu_{\boldsymbol{z}^c}, \sigma^2/w_i)$ に従うとする．なお，これらの観測値は互いに独立であるとする．ここに，$\mu_{\boldsymbol{z}}$ と $\mu_{\boldsymbol{z}^c}$ はそれぞれの領域集合における平均，σ^2 は領域全体の分散を表す．また，観測値の地域信頼性や不確実性の度合いを調整するために，領域 i に関連付けられた重み w_i を設定する．これにより，個々の観測結果ではなく，各領域内の観測値の全体的な振る舞いを反映した尺度に基づいて，空間集積性の有無を評価することが可能となる．ここで，w_i が大きいほど，その領域で観測される値の振れ幅が小さくなり，y_i は領域 i の真値に近い，信頼性の高い値と見なされる．w_i には，領域 i における観測値の分散の逆数や観測数などを充てることができる．本稿では，各領域 $i\,(i = 1, 2, \ldots, m)$ において一組の (y_i, w_i) が利用可能な状況を対象とする．なお，各領域 i において 1 つのデータ y_i しか観測されていない場合には，重み $w_i = 1$ を一律に与えるか，あるいは y_i の信頼性(重み w_i)を外生的に与える方法が考えられる．具体的には，地域ごとの自然災害リスクやインフラ整備状況などの地理的要因を指標化したり，歴史的背景といった過去の情報に基づいて重みを設定する方法が挙げられる．さらには，専門家の意見や評価に基づいてデータの信頼性を外生的に付与するアプローチも考えられる．

さてこのとき，$\boldsymbol{Z}$ に高い観測値からなる集積性(ホットスポットクラスター)が認められるか否かは，次の仮説検定問題となる．

$$
\begin{aligned}
H_0 &: \mu_{\boldsymbol{z}} = \mu_{\boldsymbol{z}^c} = \mu_0 \qquad \text{for} \quad {}^{\forall}\boldsymbol{Z} \in \mathcal{Z} \\
H_1 &: \mu_{\boldsymbol{z}} > \mu_{\boldsymbol{z}^c} \qquad\quad \text{for} \quad {}^{\exists}\boldsymbol{Z} \in \mathcal{Z}
\end{aligned}
$$

ここに，$\mathcal{Z}$ は $\boldsymbol{Z}$ の取り得る全体集合である．

空間スキャン統計量は，H_0 と H_1 の 2 つのモデルの尤度比で定義される．いま，H_1 のもとでの尤度関数は，(2.1)式で表される．

$$
\begin{aligned}
L_1(\boldsymbol{Z}) &= \prod_{i \in \boldsymbol{Z}} \sqrt{\frac{w_i}{2\pi\sigma^2}} \exp\left(-\frac{w_i}{2\sigma^2}(y_i - \mu_{\boldsymbol{z}})^2\right) \times \prod_{i \notin \boldsymbol{Z}} \sqrt{\frac{w_i}{2\pi\sigma^2}} \exp\left(-\frac{w_i}{2\sigma^2}(y_i - \mu_{\boldsymbol{z}^c})^2\right) \\
&= (\sqrt{\sigma^2})^{-m} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i \in \boldsymbol{Z}} w_i (y_i - \mu_{\boldsymbol{z}})^2 + \sum_{i \notin \boldsymbol{Z}} w_i (y_i - \mu_{\boldsymbol{z}^c})^2\right)\right] \prod_{i=1}^{m} \sqrt{\frac{w_i}{2\pi}} \qquad (2.1)
\end{aligned}
$$

このとき，重み付き平均 $\mu_{\boldsymbol{z}}$, $\mu_{\boldsymbol{z}^c}$ の最尤推定量は，それぞれ $\hat{\mu}_{\boldsymbol{z}} = \frac{\sum_{i \in \boldsymbol{Z}} (w_i y_i)}{\sum_{i \in \boldsymbol{Z}} w_i}$, $\hat{\mu}_{\boldsymbol{z}^c} = \frac{\sum_{i \notin \boldsymbol{Z}} (w_i y_i)}{\sum_{i \notin \boldsymbol{Z}} w_i}$ となる．また，重み付き分散 σ^2 の最尤推定量を $\hat{\sigma}_1^2$ とすると，$\hat{\sigma}_1^2 = \frac{\sum_{i \in \boldsymbol{Z}} w_i (y_i - \hat{\mu}_{\boldsymbol{z}})^2 + \sum_{i \notin \boldsymbol{Z}} w_i (y_i - \hat{\mu}_{\boldsymbol{z}^c})^2}{\sum_{i=1}^{m} w_i}$ と与えることができる．一方，H_0 のもとでの尤度関数は，(2.2)式で表される．

$$
\begin{aligned}
L_0 &= \prod_{i=1}^{m} \sqrt{\frac{w_i}{2\pi\sigma^2}} \exp\left(-\frac{w_i}{2\sigma^2}(y_i - \mu_0)^2\right) \\
&= (\sqrt{\sigma^2})^{-m} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{m} w_i (y_i - \mu_0)^2\right] \prod_{i=1}^{m} \sqrt{\frac{w_i}{2\pi}} \qquad (2.2)
\end{aligned}
$$

このとき，重み付き平均 μ_0 の最尤推定量は $\hat{\mu}_0 = \frac{\sum_{i=1}^{m} (w_i y_i)}{\sum_{i=1}^{m} w_i}$，重み付き分散 σ^2 の最尤推定量 $\hat{\sigma}_0^2$ は $\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{m} w_i (y_i - \hat{\mu}_0)^2}{\sum_{i=1}^{m} w_i}$ でそれぞれ与えられる．

尤度比 $LR = L_1/L_0$ は，最尤推定量を代入することで(2.3)式を導出する．

$$LR(\boldsymbol{Z}) = \begin{cases} (\hat{\sigma}_1^2/\hat{\sigma}_0^2)^{-m/2}, & \hat{\mu}_{\boldsymbol{z}} > \hat{\mu}_{\boldsymbol{z}^c} \\ 1, & \text{その他} \end{cases} \tag{2.3}$$

通常，計算を簡略化するために，LR を対数変換した $LLR = \log LR$ が使用される．なお，$\boldsymbol{Z}$ が低い観測値からなる集積性(コールドスポットクラスター)であるか否かについて検討する場合は，H_1 の不等号の向きを逆にして考えればよく，このときの尤度比は(2.3)式の上段の不等号の向きを逆にしたものが対応する．また，これ以降はホットスポットクラスターを “ホットスポット”，コールドスポットクラスターを “コールドスポット” と呼ぶこととする．

なおここに，$\hat{\mu}_{\boldsymbol{z}}, \hat{\mu}_{\boldsymbol{z}^c}, \hat{\mu}_0$ は，それぞれ $\mu_{\boldsymbol{z}}, \mu_{\boldsymbol{z}^c}, \mu_0$ の不偏推定量となるが，$\hat{\sigma}_1^2$ および $\hat{\sigma}_0^2$ は，それぞれ σ_1^2 と σ_0^2 の不偏推定量とはならない．そのため，実際に LLR を算出する際には，不偏性を考慮した分散を用いる．重み付き分散の不偏推定量にはいくつかの流儀が存在するが，本稿では SaTScan™ が採用している方法，すなわち重み付き分散の最尤推定量に $\frac{m}{m-1}$ を乗じたものをそれぞれ $\hat{\sigma}_1^2, \hat{\sigma}_0^2$ とする．

つづいて，ウィンドウ $\boldsymbol{Z}$ の全体集合 $\mathcal{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots\}$ の中から最大の対数尤度比をもつ，

$$\hat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z} \in \mathcal{Z}} LLR(\boldsymbol{Z}) \tag{2.4}$$

なるウィンドウ $\hat{\boldsymbol{Z}}$ をホットスポットの候補と考える．また，$\hat{\boldsymbol{Z}}$ が統計的に有意なホットスポットであるかどうかを評価するためには，H_0 の下での $\max_{\boldsymbol{Z} \in \mathcal{Z}} LLR(\boldsymbol{Z})$ の分布が必要だが，これを一意に定めることは解析的に困難なため，モンテカルロ法によるシミュレーションを用いて求めた p 値によって有意性を評価するのが通例となっている．具体的には，非パラメトリックな検定法である permutation test を利用することで，観測値の分布に影響を受けることなく有意性を評価する方法が行われる（Kulldorff et al., 2009; Huang et al., 2009）．これは，領域 $i\,(i = 1, 2, \ldots, m)$ における (y_i, w_i) の組み合わせを固定した上で，領域 i の配置をランダムに並び替えることで生成されるデータセット (y_i^*, w_i^*) に対してウィンドウ $\boldsymbol{Z}^*$ の集合 $\mathcal{Z}^*$ を求め，$\lambda^* = \max_{\boldsymbol{Z}^* \in \mathcal{Z}^*} LLR(\boldsymbol{Z}^*)$ を算出する．同様の手順を N 回繰り返し，得られた $\boldsymbol{\Lambda}^* = \{\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*\}$ を H_0 の下での最大対数尤度比の分布と見なして，元の観測データセット $\boldsymbol{G}$ から求めた $LLR(\hat{\boldsymbol{Z}})$ が $\boldsymbol{\Lambda}^*$ の中でどの程度極端な位置にあるかを確認し，その確率を p 値とするものである．$\boldsymbol{\Lambda}^*$ における $LLR(\hat{\boldsymbol{Z}})$ の降順での順位を R とすると，ホットスポットの候補 $\hat{\boldsymbol{Z}}$ に対する p 値は，

$$p = \frac{R}{1+N} \tag{2.5}$$

となる．

2.2 領域のスキャン

さてここで，ウィンドウ $\hat{\boldsymbol{Z}}$ をどのように求めるかが問題になる．一つ一つの $\boldsymbol{Z}$ に対し $LLR(\boldsymbol{Z})$ を算出しながら $\hat{\boldsymbol{Z}}$ を探すのは非常に手間であるし，そもそも領域数 m が極端に少ない場合を除き，一般的に $\boldsymbol{Z}$ を形成するパターンは膨大な数となるため，これらをすべて調べることは不可能である．そこで，効率的に $\boldsymbol{Z}$ のパターンを調べる(これを “スキャンする” という)ために，現実的に計算可能な M 個の $\boldsymbol{Z}$ の集合 $\mathcal{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_M\}$, $\boldsymbol{Z}_i \subset \boldsymbol{G}$ を構築するためのスキャン方式がいくつか提案されている．空間スキャン統計量を提唱した Kulldorff 自身は，円状に $\boldsymbol{Z}$ をスキャンする circular scan 法を提案している．これは，ある発生源を中心に同心円状に感染範囲が拡大していくような伝染性の疾病などを対象にする場合に高い検出力を

