

ベイズ推定を応用したスパースモデリングによる集積領域検出手法の提案

増田 亮*・井上 亮**

Bayesian spatial cluster detection based on sparse modeling

Ryo Masuda*, Ryo Inoue**

Abstract: Spatial cluster detection has been one of the focus areas in spatial analysis. Its objective is to identify clusters from the spatial distribution of point events aggregated at small area units. In recent years, sparse modeling has been utilized for cluster detection. Sparse modeling-based methods have the advantage of explicitly taking covariates observed at each area into account by employing regression models. However, the existing estimation method fails to offer information on the reliability of estimated parameters, which is useful in assessing the results. Therefore, this study extends the sparse modeling-based cluster detection method to the Bayesian framework and examines the practicability of the proposed method with simulation data.

Keywords: 集積領域検出 (spatial cluster detection), ベイズ推定 (Bayesian estimation), スパースモデリング (sparse modeling)

1. はじめに

地域の将来像を議論・立案する上で、各地域の社会経済活動の現状を把握できる情報は欠かせない。そのため、地域分析を通して、地理空間データから地域特性の把握につながる情報を抽出することは重要である。近年、各国政府のオープンデータ戦略の推進により、多種多様な地理空間データを容易に入手できる環境が実現されつつあり、これらの地理空間データを活用した、よりの確な地域分析が可能になると期待されている。

地理空間データの一つが、本研究でも着目する点事象データである。点事象データは、疾病の発症や犯罪の発生、商業施設の出退店などの事象の発生を、位置・時刻とともに記録したデータである。点事象データを活用した地域分析手法として、集積領域検出が長年検討されてきた。集積領域検出は、行政区画や地域メッシュなどの小地域単位で集計された点事象データの空間分布から、事象の発生する頻度が高い領域を特定することを目的とする。集積領域検出手法の代表例として、Kulldorff and Nagarwalla

(1995) が提案した空間スキャン統計があげられる。空間スキャン統計に基づいた手法は、隣接する複数の小地域からなる集積領域の候補を作成し、集積領域の候補を尤度比によって比較することで集積領域を特定する統計的アプローチをとる。空間スキャン統計は、集積領域の統計検定において問題となる「多重検定問題」を回避した集積領域検出が可能であることから、広く活用されてきた。しかし、空間スキャン統計を含む既存の集積領域検出手法の大半は同時に検出できる集積領域数が一領域にとどまり、また地域ごとの共変量の影響を考慮した集積領域の検出は難しい。

これらの課題を解決した手法として、スパースモデリングの一手法である Generalized fused lasso に基づく集積領域検出手法 (Choi *et al.*, 2018) が提案されている。Choi *et al.* (2018) は、小地域単位で集計された点事象データからの集積領域検出を、各小地域の集積の程度を表現する「集積性パラメータ」を組み込んだボアソン回帰モデルの最尤推定問題として定式化した。集積性パラメータの値を 0 に誘導す

* 学生会員 東北大学 大学院情報科学研究科 (Tohoku University)

〒980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-06 E-mail : ryo.masuda.t8@dc.tohoku.ac.jp

** 正会員 東北大学 大学院情報科学研究科 (Tohoku University)

る L1 正則化項に加え、隣接する地域の集積性パラメータを共通の値に誘導する L1 正則化項を導入することで、空間的自己相関を考慮した集積性パラメータの推定を可能にしている。

Choi *et al.* (2018) は、Generalized fused lasso に基づく集積領域検出問題の定式化と合わせて、MM アルゴリズムを適用したパラメータ推定法も提案している。しかしながら、MM アルゴリズムはパラメータの点推定値のみを求めるアルゴリズムであり、パラメータの不確実性に関する情報は推定されない。地域ごとの集積性パラメータの信頼性に基づく集積領域検出を行うには、パラメータの点推定値だけではなく、不確実性に関する情報も提供できる推定法を用いることが望ましい。

