統計数理 Vol. 73, No. 1

（通巻 141 号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「空間統計モデリング：理論と応用」

2025 年 6 月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）

本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 2760-2125

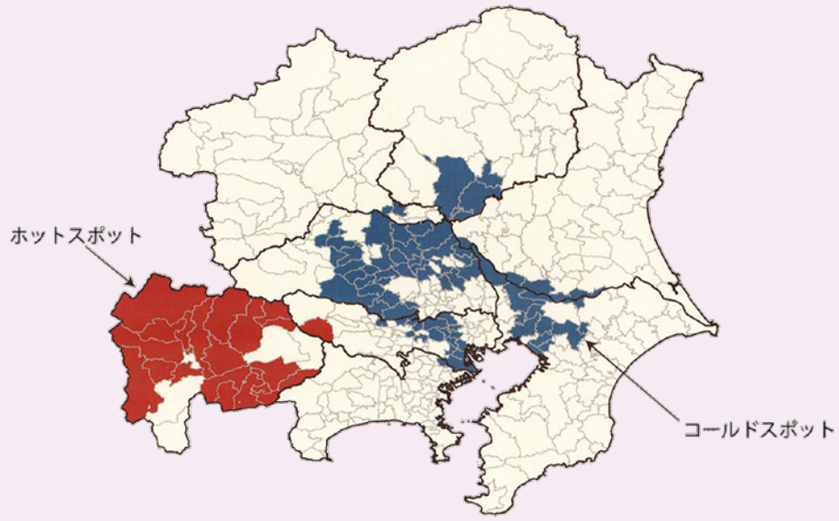
統計数理

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第73巻 第1号 2025

統計数理 Vol. 73, No. 1

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

（年 2 回発行）

編集委員長　小山　慎介
編集委員　庄　建倉
野間　久史
林　慶浩
Figueira Lourenço, Bruno
村上　隆夫
特集担当編集委員　村上　大輔
瀬谷　創（神戸大学）

編集室
池田　広樹　　川合　純華　　長嶋　昭子

「統計数理」は，統計数理研究所の研究成果を掲載する機関誌「彙報」として 1953 年に創刊され，1985 年に誌名を現在の「統計数理」に改めました．2025 年からはオンラインジャーナルとして新たな形での発行を開始しています．本ジャーナルは，統計科学全般に関する論文を広く受け付け，統計科学の深化と発展，さらに統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a. 原著論文　　b. 総合報告　　c. 研究ノート
d. 研究詳解　　e. 統計ソフトウェア　　f. 研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/

© The Institute of Statistical Mathematics 2025

印刷：笹氣出版印刷株式会社

表紙の図は本誌 95 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 73, No. 1

Contents

June, 2025
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2025)
第 73 巻 第 1 号 1–2
©2025 統計数理研究所

「特集 空間統計モデリング：理論と応用」について

村上 大輔[1]・瀬谷 創[2]（オーガナイザー）

地理情報システム(Geographic information system; GIS)や各種測位システムが発展し，大規模データを扱う基盤も日進月歩する中，自然科学から人文・社会科学に至るまで，社会・経済・環境・生物・疾病・交通・気象といった幅広い情報が，位置情報と紐づけられた(時)空間データとして高頻度かつ高解像度で観測され，日々利活用されている．

一般に，空間データには空間従属性(spatial dependence)や空間異質性(spatial heterogeneity)という性質が見られることが知られている．空間従属性とは，データが近隣のデータとより強い従属関係を持つという性質であり，空間異質性とは，場所毎にデータの特性が異なるという性質である．Tobler (1970)は前者を地理学第一定理(first law of geography)「Everything is related to everything else, but near things are more related than distant things」として位置づけ，後者はしばしば地理学第二定理として位置付けられるなど，古くから空間データの統計モデリングを行う上での重要な性質として認識されてきた．

空間データを対象とした統計モデル研究の起源は，概ね 1950 年代まで遡れるようである．Moran (1950)は空間従属性の代表的な診断統計量であるモラン I 統計量を提案した．また，Krige (1951)が鉱物資源の分布を評価する目的で開発し，フランスの数学者の Georges Matheron 氏により拡張・体型的に整理された空間予測手法は，今日ではクリギング(kriging)と呼ばれ広く用いられている．その後，1980 から 90 年代にかけていくつかの標準的なテキストが刊行され，空間データを対象とした統計モデルの研究はその裾野を広げてきた．例えば，自然科学を主な適用対象に発展してきた地球統計学(geostatistics) (Cressie, 1991)，社会科学における空間計量経済学(spatial econometrics) (Anselin, 1988)，医学における空間疫学(spatial epidemiology)，生態学における空間生態学(spatial ecology)等である．今日では，それらの分野の垣根はどんどん曖昧になっており，空間データを取り扱う統計分野は総称して空間統計学と呼ばれることが多くなっている．

空間統計学では，モデルの代表的なパラメータにナゲット，レンジ，シルがあり，共分散の要素を与える関数はコバリオグラムと呼ばれる．かつては，これら特有の概念が参入障壁となっていた部分があるが，近年ではクリギングがガウス過程(Gaussian process)の空間予測に特化した(2–4 次元の)応用例と位置づけられ，手法の見通しも格段によくなってきている．現在，空間従属性や空間異質性を柔軟にモデル化することのできる統計モデルの開発や理論的な整理が急速に進んでいる．他方，空間統計モデリングのためのソフトウェアの整備も着実に行われている．空間統計学分野では，研究で実装した手法を R のコードとして公開することが一般化しており，多数の有用なパッケージが存在する．また近年では機械学習分野との融合が進んでいることから，Python のコードとして公開される例も増加している．以上より，空間統計学は，学問分野としてある程度成熟してきたといえる．

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 神戸大学 大学院工学研究科：〒 657–8501 兵庫県神戸市灘区六甲台町 1–1

その一方で，空間データを取り巻く状況は近年目まぐるしく変化している．具体的には，観測技術や Web 情報などの自動収集技術の進展に伴い，空間データの大規模化，超高頻度化，多様化が進んでいる．多様化については，低質なデータ(画像やテキストなどの未成形の情報からなるデータ等)の普及も顕著である．また，社会科学分野では因果推論が重要視され，空間データモデリングも，因果推論への貢献を求められつつある．このような中，データの大規模化・超高頻度化・多様化に対応するような新たなモデルがベイズ推論や機械学習を駆使しながら盛んに開発され，空間データの因果推論に関する研究も進みつつある．

以上述べたような最新の動向を踏まえ，本特集号「空間統計モデリング：理論と応用」では，空間統計モデリングに関連する 7 編の論文を掲載する．最初の 2 編は，地球統計学と空間計量経済学の研究動向についてのレビューであり，続く 2 編は近年の興味深い手法についてのよりフォーカスした紹介となっている．残り 3 本はそれぞれ賃貸物件市場，COVID-19，観光客の動向に着目した手法開発と応用に関する論文となっている．以上のように，空間統計学に関する文献レビューから理論・応用までをカバーする幅広い内容となっている．空間統計学に関する和文文献は，特に最近の研究動向に関するものは少ないと思われる．本特集号が空間統計学の学習や情報収集，今後の研究や実務の参考として役立てば幸いである．

最後に，本特集「空間統計モデリング：理論と応用」の執筆ならびに査読にご協力いただいた皆様にこの場を借りて感謝を申し上げます．

参 考 文 献

Anselin, L. (1988). *Spatial Econometrics*: *Methods and Models*, Kluwer Academic Publishers, Dordrecht.

Cressie, N. (1991). *Statistics for Spatial Data*, Wiley, New York.

Krige, Danie G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand, *The Journal of the Chemical*, *Metallurgical and Mining Society of South Africa*, **52**(6), 119–139.

Moran, P. A. (1950). Notes on continuous stochastic phenomena, *Biometrika*, **37**(1/2), 17–23.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region, *Economic Geography*, **46**, 234–240.

統計数理(2025)
第 73 巻 第 1 号 3–17
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
［原著論文］

空間統計学の研究動向と今後の展望

村上 大輔†

（受付 2024 年 6 月 30 日；改訂 9 月 8 日；採択 9 月 9 日）

要　　旨

(時)空間データを対象とした統計解析手法の研究が空間統計学で進められてきた．とりわけ空間データの大規模化や多様化の進む近年では，計算効率と柔軟性を両立するような解析手法が発展を遂げてきている．本研究の目的は，以上のような空間統計学の近年の研究動向を整理することである．そのために，まずは基礎的な空間統計モデルとその課題を紹介する．次に，同モデルを大規模データに応用するための近似手法を低ランク近似，共分散行列の近似，精度行列の近似に分けて，それぞれについて既往研究を整理する．その後，空間統計モデルの柔軟性を高めるための拡張手法を，共分散カーネルに基づく方法と，そうでない方法に分け，ニューラルネットワークの応用などの直近の動向も踏まえて整理する．その後，以上の手法の時空間データへの応用研究について紹介する．最後に，当該分野の手法を実装するためのソフトウェアについて解説したうえで，今後の空間統計学の課題について議論する．

キーワード：空間統計学，時空間モデリング，ガウス過程，空間相関，ニューラルネットワーク．

1. はじめに

地上・衛星からの観測技術や位置測位技術の発展に伴い，位置・時間情報を伴う時空間データが幅広く収集されている．時空間データの収集や公開は国などの公的機関だけでなく，研究グループなどの有志によっても活発に進められるようになってきている．オープンデータ化の進む今日，収集された空間データは無償公開されることも多く，例を挙げるならば全球の道路網や建物データを公開する OpenStreetMap（Vargas-Muñoz et al., 2020)や Microsoft Building Footprints（Huang et al., 2021)，空間詳細な推計人口統計を公開する WorldPop（Stevens et al., 2015)，幅広い衛星画像を直ちに分析可能な形で公開する GoogleEarthEngine（Gorelick et al., 2017)などがある．今後も利用可能な時空間データは増えていき，それらの統計解析はより一層重要となっていくと見込まれる．

時空間データを統計的にモデリングすることで予測，要因分析，不確実性評価などの様々な分析を行うことができる．特に，時空間データの得られる地点，集計単位，時点などは一般に限られているため，空間統計学(狭義には地球統計学)では時空間予測／補間に着目した研究が盛んに進められてきた（Cressie, 2015; Cressie and Wikle, 2015)．その中で，空間データの一般的性質とされる空間相関(近所と強く相関)や時系列相関(直前／直後と強く相関)を，その不確実性とともにモデル化する手法が確立されてきた．また近年では，時空間データの大規模化や多様化に対応するための手法の幅広い拡張が試みられてきた．しかしながら，空間統計学に関

† 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

するレビュー論文は，特に和文では，筆者の知る限り堤・瀬谷 (2012)のみであり，近年の研究動向について把握することが困難となっている．

そこで本研究では空間統計学の近年の研究動向を整理する．以降の章立ては次のとおりである．2 章では基礎的な空間統計モデルとその課題を紹介する．3，4 章では，同モデルの肝である空間過程のモデル化に関する研究動向を，計算効率化と柔軟化の 2 軸に分けてそれぞれレビューする．5 章では時空間過程に対するモデル化について整理する．6 章では空間統計モデルを様々なデータに応用するための手法についてレビューし，7 章では以上の手法を実装するためのソフトウェアについて整理する．8 章では残された課題を紹介する．

2. 空間統計モデル

2.1 基礎

地点・時点毎の観測データのベクトル $\boldsymbol{y} = [y(s_1, t_1), \ldots, y(s_N, t_N)]'$ を考える．$s_1, \ldots, s_N$ は対象地域 $D \subset \mathbb{R}^d$(典型的には $d = 2$)上の観測地点，$t_1, \ldots, t_N$ は時点を表す．具体例としては，観測所毎の気温や調査地点毎の地価などが挙げられる．空間統計学では時点数が 1 つ ($t_1 = t_2 = \cdots = t_N$)と仮定される場合も多く，本研究でも 2–4 章では時点は 1 つと仮定して明示的には扱わない．N はサンプルサイズであり 2–4 章では観測地点数に一致する(地点の重複がない場合)．複数時点を考慮する 5 章では，時点 t のサンプルサイズは N_t と，全時点についてのサンプルサイズ N は区別することとする．

基礎的な空間統計モデルは観測データが下式に従うと仮定する：

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I}), \tag{2.1}$$

$\boldsymbol{X}$ は説明変数行列，$\boldsymbol{\beta}$ は回帰係数ベクトル，$\boldsymbol{I}$ は単位行列である．空間統計学では行列 $\boldsymbol{K}_r$ の第 (i, j) 要素を直線距離 $d_{i,j}$ のカーネル $k_r(d_{i,j})$ で与えることで空間相関をモデル化する．例えば指数カーネル：$\exp(-\frac{d_{i,j}}{r})$，ガウスカーネル：$\exp(-(\frac{d_{i,j}}{r})^2)$，それらを包含する Matérn カーネル：$\frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\frac{d_{i,j}}{r})K_\nu(\sqrt{2\nu}\frac{d_{i,j}}{r})$ ($\Gamma(\cdot)$ はガンマ関数，$K_\nu(\cdot)$ は次数 ν の第 2 種の修正ベッセル関数)は標準的に用いられている．r は空間相関の距離減衰の速さを決めるパラメータであり，その値が大きいことは大域的な空間相関パターンが誤差項にみられることを意味する．ν は空間過程のなめらかさを決めるパラメータであるが，r との識別の問題から固定値で与えられる場合も多く，本研究でも固定値とする．τ^2 と σ^2 は誤差項に対する分散パラメータであり，τ^2 が大きいことは空間相関パターンが支配的であることを，σ^2 が大きいことは空間相関では説明されないノイズパターンが支配的であることを，それぞれ意味する．なお，(2.1)式は期待値と共分散の関数が移動不変という 2 次定常性の仮定に基づいている点に注意されたい．

多変量ガウス分布の性質から(2.1)式の対数尤度は以下となる：

$$\log L = -\frac{1}{2}[\log(|\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I}|) - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - N\log(2\pi)], \tag{2.2}$$

M は説明変数の数である．パラメータ $\{\boldsymbol{\beta}, \tau^2, \sigma^2, r\}$ は尤度最大化により容易に推定できる．回帰係数の最尤推定量は $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}$ となり，$\{\tau^2, \sigma^2, r\}$ は $\boldsymbol{\beta}$ を $\hat{\boldsymbol{\beta}}$ に置き換えた(2.2)式を最大化することで推定できる．

いま，未観測点 (s_0, t_0) のデータ(未知)を $y(s_0, t_0)$ とする．2 次定常性を仮定したため，(2.1)式と同様，その期待値は $\boldsymbol{x}_0'\boldsymbol{\beta}$，各観測点との共分散は $\boldsymbol{k}_{r,0} = [k_r(d_{1,0}), \ldots, k_r(d_{N,0})]'$ で与えられることとなる．ただし $\boldsymbol{x}_0$ は同地点の説明変数ベクトル，$d_{i,0}$ は地点 i と地点 0 の間の直線距離である．以上の仮定を用いると $y(s_0, t_0)$ の最良線形不偏予測量は以下となる：

$$\hat{y}(s_0, t_0) = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + \tau^2\boldsymbol{k}_{r,0}'(\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{2.3}$$

空間統計学では同式を用いて(時)空間予測／補間が行わる．また，同予測量の不確実性を表す期待二乗誤差 $E[(y(s_0,t_0)-\hat{y}(s_0,t_0))^2]$ も解析的に得られる．任意地点における予測値をその不確実性とともに評価できる点は，空間統計モデルの利点の一つである．

なお，空間統計モデルは機械学習分野におけるガウス過程（Williams and Rasmussen, 2006）と同一である．従って，以降のレビューは空間統計分野におけるガウス過程の研究動向を整理したものとみなすこともできる．

2.2 課題

(2.1)-(2.3)式からは空間統計モデルの課題がみえてくる．まず $(\tau^2\boldsymbol{K}_r+\sigma^2\boldsymbol{I})^{-1}=\boldsymbol{Q}$ の計算量は $O(N^3)$ のオーダーで指数的に増大するため大規模な空間データの解析には向かない．ビッグデータ時代である昨今においてこの課題は深刻である．そのため空間統計学では大規模データを扱うための計算コストの削減が近年のホットトピックとなっている．なお，$\boldsymbol{Q}$ は精度行列と呼ばれる．(2.2)-(2.3)式に $\boldsymbol{Q}$ を代入すると $\boldsymbol{K}_r$ を含まない形で表せるため，その計算効率化のためには，$\boldsymbol{K}_r$ に対する近似以外に，$\boldsymbol{Q}$ を直接近似する方法も研究されている(3 章参照)．

また，ガウス分布を仮定する(2.1)式では扱える問題が限られるという課題もある．この点に対処するために空間統計モデルは幅広く拡張されてきた．拡張されたモデルの多くは下式のように階層的に表せる（Cressie and Wikle, 2015 参照）：

$$\boldsymbol{y}|\boldsymbol{z}\sim f_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{z}),\quad \boldsymbol{z}\sim g_{\boldsymbol{\theta}}(\boldsymbol{K}_r) \tag{2.4}$$

ここでパラメータは $\boldsymbol{\theta}$，$\boldsymbol{z}$ は(時)空間過程である．$f_{\boldsymbol{\theta}}(\cdot)$ は観測データの分布，$g_{\boldsymbol{\theta}}(\cdot)$ は(時)空間過程の分布を表す．例えば $f_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{z})=N(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{z},\sigma^2\boldsymbol{I})$，$g(\boldsymbol{K})=N(0,\tau^2\boldsymbol{K}_r)$ とすると(2.1)式となる．

一般に空間統計学では，(a)大規模データを扱うために(時)空間過程 $\boldsymbol{z}\sim g_{\boldsymbol{\theta}}(\boldsymbol{K}_r)$ が近似され，(b)複雑な時空間過程を扱うために $g_{\boldsymbol{\theta}}(\boldsymbol{K}_r)$ が拡張され，(c)幅広いデータを扱うために $f_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{z})$ が拡張されてきた．以降では，(時)空間過程に対するモデリング(a)-(b)については 3–5 章で，観測データに対するモデリング(c)について 6 章で，それぞれ研究動向を整理する．

3. 大規模データに対する空間過程の近似

空間統計モデルの計算効率を高めるための代表的な近似に，(i)有限個の基底関数で空間過程を近似しようという低ランク近似，(ii)密行列である $\boldsymbol{K}_r$ を疎な共分散行列に置き換える近似，(iii)$\boldsymbol{Q}$ を疎な精度行列に置き換える近似がある．本章では(i)-(iii)に関する手法を紹介する．

3.1 低ランク近似

下式のように L 個($L<N$)の基底関数の線形和で空間過程を近似する方法である：

$$\boldsymbol{z}=\boldsymbol{\Phi}\boldsymbol{a},\quad \boldsymbol{a}\sim N(0,\boldsymbol{V}_{\boldsymbol{a}}) \tag{3.1}$$

$\boldsymbol{\Phi}$ は L 個の基底関数を並べた行列($N\times L$)である．各基底関数を 2 階微分可能な位置座標の関数で与え，そのランダム係数ベクトル $\boldsymbol{a}$ を観測データから推定することで空間過程 $\boldsymbol{z}$ を近似する．基底関数には radial basis function（RBF; Cressie and Johannesson, 2008; Nychka et al., 2015），有限要素法に基づく基底（Lindgren et al., 2011），multi-resolution splines（Tzeng and Huang, 2018），random Fourier features（Miller and Reich, 2022）を含む所与の関数が用いられることが多い一方で，未知の関数とみなして期待二乗誤差の最小化に基づいて与えることもできる（例：Banerjee et al., 2008）．また気象分野では複数ケースで生成した大気拡散等のシミュ

レーションの結果を基底として用いることもある（Zammit-Mangion et al., 2022a 参照）．

低ランク近似の計算効率は複数の研究で確認されている（例：Heaton et al., 2019）．その一方で，過度な平滑化を招きやすい点や（Datta et al, 2016）ノイズ分散が小さい場合に精度が低下するなど（Stein, 2014），近似精度には課題が残されている．近似精度を高めるために，空間解像度毎の基底関数を用いて空間過程を近似する Nychka et al.（2015）や Katzfuss and Hammerling（2017）が提案されたほか，後述の疎な共分散関数と組み合わせることで表現能力を高めようという full scale approximation（Sang and Huang, 2012），multiresolution approximation（Katzfuss, 2017），modified linear projection（Hirano, 2017）なども提案されてきた．

低ランク近似は，基底を投入するだけで空間相関が考慮できるという簡便さなどの理由から，時空間モデリングや機械学習などでも広く用いられている（4–5 章参照）．

3.2 疎な共分散行列を用いる近似

$\boldsymbol{K}_r$ を疎な行列で置き換えようという近似である．$\boldsymbol{K}_r$ が疎のとき $\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I}$ も疎になるため，コレスキー分解を用いた効率的な逆行列計算が使用可能となる．

具体的な方法は，共分散行列の半正定値性を満たす形で提案されてきた．対象地域をボロノイ領域やグリッド毎に分割する方法はその一つである．分割により得られた空間領域間に独立性を仮定する近似（Stein et al., 2004; Hazra et al., 2024）や，隣接以外の領域との条件付き独立性を仮定する近似（Eidsvik et al., 2014）がある．いずれも共分散行列 $\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I}$ はブロック対角行列またはそれに類する疎行列となるため，その逆行列は効率よく評価できる．また，空間領域毎にモデルを推定し，それらの予測分布を合成することで全域の予測分布を得ようという nested kriging（Rullière et al., 2018）や，各領域から得られた予測分布の加重和を最適化しようという cluster kriging（Van Stein et al., 2020）など，関連したアンサンブル学習手法もある．

別の方法に，共分散カーネルに taper 関数と呼ばれる距離減衰の速い関数をかけることで共分散を疎にする covariance tapering がある（Furrer et al., 2006）．同手法には共分散カーネルにのみ taper 関数をかける one-taper 法と，データの共分散行列にも taper 関数を書ける two-taper 法があり，前者は，特に taper 関数の減衰が早すぎる場合にパラメータにバイアスがかかることが知られている（Furrer et al., 2016）．一方で，後者は不偏となるものの計算コストが増大するため一長一短である．関連手法に，ある地点の空間構造をモデル化するために，近隣の観測地点を動的に選択していくことで疎な共分散構造を実現しようという local approximate Gaussian process（laGP; Gramacy, 2016）がある．同手法は全域に対する半正定値な共分散構造を与えるものではないという批判はあるが，地点毎にパラメータが推定でき，非定常性を捉える実用手法としては便利である．

3.3 疎な精度行列を用いる近似

精度行列 $\boldsymbol{Q} = (\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}$ を直接疎にする方法である．(i) Stochastic partial differential equation（SPDE）を用いる方法と，(ii) Vecchia 近似を用いる方法が注目を浴びている．

(i) に関し，Lindgren et al.（2011）は，SPDE がグラフ上の確率場である Gauss-Markov random fields（GMRF）で近似できること，ならびに特定の SPDE の解がガウス過程（Matern カーネル）に一致することを利用して，GMRF とガウス過程が一対一対応することを示した．SPDE 法（Krainski et al., 2018 参照）は，この対応関係を利用して空間グラフ上の GMRF の線形和で空間過程を近似する方法であり，ガウス過程の代わりに GMRF を推定すればよい．GMRF の精度行列 $\boldsymbol{Q}_{GMRF}$ は空間グラフの隣接構造から自動的に決まるため，明示的な逆行列計算は不要である．結果として，GMRF で近似されるガウス過程の精度行列もまた自動的に得られ，計算コストは大幅に削減される．