示し，また，アルゴリズムが簡便で計算コストが低いといった長所がある反面，川や道路に沿うような線状やその他の任意の形状をした空間集積性の検出には適していないことが指摘されている．

3. 位相的階層構造に基づいたスキャン方式

3.1 エシェロン解析法

栗原 (2003)や石岡・栗原 (2012)は，エシェロン解析法に基づいてデータの位相的な階層構造を得て，その構造を利用してスキャンを行うエシェロンスキャン法を提案した．エシェロン解析法 (Myers et al., 1997)は，空間データを構成する各領域の構造を客観的に表現するための手法である．図 1 は，3×3 の格子状のデータに対するエシェロン解析法の適用例を示している．この例では，9 つの領域の空間的な位置を表面上の 1 変量値 x の大小(高低)に基づいて分類し，エシェロンデンドログラム(これ以降，デンドログラムと呼ぶ)と呼ばれるグラフによってデータの位相的な階層構造を視覚化している．デンドログラムから，このデータは 5 つの階層に分類されることがわかる．

デンドログラムの任意の階層 k について，k の上位に他に階層が存在しない場合，k を「ピークの階層」とみなす．また，ある階層 l の上位に 2 つ以上の別の階層が存在する場合，l を「ファウンデーションの階層」とみなす．図 1 のデンドログラムにおいて，階層 1，階層 2，階層 3 はピークであり，階層 4 と階層 5 はファウンデーションである．エシェロン解析法の最大の利点は，空間データの各領域に対し近傍情報が与えられれば，次元を問わずその構造が客観的に階

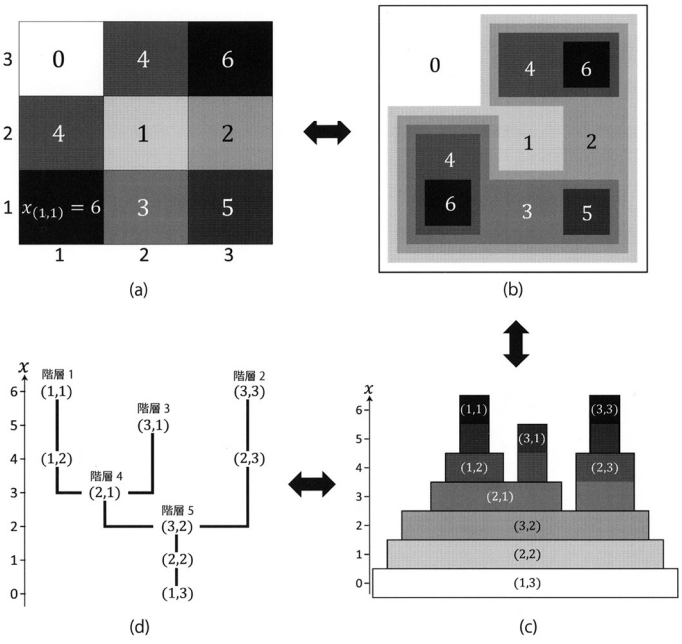

図 1. 3×3 の格子状のデータに対するエシェロン解析法の適用例．(a)3×3 の格子データ．各領域が 1 変量値 x を有している．ここに，$x_{(1,1)}$ は座標 $(1,1)$ の値を示している．(b)x の値に基づいて，データを同じ位相領域に分類する様子．(c)各位相領域の相対関係を，x を縦軸にとって表現したもの．ここに，$(1,1)$ は座標 $(1,1)$ の領域を指す．(d)デンドログラムの完成．

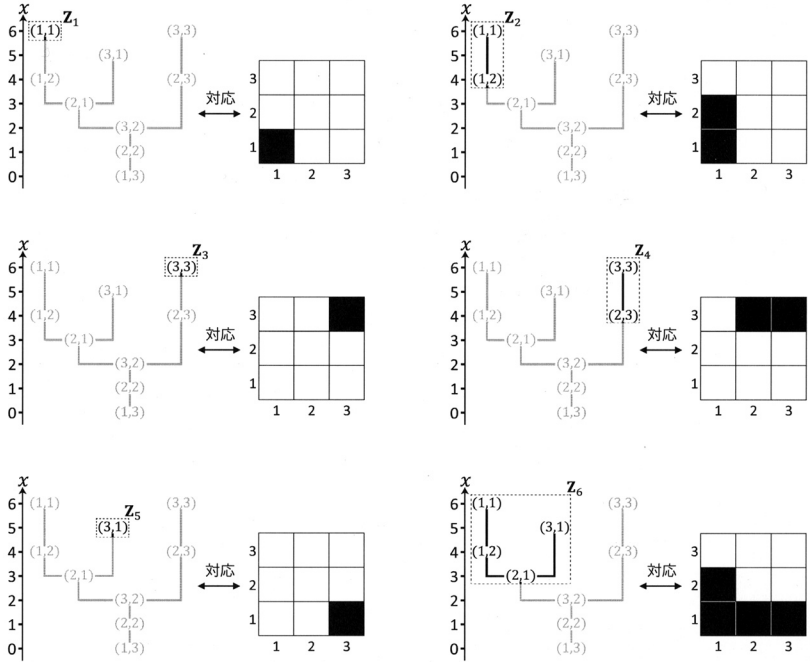

図 2. 3×3 の格子状のデータに対しエシェロンスキャン法を適用する様子．ここに，ウィンドウ内に許容する最大領域数は 4 としている．このとき，$\boldsymbol{Z}$ は全部で 6 つのパターンが存在する．

層化できることである．これは，データの分布や構造を調べるために，ヒストグラム，箱ひげ図，散布図などを使用するのと同様に，空間データを視覚的に記述するための有用なツールとなる．

3.2 エシェロンスキャン法

エシェロンスキャン法では，デンドログラムのピークとなる階層から順に，その階層に含まれる領域を上から順に $\boldsymbol{Z}$ に取り込みながら $\mathcal{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots\}$ を構築していく．なおここに，解析者が予め定めた「ウィンドウ内に許容する最大領域数」に達するか，またはその階層に含まれるすべての領域をスキャンし終えたとき，スキャンの対象が次の階層に移行する．また，ファウンデーションの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すべて含めた状態で，そのファウンデーションの階層の領域を上から順に $\boldsymbol{Z}$ に含めていく．図 2 は，図 1 のデンドログラムから構築される $\mathcal{Z}$ の結果を示している．なお，エシェロンスキャン法によって選ばれる $\boldsymbol{Z}$ は，必ず互いに連結する領域群になる．エシェロン解析法やエシェロンスキャン法のより詳細なアルゴリズムについては，Kurihara et al. (2020)や栗原・石岡 (2021)を参照されたい．

4. 数値実験

4.1 実施手法の概要

本節では，「Normal モデルに基づく空間スキャン統計量」と「エシェロンスキャン法」を組み合わせた新たな空間集積性検出手法の特性について，数値実験を通じて検証する．具体的に

表 1．数値実験で用いる手法の概要．

	従来法	提案法
方式	Normal モデル ＋ Circular scan 法	Normal モデル ＋ エシェロンスキャン法
スキャンの際に必要となる情報	各領域の座標情報	各領域間の近傍情報， 各領域の 1 変量値 x
使用したソフトウェア	SaTScan™(ver.10.1.3), rsatscan (ver.1.0.7)	R による独自プログラム

は，従来広く用いられている circular scan 法との比較を行い両手法の特徴的な違いを明らかにするとともに，母集団分布が正規分布から逸脱した場合での提案法の頑健性についても確認する．従来法と提案法の概要を表 1 に示す．

従来法である circular scan 法は，それぞれ領域 $i\,(i=1,2,\ldots,m)$ を起点として同心円状にスキャンを行う．言い換えれば，circular scan 法は特定の領域 i を中心として，その周囲の領域を順に調査していくものと言える．ここで，領域間の距離が重要な役割を果たし，i と距離が近い領域から順にウィンドウ $\boldsymbol{Z}$ に取り込みながら $\mathcal{Z}=\{\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots\}$ を構築していく．領域間の距離の算出には，各領域の代表点の座標情報を利用する．従来法の適用に際し，Kulldorff et al.（2024）が開発したソフトトウェア SaTScan™ および，SaTScan™ を R 上で実行可能にする rsatscan パッケージ（Kleinman, 2023）を使用する．

提案法では，まずデータに対してデンドログラムを作成するために，各領域間の近傍情報と各領域の 1 変量値 $x_i\,(i=1,2,\ldots,m)$ を準備する必要がある．ホットスポットが比較的大きな値を持つ x で形成される領域群の場合，それらの各領域はデンドログラムの上位階層に位置することになるため，優先的に $\boldsymbol{Z}$ に取り込まれる．このプロセスにより，エシェロンスキャン法ではある特定の形状に囚われない柔軟な形状をした集積領域群の検出を可能にする．x の定義において最も単純な方法は，領域 i における観測値 y_i をそのまま用いることが考えられる．

$$x_i = y_i,\ i=1,2,\ldots,m \tag{4.1}$$

一方で，エシェロンスキャン法は，近傍情報と 1 変量値に基づいて領域を階層化していく際に，高い観測値を持つ領域の近傍にある領域を次々と同じ階層に取り込んでしまい，結果として集積性の観点からは不自然な，巨大でいびつなホットスポットが検出される場合がある．この問題は，空間スキャン統計量のモデルに依存するものではなく，エシェロンスキャン法自体に起因するものである．なお，この問題に対処するため，竹村 他（2021）や Takemura et al.（2022）は AESM（Adjusted echelon scan method）を提案している．その着想に基づいて，本稿では，解析対象領域全体の平均 $\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i$ より低い値を持つ領域 i については，x_i を y の最小値に置き換えることを試みる．これにより，$\bar{y}$ を下回る値を持つような，本来ならばホットスポットとして適切でない領域 i はデンドログラムの底に移動し，スキャン対象から除外されることとなる（図 3）．このアプローチは，スキャンされるウィンドウ数の削減にも繋がり，結果的に計算コストの改善にもなる．