パラメータの不確実性の取り扱いに優れた推定法として、ベイズ推定が広く用いられている。ベイズ推定は、事前分布と尤度関数から、ベイズの定理によりパラメータの確率分布を推定する手法である。ベイズ推定の枠組みにおいても、パラメータのスパース性を誘導する事前分布が提案されている。Tibshirani (1996) は、線形回帰モデルにおいて、回帰係数パラメータの事前分布に二重指数分布を指定した場合、回帰係数パラメータ事後分布の最頻値が Lasso の推定値に理論上一致することを示した。これを受けて、Park and Casella (2008) は二重指数分布を事前分布に指定したベイズモデルのギブスサンプリングの構成を提案した。Kyung *et al.* (2010) は、Park and Casella (2008) を拡張し、隣接パラメータの差にも L1 正則化項を導入した Fused lasso のベイズ推定を定式化している。また、Lasso の推定値はバイアスを持つことが指摘されており (Fan and Li, 2001)、このバイアスを軽減したスパース推定が可能な事前分布も提案されている (例えば、Carvalho *et al.*, 2010; Shimamura *et al.*, 2019)。

ベイズ推定を用いた空間分析の観点からは、パラメータの空間的自己相関を表現することができる Conditional Autoregressive 事前分布 (Besag *et al.*, 1991) が提案されており、疾病地図の作成 (Lee, 2011) などに応用されている。しかし、スパースモデリングに基づく集積領域検出をベイズ推定に拡張した手法

はこれまで提案されていない。

そこで本研究は、スパースモデリングの手法である Generalized fused lasso に基づく集積領域検出手法を、空間的自己相関を考慮した事前分布を応用してベイズ推定の枠組みに拡張した手法を提案する。また、提案手法をシミュレーションデータに適用し、その実行可能性を確認する。

2. Generalized fused lasso に基づく集積領域検出

Choi *et al.* (2018) が提案した集積領域検出手法は、スパースモデリングの手法である Generalized fused lasso (Tibshirani and Taylor, 2011) に基づいている。本節では、まず Generalized fused lasso を紹介し、続いて Generalized fused lasso に基づく集積領域検出手法を説明する。

2.1. Generalized fused lasso

Generalized fused lasso は、パラメータの値自体と隣接するパラメータの差に L1 正則化項を導入し、変数選択とパラメータの隣接関係に基づく変化点抽出を同時に行うことができる手法である。線形回帰モデルにおける Generalized fused lasso の最小化問題は、式(1)で表現される。

$$\min_{\beta} \left[\|y - X\beta\|_2^2 + \lambda \sum_{(i,j) \in C} |\beta_i - \beta_j| + \gamma \lambda \sum_{i=1}^p |\beta_i| \right] \quad (1)$$

ただし、 $\|\cdot\|_2$ は L2 ノルムを表し、 $y = (y_1, \dots, y_n)^T$ は被説明変数ベクトル、 $X = (x_1, \dots, x_p)$ は説明変数、 $\beta = (\beta_1, \dots, \beta_p)^T$ は回帰係数ベクトル、 C は隣接したパラメータの組を表す集合である。また、 λ, γ は L1 正則化の強さを規定するハイパーパラメータである。 λ の値が大きくなるにつれて、より多くの隣接するパラメータが共通の値に推定される。また、 γ の値が大きくなるにつれて、より多くのパラメータが 0 に推定される。 λ, γ は先験的に与える必要があるが、通常、分析前に正則化の強さの程度に関する情報は入手できない。そのため、交差検証法や AIC などの情報量規準を用いて λ, γ の値を選択することが多い。

2.2. Generalized fused lasso に基づく集積領域検出

Choi *et al.* (2018) は、小地域ごとに集計されたカ

ウンデータを用いた集積領域検出問題を、共変量を含むポアソン回帰モデルに Generalized fused lasso の正則化項を導入し、定式化した。

まず、地域 i で記録された点事象データの件数 y_i を、式(2)のポアソン回帰モデルによって表す。

$$\begin{aligned} y_i &\sim \text{Poisson}(\mu_i) \\ \log \mu_i &= \log e_i + \alpha_i + \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned} \quad (2)$$