この方法は単に計算効率が良くなるだけでなく，SPDE の特性を活かすことで様々な非定常・非ガウス空間過程への拡張が可能である（Bakka et al., 2018; Lindgren et al., 2022; 本論文 4 章参照）．

(ii)では観測点 $\boldsymbol{s}_{1:\,N} = [s_1, \ldots, s_N]$ を順位づけたうえで空間過程 $\boldsymbol{z} \sim N(0, \boldsymbol{K}_r)$ の分布が $p(\boldsymbol{z}) = p(z(s_1)) \prod_{i=2}^{N} p(z(s_i)|\boldsymbol{s}_{1:\,(i-1)})$ のように表せることを利用する．$p(z(s_i)|\boldsymbol{s}_{1:\,(i-1)})$ は地点群 $\boldsymbol{s}_{1:\,(i-1)} = [s_1, \ldots, s_{i-1}]$ における空間過程 $z(s_1), \ldots, z(s_{i-1})$ が与えられた下での空間過程 $z(s_i)$ の条件付き分布を表す．General Vecchia approximation（Katzfuss and Guinness, 2021）では，上位の地点群 $\boldsymbol{s}_{1:\,(i-1)}$ をより少数のサブサンプル $\boldsymbol{s}_{g(i)}$ で置き換えた下式で空間過程を近似する：

$$p(\tilde{\boldsymbol{z}}) = p(z(s_1)) \prod_{i=2}^{N} p(z(s_i)|\boldsymbol{s}_{g(i)}) \tag{3.2}$$

$\tilde{\boldsymbol{z}} \sim N(0, \tilde{\boldsymbol{K}}_r)$ の精度行列をコレスキー分解すると $\tilde{\boldsymbol{K}}_r^{-1} = \boldsymbol{U}\boldsymbol{U}'$ となり，$\boldsymbol{U}$ は地点 s_i と地点 $s_j \in \boldsymbol{s}_{g(i)}$ のペアについての要素のみが非ゼロの疎な上三角ブロック行列となる．この特性を活かすと，尤度と予測量が線形時間で評価できる．

既存の高速近似手法には(3.2)式の応用と見做せるものも多い（Katzfuss and Guinness, 2021）．例えば nearest neighbor Gaussian process（Datta et al., 2016）は $\boldsymbol{s}_{g(i)}$ を近隣地点のみで与え，modified predictive process（Finley et al., 2009）は $\boldsymbol{s}_{g(i)}$ を対象地域に均等に配置する地点群で与えるほか，3.1 節で紹介した full scale approximation（Sang and Huang, 2012）や multiresolution approximation（Katzfuss, 2017）も同近似の一種とみなせる．(3.2)式を用いた近似の精度は観測点の順位とサブサンプルの選び方 $\boldsymbol{s}_{g(i)}$ に依存するが，それらの決定方法については Guinness (2018; 2021)を参照されたい．Katzfuss and Guinness (2021)は精度行列のスパース性を保持しながら Kullback-Leibler ダイバージェンスの意味で精度を改良させた Sparse General Vecchia (SGV)を提案した．その精度の良さは比較研究でも確認している（Hazra et al., 2024）．

4. 複雑な空間過程を表現するための拡張

複雑な空間過程を表現するために共分散カーネルの拡張が行われてきた．それに加え，近年では共分散カーネルの代わりに neural network(NN)を用いて柔軟性を高めようという手法の研究も盛んである．そこで本章ではそれぞれについて簡単にレビューする．

4.1 共分散に基づく方法

基礎的な空間統計モデルの共分散カーネルは，どの方向も空間相関の及ぶ範囲は同じという等方性や，どの場所の空間相関も同じカーネルで記述されるという定常性の仮定に基づいているが，現実にはそれらが満たされない場合もある．例えば CO2 濃度の場合，地形や風向きの影響で空間相関の及ぶ範囲が方向毎に異なるかもしれないし，陸上と海上では空間相関パターンが異なるかもしれない．以上のような等方性や定常性を満たさない空間過程を扱うために共分散カーネルが拡張されてきた．

例えば，方向毎に空間相関パターンが異なるという異方性(anisotropy)が考慮されてきた．例えば Tsutsumi and Seya (2009)はつくばエクスプレス開業に伴う沿線の地価上昇に，同路線に沿った異方性がみられることを示した．さらに，Nychka et al. (2018)と Wiens et al. (2020)は局所尤度を用いて場所毎の局所異方性を計算効率良く推論する手法を提案した．同手法を用いることで，海岸線や山脈などによる空間相関パターンの歪みが表現できる．彼らは同手法を気候モデルのエミュレーションに応用した．

一方，SPDE 法(3 章参照)は非定常な空間過程のモデリングに広く用いられている（Krainski et al., 2018; Lindgren et al., 2022）．同手法は空間過程と SPDE を関連付けるものであり，SPDE のパラメータを場所毎に変えたり（Lindgren et al., 2011; Fuglstad et al., 2015)説明変数の関数で与えたりすることで（Ingebrigtsen et al., 2015），非定常な空間過程が表現できる．また，SPDE を多様体上で定義することで，例えば球面上の空間過程や，川や国境といった障害物を考慮した空間過程などが表現できるため（Barrier model; Bakka et al., 2019），全球を対象とした気象の分析や入り江などを対象とした生態の分析などで用いられてきた（例：Lezama-Ochoa et al., 2020; Fioravanti et al., 2023）．また，SPDE 法で扱うモデルには空間的な滑らかさを表すパラメータ ν があり，その値は Matern カーネルに基づく空間過程を近似する整数値で与えられることが多い．一方，ν に整数値以外の実数値も許容することで，より柔軟に空間過程をモデル化しようという fractional SPDE 法も提案されている（Bolin et al., 2024）．

4.2 共分散に基づかない方法

前節では共分散カーネルに基づく非定常モデルを紹介したが，現実には共分散では捉えきれない複雑なパターンがみられる場合もある．そのような場合に対処するために，近年では NN の応用が空間統計分野で活発化してきている．特に deep NN(DNN)は N の増加に対して高いスケーラビリティを有することや，空間過程が非連続となる特異性を有する場合に他手法に優ることが知られており（Imaizumi and Fukumizu, 2022），大規模で複雑な現象の解析に役立つ．

位置情報に基づく説明変数(特徴量)の投入は，NN で空間パターンを学習するための代表的な方法の一つである．これまでに用いられてきた説明変数には位置座標（Cracknell and Reading, 2014），経験バリオグラム（Gerber and Nychka, 2021），kriging 予測量（Wang et al., 2019），空間基底（Yoshida et al., 2022; Zammit-Mangion et al., 2022a; Chen et al., 2024)などがある．一方，この方法では空間パターンの非線形が間接的にしか学習されず，その柔軟性には課題もある（Zhan and Datta, 2024 参照）．

より柔軟性を高めるために，NN 内部の変換関数への空間情報の導入も試みられてきた．例えば，Zammit-Mangion et al.（2022b; 2024a)は変換関数のパラメータを空間可変とすることで，非定常性や異方性を有する複雑なパターンを捉える spatial Bayesian NN を提案した．また，Katzfuss and Schäfer（2023)は，空間相関情報を織り込んだ変換関数を用いたベイズ最適輸送により，複雑な空間パターンを計算効率よく学習する生成モデルを提案した．いずれも急激な変化を伴うような複雑なパターンが表現できる．

グラフを考慮する NN である graph NN（Wu et al., 2020 参照)の応用も盛んであり，隣接グラフで繋がれたゾーン/地点間の空間相関パターンが学習されてきた（例：Tonks et al., 2024; Cisneros et al., 2024）．また，Zhan and Datta（2024)は，疎な精度行列を用いる空間過程の近似手法である nearest neighbor Gaussian process と graph NN の理論的関係を明らかとしたうえで，両者を統合した空間過程モデル NN-GLS を提案した．

モデルの柔軟性を高める以外の NN の利点に，明記的に尤度を評価することなく尤度が近似できる点がある(likelihood-free)．この性質により，NN は，既存手法では計算コスト等の観点で解析困難であった，複雑な分布の解析を容易にした．例えば，max stable process は極端現象(例：異常気象)を扱う上で用いられるが，N の増加に伴う計算量の増大が速く大規模データには応用できなかった．そこで Sainsbury-Dale et al.（2024)は同確率過程を NN で近似することでこの課題に対処した．

以上に加え，時空間データに対する NN の応用も盛んである．この点については次章で紹介する．

5. 時空間過程に対するモデルの発展

時空間過程のモデリング方法は，(i)時空間上のカーネルを用いて全時点を同時にモデリングしようという静的な方法と，(ii)過去の状態が与えられた下で現在の状態を推論しようという動的な方法がある．

5.1 静的な方法

空間統計モデル((2.1)式)の $\boldsymbol{K}_r$ を，時差と距離に依存して減衰する時空間カーネルに基づいて与えることで時空間過程がモデリングする方法である（Cressie and Wikle, 2015）．Porcu et al.（2021）でレビューされたように，これまでに数多くの時空間カーネルが，半正定値性を満たすという条件の下で提案されてきた．それらは時間方向のカーネルと空間方向のカーネルに線形分離可能な separable カーネルとそうでない non-separable カーネルに分類できる．separable カーネルは時間・空間カーネルの和または積をとることで得られる．それらを用いることで計算コストは抑えられるが（Porcu et al., 2021），時間と空間の相互作用は無視されるため柔軟性には難がある．そのため，product-sum カーネル（De Cesare et al., 2001），Gneiting カーネル（Gneiting, 2002），scale mixture に基づくカーネル（Porcu and Zastavnyi, 2011）を含む，non-separable カーネルが数多く提案されてきた．

静的な方法は，全時点についての大きなカーネル行列 $\boldsymbol{K}_r$ を扱うため，逆行列計算 $(\tau^2\boldsymbol{K}_r + \sigma^2\boldsymbol{I})^{-1}$ の計算コストが増大する．そのため separable カーネルを用いるか，3 章で説明した近似手法を応用することで，計算コストが軽減されてきた（例えば Wood et al., 2017; Wikle et al., 2019）．静的な方法について，詳しくは Porcu et al.（2021）を参照されたい．

5.2 動的な方法

過去からの動的な変化を明示的に記述する方法である．例えば，離散時間 $t \in \{1, \ldots, T\}$ 毎に同一の複数地点でデータが観測されている場合，時点 t のデータベクトルは状態空間モデルを用いて次のように表現できる：$\boldsymbol{y}_t = \boldsymbol{z}_t + \boldsymbol{e}_t, \boldsymbol{e}_t \sim N(0, \sigma^2\boldsymbol{I})$, $\boldsymbol{z}_t = \rho\boldsymbol{W}_t\boldsymbol{z}_{t-1} + \boldsymbol{u}_t, \boldsymbol{u}_t \sim N(0, \tau^2\boldsymbol{K}_t)$．$\boldsymbol{z}_t$ は動的に変化する時空間過程を表し，パラメータ ρ が 0 の場合，$\boldsymbol{z}_t$ は時点毎に独立な空間過程となり，ρ が大きくなるにつれて $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ の時系列相関が強まる．また $\boldsymbol{W}_t$ は遷移行列であり，同行列を通して時空間過程の移流や拡散などの流れを柔軟に表現することができる（Integro-difference equation（IDE）モデル; Wikle et al., 2019 参照）．一般に，動的なモデルでは時点毎の共分散行列 $\boldsymbol{K}_t$ が逐次的に処理されるため，全時点についてのカーネルを用いる静的な方法に比べて計算効率が良い．また，現象の動きが明示的に表現できる点や，急激な変化をとらえやすい点，将来予測に応用しやすい点なども動的な方法の利点である（Cressie and Wikle, 2015 参照）．

一方で，時点毎の観測データが多い場合は動的な方法でも計算量が増大するため，データベクトルを $\boldsymbol{y}_t = \boldsymbol{\Phi}_t\boldsymbol{z}_t + \boldsymbol{e}_t$ のように与えて，低ランク近似した時空間過程 $\boldsymbol{\Phi}_t\boldsymbol{z}_t$ を用いる場合も多い（$\boldsymbol{z}_t(L \times 1)$：潜在変数ベクトル，$\boldsymbol{\Phi}_t(N_t \times L)$：基底関数行列．ただし N_t は時点 t のサンプルサイズ）．fixed rank filter（Cressie et al., 2010）や multiresolution filter（Jurek and Katzfuss, 2021）はその一例である．低ランク近似は自然に時空間モデルに組み込むことができ，計算効率が良く拡張もしやすいため，動的な時空間モデリングでも幅広く用いられている（Cressie et al., 2022 参照）．

状態空間モデルは，観測データ $\boldsymbol{y}_t$ と潜在変数 $\boldsymbol{z}_t$ に対する 2 層のモデルだが，これを多層化する試みもある．例えば Katzfuss et al.（2019）は変換，観測，時間発展，パラメータを表す 4 層の階層モデルを提案した．Wikle（2019）も関連する階層モデルを提案して DNN 等との

比較を行っている．より近年は，4.2 章でも述べた通り DNN などの機械学習手法の応用が盛んとなっている．特に時空間データに関しては，複雑な時間変化が学習可能な Recurrent NN（Yu et al., 2019 参照）や echo state network（Jaeger, 2001）などが，非定常・非ガウスの時空間過程をモデリングするために応用されてきた（Nag et al., 2023; McDermott and Wikle, 2017; 2019）．（D）NN は，一度学習すれば，以降は新規データを学習済みモデルに入力するだけで，極めて小さな計算コストで解析結果が得られるという計算上の利点がある（Amortized inference; Zammit-Mangion et al., 2024b）．高頻度に観測される時空間データが分野を問わず増えてきていることを踏まえると，（D）NN は時空間データに対してもより一層重要となると考えられる．空間統計分野における（D）NN の応用・拡張に関してより詳しくは Wikle and Zammit-Mangion（2023）を参照されたい．

6. 観測データに対するモデルの発展

5 章までは空間・時空間過程をモデル化する方法についてレビューした．本章では，観測データを柔軟にモデル化するための代表的な方法である generalized linear model（GLM）を応用する方法と，変換を行う方法について紹介する．

GLM とは指数分布族であるポアソン分布や二項分布などに従うデータを扱う回帰である．GLM は空間統計分野で古くから応用されており（Gotway and Stroup, 1997），今日では GLM の潜在変数として（時）空間過程を導入した階層ベイズモデルが Stan や INLA といった R パッケージに実装されている．例えば INLA パッケージでは 2024 年 6 月現在で約 100 もの確率分布がデータに仮定できるなど（Bakka et al., 2018），幅広いデータの空間統計モデリングが誰でも実装できるようになっている．以上のように GLM に基づく空間統計モデルはすでに確立されている一方で，計算効率を改善するための推定法は今なお改良されており，例えば上述の INLA パッケージで用いられている近似ベイズ推論法 integrated nested Laplace approximation（INLA 法）に関しては，効果識別のために導入していたノイズ項を除去したより精緻で計算効率の良い定式化（Van Niekerk et al., 2023），ラプラス近似部分を変分ベイズ推論で置き換えることによる近似精度改善（Van Niekerk and Rue, 2024）などが試みられてきた．INLA 法以外では，自動微分（automatic differentiation）を応用したモデル推定の高速化も試みられており，（時）空間モデリングに対する精度と計算効率が INLA 法を含む既存手法を上回ることなどを確認している（Anderson et al., 2022）．なお，以上の研究の多くは，空間相関だけでなく時系列相関，非線形効果，グループ別の効果なども扱うことのできるセミパラメトリックモデルの枠組みを扱っており，単に精度や計算効率が良いだけでなく，汎用性も高い統計手法が確立されてきている．

一方，変換を行う方法とは，観測データを対数変換や Box-Cox 変換などで変換することで非ガウスデータを扱おうという方法である．近年は NN による変換が盛んに試みられている．5 章でも述べた通り NN は likelihood-free であり，様々な分布に従うデータが近似できるため，これまで扱うことが困難だったような複雑な分布に従うデータも分析可能である．この特性を活かしたベイズ推論である neural Bayes 推定（Sainsbury-Dale et al., 2024）は今後幅広く利用される可能性がある．また，NN をはじめとした機械学習手法を用いれば複数ドメイン（例：地域）のデータが同時に学習でき，例えば転移学習やメタ学習といったこれまで空間統計分野であまり扱われてこなかったような問題設定にも応用できる可能性がある．

7. ソフトウェア

最後に以上で紹介した手法を実装するためのソフトウェア・パッケージを R を中心に紹介す

る．まず，3 章で紹介した大規模データモデリング手法に関しては，Wikle et al.（2019）でも一部解説されている通り，低ランク近似法は FRK，LatticeKrig，mgcv パッケージなどで，疎な共分散を用いる方法は spam，laGP パッケージなどで実装できる．また，SPDE を用いて精度行列を疎にする方法は INLA パッケージで，Vecchia 近似を用いる方法は GpGp や GPvecchia で実装できるほか，spNNGP，spBayes などでも本文中で紹介した関連手法が実装できる．各空間モデルの精度や計算時間の比較については Heaton et al.（2019）や Hazra et al.（2024）を参照されたい．

4 章で紹介した，共分散カーネルに基づいて柔軟な時空間過程を扱う方法に関しては，大域的な異方性は gstat で，局所異方性は LatticeKrig でそれぞれ実装できる．また，SPDE を用いて非定常パターンを捉える方法に関しては INLA パッケージで実装できる（Krainski et al., 2018 参照）．NN を用いる方法に関しては deepspat（https://github.com/andrewzm/deepspat）パッケージで実装できるほか，Python であれば Tensorflow，PyTorch，Pyro ライブラリなどでも幅広い NN の実装が可能である．また，GPyTorch では，PyTorch の自動微分や GPU（Graphics Processing Unit）による並列計算を活かすことで，柔軟・高速なガウス過程の実装を実現している．

5 章で紹介した時空間モデリングに関しては，静的な方法は gstat，FRK，mgcv パッケージなどで実装でき，動的な方法は INLA，KFAS などで実装できるもののパッケージがやや少ない印象である．

6 章で紹介した GLM に基づく方法に関しては INLA，sdmTMB，mgcv パッケージなどで実装できる．それらの比較については Anderson et al.（2022）を参照されたい．変換／ NN に基づく方法に関してはすでに述べたとおりである

8. 今後の展望

以上で空間統計分野の最近の動向について整理した．無論，本研究で紹介できたのは当該分野の一部のトピックのみであり，例えば空間領域別の離散データのモデリング（例：Lee, 2013），多変量モデリング（例：Zhang et al., 2021），空間交絡（例：Hughes and Haran, 2013）など，活発に議論されているトピックは他にも複数存在する点には注意されたい．

空間統計学はまだまだ課題が残されている．例えば NN の応用は活発化してきているが，NN と空間統計モデルの融合モデルの統計的性質はほぼ明らかとされておらず（Zhan and Datta, 2024），今後，両分野をまたぐ統計理論の確立が必要である．

また，地球物理学を中心に発展を遂げてきたデータ同化手法の空間統計学での応用が，基礎的なカルマンフィルタ以外は進んでいないという課題もある．実際，代表的なデータ同化手法の一つである ensemble Kalman filter（Evensen, 2003）は複雑な時空間パターンを捉えるうえで有用であるものの Katzfuss et al.（2016）によれば統計コミュニティでは largely unknown な手法であり適用例が少なかった．数多くのデータ同化手法が提案されてきているとともに（Carrassi et al., 2018 参照），近年は S4（Structured State Spaces for Sequence Modeling; Gu et al., 2022）や Mamba（Gu and Dao, 2023）といったより柔軟な時空間モデルが機械学習分野でも登場しており，空間統計モデルとの融合の余地がある．

また，大量で多様なデータで学習され，幅広いタスク（例えば気温の将来予測，風速の空間内挿，土地被覆の分類）に適応可能な大規模モデルである基盤モデル（Bommasani et al., 2021）が，空間統計分野の主要応用分野である気象・環境分野や都市分野で活用され始めている（Nguyen et al., 2023; Zhang et al., 2024 参照）．以上のような機械学習分野の研究動向も踏まえながら，より大規模で複雑なモデリングも扱えるように，空間統計学の手法の高度化や実用化を進めて

いくことが今後期待される.

参 考 文 献

Anderson, S. C., Ward, E. J., English, P. A. and Barnett, L. A. (2022). sdmTMB: An R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields, *BioRxiv*, https://doi.org/10.1101/2022.03.24.485545.

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D. and Lindgren, F. (2018). Spatial modeling with R-INLA: A review, *Wiley Interdisciplinary Reviews*: *Computational Statistics*, **10**(6), e1443.

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D. and Rue, H. (2019). Non-stationary Gaussian models with physical barriers, *Spatial Statistics*, **29**, 268–288.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**(4), 825–848.

Bolin, D., Simas, A. B. and Xiong, Z. (2024). Covariance–based rational approximations of fractional SPDEs for computationally efficient Bayesian inference, *Journal of Computational and Graphical Statistics*, **33**(1), 64–74.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models, ArXiv, https://doi.org/10.48550/arXiv.2108.07258.

Carrassi, A., Bocquet, M., Bertino, L. and Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives, *Wiley Interdisciplinary Reviews*: *Climate Change*, **9**(5), e535.

Chen, W., Li, Y., Reich, B. J. and Sun, Y. (2024). Deepkriging: Spatially dependent deep neural networks for spatial prediction, *Statistica Sinica*, **34**, 291–311.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L. and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme Australian wildfires, *Spatial Statistics*, https://doi.org/10.1016/j.spasta.2024.100811.

Cracknell, M. J. and Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, *Computers and Geosciences*, **63**, 22–33.

Cressie, N. (2015), *Statistics for Spatial Data*, John Wiley, New York.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**(1), 209–226.

Cressie, N. and Wikle, C. K. (2015), *Statistics for Spatio-Temporal Data*, Wiley, New Jersey.

Cressie, N., Shi, T. and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data, *Journal of Computational and Graphical Statistics*, **19**(3), 724–745.

Cressie, N., Sainsbury-Dale, M. and Zammit-Mangion, A. (2022). Basis-function models in spatial statistics, *Annual Review of Statistics and Its Application*, **9**(1), 373–400.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets, *Journal of the American Statistical Association*, **111**(514), 800–812.

De Cesare, L., Myers, D. E. and Posa, D. (2001). Product-sum covariance for space-time modeling: An environmental application, *Environmetrics*: *The Official Journal of the International En-*

vironmetrics Society, **12**(1), 11–23.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods, *Journal of Computational and Graphical Statistics*, **23**(2), 295–315.

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343–367.

Finley, A. O., Sang, H., Banerjee, S. and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets, *Computational Statistics and Data Analysis*, **53**(8), 2873–2884.

Fioravanti, G., Martino, S., Cameletti, M. and Toreti, A. (2023). Interpolating climate variables by using INLA and the SPDE approach, *International Journal of Climatology*, **43**(14), 6866–6886.