また，重み w_i は各観測値の不確実性を示す指標であり，w_i が大きいほど，その領域 i における観測値 y_i の変動が小さくなり，y_i は真値に近い信頼性の高い値と見なすことができる．この特性を踏まえ，重みの大きい，すなわち信頼性の高い観測値を持つ領域ほどデンドログラムの上位に配置されやすく，逆に重みが小さい観測値を持つ領域ほどデンドログラムの下位に配置されやすくするアプローチを試みる．その一環として，ここでは（4.2）式について検討する．

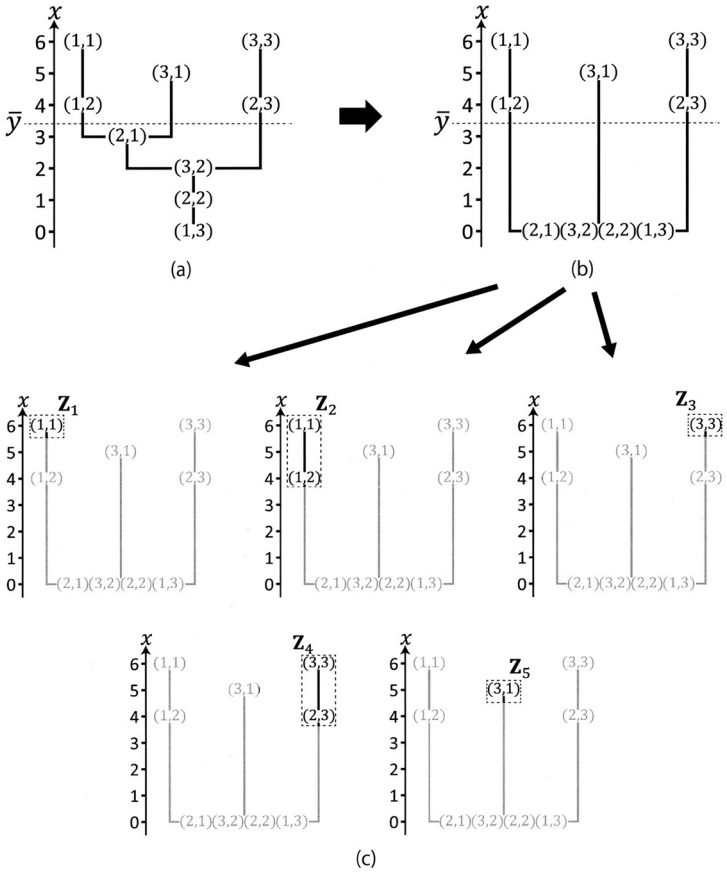

図 3. (a)図 1 の 3×3 の格子状のデータのデンドログラム．(b)$\bar{y}$ よりも低い値を持つ領域の値 x を y の最小値に変換した際に作成されるデンドログラム．(c)変換後のデンドログラムに対するスキャン．ここに，ウィンドウ内に許容する最大領域数は 4 としている．

これにより，例えば複数の高い観測値を持つ領域が存在し，それらの重みが異なる場合に，より信頼性の高い領域がホットスポットして検出されやすくなることが期待される．

$$x_i = \begin{cases} w_i \cdot y_i, & y_i > \bar{y} \\ w_i \cdot \min\{y_1, y_2, \ldots, y_m\}, & \text{その他} \end{cases} \tag{4.2}$$

なお，提案法を適用するためのソフトウェアは存在しないが，デンドログラムの階層構造の取得には，R の echelon パッケージ (Ishioka, 2024)が利用できる．

4.2 設計

数値実験の設計においては，先行研究 (Huang et al., 2009)を一部参考にし，以下の方法に従った．使用するデータは，10×10 の格子データ ($m = 100$)とし，そのデータ上に図 4 のような 3 つの形状からなる真のホットスポット $\boldsymbol{Z}^+$ を想定する．これら各形状に対して従来法と提案法を適用し，$\boldsymbol{Z}^+$ の形状をどれだけ正確に検出できるかを検証する．

円状を想定した $\boldsymbol{Z}^+$ として，座標 $(3, 6)$ を中心に距離が 2 以内に含まれる領域を設定した．これは従来法でも十分に検出可能と考えられる．また，直線状に伸びた線状の $\boldsymbol{Z}^+$ と，輪のよ

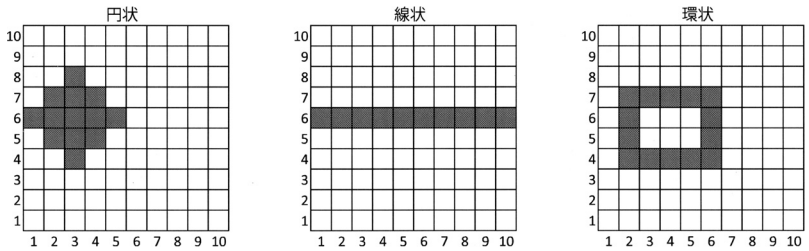

図 4．想定した真のホットスポット $\boldsymbol{Z}^+$．

表 2．$\boldsymbol{Z}^+$ 内外の各領域の y と w．ケース 1，ケース 2，ケース 3 は従来法と提案法を適用．ケース 4 は(4.2)式に基づく提案法を適用．

		y（乱数生成）				
	$\mathbf{Z}^+$ の形状	$i \in \mathbf{Z}^+$	$i \notin \mathbf{Z}^+$	重み w	c	η
ケース 1	円状	$N(c\sqrt{2},1)$	$N(0,1)$	$w_i=\eta,\ i\in\mathbf{G}$	1.0 2.0 3.0	1 1 1
ケース 2	円状， 線状， 環状	$N(c\sqrt{2},1)$	$N(0,1)$	$w_i=U[1-\eta,1+\eta],$ $i\in\mathbf{G}$	2.0 2.0 2.0 2.0 2.0	0.00 0.05 0.10 0.25 0.50
ケース 3	円状， 線状， 環状	$N(c\sqrt{2},1)$	$N(0,1)$	$w_i=\eta,\ i\in\mathbf{Z}^+$ $w_i=1,\ i\notin\mathbf{Z}^+$	2.0 2.0 2.0 2.0 2.0	1 2 5 10 100
ケース 4	円状	$D(c\sqrt{2},1)$ D：正規分布，ラプラス分布， ロジスティック分布，一様分布 $D(1+c\sqrt{2},1)$ D：対数正規分布 $D(1+c\sqrt{2})$ D：ポアソン分布	$D(0,1)$ $D(1,1)$ $D(1)$	$w_i=U[1-\eta,1+\eta],$ $i\in\mathbf{G}$	2.0 2.0 2.0 2.0 2.0	0.00 0.05 0.10 0.25 0.50

うな環状の $\boldsymbol{Z}^+$ を想定した場合についても検証する．これらは従来法ではその形状を正確に捉えることは困難と予想される．次に，$\boldsymbol{Z}^+$ 内外の各領域の y と w をそれぞれ表 2 のように与える．

ケース 1 では，$\boldsymbol{Z}^+$ 内外での平均差が，従来法と提案法の違いに及ぼす影響を検証する．ケース 2 では，ホットスポットの有無にかかわらず，領域全体を対象として，各領域で得られる観測値のばらつき(不確実性)の違いが結果に与える影響を確認する．このとき，重み w は一様乱数によって生成し，具体的には，$w=1.0$ を中心に幅 0.1 から幅 1.0 までの範囲で生成した場合を検証する．さらに，ケース 2 では $\boldsymbol{Z}^+$ の形状を円状，線状，環状に設定した場合の違いについても検証する．

ケース 3 では，ホットスポット内における観測値の不確実性の違いが結果に与える影響を確認する．ここで，η が大きくなるにつれて，ホットスポット内で観測される y の不確実性が低下することを意味する．なお，2.1 節で示した y の従う分布からも明らかなように，$\boldsymbol{Z}^+$ の内外

で重みが異なる状況は特殊なケースである点に留意する必要がある．具体例として，都市部における感染症の拡大が挙げられる．人口密度の高い都市部では感染症が急速に広がりやすく，クラスター(ホットスポット)が形成されやすい状況が存在する．実際，COVID-19 のパンデミック時には，都市部での感染拡大が顕著であった．そして都市部には大規模な病院や医療機関が多く存在し，感染率をより正確に推定するのに必要な情報が集まりやすいと考えられる．また，ケース 3 においても同様に，$\boldsymbol{Z}^+$ の形状を円状，線状，環状と変化させた場合について検証する．

ケース 4 では，正規母集団ではない状況における提案法の頑健性について確認する．ここでは，Huang et al. (2009)の先行研究に倣い，母集団分布として正規分布，ラプラス分布，ロジスティック分布，一様分布，対数正規分布，ポアソン分布を仮定し，これら各々に対して(4.2)式に基づく提案法の検出精度を評価する．このとき，$\boldsymbol{Z}^+$ 内外で生成したデータの分布 $D(\mu, \sigma^2)$ は，表 2 に示す通りである．なお，対数正規分布においては平均を 0 にすることができないため，$\boldsymbol{Z}^+$ 内外において平均を $\mu + 1$ に設定している．また，ポアソン分布のパラメータ λ は平均と分散の両方を表し，分散が 0 より大きい値を持つことから，$\boldsymbol{Z}^+$ 内では $\lambda = 1 + c\sqrt{2}$，$\boldsymbol{Z}^+$ 外では $\lambda = 1$ と設定している．さらに，ケース 4 ではケース 2 と同様に，領域全体で重みを一様乱数で変動させている．

ここで，circular scan 法で用いる座標情報には領域の中心座標を使用し，エシェロンスキャン法で用いる近傍情報には上下左右の 4 近傍(ルーク型の隣接関係)を使用する．また，ウィンドウ内に許容する最大領域数は，両手法とも全領域の半数に相当する 50 とした．