ただし、 e_i は地域 i のオフセット（面積など）、 $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ は共変量ベクトル、 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ は対象領域全域で共通の共変量パラメータである。 α_i は地域 i の集積性の程度を示す集積性パラメータであり、 α_i が 0 ならば地域 i は非集積領域、 α_i が 0 より大きい値ならば地域 i は集積領域に属していることを示す。このとき、ポアソン尤度関数 $L(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta})$ およびポアソン対数尤度関数 $l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta})$ は、式(3)・(4)と書ける。

$$L(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) = \mu_i^{y_i} \frac{1}{y_i!} \exp(-\mu_i) \quad (3)$$

$$l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) = y_i \log \mu_i - \mu_i - \log y_i! \quad (4)$$

ポアソン対数尤度関数に L1 正則化項を導入し、Choi *et al.* (2018) による集積領域検出問題は式(5)で表される。

$$\begin{aligned} \min_{\alpha, \boldsymbol{\beta}} \left[\sum_{i=1}^n l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) + \lambda \sum_{(i,j) \in C} |\alpha_i - \alpha_j| + \gamma \lambda \sum_{i=1}^n |\alpha_i| \right. \\ \left. + \delta \sum_{i=1}^p |\beta_i| \right] \quad (5) \end{aligned}$$

ここで、 λ, γ, δ は正則化パラメータである。集積性パラメータの値が 0 に推定されることと、隣接地域の集積性パラメータの値が共通に推定されることの 2 点を誘導する正則化項を導入することにより、空間的自己相関を考慮した集積領域検出を実現している。

Choi *et al.* (2018) は、集積領域検出問題の定式化とあわせて、MM (Majorize-Minimization) アルゴリズムを適用したパラメータ推定方法も提案している。MM アルゴリズムは、目的関数を近似した代理関数を用いた繰り返し計算により、パラメータを推定する最適化手法である。MM アルゴリズムでは、目的関数を直接的に最適化できない問題からパラメータ

を推定することができる一方で、パラメータの不確実性など、点推定値以外の情報を提供することはできない。

3. Conditional Autoregressive 事前分布

Conditional Autoregressive 事前分布 (Besag *et al.*, 1991) は、空間的自己相関を表現した事前分布である。Conditional Autoregressive 事前分布には様々な拡張が提案されているが、本節では、Conditional Autoregressive 事前分布の最も単純なクラスである Intrinsic Conditional Autoregressive (ICAR) 事前分布を紹介する。

対象領域が n 地域に分割されており、地域ごとにパラメータ $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$ が設定されているとする。パラメータが空間的自己相関を有する場合、隣接地域のパラメータは互いに類似した値をとる。このとき、式(6)で表される ICAR 事前分布を地域ごとのパラメータの事前分布として用いることで、隣接地域の推定値を取り入れた事後分布の推定を誘導することができる。

$$\phi_i | \boldsymbol{\phi}_{-i}, \tau \sim N \left(\frac{\sum_{j=1}^n w_{ij} \phi_j}{\sum_{j=1}^n w_{ij}}, \frac{1}{\tau \sum_{j=1}^n w_{ij}} \right) \quad (6)$$

ここで、 $\boldsymbol{\phi}_{-i} = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)$ であり、 τ は空間的自己相関の強さを制御するパラメータである。また、 C を隣接したパラメータの組を表す集合としたとき、 w_{ij} は $(i, j) \in C$ ならば1、 $(i, j) \notin C$ ならば0を返す指示変数である。

ICAR 事前分布は、式(7)で示す Pairwise Difference 形式の等価な同時分布に変換される。式(7)から、ICAR 事前分布は隣接パラメータの差に対する L2 正則化に相当すると解釈できる。

$$p(\boldsymbol{\phi} | \tau) \propto \exp \left\{ -\frac{\tau}{2} \sum_{(i,j) \in C} (\phi_i - \phi_j)^2 \right\} \quad (7)$$