Fuglstad, G. A., Simpson, D., Lindgren, F. and Rue, H. (2015). Does non-stationary spatial data always require non-stationary random fields?, *Spatial Statistics*, **14**, 505–531.

Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets, *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.

Furrer, R., Bachoc, F. and Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction, *Journal of Multivariate Analysis*, **149**, 177–191.

Gerber, F. and Nychka, D. (2021). Fast covariance parameter estimation of spatial Gaussian process models using neural networks, *Stat*, **10**(1), e382.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data, *Journal of the American Statistical Association*, **97**(458), 590–600.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone, *Remote Sensing of Environment*, **202**, 18–27.

Gotway, C. A. and Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction, *Journal of Agricultural, Biological, and Environmental Statistics*, **2**(2), 157–178.

Gramacy, R. B. (2016). laGP: large-scale spatial modeling via local approximate Gaussian processes in R, *Journal of Statistical Software*, **72**, https:/doi.org/10.18637/jss.v072.i01.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces, ArXiv, https://doi.org/10.48550/arXiv.2312.00752.

Gu, A., Goel, K., Gupta, A. and Ré, C. (2022). On the parameterization and initialization of diagonal state space models, *Advances in Neural Information Processing Systems*, **35**, 35971–35983.

Guinness, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations, *Technometrics*, **60**(4), 415–429.

Guinness, J. (2021). Gaussian process learning via Fisher scoring of Vecchia's approximation, *Statistics and Computing*, **31**(3), 25.

Hazra, A., Nag, P., Yadav, R. and Sun, Y. (2024). Exploring the efficacy of statistical and deep learning methods for large spatial datasets: A case study, *Journal of Agricultural, Biological and Environmental Statistics*, https://doi.org/10.1007/s13253-024-00602-4.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F. and Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data, *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 398–425.

Hirano, T. (2017). Modified linear projection for large spatial datasets, *Communications in Statistics-Simulation and Computation*, **46**(2), 870–889.

Huang, X., Wang, C., Li, Z. and Ning, H. (2021). A 100 m population grid in the CONUS by disag-

gregating census data with open-source Microsoft building footprints, *Big Earth Data*, **5**(1), 112–133.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **75**(1), 139–159.

Imaizumi, M. and Fukumizu, K. (2022). Advantage of deep neural networks for estimating functions with singularity on hypersurfaces, *Journal of Machine Learning Research*, **23**(111), 1–54.

Ingebrigtsen, R., Lindgren, F., Steinsland, I. and Martino, S. (2015). Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field, *Spatial Statistics*, **14**, 338–364.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks—with an erratum note, GMD Technical Report, 148, German National Research Center for Information Technology, Bonn, Germany.

Jurek, M. and Katzfuss, M. (2021). Multi-resolution filters for massive spatio-temporal data, *Journal of Computational and Graphical Statistics*, **30**(4), 1095–1110.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets, *Journal of the American Statistical Association*, **112**(517), 201–214.

Katzfuss, M. and Guinness, J. (2021). A general framework for Vecchia approximations of Gaussian processes, *Statistical Science*, **36**(1), 124–141.

Katzfuss, M. and Hammerling, D. (2017). Parallel inference for massive distributed spatial data using low-rank models, *Statistics and Computing*, **27**, 363–375.

Katzfuss, M. and Schäfer, F. (2023). Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields, *Journal of the American Statistical Association*, **119**, 1409–1423, https://doi.org/10.1080/01621459.2023.2197158.

Katzfuss, M., Stroud, J. R. and Wikle, C. K. (2016). Understanding the ensemble Kalman filter, *The American Statistician*, **70**(4), 350–357.

Katzfuss, M., Stroud, J. R. and Wikle, C. K. (2019). Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models, *Journal of the American Statistical Association*, **115**(530), 866–885, https://doi.org/10.1080/01621459.2019.1592753.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. (2018), *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*, Chapman and Hall/CRC, Boca Raton.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors, *Journal of Statistical Software*, **55**(13), https://doi.org/10.18637/jss.v055.i13.

Lezama-Ochoa, N., Pennino, M. G., Hall, M. A., Lopez, J. and Murua, H. (2020). Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular mobular), *Scientific Reports*, **10**(1), 18822.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **73**(4), 423–498.

Lindgren, F., Bolin, D. and Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running, *Spatial Statistics*, **50**, https://doi.org/10.1016/j.spasta.2022.100599.

McDermott, P. L. and Wikle, C. K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting, *Stat*, **6**(1), 315–330.

McDermott, P. L. and Wikle, C. K. (2019). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting, *Environmetrics*, **30**(3), e2553.

Miller, M. J. and Reich, B. J. (2022). Bayesian spatial modeling using random Fourier frequencies,

Spatial Statistics, **48**, https://doi.org/10.1016/j.spasta.2022.100598.

Nag, P., Sun, Y. and Reich, B. J. (2023). Spatio-temporal DeepKriging for interpolation and probabilistic forecasting, *Spatial Statistics*, **57**, https://doi.org/10.1016/j.spasta.2023.100773.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. and Grover, A. (2023). ClimaX: A foundation model for weather and climate, ArXiv, https://doi.org/10.48550/arXiv.2301.10343.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets, *Journal of Computational and Graphical Statistics*, **24**(2), 579–599.

Nychka, D., Hammerling, D., Krock, M. and Wiens, A. (2018). Modeling and emulation of nonstationary Gaussian fields, *Spatial Statistics*, **28**, 21–38.

Porcu, E. and Zastavnyi, V. (2011). Characterization theorems for some classes of covariance functions associated to vector valued random fields, *Journal of Multivariate Analysis*, **102**(9), 1293–1301.

Porcu, E., Furrer, R. and Nychka, D. (2021). 30 Years of space–time covariance functions, *Wiley Interdisciplinary Reviews*: *Computational Statistics*, **13**(2), e1512.

Rullière, D., Durrande, N., Bachoc, F. and Chevalier, C. (2018). Nested Kriging predictions for datasets with a large number of observations, *Statistics and Computing*, **28**, 849–867.

Sainsbury-Dale, M., Zammit-Mangion, A. and Huser, R. (2024). Likelihood-free parameter estimation with neural Bayes estimators, *The American Statistician*, **78**(1), 1–14.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **74**(1), 111–132.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data, *Spatial Statistics*, **8**, 1–19.

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **66**(2), 275–296.

Stevens, F. R., Gaughan, A. E., Linard, C. and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data, *PLoS ONE*, **10**, e0107042.

Tonks, A., Harris, T., Li, B., Brown, W. and Smith, R. (2024). Forecasting West Nile virus with graph neural networks: Harnessing spatial dependence in irregularly sampled geospatial data, *GeoHealth*, **8**(7), e2023GH000784, https://doi.org/10.1029/2023GH000784.

Tsutsumi, M. and Seya, H. (2009). Hedonic approaches based on spatial econometrics and spatial statistics: Application to evaluation of project benefits, *Journal of Geographical Systems*, **11**, 357–380.

堤盛人, 瀬谷創 (2012). 応用空間統計学の二つの潮流：空間統計学と空間計量経済学, 統計数理, **60**(1), 3–25.

Tzeng, S. and Huang, H. C. (2018). Resolution adaptive fixed rank kriging, *Technometrics*, **60**(2), 198–208.

Van Niekerk, J. and Rue, H. (2024). Low-rank variational Bayes correction to the Laplace method, *Journal of Machine Learning Research*, **25**(62), 1–25.

Van Niekerk, J., Krainski, E., Rustand, D. and Rue, H. (2023). A new avenue for Bayesian inference with INLA, *Computational Statistics and Data Analysis*, **181**, https://doi.org/10.1016/j.csda.2023.107692.

Van Stein, B., Wang, H., Kowalczyk, W., Emmerich, M. and Bäck, T. (2020). Cluster-based Kriging approximation algorithms for complexity reduction, *Applied Intelligence*, **50**(3), 778–791.

Vargas-Muñoz, J. E., Srivastava, S., Tuia, D. and Falcão, A. X. (2020). OpenStreetMap: Challenges and opportunities in machine learning and remote sensing, *IEEE Geoscience and Remote Sensing*

Magazine, **9**, 184–199.

Wang, H., Guan, Y. and Reich, B. (2019). Nearest-neighbor neural networks for geostatistics, *Proceedings of International Conference on Data Mining Workshops*, 196–205, https://doi.org/10.1109/ICDMW.2019.00038.

Wiens, A., Nychka, D. and Kleiber, W. (2020). Modeling spatial data using local likelihood estimation and a Matérn to spatial autoregressive translation, *Environmetrics*, **31**(6), e2652.

Wikle, C. K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data, *Journal of Agricultural, Biological and Environmental Statistics*, **24**(2), 175–203.

Wikle, C. K. and Zammit-Mangion, A. (2023). Statistical deep learning for spatial and spatiotemporal data, *Annual Review of Statistics and Its Application*, **10**(1), 247–270.

Wikle, C. K., Zammit-Mangion, A. and Cressie, N. (2019). *Spatio-temporal Statistics with R*, Chapman and Hall/CRC, Boca Raton.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Massachusetts.

Wood, S. N., Li, Z., Shaddick, G. and Augustin, N. H. (2017). Generalized additive models for gigadata: modeling the UK black smoke network daily data, *Journal of the American Statistical Association*, **112**(519), 1199–1210.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S. Y. (2020). A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, **32**(1), 4–24.

Yoshida, T., Murakami, D. and Seya, H. (2022). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset, *The Journal of Real Estate Finance and Economics*, **69**, 1–28.

Yu, Y., Si, X., Hu, C. and Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures, *Neural Computation*, **31**(7), 1235–1270.

Zammit-Mangion, A., Bertolacci, M., Fisher, J., Stavert, A., Rigby, M. L., Cao, Y. and Cressie, N. (2022a). WOMBAT v1.0: A fully Bayesian global flux-inversion framework, *Geoscientific Model Development*, **15**(1), 45–73.

Zammit-Mangion, A., Ng, T. L. J., Vu, Q. and Filippone, M. (2022b). Deep compositional spatial models, *Journal of the American Statistical Association*, **117**(540), 1787–1808.

Zammit-Mangion, A., Kaminski, M. D., Tran, B. H., Filippone, M. and Cressie, N. (2024a). Spatial Bayesian neural networks, *Spatial Statistics*, https://doi.org/10.1016/j.spasta.2024.100825.

Zammit-Mangion, A., Sainsbury-Dale, M. and Huser, R. (2024b). Neural Methods for Amortised Parameter Inference, ArXiv, https://doi.org/10.48550/arXiv.2404.12484.

Zhan, W. and Datta, A. (2024). Neural networks for geospatial data, *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2024.2356293.

Zhang, L., Banerjee, S. and Finley, A. O. (2021). High-dimensional multivariate geostatistics: A Bayesian matrix-normal approach, *Environmetrics*, **32**(4), e2675.

Zhang, W., Han, J., Xu, Z., Ni, H., Liu, H. and Xiong, H. (2024). Towards urban general intelligence: A review and outlook of urban foundation models, ArXiv, https://doi.org/10.48550/arXiv.2402.01749.

Recent Advances in Spatial Statistics

Daisuke Murakami

The Institute of Statistical Mathematics

Statistical methods for geospatial data have been studied in spatial statistics. Especially in recent years, when spatio-temporal data have become larger and more diverse, methods that are both computationally efficient and flexible have been developed rapidly. This study therefore summarizes recent advances in spatial statistics. We first introduce basic spatial statistical models and their challenges. Next, approximations for large samples are classified into low-rank approximation, covariance approximation, and approximation of the precision matrix, and studies are reviewed for each. Then, methods for flexibility modeling spatial processes and observations are reviewed respectively. Software packages for implementing spatial statistical methods are explained after that. Finally, future directions in spatial statistics are discussed.

Key words: Spatial statistics, spatiotemporal modeling, Gaussian process, spatial correlation, neural networks.

統計数理(2025)
第 73 巻 第 1 号 19–34
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
［原著論文］

空間計量経済学における近年の方法論的な発展

瀬谷 創†・泊 将史†

（受付 2024 年 7 月 3 日；改訂 8 月 21 日；採択 8 月 26 日）

要　　旨

空間計量経済学という分野が産声を上げてから，もうすぐ 50 年が経過しようとしている．標準的な方法論は 2000 年代までにおおむね整備され，現在までに多くの実証研究が行われてきた．本稿では，空間計量経済学分野の近年の方法論的な発展に関する個人的な見解を論じる．具体的には，1)識別と因果推論，2)空間重み行列の特定化，3)空間的自己相関構造の柔軟なモデル化，4)時空間データのモデリング，5)ダーティデータのモデリングの観点から，近年の重要な方法論的発展をレビューすることを試みる．

キーワード：空間計量経済学，空間重み行列，統計的因果推論，機械学習，ダーティデータ，時空間パネルデータ．

1. はじめに

1960〜70 年代，計量地理学の分野では，空間的自己相関(Spatial autocorrelation)が，最も基礎的かつ重要な問題のひとつと位置づけられ，空間データの分析モデリングのための知見が積み重ねられていった．その後，計量地理学の流れをくみながら，離散的な空間(市区町村等のゾーンなど)上の格子データ(Lattice data)のモデリングを扱う空間計量経済学(Spatial econometrics)と呼ばれる分野が，空間統計学と相互に影響を与えながらも異なる分野として発展した（Anselin, 2010; Arbia, 2011）．Anselin and Bera（1998)によると，「空間計量経済学」という用語は，1970 年代初頭にベルギーの経済学者である Jean Paelinck が用い始めたものであり，Paelinck and Klaassen（1979)は空間計量経済学の分野における初のテキストとされている．その後，空間計量経済学の発展の一つの大きなきっかけになったのが，Anselin（1988)や，LeSage and Pace（2009)［各務・和合 訳, 2020］の出版であり，Anselin（1988)では，空間計量経済学が，「空間的自己相関と，空間的異質性(Spatial heterogeneity)を扱う計量経済学の一分野」と定義されている．

格子データのモデリングは，大きく Besag（1974)流のマルコフ確率場・条件付き分布に基づく方法と，Anselin（1988)流の非マルコフ確率場・同時分布に基づく方法に分類できる．モデルで言えば，前者は Conditional autoregressive(CAR)モデルを用い，後者は Simultaneous autoregressive モデルを用いることが多い．本稿は特に空間計量経済学に着目するため，説明は後者が中心となることをご容赦願いたい．なお，モデルの呼称については，統計コミュニティと空間計量経済学のコミュニティで異なる点があるので，注意が必要である（瀬谷・堤, 2014）．

空間計量経済学は 2000 年代に大きく花開いた分野である．その要因はいくつか考えられるが，私見では，1)計量経済学者や統計学者の参入により，代表的なモデルの推定量の漸近的性

† 神戸大学 大学院工学研究科：〒 657–8501 兵庫県神戸市灘区六甲台町 1–1

質が整理され（Lee, 2023），モンテカルロ実験等によって小標本特性についても理解が進んだこと，2)空間データの普及により，空間的自己相関のモデル化のニーズが高まったこと，3)計算ソフトの普及（Anselin and Rey, 2022)や計算機性能の向上により手法実装の敷居が低下したこと，4)地理情報システム(Geographic information system（GIS))の普及によって空間データの処理にかかるコストが大きく低下したこと，5)Spatial Econometrics Association の立ち上げ(Arbia, 2011)により，集中的な議論ができる場が設けられたことなどが重要であったように思う．当時，数多くの国際誌で特集号が組まれ，Anselin（2010)は，それまでの 30 年間の回想論文において，応用計量経済学や社会科学の方法論として，マージンからメインストリームへ移行したと述べている．

しかしその後，Gibbons and Overman（2012)が “Mostly pointless spatial econometrics?” という刺激的なタイトルの論文を地域科学分野のトップジャーナルの一つである Journal of Regional Science 誌の Symposium on spatial econometrics と題された特集号に発表した．これは，因果の識別問題の観点から，空間計量経済学の「モデルの関数形が既知であると仮定して推定値を計算し，モデル比較の手法を用いてモデルをデータに選択させるアプローチ」を批判したものである．経済学では，1990 年代のいわゆる信頼性革命以降，因果や識別が重要なトピックとなった．一方空間計量経済学は，そのような動きに対して鈍感であったと言える．Mur(2013)は，空間計量経済学が識別や因果関係に注意を払ってこなかったのは，おそらく初期の空間計量経済学の進展が，時系列解析の進展の模倣によるものだったという歴史に関係していると指摘している．また，Pinkse and Slade（2010)は，“The future of spatial econometrics” と題するレビュー論文において，1 つの空間パラメータのみで空間的自己相関をモデル化する標準的な空間計量経済モデルについて，“laughable notion” と問題視している．

このように，伝統的な空間計量経済学のアプローチについて，特に 2000 年代後半頃から様々な課題が指摘されてきた．一方で，それらに対処するためのアプローチの提案や，モデルの高度化も大きく進展した．しかし，伝統的なアプローチを整理した堤・瀬谷（2012a)，土木や地域科学関連分野の応用研究を整理した堤・瀬谷（2012b)では，このような進展が反映されておらず，近年の方法論的発展を和文で整理した文献は筆者らの知る限り存在しない．以上の背景から本研究では，1)識別と因果推論，2)空間重み行列の特定化，3)空間的自己相関構造の柔軟なモデル化，4)時空間データのモデリング，5)ダーティデータのモデリングの観点から，近年の重要な方法論的発展をレビューすることを試みる．

2. 空間計量経済モデルの概要

本研究では，基本的にはクロスセクショナルな格子データにおける線形の空間計量経済モデルを対象とするが，一部パネルデータも対象とする．

クロスセクションにおける代表的な空間計量経済モデルは，

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} = \lambda \boldsymbol{W}\boldsymbol{u} + \boldsymbol{\varepsilon} \tag{2.2}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2.3}$$

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2.4}$$

と定式化できる．ここで，サンプルサイズを N としたとき，$\boldsymbol{y}$ は $N \times 1$ の被説明変数ベクトル，$\boldsymbol{W}$ は $N \times N$ の対角要素を 0 とする空間重み行列，$\boldsymbol{X}$ は $N \times K$ の説明変数行列，$\boldsymbol{\varepsilon}$ は $N \times 1$ の平均 0，均一分散の independent and identically distributed (i.i.d.) 誤差のベクトル（$\boldsymbol{X}$ は定数項

を含むが，$\boldsymbol{WX}$ は定数項を含まないとする），$\boldsymbol{\beta}$，$\boldsymbol{\gamma}$ は回帰係数に対応する $K \times 1$，$(K-1) \times 1$ のパラメータベクトル，λ と ρ は空間的自己相関の度合いを表すパラメータである．これらのモデルはそれぞれ空間ラグモデル（Spatial lag model）あるいは空間自己回帰モデル（Spatial autoregressive (SAR) model）（2.1），空間誤差モデル（Spatial error model）（2.2），SLX（Spatial lag of X）モデル（2.3），空間ダービンモデル（Spatial Durbin model）（2.4）と呼称されることが多い．

3. 空間計量経済モデルの方法論的発展

空間計量経済学分野では，モデルの関数形と空間重み行列 $\boldsymbol{W}$ が既知であると仮定して，ラグランジュ乗数検定や情報量規準などを用いてモデルをデータに選択させるアプローチが伝統的に採用されてきた（Anselin, 1988）．しかし，このようなアプローチに対して，識別の観点（Gibbons and Overman, 2012）やモデル構造の柔軟性の観点（Pinkse and Slade, 2010）から，批判が寄せられてきた（Debarsy and Le Gallo, 2024）．これらの概要について，それぞれ 3.1 節と 3.3 節で論じる．また，近年発展の著しい，空間重み行列の特定化（3.2 節），時空間データのモデリング（3.4 節），ダーティデータのモデリング（3.5 節）について，近年の方法論的な重要な発展をレビューする．

3.1 識別と因果推論

一定の条件を満たす $\boldsymbol{W}$（例えば，Kelejian and Piras, 2017）において，$|\rho| < 1$ のとき，SAR モデルは，

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \tag{3.1}$$

と誘導形で表現することができる．冪級数展開（Power series expansion）により，$(\boldsymbol{I} - \rho\boldsymbol{W})^{-1} = \boldsymbol{I} + \rho\boldsymbol{W} + \rho^2\boldsymbol{W}^2 + \cdots$ が得られるため，$\boldsymbol{y}$ の条件付き期待値は，

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} + \rho\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \rho^2\boldsymbol{W}^2\boldsymbol{X}\boldsymbol{\beta} + \cdots \tag{3.2}$$

と表現できる．ここで，SEM は係数に $\boldsymbol{\gamma} = -\lambda\boldsymbol{\beta}$ という制約を置けば SDM に帰着する（瀬谷・堤, 2014）．したがって，第 2 章で導入したすべての空間計量経済モデルは，

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\pi}_1 + \boldsymbol{W}^2\boldsymbol{X}\boldsymbol{\pi}_2 + \cdots \tag{3.3}$$

と誘導形で表現できることとなり（$\boldsymbol{\pi}_r (r = 1, 2, \ldots)$ はパラメータベクトル），モデル間の差異は，$\boldsymbol{X}$ の何次の空間ラグ項を導入するかの違いに他ならなくなる．前述のとおり空間計量経済学では，LM 検定や情報量規準でモデル選択を行うアプローチをとってきた．そこでは $\boldsymbol{W}$ が正しく特定化されているという前提が置かれている．しかし実際には真の $\boldsymbol{W}$ は不明であり，$\boldsymbol{X}$ の高次の空間ラグ変数同士に（多くの場合比較的強い）共線性がある中で，どのモデルがデータを発生させたかを識別することは困難である．これが Gibbons and Overman（2012）が批判した点の 1 つである．論文では，“these different specifications are generally impossible to distinguish without assuming prior knowledge about the true data generating process that we often do not possess in practice” と表現されている．

一方で，第 2 章で導入した空間計量経済モデル（式（2.1）-（2.4））は，それぞれ実証的なインプリケーションが大きく異なる．例えば，SAR モデルでは，k 番目の説明変数 $\boldsymbol{x}_k$ の限界的な変化の影響が，他地域に $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$ を通じて波及（スピルオーバー）するが，SEM では通常の線形回帰モデル同様にこのような効果はない．SLX モデルは，1 次の空間ラグ項のみが導入されているため，波及効果はローカルである．したがって，本来はモデル特定化の背景として

理論的な考察が必要であるが，空間計量経済モデルを用いた実証研究では，その部分が捨象され，データによる統計的なモデル選択のみに頼ったモデル特定化が行われている例が少なくない（SAR モデルのゲーム理論的な解釈については Lee, 2023 を，社会的相互作用的な解釈については力石 他, 2018）を参照されたい）．Pointless という誹りを免れるためには，モデル特定化のための考察を厳密に行う必要があり，少なくとも目的に応じた使い分けを行う必要があるといえよう．