4.3 評価指標

各ケースにおいて，従来法と提案法の検出精度は，*Sensitivity* と *PPV* (positive predictive value)の 2 つ指標によって評価する．*Sensitivity*，*PPV* ともに，空間集積性の検出精度を議論する際に広く用いられている（Huang et al., 2007; Ma et al., 2016; Lee et al., 2021; 竹村 他, 2021）．*Sensitiviry* は，$\boldsymbol{Z}^+$ 内の領域のうち，$\hat{\boldsymbol{Z}}$ によって捉えることのできた割合のことで，次式で定義される．

$$Sensitivity(\hat{\boldsymbol{Z}}) = \frac{\mathrm{length}(\boldsymbol{Z}^+ \cap \hat{\boldsymbol{Z}})}{\mathrm{length}(\boldsymbol{Z}^+)} \tag{4.3}$$

ここに，length() は領域数である．また，$0 \leq Sensitivity \leq 1$ であり，1 に近いほど「$\boldsymbol{Z}^+$ をホットスポットとして検出できた」と判断できる．一方，*PPV* は，$\hat{\boldsymbol{Z}}$ 内の領域のうち，$\boldsymbol{Z}^+$ を含んでいる割合であり，次式で定義される．

$$PPV(\hat{\boldsymbol{Z}}) = \frac{\mathrm{length}(\boldsymbol{Z}^+ \cap \hat{\boldsymbol{Z}})}{\mathrm{length}(\hat{\boldsymbol{Z}})} \tag{4.4}$$

Sensitivity と同様に $0 \leq PPV \leq 1$ となり，1 に近いほど「$\boldsymbol{Z}^+$ 以外をホットスポットとして検出しなかった」と判断できる．さらに，ケース 1，ケース 2，ケース 3 では両手法における対数尤度比 $LLR(\hat{\boldsymbol{Z}})$ についても確認する．$LLR(\hat{\boldsymbol{Z}})$ が大きいほど，高尤度のホットスポットが検出できたことになる．なお，今回の数値実験においては，両手法ともに(2.4)式を満たすウィンドウ $\hat{\boldsymbol{Z}}$ のみを「検出されたホットスポット」と見なし，いわゆる第 2 ホットスポットなどについては考慮しない．また，検出された $\hat{\boldsymbol{Z}}$ の形状や $LLR(\hat{\boldsymbol{Z}})$ のみに焦点を当て，有意性については言及しない．

4.4 結果

表 2 の各ケースについて，繰り返し数を 1000 回として y を乱数によって生成した．図 5，図

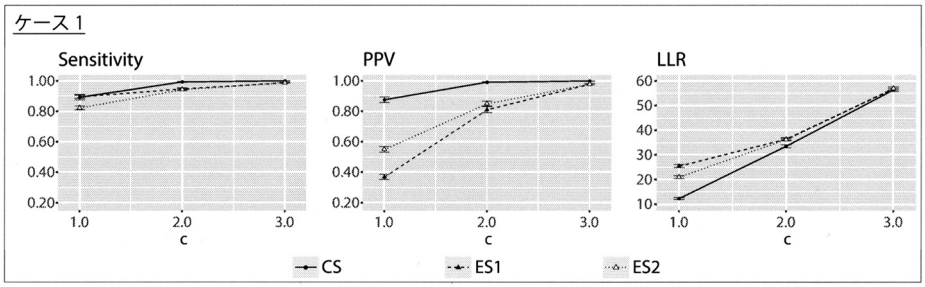

図 5．従来法(CS)と提案法(ES1，ES2)に対する *Sensitivity*，*PPV*，*LLR* の結果(ケース 1)．

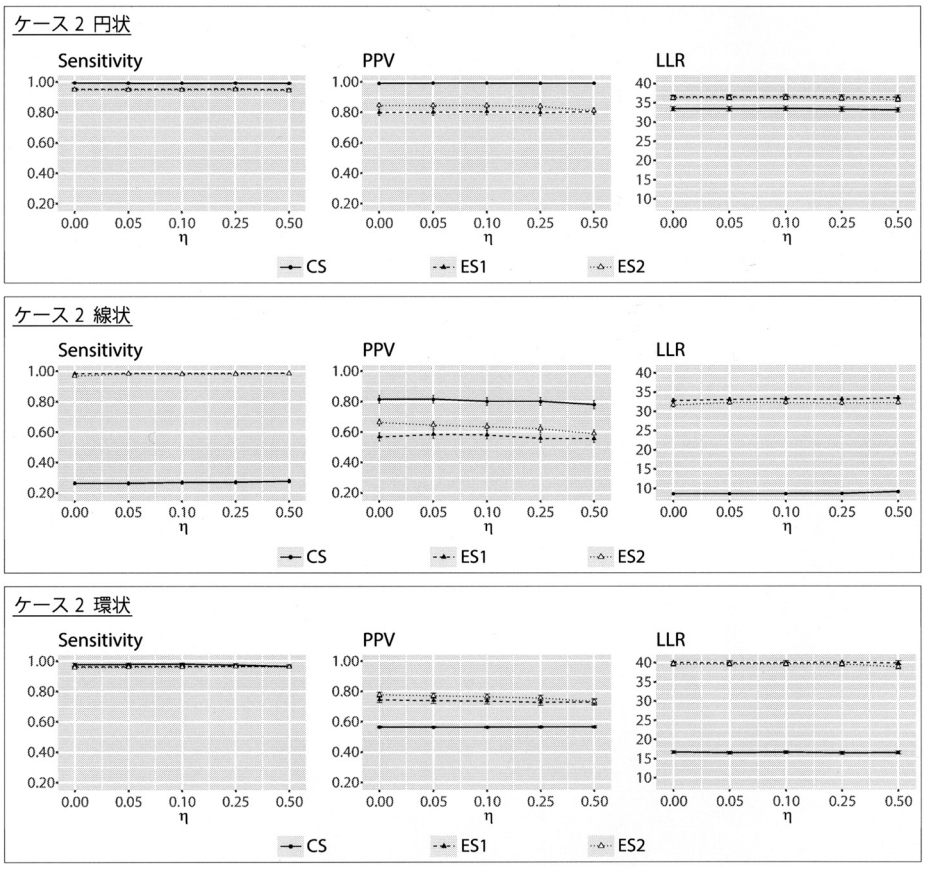

図 6．従来法(CS)と提案法(ES1，ES2)に対する *Sensitivity*，*PPV*，*LLR* の結果(ケース 2)．

6，図 7，図 8 に，それぞれ各指標の 平均 $\pm 3 \times$ 標準誤差 の結果を示す．これらの図において，CS は circular scan 法(従来法)，ES1 と ES2 はそれぞれ(4.1)式と(4.2)式に基づいたエシェロンスキャン法(提案法)の結果を示している．

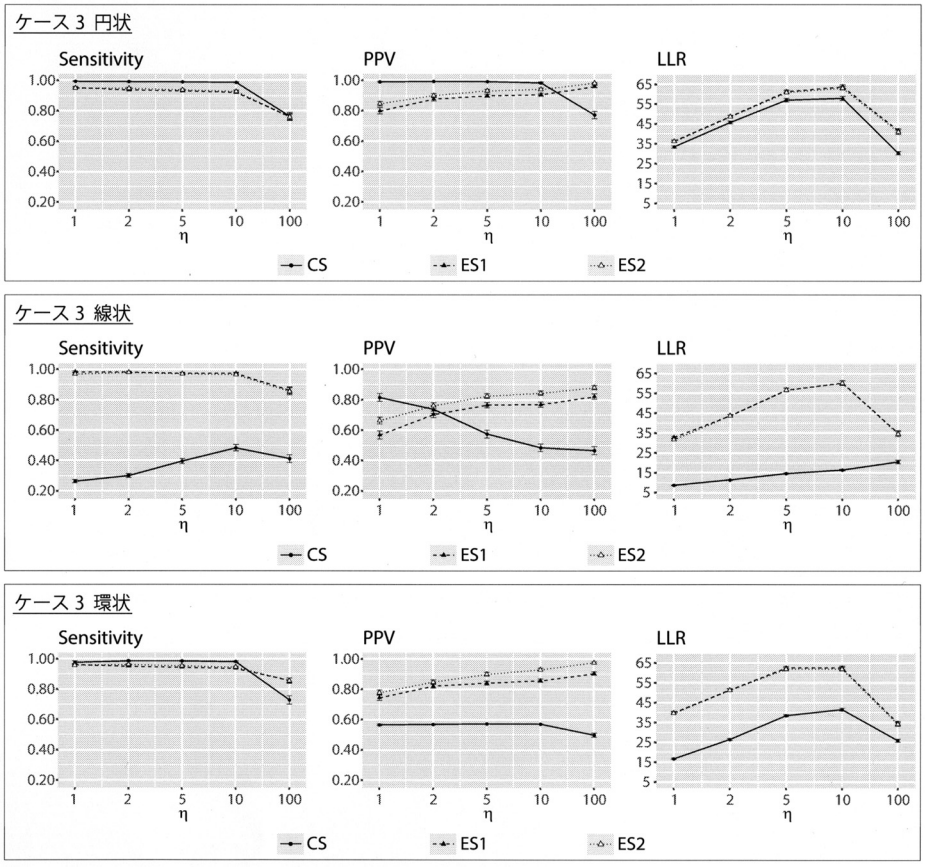

図 7. 従来法(CS)と提案法(ES1, ES2)に対する *Sensitivity*, *PPV*, *LLR* の結果(ケース 3).