Besag *et al.* (1991)は、Pairwise Difference 形式の CAR 事前分布を基に、隣接パラメータの差に対する L1 正則化に相当する式(8)の分布も提案している。

$$p(\boldsymbol{\phi} | \tau) \propto \exp \left\{ -\frac{\tau}{2} \sum_{(i,j) \in C} |\phi_i - \phi_j| \right\} \quad (8)$$

ICAR 事前分布を組み込んだ階層ベイズモデルは、解析的に事後分布を導出することが困難である。そこで、事後分布の推定には、事後分布から乱数列を生成することができるマルコフ連鎖モンテカルロ (MCMC) 法が用いられる。

4. ベイズ推定を応用した Generalized fused lasso による集積領域検出手法の提案

4.1. 事前分布・尤度関数の設定

本研究では、Conditional Autoregressive 事前分布を応用し、Choi *et al.* (2018) の Generalized fused lasso に基づく集積領域検出手法をベイズ推定の枠組みに拡張した手法を提案する。

ここで、 n 地域で構成される対象領域を考える。地域 $i (i = 1, \dots, n)$ で記録された点事象データの件数 y_i を、オフセット e_i 、集積性パラメータ α_i 、共変量ベクトル $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ 、対象領域全域で共通の共変量パラメータ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T = (\beta_0, \tilde{\boldsymbol{\beta}}^T)^T$ を用いたポアソン回帰モデルで表現する。このポアソン回帰モデルは式(9)で、ポアソン尤度関数・対数尤度関数は式(10)・式(11)で書ける。

$$\log E(y_i) = \log \mu_i = \log e_i + \alpha_i + \mathbf{x}_i^T \boldsymbol{\beta} \quad (9)$$

$$L(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) = \mu_i^{y_i} \frac{1}{y_i!} \exp(-\mu_i) \quad (10)$$

$$l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) = y_i \log \mu_i - \mu_i - \log y_i! \quad (11)$$

$\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}}$ の事前分布には、式(12)の同時分布を設定する。
 $\pi(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}} | \lambda, \gamma, \delta) =$

$$\exp \left\{ -\lambda \sum_{(i,j) \in C} |\alpha_i - \alpha_j| - \gamma \lambda \sum_{i=1}^n |\alpha_i| - \delta \sum_{i=1}^p |\beta_i| \right\} \quad (12)$$

ただし、 C は隣接したパラメータの組を表す集合、 λ, γ, δ は正則化パラメータである。なお、本研究では互いに境界を共有する 2 地域を隣接地域と定義している。また、定数項パラメータ β_0 の事前分布 $\pi(\beta_0)$ には無情報事前分布を設定する。

ここで、ベイズの定理による事後分布の導出を考える。事前分布 $\pi(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}} | \lambda, \gamma, \delta) \cdot \pi(\beta_0)$ と、地域ごとのポアソン尤度関数 $L(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta})$ ・ポアソン対数尤度関数 $l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta})$ を用いると、事後分布の同時分布に関して式(13)が成り立つ。

$$\begin{aligned} & \pi(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{x}_i, y_i, \lambda, \gamma, \delta) \\ & \propto \pi(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}} | \lambda, \gamma, \delta) \pi(\beta_0) \prod_{i=1}^n L(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) \\ & \propto \pi(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}} | \lambda, \gamma, \delta) \prod_{i=1}^n \exp\{l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta})\} \\ & \propto \exp \left\{ -\lambda \sum_{(i,j) \in C} |\alpha_i - \alpha_j| - \gamma \lambda \sum_{i=1}^n |\alpha_i| - \delta \sum_{i=1}^p |\beta_i| \right. \\ & \quad \left. + \sum_{i=1}^n l(\mathbf{x}_i, y_i | \alpha_i, \boldsymbol{\beta}) \right\} \quad (13) \end{aligned}$$