まず，目的が除外変数バイアスの緩和にある場合，誤差相関を考慮する SEM を用いることが自然である．実証研究では，空間的な固定効果（例えば地価のモデリングにおける市区町村ダミー）の導入により空間的な相関を持つ除外変数に対処することが多いが（Kuminoff et al., 2010），どのような空間単位を選定するか（例えば，町丁目 or 市区町村）に結果が依存するという課題がある．空間単位が粗すぎるとコントロールとして不十分であるが，詳細に取りすぎると興味のある効果が抽出できなくなってしまう（Abbott and Klaiber, 2011; von Graevenitz and Panduro, 2015）．したがって，SEM によるコントロールは（$\boldsymbol{W}$ の特定化の問題は残るが）実証研究において使いやすい（Anselin and Arribas-Bel, 2013）．他のアプローチとして，von Graevenitz and Panduro（2015）は，Generalized additive model（GAM）のフレームで緯度経度の関数を導入し，感度分析的な考察を行うことを提案している．興味のある変数が地理的な境界で非連続な場合は（例えば税率が異なるなど），境界でモデルの空間的な差分を取ることで除外変数を difference-out する Spatial differencing と呼ばれる方法も使用可能である（例えば，Holmes, 1998; Belotti et al., 2018; Klein and Tchuente, 2023）．

次に，隣接地域における説明変数からのスピルオーバー効果を分析したい場合を考える．空間計量経済学の普及に大きな影響を果たした LeSage and Pace（2009）が推奨したこともあり，現在までに SDM を用いた数多くの実証研究が行われてきた（例えば，Seya et al., 2012）．しかし，Gibbons and Overman（2012）は，前述した識別問題のため，SDM や SAR ではなく，誘導系の SLX を用いることを推奨している．Halleck Vega and Elhorst（2015）は，SAR と異なる SLX の利点として，$\boldsymbol{W}$ をパラメタライズして，局所最適に陥らずに最尤法によりパラメータを推定できる点を挙げている（Olmo and Sanso-Navarro, 2023 はさらに柔軟なセミパラメトリックな SLX モデルを提案している）．ただし，$\boldsymbol{WX}$ が外生性を満たすか否かについても，別途慎重な検討が必要である．例えば，Lyytikäinen（2012）は，政策の変更を外生的変動として，$\boldsymbol{WX}$ の操作変数に用いている．

内生効果 ρ 自体に興味がある場合，SDM や SAR モデルを推定する必要がある．ここで $\boldsymbol{Wy}$ は内生変数であるため，（疑似）最尤法（Lee, 2004）や操作変数法（Kelejian and Prucha, 1998）を用いてパラメータ推定を行うが，$\boldsymbol{X}$ や $\boldsymbol{WX}$ の内生性を考慮する場合には後者が用いられる．$\boldsymbol{Wy}$ に対する操作変数（instrumental variable（IV））としては，$\boldsymbol{X}$ の高次の空間ラグ変数 $[\boldsymbol{WX}, \boldsymbol{W}^2\boldsymbol{X}, \boldsymbol{W}^3\boldsymbol{X}, \ldots]$ が用いられることが多い．このようなアプローチについて，Gibbons and Overman（2012）は，1）$\boldsymbol{X}$ の空間ラグ変数が除外制約を満たさない場合，IV として適切でないこと，2）$\boldsymbol{X}$ の空間ラグ変数は共線性の問題から，独立した外生的変動をわずかしか提供できず（特に SDM の場合），いわゆる weak instruments の問題が存在することを指摘した．実際に空間計量経済モデルの実証研究では，$\boldsymbol{X}$ の空間ラグ変数が除外制約を満たすか否かに関する考察が行われることは稀である．$\boldsymbol{X}$ の空間ラグ変数が IV として適切でない場合，別の IV を準備する必要がある．例えば，政策の変更を外生的変動として利用した前述の Lyytikäinen（2012）や，制度上の閾値を利用した Fruehwirth（2013）等の取り組みが参考になると思われる．

さて，経済学では，1990 年代のいわゆる「信頼性革命」以降，統計的な意味での因果推論が重要になった．因果推論の方法論は，大きく Pearl 流の Directed acyclic graph（DAG）アプローチ（Pearl and Mackenzie, 2018［夏目 訳, 2022］）と，Rubin 流の潜在的結果（Potential outcome）ア

プローチ（Imbens and Rubin, 2015）に大別できる．経済学では，特に後者を実証分析に用いることが一般化した．本稿では，後者のアプローチにおける空間計量経済モデルの可能性について論じる（前者と空間モデルの関係については，Akbari et al., 2023 等を参照されたい）．

潜在的結果アプローチで平均処置効果を求める上では，Stable Unit Treatment Value Assumption（SUTVA）条件が満たされる必要がある（Rubin, 1980）．SUTVA 条件は，1）No interference between units（潜在的結果が他の主体（地域）の処置状態に依存しない）と 2）No multiple versions of treatments（ある処置に異なる形（バージョン）は存在しない）からなる．ここで仮定 1 は，スピルオーバー効果や一般均衡的効果を除外する条件であるが，現実の処置においてはこれが成り立たない場合も少なくない．例えば，都市の文脈では社会資本が隣接自治体に与える正の影響や（要藤・吉村, 2016），都市計画の用途地域の変更が近隣地域の住宅建設数に与える負の影響などが挙げられる（Greenaway-McGrevy and Phillips, 2023）．相互干渉（Interference）が存在する場合，ある主体は実際には処置を受けていないにも関わらず処置効果を間接的に受けることになるため，これを無視して通常の Difference-in-differences（DID）推定を行うと，平均処置効果の推定値にバイアスが生じる．

相互干渉をモデルで扱う標準的なアプローチは，処置のスピルオーバー効果を低次元の十分統計量に変換する Exposure mapping と呼ばれる方法であり（Aronow and Samii, 2017），処置された隣接主体のシェア（Hudgens and Halloran, 2008）などがよく用いられる．また，空間的なスピルオーバーの場合，隣接地域，隣々接地域における処理の有無（要藤・吉村, 2016）や隣接地域へのダミー変数の導入（Butts, 2023）などが用いられている．Exposure mapping の方法の一つとして，空間重み行列を用いることができる（Delgado and Florax, 2015; Bardaka et al., 2019）．今，$\boldsymbol{Y}$ を $NT \times 1$ の被説明変数のベクトル（N は地域数，T は観測時点数），$\boldsymbol{D}$ を $NT \times 1$ の処置群ベクトル（処置群であれば 1，対照群であれば 0），$\boldsymbol{T}$ を $NT \times 1$ の処置後ベクトル（処置後であれば 1，処置前であれば 0），$\boldsymbol{I}_{NT}$ を NT 次元の単位行列，$\boldsymbol{\iota}$ を 1 からなる $NT \times 1$ のベクトル，$\boldsymbol{\varepsilon}$ を $NT \times 1$ の平均 0，均一分散の i.i.d. 誤差のベクトルとする．空間重み行列が時間不変であるとき，NT 次元の空間重み行列 $\boldsymbol{W}_{NT}$ は，$\boldsymbol{W}_{NT} = \boldsymbol{I}_{\boldsymbol{T}} \otimes \boldsymbol{W}_N$ と与えることができる（$\otimes$ はクロネッカー積）．このとき，SLX 型の Spatial DID モデル（Delgado and Florax, 2015）は，

$$\boldsymbol{Y} = \alpha_0 \boldsymbol{\iota} + \alpha_1 \boldsymbol{D} + \alpha_2 \boldsymbol{T} + \alpha_3 (\boldsymbol{I}_{NT} + \phi \boldsymbol{W}_{NT}) \boldsymbol{D} \circ \boldsymbol{T} + \boldsymbol{\varepsilon} \tag{3.4}$$

と定式化できる．ここで，$\alpha_k (k = 1, 2, 3)$ はパラメータ，ϕ は空間パラメータ，$\circ$ はアダマール積である．$\alpha_4 = \alpha_3 \phi$ と置き直したとき，α_3 は直接効果（Average treatment effects on the treated（ATT））を，α_4 は間接効果（スピルオーバー効果）を表す（Hudgens and Halloran, 2008）．Bardaka et al.（2019）は，Delgado and Florax（2015）を複数種類の処置がある場合に，Chagas et al.（2016）は複数の空間重み行列がある場合に，それぞれ拡張している．Hudgens and Halloran（2008）における処置された隣接主体のシェアは，$\boldsymbol{W}_N$ が行規準化された場合に相当する．また，Chagas et al.（2016）のように複数の重みを用いれば，要藤・吉村（2016）のように隣接地域，隣々接地域を考慮することもできる．

因果推論における空間モデルの活用に関するさらなる話題については，Kolak and Anselin（2020），Reich et al.（2021），Debarsy and Le Gallo（2024）などを参照されたい．

3.2 空間重み行列の特定化

前節では，識別の観点から，SLX モデルの有用性について議論した．しかし，どのモデルを用いるにせよ，基本的には空間重み行列 $\boldsymbol{W}$ の特定化が必要である．Stakhovych and Bijmolt（2009）は，$\boldsymbol{W}$ の与え方を，（1）完全に外生とする，（2）データから決定する，（3）推定するとい

う 3 つに分類した．(1)は，地域の境界が接しているか否か(隣接行列)や，距離の逆数，ドローネ三角網等で与える典型的な方法である．(2)には，情報量規準 (Seya et al., 2013; Agiakloglou and Tsimpanos, 2023)やベイズモデル平均化 (Debarsy and LeSage, 2022)等のアプローチが存在する($\boldsymbol{W}$ の特定化については，Griffith, 2020 等を参照されたい)．

ここで，(1)のアプローチがよく用いられる理由としては，「距離が近い事物はより強く影響しあう」という地理学の第一法則(first law of geography)が持ち出されることが多いが (Anselin, 1988)，$\boldsymbol{W}$ の外生性が自然に担保されるという計量経済学的な側面も重要である．すなわち，SAR モデルにおける k 番目の説明変数の限界効果は，$\partial \boldsymbol{y}/\partial \boldsymbol{x}_k' = \beta_k(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$ で与えられるため，処置が $\boldsymbol{W}$ を変化させるような内生性が存在する場合，限界効果にバイアスが発生する．しかし，経済理論的な理由から $\boldsymbol{W}$ が社会経済的な距離として与えられる場合もあり (例えば，Behrens et al., 2012)，$\boldsymbol{W}$ の内生性を考慮した推定法が必要である．

Qu and Lee (2015)は，重み行列の内生性を考慮する実用的な方法を提案した．まず，p 個の内生変数 $\boldsymbol{z} = (z_1, \ldots, z_p)'$ を用いて，距離測度を $f(.)$ としたとき，重み行列を $w_{ij} = f(\boldsymbol{z}_i, \boldsymbol{z}_i)$ と表現する(i, j は地域を表す添字)．例えば，$p = 2$ で z_1 が GDP，z_2 が人口だとすると，重み行列は GDP および人口の類似度を意味する．ここで，$\boldsymbol{z}$ が 外生説明変数と誤差項の線形モデルで与えられると仮定すると，

$$\boldsymbol{Z} = \boldsymbol{X}_2 \boldsymbol{B} + \boldsymbol{E} \tag{3.5}$$

とおける．ここで $\boldsymbol{Z}$ は $N \times p$ の内生変数の行列，$\boldsymbol{X}_2$ は $N \times q$ の外生説明変数からなる行列，$\boldsymbol{B}$ は $q \times p$ の回帰係数行列，$\boldsymbol{E}$ は $N \times p$ の誤差項の行列である．Qu and Lee (2015)は，SAR 方程式(式(2.1))の誤差項 $\boldsymbol{\varepsilon}$ と，$\boldsymbol{W}$ の要素に関する方程式の誤差項 $\boldsymbol{E}$ が相関を持つことを許容し，QML 等による複数のパラメータ推定方法を提案した．結果として，SAR 方程式は，残差 $\boldsymbol{Z} - \boldsymbol{X}_2\boldsymbol{B}$ をコントロール変数として導入する形で再構成される．また，$\boldsymbol{\varepsilon}$ と $\boldsymbol{E}$ の相関を調べれば，$\boldsymbol{W}$ の内生性についての仮説検定も可能である．本アプローチは，w_{ij} がバイラテラル(bilateral)変数(例えば地域間流動など)で与えられる場合や (Qu et al., 2021)，時間変化する $\boldsymbol{W}_t$ の場合にも拡張されている (Qu et al., 2017)．

瀬谷・堤 (2014)の時点では，「$\boldsymbol{W}$ の要素の推定に関する研究は，まだまだ発展途上であるため，現状では(1)や(2)のアプローチを用いて作成された複数の $\boldsymbol{W}$ から，何らかの基準により，最適なものを選択するというアプローチが重要になろう.」と指摘した．しかし近年，$\boldsymbol{W}$ の推定に関する研究も蓄積されつつある．ただし，サンプルサイズ N に対して $\boldsymbol{W}$ の要素数は(対角項を除けば)$N^2 - N$ 存在するため，多くの場合推定には $T \gg N$ が要求され，T が小さい場合への適用性をモンテカルロ実験で調べる形がとられている．既往研究のアプローチは大きく，1)機械学習の方法と 2)ベイズ推定を用いた方法に区分できる．前者は，モデルの適合度と $\boldsymbol{W}$ の疎性のバランスを調整する方法であり，Lam and Souza (2020)の Adaptive LASSO, de Paula et al. (2024)の Adaptive elastic net generalized methods of moments(GMM) (Caner and Zhang, 2014)を用いた研究が挙げられる．交通分野の交通状態推定も，LASSO による $\boldsymbol{W}$ の推定と類似するところがある (原 他, 2016; Hara et al., 2018)．一方後者は，$\boldsymbol{W}$ に事前分布を与えるアプローチである．Krisztin and Piribauer (2023)，Piribauer et al. (2023)は，$\boldsymbol{W}$ の非対角要素に事前分布としてベルヌーイ分布を仮定し，また近傍となる主体の数がベータ二項分布に従うと仮定することで，ハイパーパラメータの設定によって $\boldsymbol{W}$ の疎性がコントロールできることを示した．この手法は，空間データ分析にとって標準的な，$N > T$ の状況での適用も可能である．事前分布を用いた方法には，他にも CAR モデルをベースとした Gao and Bradley (2019)等が存在する．このように機械学習やベイズ推定を応用したいくつかの手法が提案されてきているが，$\boldsymbol{W}$ の推定方法は未だ確立されてはいない．

3.3 空間的自己相関構造の柔軟なモデル化

典型的な空間計量経済モデルのシンプルなモデル構造については，解釈の容易さ等の利点はあるが，批判も少なくない（Pinkse and Slade, 2010）．しかし近年，非線形性や異質性などを考慮した，より柔軟なモデルが，機械学習アプローチを取り入れながら提案されている．本節では，それらの研究を概観する．

このような方向での萌芽的な取り組みとして，Pinkse et al.（2002）が挙げられる．Pinkse et al.（2002）は，SAR タイプのモデル

$$\boldsymbol{y} = \boldsymbol{W}(g)\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.6}$$

において，$\boldsymbol{W}$ の要素を $w_{ij} = g(d_{ij})$ と与えた．$g(\cdot)$ は，距離 d_{ij} の影響度合いを表す未知の関数であり，例えばフーリエ基底関数や多項式基底関数で近似すればよい．Sun（2016）は，Pinkse et al.（2002）のアプローチを，$w_{ij} = g(z_{ij})$ と内生変数に依存する場合に拡張した．ただし，説明変数が複数の場合には，前述の Qu and Lee（2015）アプローチ等を用いる必要がある．若干異なるアプローチとして，塚井・奥村（2007），Sun（2024）は，$\rho(\boldsymbol{z})\boldsymbol{W}\boldsymbol{y}$ と functional coefficient 型の定式化を行った．前者は $\boldsymbol{z}$ が外生変数と仮定している一方，後者では $\boldsymbol{z}$ の内生性が考慮されている．Hoshino（2022）は，$\boldsymbol{W}g(\boldsymbol{y})$ と非線形の内生効果を考慮し，sieve IV によるパラメータ推定法を提案している．Mínguez et al.（2022）は，SAR タイプの空間パネルモデルにおいて，説明変数の非線形性を考慮しつつ，空間座標の非線形効果を導入するモデルを考案している．Geniaux and Martinetti（2018）は，Geographically weighted regression（GWR）モデルと SAR モデルを結合し，空間パラメータを場所によって変えるモデルを構築した（ただし，GWR モデルはローカルな多重共線性に弱いため注意が必要である，村上, 2024）．

さらに柔軟な関数系として，グラフニューラルネットワークを用いた研究も行われている．Zhu et al.（2022）は，活性化関数を σ としたとき，$\boldsymbol{y} = \sigma(\tilde{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\Theta})$ とするモデル化を行っている（$\boldsymbol{\Theta}$ はニューラルネットワークにおける重み，$\tilde{\boldsymbol{W}}$ は $\boldsymbol{W}$ から構築されたグラフラプラシアン行列である．簡単のため中間層とバイアス項は無視している）．Yang et al.（2022）は，不動産価格予測の文脈で，予測地点を含む形で $\tilde{\boldsymbol{W}}$ と $\boldsymbol{X}$ を構築し，予測地点においては $\boldsymbol{y}$ が欠損しているとするモデル化のほうが，学習地点のみでモデルを構築し（$\hat{\boldsymbol{\Theta}}$ を得て），予測地点のみで構築された $\tilde{\boldsymbol{W}}$ と $\boldsymbol{X}$ に適用するよりも予測正確度が高いと指摘している．類似の取り組みには，Wang and Song（2023）と Xiao et al.（2023）があり，それぞれモデル化の方法が若干異なる．

3.4 時空間データのモデリング

Elhorst（2022）は，次式に示す一般化された空間パネル（データ）モデルを導入し，各項の役割について説明を行っている．

$$\begin{aligned} \boldsymbol{y}_t &= \tau\boldsymbol{y}_{t-1} + \rho\boldsymbol{W}\boldsymbol{y}_t + \eta\boldsymbol{W}\boldsymbol{y}_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma} + \Sigma_r\boldsymbol{\Gamma}_r^{'}\boldsymbol{f}_{rt} + \boldsymbol{u}_t \\ \boldsymbol{u}_t &= \lambda\boldsymbol{W}\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t \end{aligned} \tag{3.7}$$

ここで，$\boldsymbol{y}_t$ は被説明変数 y_{it} からなる $N \times 1$ ベクトル，$\boldsymbol{y}_{t-1}$，$\boldsymbol{W}\boldsymbol{y}_t$，$\boldsymbol{W}\boldsymbol{y}_{t-1}$ はそれぞれ $\boldsymbol{y}_t$ の時間，空間，時空間ラグ変数，$\boldsymbol{X}_t$ は $N \times K$ の説明変数行列，$\boldsymbol{W}\boldsymbol{X}_t$ はその空間ラグ変数である．また，τ，ρ，η，λ はそれぞれの変数の感応度を示すパラメータであり，$\boldsymbol{\beta}$，$\boldsymbol{\gamma}$ はそれぞれ $K \times 1$，$(K-1) \times 1$ の回帰係数ベクトルである．モデルに $\boldsymbol{W}\boldsymbol{X}_{t-1}$ が導入されていない理由は，この項が $\boldsymbol{W}\boldsymbol{y}_{t-1}$ に含まれているため，識別の問題を引き起こすためである（Anselin et al., 2008, Anselin 2021）．式（3.7）のモデルは，時間方向のラグが入っているという意味で，動学的空間パネルモデルとなっている．瀬谷・堤（2014）では，いくつかの静学的空間パネルモデルが説明されているが，2000 年代後半以降，動学的な空間パネルモデルの推定に関する研究が大き

く発展した(詳しくは，Lee, 2023 を参照されたい)．

もう一つの近年の重要な進展が，$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ 項の存在である．空間計量経済モデルでモデル化される空間的自己相関は，$\boldsymbol{W}$ を通したローカルな相関である．一方で，例えば東京と大阪の類似性など，地理的な距離に依存しない形でのグローバルなクロスセクショナル相関の存在も考えられる (Pesaran, 2015)．$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ 項は，このようなグローバルなクロスセクショナル相関を考慮するために導入されるものである．Elhorst (2024)によれば，ローカル・グローバルの両方のクロスセクショナル相関を導入したモデルはまだまだ少ない．ベンチマーク的な研究である Shi and Lee (2017)は，式(3.7)から $\boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma}$ を除いたモデルの QML 推定量を導出し，その一致性を証明している．なお，通常の空間パネルモデルは N が大きく，T が小さいという条件を仮定したものが多いが，$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ 項を導入したモデルは，通常 T も大きいことが仮定されていることに注意されたい．

$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ における $\boldsymbol{f}_{rt} (r = 1, \ldots, R)$ は，Bai (2009)の相互作用固定効果(interactive fixed effects)モデルで用いられたものであり，因子分析における共通因子を意味する(和文の解説としては，奥井, 2014 が参考になる)．例えば，共通因子が $\boldsymbol{f}_{1t} = (1, \ldots, 1)'$，$\boldsymbol{f}_{2t} = (\xi_1, \ldots, \xi_T)'$ の 2 つであるとき，それぞれの負荷量を $\boldsymbol{\Gamma}_1 = (\delta_1, \ldots, \delta_N)'$，$\boldsymbol{\Gamma}_2 = (1, \ldots, 1)'$ で与えれれば，$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ 項は標準的な時間と主体(空間)の固定効果に等しくなる．無論，このように時間効果と空間効果を分離する必要はなく，相互作用固定効果を用いれば，因子分析により場所によって異なる時間トレンドや，クロスセクショナルな相関を捉えることができる．また，$\Sigma_r \boldsymbol{\Gamma}_r' \boldsymbol{f}_{rt}$ 項を，(空間固定効果を導入しつつ)，被説明変数やその時間ラグ，さらには必要に応じて説明変数のクロスセクション平均で置き換える簡易推定法も発展している (Pesaran, 2015; 千木良他, 2011)．このように，空間パネルデータ分析のモデルは近年，より一般化される方向で発展したが，Angrist and Pischke (2009) [大森 他 訳, 2013]で議論されている通り，実証分析において時間ラグを入れるかどうかの選択は難しい問題であり，異なる識別の仮定を用いた感度分析的な考察による頑健性のチェックが極めて重要である．

その他の近年開発された空間計量経済学に関連する時空間モデルを以下で概説する．Amba and Le Gallo (2022)は，空間パラメータが時間変化することを許容した空間パネルモデルを構築している．一方，LeSage and Chih (2018)，Aquaro et al. (2021)は，回帰係数と空間パラメータが場所によって異なる Heterogeneous coefficients 型のモデルを構築した(ただし，T が大きいことが必要)．Chen et al. (2022)は，そこに共通因子を導入し，Bao and Zhou (2023)は動学的パネルに拡張している．Wu and Matsuda (2021)は，パラメータを閾値変数で分類し，複数のレジームを考慮する動学的空間パネルモデルを構築した．Yang and Lee (2019)は，同次方程式タイプの動学的空間パネルモデルを構築し，QML 推定量と IV 推定量の漸近的性質について議論している．