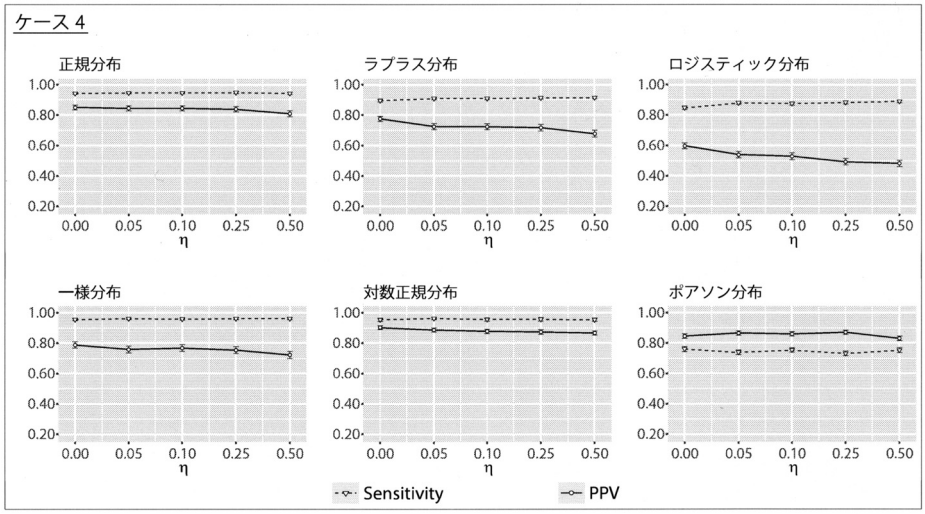

図 8. 提案法(ES2)に対する *Sensitivity*, *PPV* の結果(ケース 4).

4.5 考察

ケース 1，ケース 2，ケース 3 の結果から，全体的に提案法は従来法よりも LLR の大きい $\hat{\boldsymbol{Z}}$，つまりは高尤度なホットスポットを検出できている．このことは，提案法が形状の制限を受けず柔軟にウィンドウをスキャンすることに起因していると考えられる．

ケース 1 の結果に注目すると，従来法は $Sensitivity$，PPV ともに 0.85 以上を示しており，このことから真のホットスポット $\boldsymbol{Z}^+$ を高い精度で捉えることができている．なお，$c \geq 2$ では提案法も $Sensitivity$ と PPV の両方で 0.8 以上の高い値を示した．このことから，ホットスポットが円状である場合，従来法は提案法よりも $\boldsymbol{Z}^+$ 内外のわずかな平均差をより正確に識別できることがわかる．

次に，$c = 2$ に固定し，重みを変化させたケース 2 およびケース 3 に焦点を当てる．まず，従来法については，両ケースともに円状の $\boldsymbol{Z}^+$ において高い検出精度を示したが，それぞれ線状の $\boldsymbol{Z}^+$ では $Sensitivity$ が低下し，環状の $\boldsymbol{Z}^+$ では PPV が低下する傾向が見られた．一方，提案法ではすべての形状において高い $Sensitivity$ を維持しており，$\boldsymbol{Z}^+$ に含まれる領域をほぼ確実に捉えていることが確認できる．

また，ケース 2 において，提案法は全体的に PPV がやや低い値を示した．この原因として，4.1 節でも述べたように，エシェロンスキャン法の特性上，高い観測値を持つ領域の近傍にある領域がスキャン時にウィンドウ内に取り込まれやすいことが挙げられる．しかし，ケース 3 のように $\boldsymbol{Z}^+$ 内の重みが大きくなるにつれて，提案法の PPV が改善する傾向が見られた．なお，重みが極端に大きくなる(ケース 3, $\eta = 100$)場合，従来法の $Sensitivity$ と PPV は提案法に比べ大きく低下する様子も確認された．さらに，ケース 3 において $\boldsymbol{Z}^+$ 内の重みを 100 に設定した場合，すべての評価指標において提案法が良好な結果を示した．また，提案法におけるデンドログラムの違い(ES1 と ES2)に着目すると，$Sensitivity$ と LLR にはほとんど差異が見られなかったが，PPV に関しては ES2 において改善が見られた．

最後に，ケース 4 の結果について考察する．$Sensitivity$ に関しては，ポアソン分布を除く 5 つの連続型分布において，いずれも 0.8 以上の高い精度が確認された．ポアソン分布の $Sensitivity$ がやや低かった要因としては，他の分布と比較してデータのばらつきが大きく，観測値が離散値であるために，$\boldsymbol{Z}^+$ 内の値の変動が大きくなったことが考えらえる．なお，この傾向は Huang et al. (2009)の先行研究における従来法でも確認されている．また，ロジスティック分布において PPV が特に低かった理由として，ロジスティック分布が正規分布に比べて裾が広く，$\boldsymbol{Z}^+$ 外の領域でも比較的高い値が観測されやすいため，それらの領域がウィンドウ内に取り込まれたと推察される．一方，対数正規分布では正規分布よりも PPV が向上している．これは分布が右に長い裾を持つため，$\boldsymbol{Z}^+$ 外の領域で小さな値が観測されやすくなり，その結果，$\boldsymbol{Z}^+$ 外の領域がウィンドウ内に取り込まれにくくなったためと考えられる．

5. 実データへの応用

前節の数値実験の結果から，データが連続値で与えられる状況において，提案法は，$\boldsymbol{Z}^+$ 内外の平均に一定の差が存在する場合，その形状にかかわらず，$\boldsymbol{Z}^+$ を構成する領域を安定的に取りこぼすことなく検出できることが示唆された．また，母集団が正規分布に従わない場合であっても，提案法は一定の検出精度を保持できることが確認された．本節では，実データに対する提案法の適用例を紹介する．

5.1 マース川沿岸の亜鉛濃度のクリギング予測値への適用

5.1.1 データの概要

はじめにクリギングによる推定値に対し提案法を適用する例を紹介する．使用するデータは，

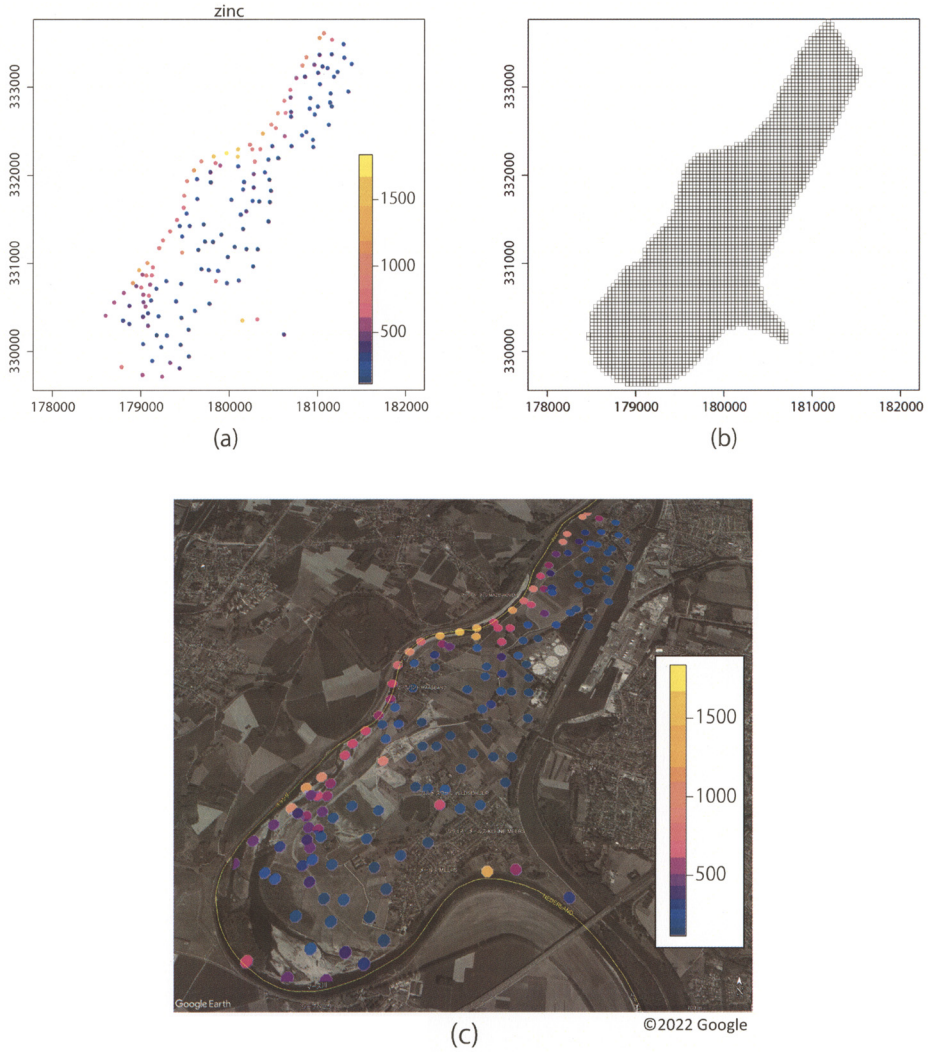

図 9. (a)元データにおける 155 の観測地点とその亜鉛濃度．(b)推定する地点を表した格子．(c)観測地点を Google Earth 上に重ねた図．マース川北部で亜鉛濃度が高くなっている様子が見て取れる．

R の sp パッケージが提供する `meuse` である．これは，オランダのマース川沿岸の 155 地点で観測された複数の重金属濃度の情報を収録しており，今回はこの中から表土の亜鉛濃度(`zinc`)のデータを扱う．このデータ対して空間補間法の一種であるクリギングを適用し，`meuse.grid` が提供する 40m 四方の中心座標における 3103 地点の亜鉛濃度を推定する．図 9 は，155 の観測地点とその亜鉛濃度，および推定する 3103 地点の格子データを示している．

5.1.2 クリギングによる推定

クリギングは，空間的に相関するデータを補間・推定するための統計的手法である．具体的には，観測データの空間的な変動パターンをモデル化し，周囲の観測データの加重平均を使