式(13)を対数変換すると、Choi *et al.* (2018) により定式化された式(5)の正則化項付き尤度関数に一致する。そのため、式(10)(12)の尤度関数・事前分布から得られるパラメータ事後分布の最頻値は、Choi *et al.* (2018) の集積領域検出問題の推定値に等しい。

4.2. WAIC によるハイパーパラメータの選択

式(12)の事前分布には、Generalized fused lasso の正則化パラメータに対応するハイパーパラメータ λ, γ, δ が含まれる。本研究では、それぞれのハイパーパラメータに複数の候補値を設定し、情報量規準 WAIC (Watanabe, 2010) が最小となるハイパーパラメータの組み合わせを探索する。WAIC は、パラメータの点推定値ではなく、事後分布全体を用いて算出される情報量規準である。そのため、ベイズ推定で得られるパラメータの不確実性を考慮したモデル選択を行うことができる。

5. シミュレーションデータによる提案手法の評価

5.1. シミュレーションデータの生成

シミュレーションデータを用いた分析を通して、提案手法の有用性を評価する。小地域ごとに集計されるカウントデータがポアソン分布に従うと仮定し、対象領域内の他の地域と比較して、平均点数が高い地域の集合を集積領域とする。本分析では、対象領域が 17×17 のメッシュに分割されており、中央に 5×5 の集積領域が存在する状況を設定する。共変量は導入しない。非集積領域の平均点数は 10 と 20 の 2 ケース、点密度比 (= 集積領域平均点数 / 非集積

領域平均点数) も 1.5 と 3.0 の 2 ケースを検討する.

5.2. 検出性能の評価

検出性能の評価項目として, 式(14)~(16)で表される検出力・誤検出率・偽陽性率を算出する.

$$\text{検出力} = \frac{\text{検出された設定集積領域数}}{\text{設定集積領域数}} \quad (14)$$

$$\text{誤検出率} = \frac{\text{検出された非集積設定領域数}}{\text{検出された領域数}} \quad (15)$$

$$\text{偽陽性率} = \frac{\text{検出された設定非集積領域数}}{\text{設定非集積領域数}} \quad (16)$$

ただし本研究では, 集積性パラメータの 95%ベイズ信用区間が 0 を含む地域を非集積領域, それ以外の地域を集積領域と判定した.

シミュレーションデータ生成・集積領域検出を 20 回繰り返した結果の平均を用い, 性能を評価する.

5.3. ベイズ推定の設定

本研究では, ハミルトニアンモンテカルロ法を用いた MCMC サンプリングによりパラメータを推定した. サンプリングには確率的プログラミング言語 Stan を利用し, 1 回の推定あたり 4 本の乱数列 (チェーン) を独立に発生させた. また, 各チェーンにおいて, パラメータの事後分布を 10,000 回サンプリングし, そのうち最初の 2,000 回を burn-in period として破棄した. サンプリング終了後, Gelman-Rubin 統計量 $\hat{R}$ (Gelman and Rubin, 1992) がすべての推定パラメータで 1.1 以下となっており, サンプリングが定常状態に収束していることを確認した.

なお, 正則化の強さを制御するハイパーパラメータは λ, γ とともに $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50\}$ を候補とし, すべての組み合わせを試行した.

5.4. 集積領域検出性能の評価

WAIC の値が最小となったハイパーパラメータ λ, γ の組み合わせにおける集積領域検出性能を表-1 に示す. また, 地域ごとの集積性パラメータ事後分布の平均値と標準偏差の例を図-1 に示す.

表-1 より, すべてのケースにおいて, 誤検出率は 0.1 以下・偽陽性率は 0.01 以下と低く抑えつつ, 一定の検出力を達成していることが確認できる. また図-1 より, 非集積領域と判定された地域の中でも, 集積領域に隣接する地域や集積性パラメータ事後分

布平均値が 0 から離れた値に推定された地域では, 集積性パラメータ事後分布の標準偏差が相対的に大きく推定される様子が確認できる. ここから, 推定された集積性パラメータの信頼性が事後分布の標準偏差に反映されていることが示唆される.