Otto et al. (2023, 2024a)，は，それぞれ空間パネルタイプの確率的ボラティリティモデルと Autoregressive conditional heteroscedasticity(ARCH)モデルを構築している (Otto et al., 2024b のレビュー参照)．塚井・小林 (2007)は，社会資本の長期記憶性を考慮するために，SLX タイプの Auto-regressive fractionally integrated moving averaged model with exogenous variables (ARFIMAX)モデルを開発した．Cho et al. (2023)は，時空間の自己回帰分布ラグ(Autoregressive distributed lag (ARDL))モデルについてのレビューを行っている．Ermagun and Levinson (2018)は，交通量の時空間予測の文脈で，Space-time autoregressive integrated moving average (STARIMA)モデルおよびその関連モデルの可能性を丹念にレビューしている．Elhorst et al. (2021)は，Vector autoregressive(VAR)モデルと空間計量経済モデルの関係性について整理している．

3.5 ダーティデータのモデリング

空間計量経済学で研究されているほとんどのモデルは(A)データ欠損がなく，(B)観測位置が正確というクリーンな空間データを仮定している．しかし，利用可能な空間データではこれらが満たされないことが多く，ダーティデータに対する空間モデルの開発が求められている(Arbia et al., 2016)．なお，本節のレビューは離散空間における格子データだけでなく，連続空間における地球統計データも対象に含んでいる．

まず，(A)のデータ欠損への対応策として，最も標準的に用いられている手法は，欠損を含むデータは捨て去るという，リストワイズ除去と呼ばれる手法である．King et al. (2001)によれば，93〜97 年に調査した政策科学系の雑誌では，94% の研究で同手法による対応が行われていた．しかし，リストワイズ除去が有効なのは，Missing completely at random が成り立つ場合，すなわち欠損するかどうかが被説明変数の値や説明変数の値に依存せず，完全にランダムであるという，限定的なケースのみである．特に空間計量経済モデルの場合，データの除去は空間重み行列を変化させるため，注意が必要である (Tsutsumi and Seya, 2009)．LeSage and Pace (2004)および Wang and Lee (2013)は，欠損が説明変数に依存する Missing at random な状況において，それぞれ Expectation maximization (EM)タイプのアルゴリズムと非線形最小二乗法を用いたパラメータ推定方法を提案している．欠損が被説明変数の値に依存する Missing not at random(あるいは Non-ignorable missing)のケースに対しては，サンプルセレクションを考慮した空間計量経済モデル(社会的相互作用モデル)が提案されている (Hoshino, 2019; Seya et al., 2021; Bao et al., 2024)．転移学習によるバイアス補正の検討も進みつつある (Zeng et al., 2024)．また，異なるアプローチとして，Arbia et al. (2022)はギブズサンプラーによる補正を，Arbia and Nardelli (2024)は，フォーマルなサンプリングを模倣するような再サンプリングを提案している．なお，Missing not at random の状況で，例えば上側の分位点を中心にサンプリングがなされている場合(例えば，インフラの維持管理において危険性の高い箇所を中心に観測が行われているなど)，空間パターンにも歪みが生じ，空間内挿結果に Granularity bias と呼ばれるバイアスが生じる (Shirota and Gelfand, 2022)．これについては，空間計量経済モデルにおいても同様の注意が必要となる (Baldoni et al., 2023)．

次に，(B)の位置誤差については，まず Global navigation satellite system(GNSS)のようにそもそも位置情報に観測誤差が含まれる場合が考えられる．また，個人情報の機密性保持のための位置情報のマスキング(ジオマスキング)により位置誤差が生じるケースもある．Santi et al. (2021b)は，空間計量経済モデルとともにジオマスキングをマーク付き点過程としてモデル化し，ジオマスキングの影響を考慮した尤度最大化を行うことを提案した．Santi et al. (2021a)は，さらに位置情報が存在しない場合でも残差の空間的自己相関の影響を緩和するために，属性の類似性を用いて空間重み行列を構築することを提案した．

4. おわりに

Debarsy and Le Gallo (2024)は，経済誌の上位 5 誌に掲載された論文のうち，「空間計量経済学」モデルを利用したものは，今やほとんどネットワーク計量経済学の文献にしかないと指摘している．その一因として，Gibbons and Overman (2012)が指摘したように，空間計量経済学が識別や因果関係に十分な注意を払ってこなかったことや，Pinkse and Slade (2010)が述べた通り，空間計量経済学の標準的なモデルが，“laughable notion” と指摘されるほどにシンプルであったことが考えられる．

一方で，本研究でレビューした通り，前者・後者両方について，2000 年代後半頃から様々な進展が見られた．本研究における整理が，上述したような状況の解消に少しでも役に立てば幸

いである．ただし本稿では，CAR タイプのモデルや，一般化線形モデル，非線形モデル，さらには実装面(ソフトウェア等)については触れることができなかった．これらについては，別の機会に整理を試みたい．

謝　辞

本研究は JSPS 科研費 24K00175，24K00997，24K01004 の助成を受けたものです．

参 考 文 献

Abbott, J. K. and Klaiber, H. A. (2011). An embarrassment of riches: Confronting omitted variable bias and multi-scale capitalization in hedonic price models, *Review of Economics and Statistics*, **93**(4), 1331–1342.

Agiakloglou, C. and Tsimpanos, A. (2023). Evaluating the performance of AIC and BIC for selecting spatial econometric models, *Journal of Spatial Econometrics*, **4**, 2, https://doi.org/10.1007/s43071-022-00030-x.

Akbari, K., Winter, S. and Tomko, M. (2023). Spatial causality: A systematic review on spatial causal inference, *Geographical Analysis*, **55**(1), 56–89.

Amba, M. C. and Le Gallo, J. (2022). Specification and estimation of a periodic spatial panel autoregressive model, *Journal of Spatial Econometrics*, **3**, 13, https://doi.org/10.1007/s43071-022-00028-5.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics*: *An Empiricist's Companion*, Princeton University Press, Princeton.（大森義明, 小原美紀, 田中隆一, 野口晴子 訳 (2013). 『「ほとんど無害」な計量経済学―応用経済学のための実証分析ガイド』, NTT 出版, 東京.）

Anselin, L. (1988). *Spatial Econometrics*: *Methods and Models*, Kluwer Academic Publishers, Dordrecht.

Anselin, L. (2010). Thirty years of spatial econometrics, *Papers in Regional Science*, **89**(1), 3–25.

Anselin, L. (2021). Spatial models in econometric research, *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press, Oxford, https://doi.org/10.1093/acrefore/9780190625979.013.643.

Anselin, L. and Arribas-Bel, D. (2013). Spatial fixed effects and spatial dependence in a single cross-section, *Papers in Regional Science*, **92**(1), 3–18.

Anselin, L. and Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics, *Handbook of Applied Economic Statistics* (eds. A. Ullah and D. E. A. Giles), Marcel Dekker, New York.

Anselin, L. and Rey, S. J. (2022). Open source software for spatial data science, *Geographical Analysis*, **54**(3), 429–438.

Anselin, L., Le Gallo, J. and Jayet, H. (2008). Spatial panel econometrics, *The Econometrics of Panel Data*: *Fundamentals and Recent Developments in Theory and Practice* (eds. B. H. Baltagi, Y. Hong, G. Koop, W. Krämer and L. Matyas), Springer, Berlin, Heidelberg.

Aquaro, M., Bailey, N. and Pesaran, M. H. (2021). Estimation and inference for spatial models with heterogeneous coefficients: An application to US house prices, *Journal of Applied Econometrics*, **36**(1), 18–44.

Arbia, G. (2011). A lustrum of SEA: Recent research trends following the creation of the spatial econometrics association (2007–2011), *Spatial Economic Analysis*, **6**(4), 377–395.

Arbia, G. and Nardelli, V. (2024). Using web-data to estimate spatial regression models, *International*

Regional Science Review, **47**(2), 204–226.

Arbia, G., Espa, G. and Giuliani, D. (2016). Dirty spatial econometrics, *The Annals of Regional Science*, **56**, 177–189.

Arbia, G., Matsuda, Y. and Wu, J. (2022). Estimating spatial regression models with sample datapoints: A Gibbs sampler solution, *Spatial Statistics*, **47**, https://doi.org/10.1016/j.spasta.2021.100568.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment, *The Annals of Applied Statistics*, **11**(4), 1912–1947.

Bai, J. (2009). Panel data models with interactive fixed effects, *Econometrica*, **77**(4), 1229–1279.

Baldoni, E., Coderoni, S. and Esposti, R. (2023). The productivity-environment nexus in space. Granularity bias, aggregation issues and spatial dependence within Italian farm-level data, *Journal of Cleaner Production*, **415**, https://doi.org/10.1016/j.jclepro.2023.137847.

Bao, Y. and Zhou, X. (2023). Heterogeneous spatial dynamic panels with an application to US housing data, *Spatial Economic Analysis*, **18**(2), 259–285.

Bao, Y., Li, G. and Liu, X. (2024). A spatial sample selection model, *Oxford Bulletin of Economics and Statistics*, **86**(4), https://doi.org/10.1111/obes.12599.

Bardaka, E., Delgado, M. S. and Florax, R. J. (2019). A spatial multiple treatment/multiple outcome difference-in-differences model with an application to urban rail infrastructure and gentrification, *Transportation Research Part A*: *Policy and Practice*, **121**, 325–345.

Behrens, K., Ertur, C. and Koch, W. (2012). 'Dual' gravity: Using spatial econometrics to control for multilateral resistance, *Journal of Applied Econometrics*, **27**(5), 773–794.

Belotti, F., Di Porto, E. and Santoni, G. (2018). Spatial differencing: Estimation and inference, *CESifo Economic Studies*, **64**(2), 241–254.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *The Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Butts, K. (2023). Difference-in-differences estimation with spatial spillovers, arXiv preprint, https://doi.org/10.48550/arXiv.2105.03737.

Caner, M. and Zhang, H. H. (2014). Adaptive elastic net for generalized methods of moments, *Journal of Business & Economic Statistics*, **32**(1), 30–47.

Chagas, A. L., Azzoni, C. R. and Almeida, A. N. (2016). A spatial difference-in-differences analysis of the impact of sugarcane production on respiratory diseases, *Regional Science and Urban Economics*, **59**, 24–36.

Chen, J., Shin, Y. and Zheng, C. (2022). Estimation and inference in heterogeneous spatial panels with a multifactor error structure, *Journal of Econometrics*, **229**(1), 55–79.

力石真, 瀬谷創, 福田大輔 (2018). 社会的相互作用に着目したエビデンスベース研究の展開と土木計画への応用可能性, 土木学会論文集 D3(土木計画学), **74**(5), I_715–I_734.

千木良弘朗, 早川和彦, 山本拓 (2011).『動学的パネルデータ分析』, 知泉書館, 東京.

Cho, J. S., Greenwood-Nimmo, M. and Shin, Y. (2023). Recent developments of the autoregressive distributed lag modelling framework, *Journal of Economic Surveys*, **37**(1), 7–32.

de Paula, A., Rasul, I. and Souza, P. (2024). Identifying network ties from panel data: Theory and an application to tax competition, *Review of Economic Studies*, https://doi.org/10.1093/restud/rdae088.

Debarsy, N. and Le Gallo, J. (2024). The empirical content of spatial spillovers: Identification issues, SSRN, http://dx.doi.org/10.2139/ssrn.4751335.

Debarsy, N. and LeSage, J. P. (2022). Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices, *Journal of Business & Economic Statistics*, **40**(2), 547–558.

Delgado, M. S. and Florax, R. J. (2015). Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction, *Economics Letters*, **137**, 123–126.

Elhorst, J. P. (2022). The dynamic general nesting spatial econometric model for spatial panels with common factors: Further raising the bar, *Review of Regional Research*, **42**(3), 249–267.

Elhorst, J. P. (2024). Raising the bar in spatial economic analysis: Two laws of spatial economic modelling, *Spatial Economic Analysis*, **19**(2), 115–132.

Elhorst, J. P., Gross, M. and Tereanu, E. (2021). Cross-sectional dependence and spillovers in space and time: Where spatial econometrics and global VAR models meet, *Journal of Economic Surveys*, **35**(1), 192–226.

Ermagun, A. and Levinson, D. (2018). Spatiotemporal traffic forecasting: review and proposed directions, *Transport Reviews*, **38**(6), 786–814.

Fruehwirth, J. C. (2013). Identifying peer achievement spillovers: Implications for desegregation and the achievement gap, *Quantitative Economics*, **4**(1), 85–124.

Gao, H. and Bradley, J. R. (2019). Bayesian analysis of areal data with unknown adjacencies using the stochastic edge mixed effects model, *Spatial Statistics*, **31**, https://doi.org/10.1016/j.spasta.2019.100357.

Geniaux, G. and Martinetti, D. (2018). A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models, *Regional Science and Urban Economics*, **72**, 74–85.

Gibbons, S. and Overman, H. G. (2012). Mostly pointless spatial econometrics?, *Journal of Regional Science*, **52**(2), 172–191.

Greenaway-McGrevy, R. and Phillips, P. C. (2023). The impact of upzoning on housing construction in Auckland, *Journal of Urban Economics*, **136**, https://doi.org/10.1016/j.jue.2023.103555.

Griffith, D. A. (2020). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models, *Practical Handbook of Spatial Statistics* (ed. S. Arlinghaus), CRC Press, Boca Raton.

Halleck Vega, S. and Elhorst, J. P. (2015). The SLX model, *Journal of Regional Science*, **55**(3), 339–363.

原祐輔, 花岡洋平, 桑原雅夫 (2016). 道路ネットワーク内の関係性に着目した長期観測プローブデータによるプローブ未観測リンクの交通状態補間, 交通工学論文集, **2**(1), 1–10.

Hara, Y., Suzuki, J., and Kuwahara, M. (2018). Network-wide traffic state estimation using a mixture Gaussian graphical model and graphical lasso, *Transportation Research Part C*, **86**, 622–638.

Holmes, T. J. (1998). The effect of state policies on the location of manufacturing: Evidence from state borders, *Journal of Political Economy*, **106**(4), 667–705.

Hoshino, T. (2019). Two-step estimation of incomplete information social interaction models with sample selection, *Journal of Business & Economic Statistics*, **37**(4), 598–612.

Hoshino, T. (2022). Sieve IV estimation of cross-sectional interaction models with nonparametric endogenous effect, *Journal of Econometrics*, **229**(2), 263–275.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference, *Journal of the American Statistical Association*, **103**(482), 832–842.

Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*: *An Introduction*, Cambridge University Press, Cambridge.

Kelejian, H.H. and Piras, G. (2017). *Spatial Econometrics*, 1st ed., Academic Press, Cambridge.

Kelejian, H.H. and Prucha, I.R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, *Journal of Real Estate Economics*, **17**(1), 99–121.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). Analyzing incomplete political science data:

An alternative algorithm for multiple imputation, *American Political Science Review*, **95**(1), 49–69.

Klein, A. and Tchuente, G. (2023). Spatial differencing for sample selection models with 'site-specific' unobserved local effects, *The Econometrics Journal*, **27**(2), 235–257.

Kolak, M. and Anselin, L. (2020). A spatial perspective on the econometrics of program evaluation, *International Regional Science Review*, **43**(1-2), 128–153.

Krisztin, T. and Piribauer, P. (2023). A Bayesian approach for the estimation of weight matrices in spatial autoregressive models, *Spatial Economic Analysis*, **18**(1), 44–63.

Kuminoff, N. V., Parmeter, C. F. and Pope, J. C. (2010). Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?, *Journal of Environmental Economics and Management*, **60**(3), 145–160.

Lam, C. and Souza, P. C. (2020). Estimation and selection of spatial weight matrix in a spatial lag model, *Journal of Business & Economic Statistics*, **38**(3), 693–710.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models, *Econometrica*, **72**(6), 1899–1925.

Lee, L. F. (2023). *Spatial Econometrics*: *Spatial Autoregressive Models*, World Scientific Publishing, Singapore.

LeSage, J. P. and Chih, Y. Y. (2018). A Bayesian spatial panel model with heterogeneous coefficients, *Regional Science and Urban Economics*, **72**, 58–73.

LeSage, J. P. and Pace, R. K. (2004). Models for spatially dependent missing data, *The Journal of Real Estate Finance and Economics*, **29**, 233–254.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, Boca Raton.（各務和彦, 和合肇 訳 (2020).『入門空間計量経済学』, 彩流社, 東京.）

Lyytikäinen, T. (2012). Tax competition among local governments: Evidence from a property tax reform in Finland, *Journal of Public Economics*, **96**(7-8), 584–595.

Mínguez, R., Basile, R. and Durbán, M. (2022). An introduction to pspatreg: A new R package for semiparametric spatial autoregressive analysis, *Region*, **9**(2), R1–R15.

Mur, J. (2013). Causality, uncertainty and identification: Three issues on the spatial econometrics agenda, *Scienze Regionali*, **12**(1), 5–27.

村上大輔 (2024).『R ではじめる地理空間データの統計解析入門』, 講談社, 東京.

奥井亮 (2014). 因子モデルに関する近年の計量経済学研究の進展, 日本統計学会誌, **43**(2), 247–273.

Olmo, J. and Sanso-Navarro, M. (2023). A nonparametric spatial regression model using partitioning estimators, *Econometrics and Statistics*, https://doi.org/10.1016/j.ecosta.2023.02.003.

Otto, P., Doğan, O. and Taşpınar, S. (2023). A dynamic spatiotemporal stochastic volatility model with an application to environmental risks, *Econometrics and Statistics*, https://doi.org/10.1016/j.ecosta.2023.11.002.

Otto, P., Doğan, O. and Taşpınar, S. (2024a). Dynamic spatiotemporal ARCH models, *Spatial Economic Analysis*, **19**(2), 250–271.

Otto, P., Doğan, O., Taşpınar, S., Schmid, W. and Bera, A. K. (2024b). Spatial and spatiotemporal volatility models: A review, *Journal of Economic Surveys*, https://doi.org/10.1111/joes.12643.

Paelinck, J. and Klaassen, L. (1979). *Spatial Econometrics*, Saxon House, Farnborough.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why*: *The New Science of Cause and Effect*, Basic Books, New York.（夏目大 訳 (2022).『因果推論の科学「なぜ？」の問いにどう答えるか』, 文藝春秋, 東京.）

Pesaran, M. H. (2015). *Time Series and Panel Data Econometrics*, Oxford University Press, Oxford.

Pinkse, J. and Slade, M. E. (2010). The future of spatial econometrics, *Journal of Regional Science*, **50**(1), 103–117.

Pinkse, J., Slade, M. E., and Brett, C. (2002). Spatial price competition: A semiparametric approach, *Econometrica*, **70**(3), 1111–1153.

Piribauer, P., Glocker, C. and Krisztin, T. (2023). Beyond distance: The spatial relationships of European regional economic growth, *Journal of Economic Dynamics and Control*, **155**, https://doi.org/10.1016/j.jedc.2023.104735.

Qu, X. and Lee, L. F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix, *Journal of Econometrics*, **184**(2), 209–232.

Qu, X., Lee, L. F. and Yu, J. (2017). QML estimation of spatial dynamic panel data models with endogenous time varying spatial weights matrices, *Journal of Econometrics*, **197**(2), 173–201.

Qu, X., Lee, L. F. and Yang, C. (2021). Estimation of a SAR model with endogenous spatial weights constructed by bilateral variables, *Journal of Econometrics*, **221**(1), 180–197.

Reich, B. J., Yang, S., Guan, Y., Giffin, A. B., Miller, M. J. and Rappold, A. (2021). A review of spatial causal inference methods for environmental and epidemiological applications, *International Statistical Review*, **89**(3), 605–634.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment, *Journal of the American Statistical Association*, **75**(371), 591–593.

Santi, F., Dickson, M. M., Espa, G., Taufer, E. and Mazzitelli, A. (2021a). Handling spatial dependence under unknown unit locations, *Spatial Economic Analysis*, **16**(2), 194–216.

Santi, F., Dickson, M. M., Giuliani, D., Arbia, G. and Espa, G. (2021b). Reduced-bias estimation of spatial autoregressive models with incompletely geocoded data, *Computational Statistics*, **36**(4), 2563–2590.

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』, 朝倉書店, 東京.

Seya, H., Tsutsumi, M. and Yamagata, Y. (2012). Income convergence in Japan: A Bayesian spatial Durbin model approach, *Economic Modelling*, **29**(1), 60–71.

Seya, H., Yamagata, Y. and Tsutsumi, M. (2013). Automatic selection of a spatial weight matrix in spatial econometrics: Application to a spatial hedonic approach, *Regional Science and Urban Economics*, **43**(3), 429–444.

Seya, H., Tomari, M. and Uno, S. (2021). Parameter estimation in spatial econometric models with non-random missing data, *Applied Economics Letters*, **28**(6), 440–446.

Shi, W. and Lee, L. F. (2017). Spatial dynamic panel data models with interactive fixed effects, *Journal of Econometrics*, **197**(2), 323–347.

Shirota, S. and Gelfand, A. E. (2022). Preferential sampling for bivariate spatial data, *Spatial Statistics*, **51**, https://doi.org/10.1016/j.spasta.2022.100674.

Stakhovych, S. and Bijmolt, T. H. (2009). Specification of spatial models: A simulation study on weights matrices, *Papers in Regional Science*, **88**(2), 389–409.

Sun, Y. (2016). Functional-coefficient spatial autoregressive models with nonparametric spatial weights, *Journal of Econometrics*, **195**(1), 134–153.

Sun, Y. (2024). Semiparametric spatial autoregressive models with nonlinear endogeneity, *Econometric Reviews*, **43**(6), 434–451, https://doi.org/10.1080/07474938.2024.2339149.

塚井誠人, 小林潔司 (2007). 長期記憶性を考慮した社会資本の生産性測定, 土木学会論文集 D, **63**(3), 255–274.

塚井誠人, 奥村誠 (2007). 空間データセットにおける内生的な空間相関構造の同定に関する考察, 土木計画学研究・講演集, **35**, http://library.jsce.or.jp/jsce/open/00039/200706_no35/pdf/119.pdf (最終アクセス日 2024 年 9 月 12 日).

Tsutsumi, M. and Seya, H. (2009). Hedonic approaches based on spatial econometrics and spatial statistics: Application to evaluation of project benefits, *Journal of Geographical Systems*, **11**, 357–380.

堤盛人, 瀬谷創 (2012a). 応用空間統計学の二つの潮流：空間統計学と空間計量経済学, 統計数理, **60**(1), 3–25.

堤盛人, 瀬谷創 (2012b). 土木計画における応用空間統計学の可能性, 土木学会論文集 D3(土木計画学), **68**(5), I_1–I_20.

von Graevenitz, K. and Panduro, T. E. (2015). An alternative to the standard spatial econometric approaches in hedonic house price models, *Land Economics*, **91**(2), 386–409.

Wang, W. and Lee, L. F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable, *The Econometrics Journal*, **16**(1), 73–102.

Wang, Z. and Song, Y. (2023). Deep learning for the spatial additive autoregressive model with nonparametric endogenous effect, *Spatial Statistics*, **55**, https://doi.org/10.1016/j.spasta.2023.100743.

Wu, J. and Matsuda, Y. (2021). A threshold extension of spatial dynamic panel model with fixed effects, *Journal of Spatial Econometrics*, **2**(3), https://doi.org/10.1007/s43071-021-00008-1.

Xiao, S., Song, Y. and Wang, Z. (2023). Nonparametric spatial autoregressive model using deep neural networks, *Spatial Statistics*, **57**, https://doi.org/10.1016/j.spasta.2023.100766.