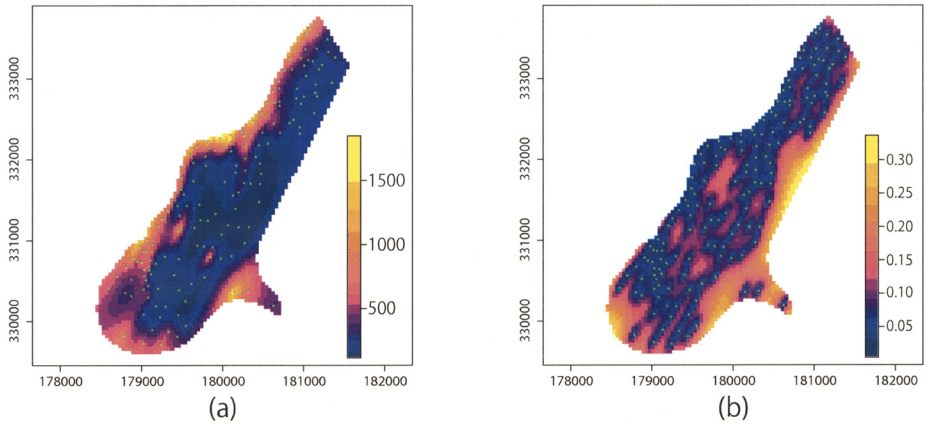

図 10. (a)クリギングによる亜鉛濃度の推定結果の可視化．なお，可視化に際し対数値を元の単位に戻している．(b)クリギング分散の可視化．図内の緑点は元データの観測地点を示す．

表 3. 今回使用したクリギング．

項目	内容
理論バリオグラムモデル	指数型モデル：$\gamma(d_{ij}) = \theta_1 + \theta_2 \left(1 - \exp\left(-\frac{d_{ij}}{\theta_3}\right)\right)$ d_{ij} は領域 i と領域 j との距離
パラメータの推定法	制限付き最尤推定法
推定されたパラメータ	$\hat{\theta}_1 = 0$，$\hat{\theta}_2 = 0.622$，$\hat{\theta}_3 = 450.00$
異方性を調整するパラメータ	座標の回転角度：0.469，異方性比：2.605
クリギングのタイプ	通常型クリギング

用して未観測地点での値を推定する．クリギング手法の詳細な説明は間瀬 (2010)や瀬谷・堤 (2014)にゆずるとして，ここでは AIC が良好な値を示した推定結果(図 10)を採用する．表 3 に，今回使用したモデルや手法，推定されたパラメータをまとめている．なお，推定に際して，亜鉛濃度を対数変換したデータを使用することで，クリギングの精度を向上させている．また，実際のクリギングの適用には R の geoR パッケージ (Paulo et al., 2024)を使用した．

このように，クリギングを活用することで，限られた少数のデータから対象領域全体の状況を推定し，その結果を色分けした図として視覚的に確認することが可能となる．しかしながら，その図から受ける印象は，階級区分の設定方法に大きく依存し，例えば濃い色で示された部分がホットスポットであるかのような誤解を招く恐れがある．今回のように，図 10(a)で示した 3103 地点の亜鉛濃度の推定値に対し，特に高亜鉛濃度となる領域群の存在の有無を明らかにすることは，データの未観測領域に跨るホットスポットの挙動に関する知見を得るための一つのアプローチとして位置付けることができると考える．

また，クリギングによって推定された値の不確実性を評価するために，クリギング分散と呼ばれる指標が用られることがある．クリギング分散は，未知の位置での値の推定誤差の大きさを表すもので，推定値の信頼性を評価する重要な尺度となる．通常，クリギング分散が小さいほど，推定された値の信頼性が高いと判断する．今回求まったクリギング分散を図 10(b)に示す．元々の観測地点が少なかった南西部や，東部から南東部にかけてのエリアでは，クリギン

表 4. 従来法と提案法によって検出されたホットスポットのウィンドウと，そのウィンドウ内外の重み付き平均，重み付き分散，対数尤度比．p 値については，それぞれスキャン方式に準じた 3 通りの H_0 の分布から算出している．

スキャン方式	$\hat{\mathbf{Z}}$	$\hat{\mu}_{\boldsymbol{z}}$	$\hat{\mu}_{\boldsymbol{z}^c}$	$\hat{\sigma}_1^2$	$LLR(\hat{\mathbf{Z}})$	p 値 $\boldsymbol{\Lambda}^*_{\mathrm{CS}}$	$\boldsymbol{\Lambda}^*_{\mathrm{ES2}}$	$\boldsymbol{\Lambda}^*_{\mathrm{CS}} \cup \boldsymbol{\Lambda}^*_{\mathrm{ES2}}$
従来法	$\hat{\mathbf{Z}}_1$	1319.67	377.40	69403.06	346.18	0.0001	0.0001	0.00005
	$\hat{\mathbf{Z}}_2$	762.63	376.10	79393.40	137.53	0.0001	0.0041	0.00205
	$\hat{\mathbf{Z}}_3$	1015.17	388.17	81795.20	91.29	0.0125	0.1827	0.09755
提案法	$\hat{\mathbf{Z}}_1$	801.82	285.74	41941.35	1127.61	0.0001	0.0001	0.00005
	$\hat{\mathbf{Z}}_2$	1008.00	387.94	81711.78	92.88	0.0116	0.1669	0.08920
	$\hat{\mathbf{Z}}_3$	152.45	402.02	85327.54	0.000	1.0000	1.0000	1.00000

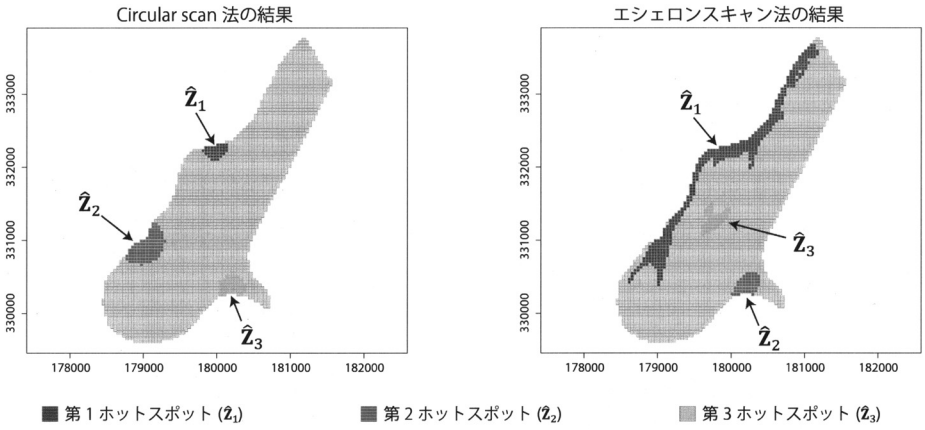

図 11. 亜鉛濃度の予測値に対するホットスポットの検出結果．

グ分散の値が高くなっていることが見て取れる．

今回，空間スキャン統計量で使用する観測値 $y_i\,(i = 1, 2, \ldots, 3103)$ には，図 10(a) の亜鉛濃度の推定値を使用し，重み $w_i\,(i = 1, 2, \ldots, 3103)$ には，図 10(b) のクリギング分散の逆数を用いた．このとき，$\hat{\mu}_0 = 396.18, \hat{\sigma}_0^2 = 86752.57$ となる．

5.1.3 ホットスポットの検出

それぞれ circular scan 法(従来法)とエシェロンスキャン法(提案法)によって検出されたホットスポットを表 4 と図 11 に示す．なおここに，ウィンドウ内に許容する最大領域数は，両手法ともに全領域数の約半数に相当する 1551 領域としている．空間情報については，前節の数値実験のときと同様に，circular scan 法で用いる座標情報には領域の中心座標を使用し，エシェロンスキャン法における各領域の近傍情報は上下左右の 4 近傍とした．また，デンドログラム作成のための x には(4.2)式を用いた．さらにここで，(2.4)式において得られた $\hat{\boldsymbol{Z}}$ を第 1 ホットスポット $\hat{\boldsymbol{Z}}_1$ とし，$\hat{\boldsymbol{Z}}_1$ と領域が重複しない $\boldsymbol{Z} \in \mathcal{Z}$ のうちで，$LLR(\hat{\boldsymbol{Z}}_1)$ に次いで大きい対数尤度比 $LLR(\hat{\boldsymbol{Z}}_2)$ を有するウィンドウ $\hat{\boldsymbol{Z}}_2$ を第 2 ホットスポットと定義した．同様に，$\hat{\boldsymbol{Z}}_1 \cup \hat{\boldsymbol{Z}}_2$ と領域が重複しない $\boldsymbol{Z} \in \mathcal{Z}$ のうちで，$LLR(\hat{\boldsymbol{Z}}_2)$ に次いで大きい対数尤度比 $LLR(\hat{\boldsymbol{Z}}_3)$ を有するウィンドウ $\hat{\boldsymbol{Z}}_3$ を第 3 ホットスポットとして定義している．

北西部で検出されたホットスポットに注目すると，従来法ではそれを 2 つの異なる領域群として検出したが，提案法では 1 つの大規模な領域群として検出した．提案法では，マース川沿岸に沿うように南北に伸びた細長いエリアをホットスポットとして捉えており，マース川北部

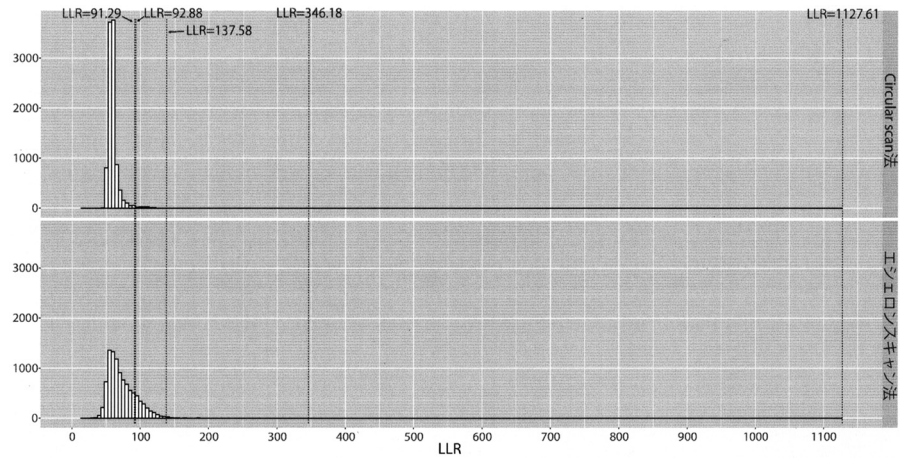

図 12. 2 つのスキャン方式による H_0 の下での最大対数尤度比の分布，ならびに今回検出された各ウィンドウにおける対数尤度比の位置.