6. おわりに

本研究では, 空間的自己相関を考慮した事前分布を応用し, スパースモデリングの手法である Generalized fused lasso に基づく集積領域検出手法をベイズ推定の枠組みに拡張した手法を提案した.

シミュレーションデータを用いた有用性の検討の結果, 集積性パラメータの不確実性を活用することで, 誤検出や偽陽性の発生を抑制した集積領域の検出が可能であることが示された. また, 集積領域の検出に加えて, 地域ごとの集積性パラメータの不確実性に関する情報も推定できることが確認された.

謝辞

本研究は JSPS 科研費 18H01552 の助成を受けた.

参考文献

- Besag, J., York, J., and Mollié, A. (1991) A Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–59.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.
- Choi, H., Song, E., Hwang, S. S., and Lee, W. (2018) A modified generalized lasso algorithm to detect local spatial clusters for count data. *AStA Advances in Statistical Analysis*, **102**(4), 537–563.
- Fan, J., and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **95**(456), 1348–1360.
- Gelman, A., and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472.

- Kulldorff, M., and Nagarwall, N. (1995) Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14**, 799–810.
- Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5**(2), 369–412.
- Lee, D. (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-Temporal Epidemiology*, **2**(2), 79–89.
- Park, T. and Casella, G. (2008) The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.
- Shimamura, K., Ueki, M., Kawano, S., and Konishi, S. (2019) Bayesian generalized fused lasso modeling via NEG distribution. *Communications in Statistics - Theory and Methods*, **48**(16), 4132–4153.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**(1), 267–288.
- Tibshirani, R., and Taylor, J. (2011) The solution path of the generalized lasso. *Annals of Statistics*, **39**(3), 1335–1371.
- Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

表-1 集積領域検出性能

非集積領域平均点数	点密度比	検出力	誤検出率	偽陽性率
10	1.5	0.154	0.021	0.001
	3.0	0.996	0.055	0.006
20	1.5	0.588	0.046	0.003
	3.0	1.000	0.012	0.001

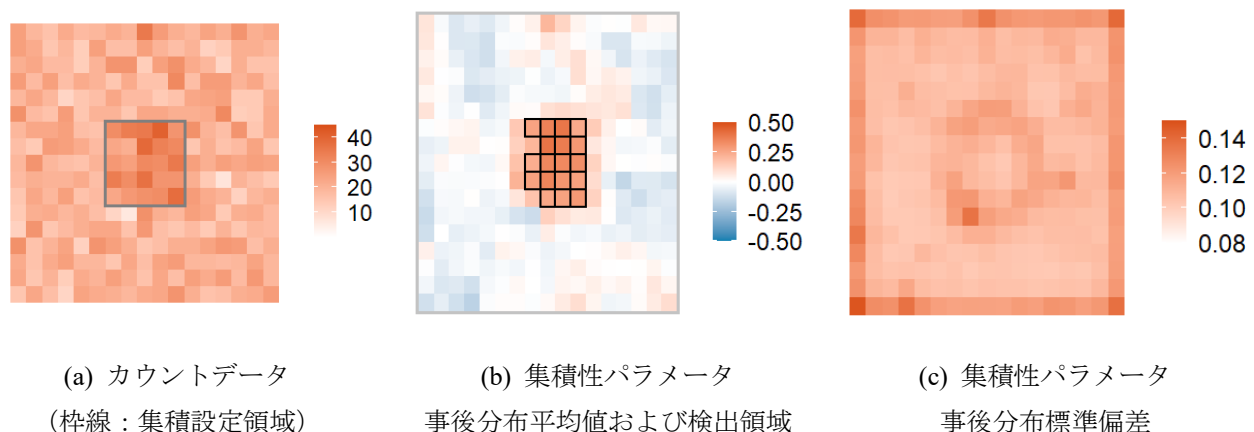


図-1 集積性パラメータ推定結果（非集積領域平均点数 20，点密度比 1.5）