Yang, K. and Lee, L. F. (2019). Identification and estimation of spatial dynamic panel simultaneous equations models, *Regional Science and Urban Economics*, **76**, 32–46.

Yang, Z., Hong, Z., Zhou, R. and Ai, H. (2022). Graph convolutional betwork-based model for megacity real estate valuation, *IEEE Access*, **10**, 104811–104828.

要藤正任, 吉村有博 (2016). 社会資本によるスピルオーバー効果と地域経済成長：市町村データを用いた高速道路整備効果の実証分析, KIER Discussion Paper, No. 1603, Institute of Economic Research, Kyoto University, 京都.

Zeng, H., Zhong, W. and Xu, X. (2024). Transfer learning for spatial autoregressive models, arXiv preprint, https://doi.org/10.48550/arXiv.2405.15600.

Zhu, D., Liu, Y., Yao, X. and Fischer, M. M. (2022). Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions, *GeoInformatica*, **26**(4), 645–676.

Recent Methodological Developments in Spatial Econometrics

Hajime Seya and Masashi Tomari

Department of Civil Engineering, Graduate School of Engineering, Kobe University

Almost 50 years have passed since the emergence of the field of spatial econometrics. Standard methodologies were largely established by the 2000s, and a vast amount of empirical research has been conducted to date. This paper discusses our personal views on recent methodological developments in the field of spatial econometrics. Specifically, we review important recent methodological developments in terms of 1) identification and causal inference, 2) specification of spatial weight matrices, 3) flexible model structures, 4) spatiotemporal data modeling, and 5) dirty data modeling.

Key words: Spatial econometrics, spatial weight matrix, statistical causal inference, machine learning, dirty data, spatial panel data.

統計数理(2025)
第 73 巻 第 1 号 35–51
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
［総合報告］

ガウス過程に基づく時空間データ解析

田中 佑典[1]・上田 修功[2]

（受付 2024 年 6 月 28 日；改訂 11 月 22 日；採択 11 月 26 日）

要　　旨

ガウス過程はデータから関数を推定するための確率モデルの一つであり，古くから時空間データのモデリングに広く使われてきた．本稿では，時空間データ解析に関連する，二つの重要な問題を取り上げ，ガウス過程を用いたモデリングについて述べる．一つ目の問題は，集約データのモデリングである．都市などで取得される時空間データ(貧困度など)は，座標や時刻など「点」に紐づくデータのことを指すが，プライバシー保護や行政上の理由で，行政区画など「領域」に紐づくデータとして扱われることが少なくない．本稿では，ガウス過程の領域に対する積分を考えることによって，集約データを自然に扱う方法について紹介する．二つ目の問題は，常微分方程式で表される力学系のモデリングである．力学系のダイナミクスは状態の時間微分によって定義され，相空間上における「ベクトル場」として表現される．素朴には多出力ガウス過程を用いることによってベクトル場をモデリングできるが，物理法則を満たすような学習結果が必ずしも得られるとは限らない．本稿では，解析力学の一形式である「ハミルトン力学」を融合したガウス過程により，エネルギーの保存・散逸則を満たすベクトル場をモデリングする方法について紹介する．

キーワード：ガウス過程，時空間データ，集約データ，力学系，ハミルトン力学．

1. はじめに

機械学習の研究はますます発展を続けており，近年では様々な分野に応用されている．機械学習とは，一言でいうと，解きたい問題に応じた入出力関係を満たす関数を，大量のデータから自動的に獲得する方法である．本稿では，関数に対する確率分布であるガウス過程(Gaussian process: GP)（Rasmussen and Williams, 2006）を用いた，時空間データのモデリングについて述べる．ガウス過程は，関数の事前分布として用いられ，ベイズ推論の枠組みに基づいて学習および予測を行う．ガウス過程は，関数形を陽に仮定しないノンパラメトリックな手法であり，共分散関数を指定することにより柔軟な関数を表現できる．また，データが与えられたときの事後分布を用いて，予測分散を評価できることもガウス過程の利点の一つである．

本稿では，時空間データ解析のためのガウス過程について，比較的新しい二つの重要な研究トピックを紹介する．一つ目は，**集約データのためのガウス過程**である．標準的な時空間回帰では，点(座標や時刻)と属性の組でサンプルが表され，その背後にある関数を推定する．一方，ここで扱う問題設定では，サンプルが領域(行政区画や時間帯)と属性の組で表されることを仮定する．例えば，都市において収集された貧困度や犯罪数のようなデータは，プライバ

[1] NTT コミュニケーション科学基礎研究所：〒 619–0237 京都府相楽郡精華町光台 2–4
[2] 理化学研究所 革新知能統合研究センター：〒 103–0027 東京都中央区日本橋 1–4–1

シー保護や行政上の理由から，行政区画や管轄区域において平均化などの統計処理がなされ，公開されることが多い．従来のガウス過程では，このような集約データを自然に扱うことが困難であった．Smith et al. (2018)において，単一の集約データを，ガウス過程の領域における積分としてモデリングする方法が提案された．これにより，領域の形や大きさを反映した共分散関数を表現でき，集約データから背後にある連続関数の推定を可能とした．その後，複数の集約データを同時に解析することを目的として，重回帰分析に基づく拡張 (Law et al., 2018; Tanaka et al., 2019a)や，多出力ガウス過程(Multi-output GP)に基づく拡張 (Tanaka et al., 2019b; Yousefi et al., 2019; Hamelijnck et al., 2019)がなされ，様々な解像度を持つ集約データ間の相関を考慮してデータの予測が可能な手法へと発展している．本稿では，Tanaka et al. (2019b)の内容を中心に解説を行う．

二つ目のトピックは，**力学系のためのガウス過程**である．多くの力学系は常微分方程式で表され，状態の時間微分をモデリングすることで系のダイナミクスを記述することができる．この時間微分を相空間(状態空間と同義)上に表したものを「ベクトル場」と呼び，機械学習モデルを用いてベクトル場をパラメタライズすることで，データからダイナミクスの学習を可能とする (Chen et al., 2018)．素朴には多出力ガウス過程を用いることによりベクトル場のモデリングが可能であるが，物理法則を満たすような学習結果が必ずしも得られるとは限らない．そこで，物理ダイナミクスをより適切にモデリングするために，物理学に由来する「数理構造」を組み込むことが重要である．ダイナミクスを表すベクトル場は，物理法則の下である種の普遍性を持つ．例えば，質量保存則に基づき，非圧縮流体の速度場は発散しない(Divergence-free)ことが知られている．また，電磁気学においては，エネルギーの保存則に基づき，磁場が時間に対して不変なとき電場は回転をもたない(Curl-free)ことが知られている．このような数理構造を，ベクトル解析の知識を利用して共分散関数に組み込む方法が提案されている (Narcowich and Ward, 1994; Macêdo and Castro, 2010)．さらに，ハミルトン力学の理論 (Goldstein, 1980; 畑, 2014)に基づき，エネルギーの保存・散逸則を満たすベクトル場に対するガウス過程が提案されている (Rath et al., 2021; Tanaka et al., 2022; Tanaka, 2024; Boffi et al., 2022)．これらのモデリングにおいては，Greydanus et al. (2019)におけるアイデアを援用し，ベクトル場を直接モデリングする代わりに，系の全体エネルギーである「ハミルトニアン(Hamiltonian)」をガウス過程によってモデリングする．ベクトル場は，ハミルトンの運動方程式に基づき，ハミルトニアンの勾配を用いて導出される．ハミルトン力学に基づくモデリングは，エネルギーの保存・散逸則を導入可能であることに加え，様々な拡張の可能性を持つという点において重要である．例えば，Beckers et al. (2022)では，制御入力を持つポート・ハミルトン系(Port-Hamiltonian system)への拡張がなされている．本稿では，Tanaka et al. (2022)や Tanaka (2024)の内容を中心に解説を行う．

本稿の構成は以下のとおりである．2 章では，ガウス過程の基本事項について整理し，本稿を理解する上で重要な多出力ガウス過程について述べる．3 章では，集約データのためのガウス過程について，4 章では，力学系のためのガウス過程について，問題設定，手法，および，実験結果を述べる．最後に 5 章で結論を述べる．

2. ガウス過程

2.1 ガウス過程回帰

本節では，1 次元の回帰問題を例にして，ガウス過程 (Rasmussen and Williams, 2006)の基本事項をまとめる．$x, y \in \mathbb{R}$ とし，回帰モデル

$$y = f(x) + \epsilon \tag{2.1}$$

を考えよう．ここで，ϵ は平均 0 で分散 σ^2 のガウスノイズとする．N 個のサンプル $\{(x_n, y_n) \mid n = 1, \ldots, N\}$ が与えられたとき，関数 $f(x) : \mathbb{R} \to \mathbb{R}$ を推定したいとする．

ガウス過程は「関数を生成する確率分布」であり，推定したい関数 $f(x)$ の事前分布として用いられる．$f(x)$ がガウス過程に従うことを

$$f(x) \sim \mathcal{GP}\left(0, k(x, x')\right) \tag{2.2}$$

と表す．ここで，簡単のため $\mathbb{E}[f(x)] = 0$ を仮定した（$\mathbb{E}[\cdot]$ は期待値を表す）．また，$k(x, x') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ はガウス過程のパラメータであり，共分散関数（Covariance function）と呼ばれ，

$$k(x, x') = \mathbb{E}[f(x)f(x')] \tag{2.3}$$

を意味する．共分散関数には，正定値カーネル（Positive definite kernel）（福水, 2010）が用いられる．正定値カーネルは以下のように定義される．

定義（正定値カーネル）． 以下の二つの条件を満たすとき $k(x, x')$ は正定値カーネルである．

- 対称性：$k(x, x') = k(x', x)$ が成り立つ．
- 正定値性：任意の正の整数 N，任意の点 $x_1, \ldots, x_N \in \mathbb{R}$ および任意の実数 $c_1, \ldots, c_N \in \mathbb{R}$ に対して以下が成り立つ．

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0. \tag{2.4}$$

この条件は，後述のグラム行列（式(2.7)）が半正定値であることを意味する．

最もよく使われるものの一つは二乗指数カーネル（Squared exponential kernel）であり，

$$k(x, x') = \alpha^2 \exp\left(-\frac{1}{2\beta^2}(x - x')^2\right) \tag{2.5}$$

と表される．ここで，$\alpha^2 \in \mathbb{R}_{>0}$ は分散パラメータであり，$\beta^2 \in \mathbb{R}_{>0}$ はスケールパラメータである．正定値カーネルを用いてガウス過程は以下のように定義される．

定義（ガウス過程）． N 個の点 $\boldsymbol{X} = (x_1, \ldots, x_N)^\top$ における関数の値を $\boldsymbol{f} = (f(x_1), \ldots, f(x_N))^\top$ とする．任意の自然数 N に対して，$\boldsymbol{f}$ が N 次元ガウス分布

$$\boldsymbol{f} \sim \mathcal{N}\left(\boldsymbol{0}, k(\boldsymbol{X}, \boldsymbol{X})\right) \tag{2.6}$$

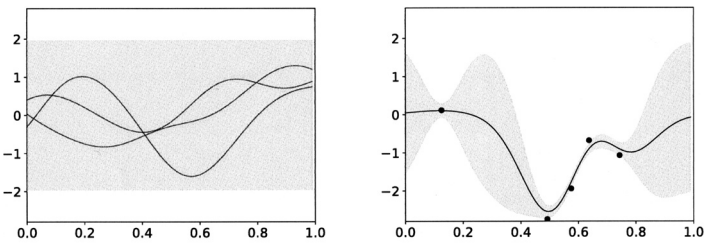
に従うとき，$\boldsymbol{f}$ はガウス過程である．ここで，$k(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{N \times N}$ は正定値カーネルで計算される対称行列（グラム行列（Gram matrix）と呼ばれる）であり，

$$k(\boldsymbol{X}, \boldsymbol{X}) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} \tag{2.7}$$

と表される．

式(2.7)からわかるように，カーネル $k(x, x')$ は任意の二点間における関数値の類似度を表していると解釈できる．したがって，カーネルを適切に設計することにより，関数の滑らかさについての事前知識を導入することができる．

次に，観測 $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ が与えられたときの事後分布を計算することにより，予測



(a) 事前分布 (b) 事後分布

図 1. ガウス過程の事前分布と事後分布．網掛け部分は 95% 信頼区間を表す．(a)カーネルパラメータを $\alpha^2 = 1$ および $\beta^2 = 0.2$ に設定し，事前分布から 3 つの関数を生成した．(b)事後分布は黒点で示す 5 つのサンプル($\sigma^2 = 0.1$)から学習した．黒線は予測平均 $m^*(x)$ を表す．

を行う方法について述べる．図 1 に，ガウス過程の事前分布と事後分布を可視化する．上記の回帰問題では，尤度はノイズ分散を σ^2 とするガウス分布であり，尤度と事前分布が共役(Conjugate)となり，事後分布が解析的に計算できる．事後分布もガウス過程となり，

$$f^*(x) \mid \boldsymbol{y} \sim \mathcal{GP}\left(m^*(x), k^*(x, x')\right) \tag{2.8}$$

と表される．事後分布の平均関数 $m^*(x) : \mathbb{R} \to \mathbb{R}$ と共分散関数 $k^*(x, x') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ は，

$$m^*(x) = k(x, \boldsymbol{X})^\top \boldsymbol{C}^{-1} \boldsymbol{y} \tag{2.9}$$

$$k^*(x, x') = k(x, x') - k(x, \boldsymbol{X})^\top \boldsymbol{C}^{-1} k(x, \boldsymbol{X}) \tag{2.10}$$

と表される．ここで，$k(x, \boldsymbol{X}) = (k(x, x_1), \ldots, k(x, x_N))^\top$ および $\boldsymbol{C} = k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}$ である．$\boldsymbol{I}$ は単位行列を表す．ガウス過程回帰では，式(2.9)を用いて任意の点における関数値の予測を行い，式(2.10)を用いて予測に対する分散を評価する．

最後に，ノイズ分散 σ^2 およびカーネルパラメータ α^2, β^2 の決め方について述べる．最もよく使われるのは，周辺尤度 $p(\boldsymbol{y})$ が最大となるようにパラメータを最適化する方法である．この方法は第二種最尤推定(Type Ⅱ maximum likelihood)と呼ばれる．上記の回帰問題では，事後分布と同様に周辺尤度も解析的に計算でき，$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ と表される．

2.2 ベクトル値関数への拡張

前章の議論を，D 次元ドメインにおいて M 個の出力を持つベクトル値関数(Vector-valued function) $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^M$ へと拡張する．$\boldsymbol{x} \in \mathbb{R}^D$ および $\boldsymbol{y} \in \mathbb{R}^M$ とし，回帰モデル

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} \tag{2.11}$$

を考える．ここで，$\boldsymbol{\epsilon}$ は平均 0 で共分散行列 $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)$ のガウスノイズとする．観測データ $\{(\boldsymbol{x}_{m,n}, y_{m,n}) \mid m = 1, \ldots, M; n = 1, \ldots, N_m\}$ が与えられたとき，関数 $\boldsymbol{f}(\boldsymbol{x})$ を推定したいとする．ここで，N_m は m 番目の出力に対するサンプル数を表す．

ベクトル値関数に対するガウス過程は，多出力ガウス過程(Multi-output GP)と呼ばれ，

$$\boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{GP}\left(\boldsymbol{0}, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{2.12}$$

と表される．ここで，$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{M \times M}$ は行列値カーネル(Matrix-valued kernel)(Álvarez et al., 2012)と呼ばれる．行列値カーネルも対称性および正定値性を満たすように構成され，以下のように定義される．

定義(行列値カーネルに対する対称性と正定値性)．以下の二つの条件を満たすとき $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')$ は正定値カーネルである．

- 対称性：$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{K}(\boldsymbol{x}',\boldsymbol{x})^\top$ が成り立つ．
- 正定値性：任意の正の整数 N，任意の点 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in\mathbb{R}^D$ および任意の実数ベクトル $\boldsymbol{c}_1,\ldots,\boldsymbol{c}_N\in\mathbb{R}^M$ に対して以下が成り立つ．

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{c}_i^\top\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\boldsymbol{c}_j\geq 0. \tag{2.13}$$

この条件は，後述のグラム行列(式(2.15))が半正定値であることを意味する．

$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')$ の具体的な構成法については後続の段落にて述べる．行列値カーネルを用いて多出力ガウス過程は以下のように定義される．

定義(多出力ガウス過程の定義)．M 個の出力に対する全サンプル数を $N=\sum_{m=1}^{M}N_m$ とする．$\boldsymbol{X}_m=(\boldsymbol{x}_{m,1},\ldots,\boldsymbol{x}_{m,N_m})$ とし，全ての入力をまとめて $\boldsymbol{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_M)^\top\in\mathbb{R}^{N\times D}$ とする．$\boldsymbol{X}_m$ における関数値を $\boldsymbol{f}_m=(f_m(\boldsymbol{x}_{m,1}),\ldots,f_m(\boldsymbol{x}_{m,N_m}))$ とし，M 個の出力からの関数値をまとめて $\boldsymbol{f}=(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_M)^\top\in\mathbb{R}^N$ とする．任意の自然数 N に対して，$\boldsymbol{f}$ が N 次元ガウス分布

$$\boldsymbol{f}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})) \tag{2.14}$$

に従うとき，$\boldsymbol{f}$ は多出力ガウス過程である．ここで，$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\in\mathbb{R}^{N\times N}$ は行列値カーネルで計算されるグラム行列であり，

$$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})=\begin{pmatrix}k_{1,1}(\boldsymbol{X}_1,\boldsymbol{X}_1) & \cdots & k_{1,M}(\boldsymbol{X}_1,\boldsymbol{X}_M)\\ \vdots & \ddots & \vdots\\ k_{M,1}(\boldsymbol{X}_M,\boldsymbol{X}_1) & \cdots & k_{M,M}(\boldsymbol{X}_M,\boldsymbol{X}_M)\end{pmatrix} \tag{2.15}$$

と表される．ここで，$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ は $M\times M$ のブロック行列であり，(m,m') 番目のブロック $k_{m,m'}(\boldsymbol{X}_m,\boldsymbol{X}_{m'})\in\mathbb{R}^{N_m\times N_{m'}}$ は，行列値カーネル $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')$ の (m,m') 番目の要素 $k_{m,m'}(\boldsymbol{x},\boldsymbol{x}')$ を $\boldsymbol{X}_m$ と $\boldsymbol{X}_{m'}$ について計算したグラム行列である．

以下では，行列値カーネル $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')$ を構成する方法について述べる．行列値カーネルを構成する最も簡単な方法は，出力間の依存関係を表す変数を直接導入することにより，

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')=k(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{Q} \tag{2.16}$$

とすることである．ここで，$k(\boldsymbol{x},\boldsymbol{x}'):\mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}$ は任意の正定値カーネルであり，$\boldsymbol{Q}\in\mathbb{R}^{M\times M}$ は出力の依存関係を表す半正定値行列である．行列 $\boldsymbol{Q}$ を半正定値とすることで，適当な行列 $\boldsymbol{A}$ を用いて $\boldsymbol{Q}=\boldsymbol{A}\boldsymbol{A}^\top$ のように分解することが可能である．これにより，式(2.16)が式(2.13)の正定値性を満たすことが示される．$k(\boldsymbol{x},\boldsymbol{x}')$ には，入力次元毎にスケールパラメータを推定可能な ARD(Automatic relevance determination)カーネル（Rasmussen and Williams, 2006）などが使われる．ARD カーネルは，

$$k(\boldsymbol{x},\boldsymbol{x}')=\alpha^2\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^\top\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}')\right) \tag{2.17}$$

と表され，$\boldsymbol{B}=\mathrm{diag}(\beta_1^2,\ldots,\beta_D^2)$ によって各入力次元に対するスケールが調節される．

この他にも，行列値カーネルを構成する発展的手法は複数提案されているが，その中でも最もよく使われる手法の一つが Linear model of coregionalization (LMC)（Teh et al., 2005; Álvarez

et al., 2012)である．LMC では，複数の潜在的なガウス過程を用意し，それらの線形結合により多出力ガウス過程を表現する．L 個の潜在ガウス過程を

$$g_\ell(\boldsymbol{x}) \sim \mathcal{GP}\left(0, k_\ell(\boldsymbol{x}, \boldsymbol{x}')\right), \quad \ell = 1, \ldots, L \tag{2.18}$$

とする．ここで，$k_\ell(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ は ℓ 番目の潜在ガウス過程の共分散関数である．$\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_L(\boldsymbol{x}))^\top$ として，重み行列 $\boldsymbol{W} \in \mathbb{R}^{M \times L}$ による線形変換を考え，

$$\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{W}\boldsymbol{g}(\boldsymbol{x}) \tag{2.19}$$

によって M 個の出力を持つ関数 $\boldsymbol{f}(\boldsymbol{x})$ を定義する．ガウス過程の線形変換は，またガウス過程になることが知られており，式(2.19)のように定義された関数 $\boldsymbol{f}(\boldsymbol{x})$ は多出力ガウス過程に従う．このとき，行列値カーネルは，

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{W}\boldsymbol{K}^{\mathrm{diag.}}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{W}^\top \tag{2.20}$$

と表される．ここで，$\boldsymbol{K}^{\mathrm{diag.}}(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{diag}(k_1(\boldsymbol{x}, \boldsymbol{x}'), \ldots, k_L(\boldsymbol{x}, \boldsymbol{x}'))$ である．$\boldsymbol{K}^{\mathrm{diag.}}(\boldsymbol{x}, \boldsymbol{x}')$ は明らかに正定値カーネルであり，式(2.20)を式(2.13)の不等式の左辺に代入すれば，上記のように定義された行列式カーネルが正定値性を満たすことが直ちにわかる．また，$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')$ の (m, m') 番目の要素は，

$$k_{m,m'}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\ell=1}^{L} w_{m,\ell} w_{m',\ell} k_\ell(\boldsymbol{x}, \boldsymbol{x}') \tag{2.21}$$

と表される．式(2.20)は，式(2.16)を一般化し，複数のカーネルの重み付き和によって行列値カーネルを構成することに対応する．これにより，関数の共分散構造に対する表現力を向上させることができ，$L < M$ を仮定することにより過学習を軽減することも可能である．

前章と同様，尤度と事前分布が共役であるため，事後分布が解析的に計算でき，

$$\boldsymbol{f}^*(\boldsymbol{x}) \mid \boldsymbol{y} \sim \mathcal{GP}\left(\boldsymbol{m}^*(\boldsymbol{x}), \boldsymbol{K}^*(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{2.22}$$

と表される．ここで，m 番目の出力に対する観測を $\boldsymbol{y}_m = (y_{m,1}, \ldots, y_{m,N_m})$ としたとき，全ての観測をまとめて $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M)^\top$ とした．また，事後分布の平均関数 $\boldsymbol{m}^*(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^M$，および，共分散関数 $\boldsymbol{K}^*(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{M \times M}$ は，

$$\boldsymbol{m}^*(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})^\top \boldsymbol{C}^{-1} \boldsymbol{y}, \tag{2.23}$$