沿岸地域と亜鉛濃度との関連性が示唆される結果となった．また，LLR の値から，提案法の方が高い尤度のホットスポットを検出できている．

5.1.4 有意性の評価

2.2 節で述べた permutation test によって，検出された $\hat{\boldsymbol{Z}}$ の有意性を確認する．ここで，H_0 の分布として想定する $\boldsymbol{\Lambda}^*$ は，当然ながらスキャン方式に大きく影響される．通常は，使用するスキャン方式に基づいて $\boldsymbol{\Lambda}^*$ を得ることになるが，ここでは circular scan 法とエシェロンスキャン法という異なる 2 つのスキャン方式を扱うことから，(I) circular scan 法に基づく $\boldsymbol{\Lambda}^*_{\mathrm{CS}}$ と，(II) (4.2) 式のエシェロンスキャン法に基づく $\boldsymbol{\Lambda}^*_{\mathrm{ES2}}$ をそれぞれ生成することとした．繰り返し数については，(I)，(II) ともに $N = 9999$ とした．図 12 は，クリギングによる亜鉛濃度の予測データから得られた (I) と (II) の分布と，検出された各ウィンドウにおける対数尤度比の位置を示している．また，(I) と (II) を合わせた $\boldsymbol{\Lambda}^*_{\mathrm{CS}} \cup \boldsymbol{\Lambda}^*_{\mathrm{ES2}}$ の分布についても確認した (このとき，$N = 9999 + 9999 = 19998$ となる)．これら各々の分布に対して p 値を算出した結果を表 4 に示す．この結果から，北西部で検出されたホットスポットは高度に有意であることがわかる．

5.2 首都圏の市区町村別合計特殊出生率のベイズ推定値への適用

5.2.1 データの概要

つづいて，首都圏[1)]の市区町村別に集計された合計特殊出生率に対し，提案法を適用する例を紹介する．合計特殊出生率とは，「15 歳から 49 歳までの女性の年齢別出生率を合計したもの」(厚生労働省, 2024) であり，1 人の女性が生涯に産む子供の数を示す重要な指標となる．近年，日本における出生率の低下は大きな社会問題となっており，政府や自治体，民間団体が様々な対策を講じている．2024 年 6 月に厚生労働省が公表したデータによると，2023 年の日本における合計特殊出生率は 1.20 であり，これは 8 年連続で前の年を下回る結果であった．また，都道府県別では東京都の合計特殊出生率が最も低く，0.99 であった (NHK WEB, 2024)．

使用するデータとして，e-Stat「人口動態統計特殊報告」における統計表「合計特殊出生率・母の年齢階級別出生率，都道府県・保健所・市区町村別」から，それぞれ平成 20〜24 年，平成 25〜29 年，平成 30 年〜令和 4 年の 3 期間のものをダウンロードした (e-Stat, 2024)．これらのデータは，国勢調査の年を中心とした 5 年間の日本における日本人データ[2)]を基に，市区町村

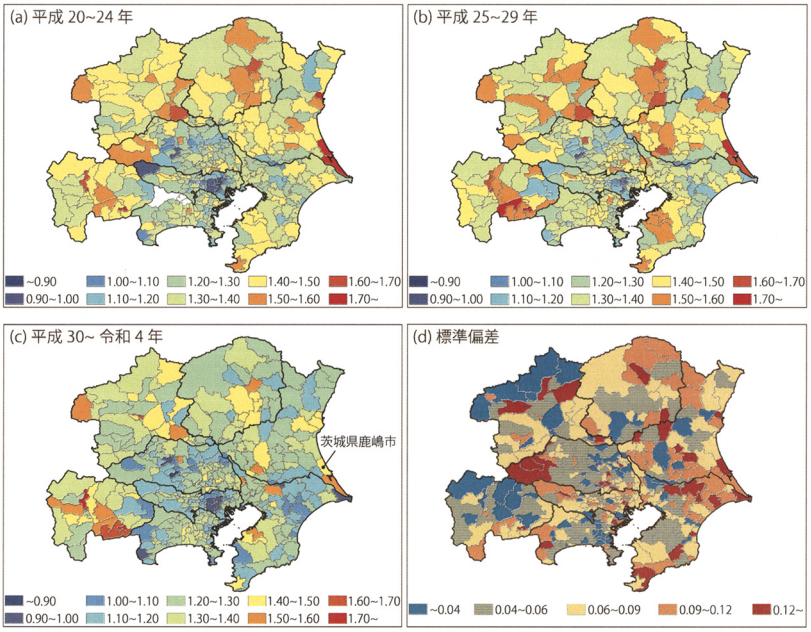

図 13. 厚生労働省公表による首都圏市区町村別の合計特殊出生率のベイズ推定値とその標準偏差を地図上に可視化した結果．(a) 平成 20〜24 年．ここに，神奈川県相模原市の緑区・中央区・南区については欠損値(白色)となっている．(b) 平成 25〜29 年．(c) 平成 30 年〜令和 4 年．(d) 3 期間での標準偏差．例えば，茨城県鹿嶋市の合計特殊出生率のベイズ推定値は，平成 20〜24 年：1.77，平成 25〜29 年：1.79，平成 30〜令和 3 年：1.49 であり，その標準偏差は 0.169 となる．

ごとで集計されたものとなっている．また，人口の少ない地域における出生数の変動を緩和するために，実際のデータは合計特殊出生率をベイズ推定した値が収録されている．図 13 は，それぞれの期間における合計特殊出生率のベイズ推定値，ならびに 3 期間でのばらつき(標準偏差)を地域別に可視化した結果である．これらのデータに対し，それぞれ合計特殊出生率の高い地域群(ホットスポット)と，逆に低い地域群(コールドスポット)の検出を試みる．ここで，3 期間のベイズ推定値の平均値を y とし，重み w には標準偏差の逆数を用いる．これらの y_i と w_i $(i = 1, 2, \ldots, 373)$ に基づいて，$\hat{\mu}_0 = 1.318, \hat{\sigma}_0^2 = 1.973 \times 10^{-2}$ となる．

5.2.2 空間集積性の検出

ホットスポットの検出については，これまでと同じくデンドログラム作成のための 1 変量値 x に(4.2)式を用い，(2.3)式に基づいた対数尤度比 LLR によって統計量を算出する．一方，コールドスポットの検出に際しては，合計特殊出生率 y_i が低い地域 i ほど，また，その重み w_i が大きいほど，デンドログラムの上位階層に位置させたいと考え，x には次の(5.1)式を用いることとした．

$$x_i = \begin{cases} -(y_i/w_i), & y_i < \bar{y} \\ -(\max\{y_1, y_2, \ldots, y_m\}/w_i), & \text{その他} \end{cases} \tag{5.1}$$

また，対数尤度比については(2.3)式の上段の不等号を逆向きにしたものに基づいて算出する．なおここで，デンドログラム作成のための近傍情報には，市区町村の境界線を共有している地

表 5. 空間集積性が認められたウィンドウ $\hat{\boldsymbol{Z}}$ に対する，ウィンドウ内外の重み付き平均，重み付き分散，対数尤度比，p 値.

集積性のタイプ	$\hat{\mu}_{\mathbf{z}}$	$\hat{\mu}_{\mathbf{z}^c}$	$\hat{\sigma}_1^2$	$LLR(\hat{\mathbf{Z}})$	p 値
ホットスポット	1.496	1.297	1.588×10^{-2}	40.478	0.0004
コールドスポット	1.215	1.358	1.563×10^{-2}	43.416	0.0001

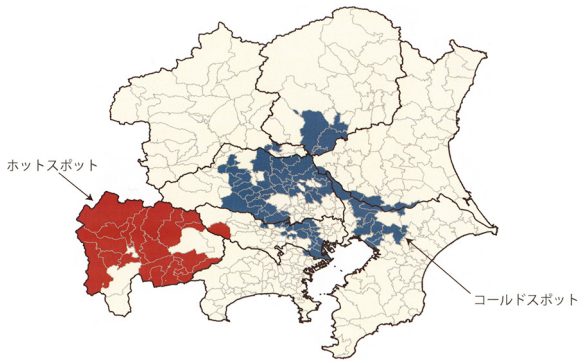

図 14. 有意な空間集積性が認められた地域群．それぞれ赤色がホットスポット，青色がコールドスポットを示す.

域を互いに近傍とした．また，スキャンに関して，ウィンドウ内に許容する最大領域数は全地域の半数に相当する 186 地域とし，モンテカルロ法の繰り返し数は $N = 9999$ とした.

上記の設定を基にエシェロンスキャン法を適用した結果，それぞれ有意なホットスポットとコールドスポットが一つずつ認められた．それらの詳細を表 5 に，地図上に可視化した結果を図 14 に示す.

今回解析対象とした首都圏の 15 年にわたる期間中で，安定して高い合計特殊出生率を示した地域群(ホットスポット)には，主に山梨県内の市町村によって構成される 23 地域が同定され，山梨県外では唯一，東京都の檜原村だけが含まれた．また，ホットスポットの範囲に着目すると，檜原村以外では県境で明確に区切られる結果となった．逆に，安定して低い合計特殊出生率を示した地域群(コールドスポット)は，山梨県を除いた都県にまたがる 78 市区町で形成され，特に東京都，埼玉県，千葉県の地域が比較的多く含まれる結果となった．全体で 3,000 万人以上の人口を擁し，日本でも有数の人口密集地域として知られる南関東エリア(東京都，埼玉県，千葉県，神奈川県)のうち，神奈川県でコールドスポットとして認められたのは川崎市中原区の 1 地域のみであった.