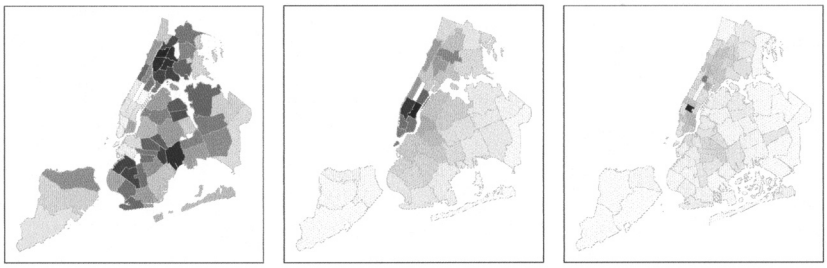
$$\boldsymbol{K}^*(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})^\top \boldsymbol{C}^{-1} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}') \tag{2.24}$$

となる．ここで，$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{R}^{N \times M}$ は，$\boldsymbol{X}$ と任意の点 $\boldsymbol{x}$ に対して行列値カーネル $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')$ の値を並べたものである．ただし，$N = \sum_{m=1}^{M} N_m$ である．また，$\boldsymbol{C} \in \mathbb{R}^{N \times N}$ は共分散行列であり，$M \times M$ のブロック行列で表され，(m, m') 番目のブロックを $\boldsymbol{C}_{m,m'}$ とすると，

$$\boldsymbol{C}_{m,m'} = k_{m,m'}(\boldsymbol{X}_m, \boldsymbol{X}_{m'}) + \delta_{m,m'} \sigma_m^2 \boldsymbol{I} \tag{2.25}$$

と表される．ここで，$k_{m,m'}(\boldsymbol{X}_m, \boldsymbol{X}_{m'}) \in \mathbb{R}^{N_m \times N_{m'}}$ は，式(2.21)の行列値カーネルの (m, m') 番目の要素を，$\boldsymbol{X}_m$ と $\boldsymbol{X}_{m'}$ に対して並べたグラム行列である．式(2.23)が M 個の出力に対する全観測 $\boldsymbol{y}$ の線形変換によって表されることからもわかるように，出力間の依存関係を考慮しつつ予測が行われる．

最後に，多出力ガウス過程回帰におけるパラメータ推定について述べる．上述の LMC に基づくモデル化において，推定すべきパラメータは，ノイズ分散 $\boldsymbol{\Sigma}$，L 個の潜在ガウス過程のカーネルパラメータ $\{\alpha_\ell, \boldsymbol{B}_\ell\}_{\ell=1}^{L}$，および，重み行列 $\boldsymbol{W}$ である．これらのパラメータをまとめ



(a) 貧困度　(b) 大気汚染度（PM2.5 濃度）　(c) 犯罪率

図 2. NYC Open Data が公開する集約データ (https://opendata.cityofnewyork.us)．属性によって都市の分割の仕方が異なることがわかる．

て $\boldsymbol{\theta}$ と書くことにすると，$\boldsymbol{\theta}$ は第二種最尤推定に基づき，以下の最適化問題を解くことにより求めることができる．

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \ \log p(\boldsymbol{y}). \tag{2.26}$$

ここで，$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ であり，解析的に計算が可能である．

3. 集約データのためのガウス過程

3.1 問題設定

前章で述べた標準的な回帰問題では，点(座標や時刻)と属性の組でサンプルが表され，その背後にある関数を推定する．一方，ここで扱う問題設定では，サンプルが領域(行政区画や時間帯)と属性の組で表されることを仮定する．例えば，都市において収集された貧困度や犯罪数のようなデータは，プライバシー保護や行政上の理由から，行政区画や管轄区域において平均化などの統計処理がなされ，公開されることが多い．図 2 にニューヨーク市が公開する集約データの例を示す．属性によって都市の分割の仕方が異なり，様々な解像度を持つことがわかる．

以下に，問題設定を数学的に定義する．本稿では，二次元空間における集約データを想定するが，時空間における集約データなども同様に扱うことが可能である．入力ドメインを $\mathcal{X} \in \mathbb{R}^2$ とし，$\boldsymbol{x} \in \mathcal{X}$ を入力変数とする．属性の数を M とし，m 番目の属性に対するサンプルを $(\mathcal{R}_{m,n}, y_{m,n})$ と表す．ここで，$\mathcal{R}_{m,n} \subset \mathcal{X}$ は，m 番目の属性における n 番目のサンプルが紐づく領域を表し，全ての領域は共通部分を持たないものとする．また，$y_{m,n} \in \mathbb{R}$ は属性の値である．m 番目の属性におけるサンプル数を N_m としたとき，観測データは $\{(\mathcal{R}_{m,n}, y_{m,n} \mid m = 1, \ldots, M; n = 1, \ldots, N_m)\}$ である．以下では，このような複数の集約データのモデリングについて考え，背後にある M 個の属性に対する関数 $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^M$ の推定問題を扱う．

3.2 手法

集約データのモデリングを難しくしているのは，図 2 のように属性によって入力ドメインの分割の仕方が異なり，また領域の形は様々であることが挙げられる．そのため，異なる領域分割に紐づいた複数のデータ間の依存関係を適切に捉える方法は自明ではない．

本稿では，Tanaka et al. (2019b)において提案された Spatially aggregated Gaussian process (SAGP) について紹介する．Yousefi et al. (2019)や Hamelijnck et al. (2019)においても，同時

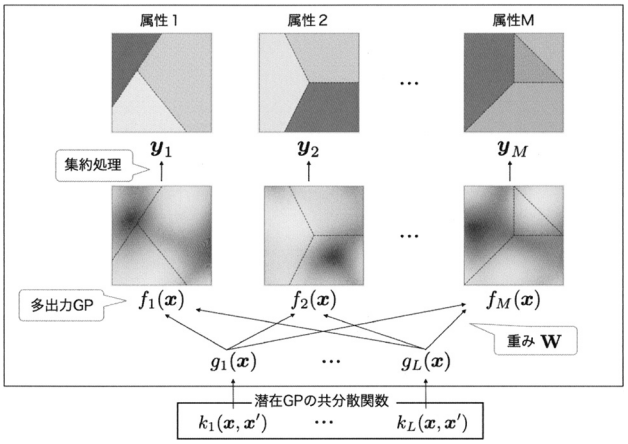


図 3. SAGP における集約データの生成モデル.

期に独立に類似手法が提案された．確率モデリングのアプローチに従い，観測データの生成プロセスを設計する．図 3 に，SAGP における集約データの生成モデルを示す．SAGP では，集約データの背後に滑らかな関数が存在することを仮定し，式(2.19)における LMC に基づく多出力ガウス過程 $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x}))^\top$ によって表す．次に，Smith et al. (2018), Burgess and Webster (1980)と同様の方法により，各集約データに対する集約処理をガウス過程の積分によってモデリングし，集約データの観測モデルを

$$\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \int_{\mathcal{X}} \boldsymbol{A}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})\, d\boldsymbol{x}, \boldsymbol{\Sigma}\right) \tag{3.1}$$

とする．ここで，ガウス過程 $\boldsymbol{f}(\boldsymbol{x})$ は積分可能であると仮定する(ガウス過程の可積分性についての議論は Tanaka et al. 2019b の付録を参照)．また，$\boldsymbol{y}_m = (y_{m,1}, \ldots, y_{m,N_m})$ として，$\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M)^\top \in \mathbb{R}^N$ とした．$\boldsymbol{A}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^{N \times M}$ は，

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{a}_1(\boldsymbol{x}) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{a}_2(\boldsymbol{x}) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{a}_M(\boldsymbol{x}) \end{pmatrix} \tag{3.2}$$

と表され，$\boldsymbol{a}_m(\boldsymbol{x}) = (a_{m,1}(\boldsymbol{x}), \ldots, a_{m,N_m}(\boldsymbol{x}))^\top$ である．ここで，各要素 $a_{m,n}(\boldsymbol{x})$ は，領域 $\mathcal{R}_{m,n}$ におけるデータ集約のための非負の重み関数であり，LMC によって表現された関数 $\boldsymbol{f}(\boldsymbol{x})$ の集約の範囲と集約の仕方を指定する．例えば，

$$a_{m,n}(\boldsymbol{x}) = \frac{\mathbb{1}(\boldsymbol{x} \in \mathcal{R}_{m,n})}{\int_{\mathcal{X}} \mathbb{1}(\boldsymbol{x}' \in \mathcal{R}_{m,n})\, d\boldsymbol{x}'} \tag{3.3}$$

とすると，$y_{m,n}$ は領域 $\mathcal{R}_{m,n}$ における $f_m(\boldsymbol{x})$ の平均値を表す．ここで，$\mathbb{1}(\cdot)$ は指示関数であり，Z が真のとき $\mathbb{1}(Z) = 1$ となり，そうでないとき $\mathbb{1}(Z) = 0$ となる．重み関数を変更することで，単純な合計や人口に対する重み付け和などを表現することも可能である．また，式(3.1)の $\boldsymbol{\Sigma}$ はノイズ分散であり，$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2 \boldsymbol{I}, \ldots, \sigma_M^2 \boldsymbol{I})$ である．

事前分布を LMC とし，尤度を式(3.1)としたときの事後分布について述べる．Tanaka et al. (2019b)において，事後分布がガウス過程として解析的に計算できることが示された．導出に

ついては（Tanaka et al., 2019b）の付録に譲り，ここでは結果だけを紹介する．事後分布は，式(2.22)と同様に表され，平均関数 $\boldsymbol{m}^*(\boldsymbol{x})$ と共分散関数 $\boldsymbol{K}^*(\boldsymbol{x}, \boldsymbol{x}')$ は，

$$\boldsymbol{m}^*(\boldsymbol{x}) = \boldsymbol{H}(\boldsymbol{x})^\top \boldsymbol{C}^{-1}\boldsymbol{y}, \tag{3.4}$$

$$\boldsymbol{K}^*(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{H}(\boldsymbol{x})^\top \boldsymbol{C}^{-1}\boldsymbol{H}(\boldsymbol{x}') \tag{3.5}$$

と表される．ここで，$\boldsymbol{H}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^{N\times M}$ は，

$$\boldsymbol{H}(\boldsymbol{x}) = \int_{\mathcal{X}} \boldsymbol{A}(\boldsymbol{x}')\boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x})\, d\boldsymbol{x}' \tag{3.6}$$

であり，任意の点 $\boldsymbol{x}$ と各領域との間の共分散を表す．また，$\boldsymbol{C} \in \mathbb{R}^{N\times N}$ は共分散行列であり，

$$\boldsymbol{C} = \iint_{\mathcal{X}\times\mathcal{X}} \boldsymbol{A}(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{A}(\boldsymbol{x}')^\top\, d\boldsymbol{x}\, d\boldsymbol{x}' + \boldsymbol{\Sigma} \tag{3.7}$$

と表される．$\boldsymbol{C}$ は $M \times M$ のブロック行列であり，(m, m') 番目のブロックを $\boldsymbol{C}_{m,m'}$ とすると，

$$\boldsymbol{C}_{m,m'} = \iint_{\mathcal{X}\times\mathcal{X}} k_{m,m'}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{a}_m(\boldsymbol{x})\boldsymbol{a}_{m'}(\boldsymbol{x}')^\top\, d\boldsymbol{x}\, d\boldsymbol{x}' + \delta_{m,m'}\sigma_m^2\boldsymbol{I} \tag{3.8}$$

である．ここで，式(3.8)における $k_{m,m'}(\boldsymbol{x}, \boldsymbol{x}')$ は，式(2.21)で表される LMC の共分散関数であり，「二点間 $(\boldsymbol{x}, \boldsymbol{x}')$ の共分散」を規定している．また，$k_{m,m'}(\boldsymbol{x}, \boldsymbol{x}')$ においては，重み $\boldsymbol{W}$ を通じて出力間 $(f_m(\boldsymbol{x}), f_{m'}(\boldsymbol{x}))$ の依存関係が導入されている．一方，式(3.8)の $\boldsymbol{C}_{m,m'}$ は，領域の各ペアに対して $k_{m,m'}(\boldsymbol{x}, \boldsymbol{x}')$ の二重積分を計算しており，これによって「二つの領域間 $(\mathcal{R}, \mathcal{R}')$ の共分散」を規定している．この二重積分は，直感的には，二つの領域に含まれる無限個の点に対する共分散を計算し，集約することを意味する．ここで，式(3.8)を実装する際には，カーネルの積分を数値的に近似計算する必要があることに注意する．

以上の定式化により，領域の形や大きさを考慮した共分散を評価可能であり，それを用いてデータの予測を実現する．集約データのための共分散行列 $\boldsymbol{C}$ を用いた周辺尤度に基づき，パラメータ最適化も可能である．

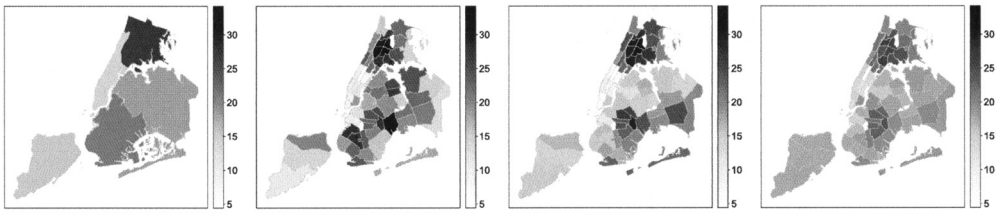
3.3 実験

本節では，2.2 節で述べた LMC と SAGP との比較実験について述べる．LMC は点に紐づくデータを前提としているため，集約データの値を領域の重心に紐付けることで適用した．潜在ガウス過程の共分散関数には ARD カーネルを用い，潜在ガウス過程の数は leave-one-out 交差検定により，$\{1, \ldots, M-1\}$ から決定した．モデルパラメータは，周辺尤度を最大にするように L-BFGS 法（Liu and Nocedal, 1989）を用いて最適化した．また，SAGP におけるカーネルの領域積分は，都市全体を 300m × 300m のグリッドで覆い，合計 9,352 個のグリッド点を用いて近似した．

実験で用いたデータは，ニューヨーク市が公開している 10 種類の集約データで，NYC Open Data（https://opendata.cityofnewyork.us）からダウンロードが可能である．表 1 に，実験で用いた集約データをまとめる．ここで，全てのデータの値を平均 0，分散 1 に正規化して使用した．本実験では，様々な解像度を持つ複数の集約データを用いて，集約データを高解像度化するタスクを解いた．高解像度な集約データの予測精度を評価するために，まずターゲットとする集約データを 1 つ決め，そのデータを低解像度化したもの（図 4(a)参照）とその他のデータを用いてモデルパラメータを学習し，学習済みモデルを用いて元の高解像度なデータを予測した．評価指標は，Mean absolute percentage error（MAPE）であり，m をターゲットの集約データの属性としたとき，

表 1. 集約データ．データは年に一度収集されており，複数年のデータがある場合は平均値を計算して実験に用いた．

データ	分割の仕方	領域数	取得年
PM2.5	UHF42	42	2009 – 2010
Poverty rate	Community district	59	2009 – 2013
Unemployment rate	Community district	59	2009 – 2013
Mean commute	Community district	59	2009 – 2013
Population	Community district	59	2009 – 2013
Recycle diversion rate	Community district	59	2009 – 2013
Crime	Police precinct	77	2010 – 2016
Fire incident	Zip code	186	2010 – 2016
311 call	Zip code	186	2010 – 2016
Public telephone	Zip code	186	2016



(a) 低解像度データ (b) 高解像度データ (c) SAGP の予測結果 (d) LMC の予測結果

図 4. Poverty rate データを高解像度化したときの可視化結果．

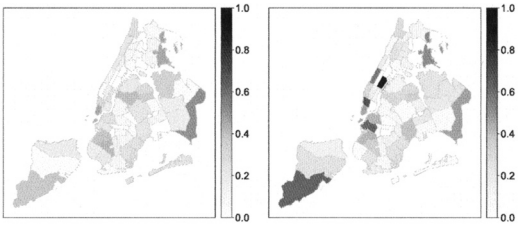
表 2. 高解像度データの予測に対する MAPE と標準誤差．括弧内の数字は交差検定によって選ばれた潜在ガウス過程の数 L を示す．太文字は t 検定(p 値は 0.05)において SAGP と LMC との間に有意差があることを示す．

	LMC	SAGP
PM2.5	0.036 ± 0.005 (6)	**0.030** ± **0.005** (5)
Poverty rate	0.207 ± 0.025 (4)	**0.177** ±**0.019** (3)
Unemployment rate	0.195 ± 0.024 (3)	**0.165** ±**0.020** (3)
Mean commute	0.057 ± 0.007 (4)	0.050 ± 0.007 (6)
Population	0.337 ± 0.039 (3)	**0.295** ± **0.033** (3)
Recycle diversion rate	0.222 ± 0.032 (4)	0.211 ± 0.029 (4)
Crime	0.401 ± 0.053 (2)	**0.379** ± **0.055** (3)
Fire incident	0.500 ± 0.052 (4)	**0.396** ± **0.038** (3)
311 call	0.061 ± 0.004 (6)	**0.052** ± **0.003** (3)
Public telephone	0.086 ± 0.008 (4)	0.080 ± 0.007 (6)

$$\frac{1}{N_m}\sum_{n=1}^{N_m}\left|\frac{y_{m,n}^{\mathrm{true}}-y_{m,n}^{*}}{y_{m,n}^{\mathrm{true}}}\right| \tag{3.9}$$

と表される．ここで，$y_{m,n}^{\mathrm{true}}$ は真の値を表し，$y_{m,n}^{*}$ は予測値を表す．SAGP の予測値は，式(3.4)の平均関数 $m^{*}(\boldsymbol{x},\boldsymbol{x}')$ を高解像度な各領域において積分することで求めた．

表 2 に SAGP と LMC に対する MAPE と標準誤差を示す．すべてのデータセットにおいて，SAGP は LMC と同等かそれ以上の精度を達成した．特に，表 2 における太文字は，t 検定の結果，SAGP と LMC との差が統計的に有意であったことを表す．また，図 4 に Poverty rate



(a) SAGP の予測誤差　　(b) LMC の予測誤差

図 5. Poverty rate データに対する予測誤差の可視化結果.

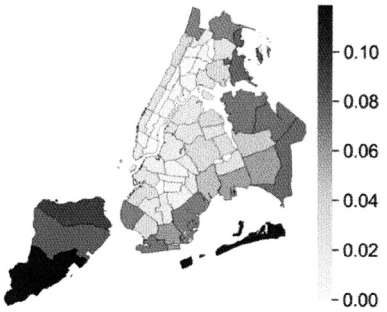


図 6. SAGP の予測分散.

データを高解像度化した際の可視化結果を示し，図 5 に SAGP と LMC に対する予測誤差の可視化結果を示す．ここで，予測誤差は Absolute percentage error であり，$\left|(y^{\mathrm{true}}_{m,n} - y^{*}_{m,n})/y^{\mathrm{true}}_{m,n}\right|$ を各領域について計算した．これらの結果より，SAGP が様々な種類の集約データを活用して，高解像度データをより正確に予測できることがわかる．

SAGP を用いると，領域に紐づくデータを予測した際の分散を評価することも可能である．図 6 に SAGP の予測分散を示す．予測分散は，式(3.5)を領域について積分することで計算できる．図 6 より，都市の端に位置する地域の分散が，都市の内部に位置する地域に比べて大きな値をとる傾向があることがわかる．一般に外挿は内挿よりも難しいので，合理的な結果であると言える．

4. 力学系のためのガウス過程

4.1 問題設定

本節では，力学系の時間発展を表すダイナミクスをガウス過程によりモデリングすることを考える．まず対象とする力学系について述べる．本稿では，解析力学の一形式である「ハミルトン力学」によって表現される系(ハミルトン系と呼ぶ)を対象とする．ハミルトン力学とは，ニュートン力学をエネルギーをベースに再定式化したものであり，ニュートン力学に基づく系を表現可能である．ハミルトン力学では，系の自由度を $M/2$ としたとき，一般化座標 $\boldsymbol{u}^{\mathrm{q}} \in \mathbb{R}^{M/2}$ と一般化運動量 $\boldsymbol{u}^{\mathrm{p}} \in \mathbb{R}^{M/2}$ を用いて，状態 $\boldsymbol{u} = (\boldsymbol{u}^{\mathrm{q}}, \boldsymbol{u}^{\mathrm{p}}) \in \mathbb{R}^{M}$ のダイナミクスを考える．$\boldsymbol{u}$ によって生成される空間を「相空間」(状態空間と同義)と呼ぶ．ハミルトン力学に基づく定式化では，系全体のエネルギー関数に対応する「ハミルトニアン」$H(\boldsymbol{u}) : \mathbb{R}^{M} \to \mathbb{R}$ を設計する．ハミルトニアンが与えられれば，ハミルトンの運動方程式にしたがって，ダイナミクスは，

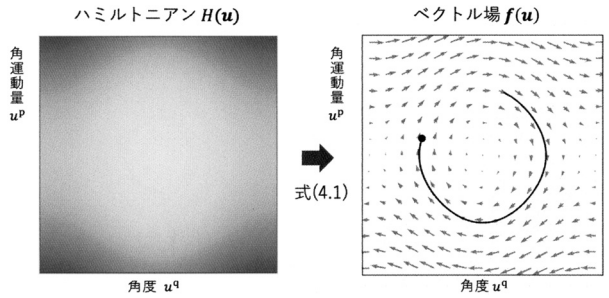


図 7. 単振り子のハミルトニアンとベクトル場．左図の色の濃淡はエネルギーの大きさを表す．式(4.1)によって定義されるベクトル場(右図)はシンプレクティック構造を持ち，任意の初期条件からの軌跡(黒線)はエネルギーの値を保存しつつ時間発展する．

$$\dot{\boldsymbol{u}} = (\boldsymbol{S} - \boldsymbol{R})\nabla H(\boldsymbol{u}) =: \boldsymbol{f}(\boldsymbol{u}) \tag{4.1}$$

と表される．ここで，$\dot{\boldsymbol{u}}$ は状態の時間微分 $d\boldsymbol{u}/dt$ を表し，$\nabla H(\boldsymbol{u}) : \mathbb{R}^M \to \mathbb{R}^M$ はハミルトニアンの状態 $\boldsymbol{u}$ についての勾配を表す．また，$\boldsymbol{S} \in \mathbb{R}^{M\times M}$ は歪対称行列

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{O} \end{pmatrix} \tag{4.2}$$

であり，$\boldsymbol{R} \in \mathbb{R}^{M\times M}$ はエネルギーの散逸を表す半正定値行列である．式(4.1)の右辺を「ベクトル場」と呼び，$\boldsymbol{f}(\boldsymbol{u}) : \mathbb{R}^M \to \mathbb{R}^M$ と定義する．$\boldsymbol{R} = \boldsymbol{O}$ のとき，ハミルトンの運動方程式によって定義されたベクトル場 $\boldsymbol{f}(\boldsymbol{u})$ は，シンプレクティック構造(Symplectic structure)という幾何学的構造を持ち，式(4.1)に従うダイナミクスはハミルトニアン $H(\boldsymbol{u})$ を保存(すなわちエネルギー保存)することが知られている．また，$\boldsymbol{R} = \mathrm{diag}(0, \ldots, 0, r_1, \ldots, r_{M/2})$ とすることで，摩擦係数を $r_d \geq 0$ とする加法的な散逸系を表すこともできる．