6. まとめと今後の課題

本稿では，連続値からなる空間データに対し，「Normal モデルの空間スキャン統計量」と「エシェロンスキャン法」を組み合わせた新たな空間集積性の検出を試みた．また，提案法の有用性を数値実験で検証した結果，従来法よりも高い尤度を持ち，柔軟な形状からなる集積領域群を検出できることが示された．特に，集積領域群内の重みが大きい(集積領域群内で観測されるデータが安定していて，ばらつきが少ない)場合は，その傾向が顕著であった．提案法の応用として，まずクリギングによる予測値に適用し，従来法では捉えられなかったホットスポットの傾向を明らかにした．また，合計特殊出生率のベイズ推定値に対する適用では，地図上で示される地理空間データに対し，ホットスポットとコールドスポットの両方の観点から空間集

積性を検討した．以上の結果から，「離散値をとらない空間データに対し，任意の形状をした集積領域群を評価できるか」という課題に対応できるツールの一つとして提案法は有益と考えられ，今後はさらなる広範な応用が期待される．なお，今回のようにモデルを用いた推定値に対して適用する際，第一段階で得られる推定値が後続の集積領域検出に大きく影響する．そのため，第一段階で用いるモデルの選定とその推定精度は，全体の結果を左右する極めて重要な要素となる．この点を踏まえ，例えば「標高や湿度，風向きといった要因の影響を除去し，純粋に気温の高い領域群を検出する」といった，観測データの傾向を適切に反映したモデルの構築が必要である．こうしたモデルを基に提案手法を適用することで，目的とする空間集積性の評価をより高い精度で行うことが可能になるだろう．

次に今後の課題について述べる．エシェロンスキャン法では，まず始めにデンドログラム作成のために各領域に対し 1 つの変量値 x を定める必要があるが，今回はこの変量値として(4.2)式や(5.1)式を用いた．これらの式は，データの平均値と最小値(または最大値)を基に簡便に算出できるという利点があるが，同時にいくつかの課題も抱えている．特に，平均値と最小値(または最大値)だけではデータの全体的な分布やばらつきを十分に反映できないため，他の統計指標や手法を取り入れることが考えられる．さらに，これらの式は重みの影響を強く受ける可能性があり，その点においても慎重な検討が求められる．デンドログラムを作成する際に，重みをどのように活用するかについては，適用する状況に応じた最適なアプローチを検討する必要があるだろう．

最後に，本文中でも触れたようにホットスポット候補となる領域の組み合わせの数は，領域数に対して指数関数的に増加する．そのため，領域の形や大きさを限定して効率的にスキャンを行う方法が数多く研究されてきた．今回提案した手法においても，真に最大尤度となる領域群を完全に検出することはできない．この点は，他のスキャン手法を含め，空間スキャン検定の限界および課題の一つであると言える．この問題に対して，Ishioka et al. (2019)では，二分決定グラフを用いた手法を導入し，領域の形や大きさを限定せずに，すべてのホットスポット候補領域の組み合わせを網羅的に算出する試みがなされた．しかし，現行の空間スキャン検定の枠組みにおいては，尤度最大化を目的とするため，結果として巨大でいびつな形状を持つホットスポットが検出されやすく，これが結果の解釈を困難にする要因となる場合がある．この問題を回避するため，Tango (2008)のように統計量自体に制約を導入した手法や，Moon et al. (2023)のようにホットスポットの大きさに対する最適化について検討がされている．上述のように，本提案法にはエシェロンデンドログラムの構築という追加のステップが含まれており，このデンドログラムの形状が結果に大きく影響する．この点はデメリットとして捉えられることもあるが，デンドログラムの情報を有効活用することで，新たなメリットを生み出せる可能性もあると考える．今後は，異なるデータセットや条件下での有用性についても検証を行い，より多角的なアプローチを取り入れながら，さらなる検出精度の向上を図ることが重要となる．

謝　辞

本研究は JSPS 科研費 JP21K11786，JP24K14856 の助成を受けたものである．

注．

[1] 東京都，神奈川県，千葉県，埼玉県，茨城県，群馬県，栃木県，山梨県の 1 都 7 県．ただし空間的な関係性の観点から東京都島嶼部は除いている．このとき，地域数 m は 373 となる．

[2] 厚生労働省によると「日本に住んでいる日本人に係る日本において発生した各事象の全数」.

参 考 文 献

Bhatt, V. and Tiwari, N. (2014). A spatial scan statistic for survival data based on Weibull distributiona, *Statistics in Medicine*, **33**, 1867–1876.

Costa, M. A., Assunção, R. A. and Kulldorff, M. (2012). Constrained spanning tree algorithms for irregularly-shaped spatial clustering, *Computational Statistics and Data Analysis*, **56**, 1771–1783.

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.

e-Stat (2024). 人口動態統計特殊報告, https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00450013（最終アクセス日 2024 年 6 月 14 日）.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data, *Biometrics*, **63**, 109–118.

Huang, L., Tiwari, R., Zuo, J., Kulldorff, M. and Feuer, E. (2009). Weighted normal spatial scan statistic for heterogeneous population data, *Journal of the American Statistical Association*, **104**, 886–898.

Ishioka, F. (2024). echelon v0.2.0: The Echelon Analysis and the Detection of Spatial Clusters using Echelon Scan Method, https://CRAN.R-project.org/package=echelon（最終アクセス日 2024 年 6 月 14 日）.

石岡文生, 栗原考次 (2012). 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, 統計数理, **60**(1), 93–108.

Ishioka, F., Kawahara, J., Mizuta, M., Minato, S. and Kurihara, K. (2019). Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting, *Japanese Journal of Statistics and Data Science*, **2**, 241–262.

掃部耀平, 竹村祐亮, 石岡文生 (2023). エシェロンスキャン法を用いた重み付き Normal モデルに基づくクラスター検出について, 日本計算機統計学会第 37 回シンポジウム講演論文集, 146–149.

Kleinman, K. (2023). rsatscan v1.0.7: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software, https://CRAN.R-project.org/package=rsatscan（最終アクセス日 2024 年 6 月 14 日）.

厚生労働省 (2024). 人口動態統計特殊報告, https://www.mhlw.go.jp/toukei/list/list58-60.html（最終アクセス日 2024 年 6 月 14 日）.

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

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, **14**, 799–810.

Kulldorff, M., Huang, L. and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model, *International Journal of Health Geographics*, **8**, https://doi.org/10.1186/1476-072X-8-58.

Kulldorff, M., Harvard Medical School and Information Management Services Inc. (2024). SaTScan™v10.1.3: Software for the Spatial and Space-Time Scan Statistics, http://www.satscan.org（最終アクセス日 2024 年 6 月 14 日）.

栗原考次 (2003). 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**, 171–183.

栗原考次, 石岡文生 (2021).『エシェロン解析：階層化して視る時空間データ』, 共立出版, 東京.

Kurihara, K., Ishioka, F. and Kajinishi, S. (2020). Spatial and temporal clustering based on the echelon scan technique and software analysis, *Japanese Journal of Statistics and Data Science*, **3**, 313–332.

Lee, S., Moon, J. and Jung, I. (2021). Optimizing the maximum reported cluster size in the spatial scan statistic for survival data, *International Journal of Health Geographics*, **20**, https://doi.org/10.1186/s12942-021-00286-w.

Ma, Y., Yin, F., Zhang, T., Zhou, X.A. and Li, X. (2016). Selection of the maximum spatial cluster size of the spatial scan statistic by using the maximum clustering set-proportion statistic, *PLoS ONE*, **11**(1), https://doi.org/10.1371/journal.pone.0147918.

間瀬茂 (2010).『地球統計学とクリギング法：R と geoR によるデータ分析』, オーム社, 東京.

Moon, J., Kim, M. and Jung, I. (2023). Optimizing the maximum reported cluster size for the multinomial-based spatial scan statistic, *International Journal of Health Geographics*, **22**, https://doi.org/10.1186/s12942-023-00353-4.

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.

NHK WEB (2024). 去年の合計特殊出生率 過去最低 厚労省「必要な取り組み加速」, https://www3.nhk.or.jp/news/html/20240606/k10014472291000.html（最終アクセス日 2024 年 6 月 14 日）.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

Paulo, J. R., Peter, D., Ole, C., Martin, S., Roger, B. and Brian, R. (2024). geoR v1.9-4: Analysis of Geostatistical Data, https://CRAN.R-project.org/package=geoR（最終アクセス日 2024 年 6 月 14 日）.

瀬谷創, 堤盛人 (2014).『空間統計学：自然科学から人文・社会科学まで』, 朝倉出版, 東京.

竹村祐亮, 石岡文生, 栗原考次 (2021). Echelon scan 法による高リスクな空間クラスター検出法の提案, 計算機統計学, **34**, 23–43.

Takemura, Y., Ishioka, F. and Kurihara, K. (2022). Detection of space-time clusters using a topological hierarchy for geospatial data on COVID-19 in Japan, *Japanese Journal of Statistics and Data Science*, **5**, 279–301.

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

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, https://doi.org/10.1186/1476-072X-4-11.

Spatial Clustering Based on Topological Hierarchical Structures for Continuous Values

Fumio Ishioka

Graduate School of Environmental and Life Science, Okayama University

One of the concerns in spatial data analysis is determining whether the subjects of interest are concentrated in a certain region (i.e., whether spatial clustering exists). In recent years, spatial scan statistics, in which each region of spatial data is scanned based on specific rules and evaluated for spatial clustering by its likelihood, have been widely used in various fields. However, many previous studies have primarily used count data (discrete values) as their objects of interest, and the shapes of the clustering regions have also been limited. In this study, we attempt to answer the question, "Can we evaluate arbitrarily shaped spatial clustering for spatial data that does not take discrete values?" by using echelon analysis. The echelon analysis method classifies spatial data into groups of regions (echelons) composed of the same phase based on the univariate value of each region and the neighborhood information between regions, representing them as a graph with a hierarchical structure. By establishing a method for detecting spatial clustering based on this echelon analysis method, it becomes possible to discuss spatial aggregation for data such as estimates based on a certain prediction model, which are often obtained as continuous values. In this paper, the effectiveness of the proposed method is verified through numerical experiments, demonstrating that the proposed method has a more stable and higher likelihood than the conventional method and can detect spatial clustering with flexible shapes. When the proposed method was applied to the predictions by kriging, it revealed trends of spatial clustering that could not be captured by the conventional method. Furthermore, when applied to Bayesian estimates of the total fertility rate, spatial clusterings were identified in terms of both hot-spot and cold-spot clusters in geospatial data presented on a map.

Key words: Spatial clustering, spatial scan statistic, echelon analysis, echelon scan method.