本稿で扱う問題を定義する．$\boldsymbol{\epsilon}$ を M 次元のガウスノイズとして，観測 $\dot{\boldsymbol{y}}$ が，$\dot{\boldsymbol{y}} = \dot{\boldsymbol{u}} + \boldsymbol{\epsilon}$ によって得られるとする．N 個の状態に対する微分観測(derivative observation)$\{(\boldsymbol{u}_n, \dot{\boldsymbol{y}}_n) \mid n = 1, \ldots, N\}$ が与えられたとしたとき，ダイナミクスを表すベクトル場 $\boldsymbol{f}(\boldsymbol{u})$ を推定する問題を扱う．推定したベクトル場と，時刻 $t = 0$ における任意の初期条件 $\boldsymbol{u}_0$ が与えられたとき，

$$\boldsymbol{u}(t) = \boldsymbol{u}_0 + \int_0^t \boldsymbol{f}(\boldsymbol{u})dt \tag{4.3}$$

を数値的に求めることで，力学系のシミュレーションが可能である．

以下にハミルトン系の例として単振り子のハミルトニアンとベクトル場について述べる．

例(単振り子)．図 7 に，単振り子を例としたときのハミルトニアンとベクトル場を可視化する．単振り子の自由度は 1 であり，u^{q} は振り子の角度を，u^{p} は振り子の角運動量を表す．単振り子のハミルトニアンは，

$$H(\boldsymbol{u}) = 2mg\ell(1 - \cos u^{\mathrm{q}}) + \frac{\ell^2 (u^{\mathrm{p}})^2}{2m} \tag{4.4}$$

と表される．ここで，g は重力定数，m は質量，ℓ は振り子の長さを表す．図 7 の可視化では，$g = 3$ および $m = \ell = 1$ とし，$\boldsymbol{R} = \boldsymbol{O}$ とした．

4.2 手法

ベクトル場 $\boldsymbol{f}(\boldsymbol{u})$ は，素朴には 2.2 節で述べた多出力ガウス過程によりモデリングすること

が可能である．しかし，単純なガウス過程では，物理法則を満たすような学習結果が必ずしも得られるとは限らない．そこで，本節では，ハミルトン力学の理論を組み込むことで，エネルギーの保存・散逸則を満たすように制約を加える方法について述べる．

Rath et al.（2021）や Tanaka et al.（2022）では，ベクトル場を直接モデリングする代わりに，ハミルトニアン $H(\boldsymbol{u})$ を単一出力を持つガウス過程を用いて，

$$H(\boldsymbol{u}) \sim \mathcal{GP}\left(0, k(\boldsymbol{u}, \boldsymbol{u}')\right) \tag{4.5}$$

とモデル化する．式(4.1)における微分作用素を $\mathcal{L} := (\boldsymbol{S} - \boldsymbol{R})\nabla$ とすると，ベクトル場は $\boldsymbol{f}(\boldsymbol{u}) = \mathcal{L}H(\boldsymbol{x})$ によって与えられる．微分作用素 $\mathcal{L}$ は線形であるため，$\boldsymbol{f}(\boldsymbol{u})$ は多出力ガウス過程で表される（Solak et al., 2003）．このとき，行列値カーネルは，

$$\boldsymbol{K}(\boldsymbol{u}, \boldsymbol{u}') = \mathcal{L}k(\boldsymbol{u}, \boldsymbol{u}')\mathcal{L}^\top \tag{4.6}$$

$$= (\boldsymbol{S} - \boldsymbol{R})[\nabla^2 k(\boldsymbol{u}, \boldsymbol{u}')](\boldsymbol{S} - \boldsymbol{R})^\top \tag{4.7}$$

と表される．ここで，∇^2 は Hessian operator を表す．式(4.7)は，シンプレクティックカーネル(Symplectic kernel)と呼ばれ（Boffi et al., 2022），シンプレクティックカーネルに基づくガウス過程回帰のことをシンプレクティックガウス過程(symplectic Gaussian process: SGP)（Rath et al., 2021）と呼ぶ．式(4.6)の右辺は，式(2.20)の LMC に基づく行列値カーネルと同様の形をしていることがわかる．LMC における線形変換 $\boldsymbol{W}$ が線形作用素 $\mathcal{L}$ に置き換わっているが，線形作用素は線形変換を無限次元に一般化したものとみなせることに注意すれば，式(4.7)のシンプレクティックカーネルは正定値カーネルであることがわかるであろう．以上より，ハミルトン力学の理論を活用し，エネルギーの保存・散逸則に従うダイナミクスを表現するのに適した多出力ガウス過程が導出された．カーネルパラメータの学習や予測は，2.2 節で述べた方法で実行可能である．

本稿では割愛するが，Tanaka et al.（2022）や Boffi et al.（2022）では，シンプレクティックカーネルをフーリエ基底によって近似するためのシンプレクティック乱択化フーリエ特徴(symplectic random Fourier feature: S-RFF)が提案されており，サンプル数が大きいときに学習の効率化を図ることができる．さらに，Tanaka et al.（2022）では，数値ソルバを活用した変分学習アルゴリズムを開発することにより，状態の軌跡だけが与えられる状況でもハミルトン系の学習を可能とした．詳細は（田中, 2023）にも解説されている．

4.3 実験

本節では，SGP の評価を行うため，ハミルトニアンニューラルネットワーク(Hamiltonian neural network: HNN)（Greydanus et al., 2019）との比較を行う．HNN は，ハミルトニアンをニューラルネットワークにより近似し，自動微分を活用してベクトル場を導出する方法である．HNN は 3 層の多層パーセプトロンを用いて実装し，ユニット数は 200，活性化関数は tanh を用いた．SGP および HNN におけるパラメータ推定には，Adam（Kingma and Ba, 2015）を用い，学習率は $1e-3$ とした．

実験では，4.1 節で述べた単振り子，および，調和振動子を対象にした．調和振動子のハミルトニアンは，

$$H(\boldsymbol{u}) = \frac{1}{2}k(u^{\mathrm{q}})^2 + \frac{(u^{\mathrm{p}})^2}{2m} \tag{4.8}$$

と表される．ここで，バネ定数 $k = 1$，質量 $m = 1$ に設定した．また，どちらの系も摩擦はないものとし，$\boldsymbol{R} = \boldsymbol{O}$ とした．学習のために $N = \{40, 60, 100\}$ 個のサンプル $\{(\boldsymbol{u}_n, \dot{\boldsymbol{y}}_n)\}$ を作成

表 3. ベクトル場に対する MSE と標準偏差．太文字は t 検定(p 値は 0.05)において SGP と HNN の間に有意差があることを示す．

(a) 調和振動子

サンプル数	40	60	100
HNN	0.600±0.098	0.476±0.048	0.368±0.034
SGP	**0.041±0.053**	**0.040±0.021**	**0.011±0.001**

(b) 単振り子

サンプル数	40	60	100
HNN	2.253±0.107	1.996±0.099	1.863±0.089
SGP	**0.323±0.161**	**0.237±0.056**	**0.200±0.030**

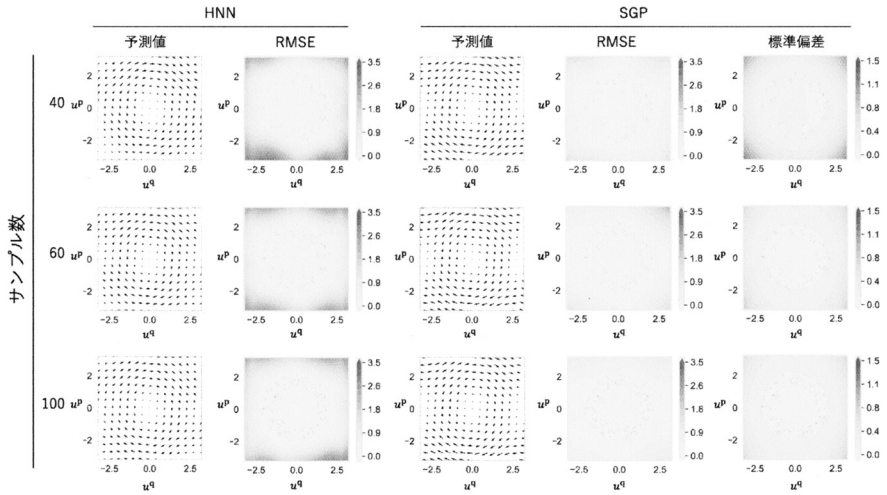


図 8. 予測されたベクトル場の可視化．灰色の点は学習点を表す．

した．ここで，状態 $\boldsymbol{u}_n$ は，エネルギーの値が $[1.3, 2.0]$ の範囲に含まれるように一様分布から生成した．また，観測 $\boldsymbol{y}_n$ は，ハミルトンの運動方程式により $\dot{\boldsymbol{u}}_n = \boldsymbol{S}\nabla H(\boldsymbol{u}_n)$ を計算し，ノイズ分散を $\sigma^2 = 0.1$ とするガウスノイズを加えることで生成した．

テスト時には，$[-3.2, 3.2]^2$ で定義された相空間を 15×15 のグリッドに区切り，テスト点を $\{\boldsymbol{x}_i^* \mid i = 1, \ldots, I\}$ とした．ここで，I はテストデータのサンプル数である．評価指標は mean squared error(MSE)であり，

$$\frac{1}{I}\sum_{i=1}^{I} \| \dot{\boldsymbol{u}}_i^{\mathrm{pred}} - \dot{\boldsymbol{u}}_i^{\mathrm{true}} \|^2 \tag{4.9}$$

と表される．ここで，$\dot{\boldsymbol{u}}_i^{\mathrm{true}}$ は各テスト点における真のダイナミクスを表し，$\dot{\boldsymbol{u}}_i^{\mathrm{pred}}$ はその予測値を表す．

表 3 に，HNN と SGP によって予測されたベクトル場に対する MSE と標準偏差を示す．この結果から，SGP がすべての場合において HNN よりも高精度にベクトル場の予測ができていることがわかる．SGP はサンプル数が少ない場合に特に有効であることがわかる．これはガウス過程が不確実性を考慮できるモデルであり，ノイズや少量データに有効であるためである．図 8 に，単振り子のデータに対して HNN と SGP が予測したベクトル場，RMSE(Root mean squared error)，および，SGP の予測に対する標準偏差を示す．真のベクトル場(図 7 の右図)

とその予測値(図 8)を比較すると，特にサンプル数が少ない場合に SGP が HNN よりも高精度であることがわかる．SGP の予測に対する標準偏差はサンプル数をよく反映しており，RMSE と比較することで，標準偏差が予測値の信頼度を正確に表現していることがわかる．さらに，図 8 の RMSE からわかるように，学習点がない領域においても，SGP は HNN よりも誤差が小さくロバストであることがわかる．

5. 結論

本稿では，ガウス過程の基礎事項，特に多出力ガウス過程について解説し，それに関連する二つの研究トピックについて紹介した．一つ目は，集約データに対するガウス過程であり，点ではなく領域に紐づくデータのモデリングである．ガウス過程の領域積分により集約データを表現することで，事後分布が解析的に求まることについて述べた．これにより，領域の形や大きさを考慮した共分散関数を定義でき，それに基づいた予測が可能となる．二つ目は，力学系のためのガウス過程であり，ダイナミクスを表現するベクトル場のモデリングである．ハミルトン力学の理論に基づき，エネルギー関数をガウス過程でモデリングし，ハミルトンの運動方程式にしたがって導出されたガウス過程は，エネルギーの保存・散逸則を満たすことについて述べた．二つのトピックにおいて共通かつ重要であることは，各トピックにおけるモデリングを通じて，多出力ガウス過程の行列値カーネルにデータ集約過程や物理法則といった事前知識が埋め込まれたことである．これにより，データと知識とを同時に活用し，高精度な予測を可能とした．近年は深層学習のブームが続いている状況であるが，データに含まれる不確実性を考慮でき，ノイズや少量データからの学習に強みがあるガウス過程は，今後も注目すべき機械学習手法の一つであると言える．

参 考 文 献

Álvarez, M. A., Rosasco, L. and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review, *Foundations and Trends® in Machine Learning*, **4**(3), 195–266.

Beckers, T., Seidman, J., Perdikaris, P. and Pappas, G. J. (2022). Gaussian process port-Hamiltonian systems: Bayesian learning with physics prior, *2022 IEEE 61st Conference on Decision and Control*, 1447–1453, https://doi.org/10.1109/CDC51059.2022.9992733.

Boffi, N. M., Tu, S. and Slotine, J.-J. E. (2022). Nonparametric adaptive control and prediction: Theory and randomized algorithms, *Journal of Machine Learning Research*, **23**(281), 1–46.

Burgess, T. M. and Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties I: The semi-variogram and punctual kriging, *Journal of Soil Science*, **31**(2), 315–331.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J. and Duvenaud, D. K. (2018). Neural ordinary differential equations, *Advances in Neural Information Processing Systems*, **31**, https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

福水健次 (2010).『カーネル法入門—正定値カーネルによるデータ解析—』, 朝倉書店, 東京.

Goldstein, H. (1980). *Classical Mechanics*, Addison-Wesley, Reading, Massachusetts.

Greydanus, S., Dzamba, M. and Yosinski, J. (2019). Hamiltonian neural networks, *Advances in Neural Information Processing Systems*, **32**, https://proceedings.neurips.cc/paper_files/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

Hamelijnck, O., Damoulas, T., Wang, K. and Girolami, M. (2019). Multi-resolution multi-task Gaussian processes, *Advances in Neural Information Processing Systems*, **32**, https://proceedings.neurips.cc/paper_files/paper/2019/file/0118a063b4aae95277f0bc1752c75abf-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

畑浩之 (2014).『解析力学，基幹講座物理学』, 東京図書, 東京.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization, *International Conference on Learning Representations*, 1–15.

Law, H. C. L., Sejdinovic, D., Cameron, E., Lucas, T. C. D., Flaxman, S., Battle, K. and Fukumizu, K. (2018). Variational learning on aggregate outputs with Gaussian processes, *Advances in Neural Information Processing Systems*, **31**, https://proceedings.neurips.cc/paper_files/paper/2018/file/24b43fb034a10d78bec71274033b4096-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, **45**(1–3), 503–528.

Macêdo, I. and Castro, R. (2010). Learning divergence-free and curl-free vector fields with matrix-valued kernels, Technical Report, Instituto Nacional de Matematica Pura e Aplicada, Brazil.

Narcowich, F. J. and Ward, J. D. (1994). Generalized hermite interpolation via matrix-valued conditionally positive definite functions, *Mathematics of Computation*, **63**(208), 661–687.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Massachusetts.

Rath, K., Albert, C. G., Bischl, B. and von Toussaint, U. (2021). Symplectic Gaussian process regression of maps in Hamiltonian systems, *Chaos*: *An Interdisciplinary Journal of Nonlinear Science*, **31**(5), https://doi.org/10.1063/5.0048129.

Smith, M. T., Álvarez, M. A. and Lawrence, N. D. (2018). Gaussian process regression for binned data, arXiv, https://arxiv.org/abs/1809.02010（最終アクセス日 2024 年 11 月 22 日）.

Solak, E., Murray-smith, R., Leithead, W., Leith, D. and Rasmussen, C. (2003). Derivative observations in Gaussian process models of dynamic systems, *Advances in Neural Information Processing Systems*, **15**, https://proceedings.neurips.cc/paper_files/paper/2002/file/5b8e4fd39d9786228649a8a8bec4e008-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

田中佑典 (2023). ガウス過程と物理現象のモデル化, 人工知能, **38**(3), 318–325.

Tanaka, Y. (2024). Learning Hamiltonian dynamics under uncertainty via symplectic Gaussian processes, *International Conference on Scientific Computing and Machine Learning*, https://scml.jp/2024/paper/15/CameraReady/scml2024.pdf（最終アクセス日 2024 年 11 月 22 日）.

Tanaka, Y., Iwata, T., Tanaka, T., Kurashima, T., Okawa, M. and Toda, H. (2019a). Refining coarse-grained spatial data using auxiliary spatial data sets with various granularities, *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 5091–5100.

Tanaka, Y., Tanaka, T., Iwata, T., Kurashima, T., Okawa, M., Akagi, Y. and Toda, H. (2019b). Spatially aggregated Gaussian processes with multivariate areal outputs, *Advances in Neural Information Processing Systems*, **32**, https://proceedings.neurips.cc/paper_files/paper/2019/file/a941493eeea57ede8214fd77d41806bc-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

Tanaka, Y., Iwata, T. and Ueda, N. (2022). Symplectic spectrum Gaussian processes: Learning Hamiltonians from noisy and sparse data, *Advances in Neural Information Processing Systems*, **35**, https://proceedings.neurips.cc/paper_files/paper/2022/file/82f05a105c928c10706213952bf0c8b7-Paper-Conference.pdf（最終アクセス日 2024 年 11 月 22 日）.

Teh, Y. W., Seeger, M. and Jordan, M. I. (2005). Semiparametric latent factor models, *International Conference on Artificial Intelligence and Statistics*, 333–340.

Yousefi, F., Smith, M. T. and Álvarez, M. A. (2019). Multi-task learning for aggregated data using Gaussian processes, *Advances in Neural Information Processing Systems*, **32**, https://proceedings.neurips.cc/paper_files/paper/2019/file/64517d8435994992e682b3e4aa0a0661-Paper.pdf（最終アクセス日 2024 年 11 月 22 日）.

Spatio-temporal Data Analysis Using Gaussian Processes

Yusuke Tanaka[1] and Naonori Ueda[2]

[1]NTT Communication Science Laboratories
[2]RIKEN Center for Advanced Intelligence Project

A Gaussian process (GP) is a probabilistic model for estimating a function from data and has long been widely used in spatio-temporal data modeling. This paper addresses two significant problems related to spatio-temporal data analysis and GP models. The first problem is the modeling of aggregate data. Due to privacy concerns and administrative reasons, spatio-temporal data (e.g., poverty rates) obtained in cities are not associated with points but with regions. This paper introduces a way to handle aggregate data by an integral of the GPs over regions. The second problem is the modeling of dynamical systems defined by ordinary differential equations. The system's dynamics are defined by the time derivative of the state and are represented as a vector field in phase space. Although vector fields can be modeled naively using multi-output GPs, the vanilla GPs do not always satisfy physical laws. In this paper, we introduce a method for modeling vector fields that adhere to energy conservation or dissipation laws by combining the theory of Hamiltonian mechanics with GPs.

Key words: Gaussian processes, spatio-temporal data, aggregate data, dynamical systems, Hamiltonian mechanics.

統計数理(2025)
第 73 巻 第 1 号 53–63
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
[研究詳解]

ベイズ的モデル統合による空間予測

菅澤 翔之助†

(受付 2024 年 6 月 30 日；改訂 9 月 8 日；採択 9 月 9 日)

要　　旨

有限地点で観測された空間データからモデルを推定し，未観測地点を予測することは空間データ分析において重要なタスクの 1 つである．近年では，古典的な地球統計学や空間計量経済学的モデルから機械学習的アプローチまで様々なモデルが利用可能である．そのため，解析対象のデータに対して適切に分析手法を取捨選択することは分析上の重要な論点である．本稿では，予測モデルが複数個得られる状況においてベイズ的にモデルを統合した空間予測の方法論についてのレビューを行う．特に，近年提案されたベイズ的空間予測統合に関して，古典的な統合方法との違いや具体的な推定アルゴリズムに関する解説を行う．

キーワード：ベイズモデル平均，スタッキング，ガウス過程，ベイズ的予測統合．

1. はじめに

近年ではセンシング技術やデータ収集技術の発展に伴い，位置情報が付随した多様な空間データを用いた分析が可能になっている．空間データ分析の主要な目的の一つとして，未観測地点(観測データが得られていない地点)における値の予測が挙げられる．そのための方法として，クリギングと呼ばれる地球統計学的な手法 (e.g. Oliver and Webster, 1990)や空間自己回帰モデルなどの空間計量経済学的方法 (e.g. Anselin, 2022)，さらには地理的加重回帰 (e.g. Fotheringham et al., 2002)や階層ベイズ的アプローチ (e.g. Banerjee et al., 2014)などが知られている．近年では，勾配ブースティング木やランダムフォレスト，深層学習などの機械学習的アプローチも採用されるようになってきた (e.g. Du et al., 2020)．

候補として複数の予測モデルが考えられる場合，大きく 2 つの方向性が考えられる．1 つは何らかの規準に基づいて 1 つの最適なモデルを選択する方法である．代表的なアプローチとして交差検証法や情報量規準があるが，このような方法の弱点として，モデルを選択するステップにおける誤差や不確実性を評価することが難しい点が挙げられる．もう 1 つの方向性として，候補モデルの中から 1 つのモデルのみを選ぶ代わりに，各モデルごとに何らかのウエイトを与えて合成するアプローチがある．これは一般的にモデル平均と呼ばれており，モデル選択に基づく方法よりも予測精度が高くなる傾向があることが知られている．本稿では，空間予測において複数モデルが得られている状況でそれらを合成するアプローチに焦点を当て，特にベイズ的な枠組みで予測モデルを統合する方法論を解説する．標準的なアプローチとして，ベイズモデル平均やスタッキングを紹介し，さらにこれらの古典的なモデル統合手法を含む一般的な枠組みとしてベイズ的予測統合(Bayesian predictive synthesis)を紹介し，空間的なモデル重要度の異質性を考慮した新しい統合手法 (Cabel et al., 2022)を解説する．本稿は，従来のモデ

† 慶應義塾大学 経済学部：〒 108–8345 東京都港区三田 2–15–45

ル統合の手法から最近の統合手法についてある程度の詳細まで含めて解説することに重きを置いており，シミュレーションや実データにおける数値的な詳細に関しては Cabel et al.（2022）などの数値例を参照されたい．

2. ベイズ的モデル統合の古典的なアプローチ

2.1 ベイズモデル平均（Bayesian model averaging）

モデル統合の古典的かつ代表的な方法としてベイズモデル平均化（Bayesian model averaging, BMA）が知られている（cf. Hoeting et al., 1999）．これはモデルに対する事後確率による重み付き平均を考える方法で，空間データに限らず一般のモデルに対して使える方法である．$M_1, \ldots, M_K$ を候補モデルとし，$y = (y_1, \ldots, y_n)$ を観測データとする．さらに，モデル M_k のパラメータを θ_k（k ごとに次元が異なる可能性がある）とし，その事前分布を $p(\theta_k|M_k)$ とする．BMA は未観測データ $\tilde{y}$ に対する予測分布を

$$p(\tilde{y}|y) = \sum_{k=1}^{K} p(\tilde{y}|y, M_k)p(M_k|y)$$

で与える．ここで，$p(M_k|y)$ はモデル M_k の事後確率

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{k'=1}^{K} p(y|M_{k'})p(M_{k'})}$$

である．また，$p(M_k)$ はモデルの事前確率で，$p(y|M_k)$ はモデル M_k の周辺尤度（marginal likelihood）

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$$

である．モデルの事前確率として一様事前分布 $p(M_k) = 1/K$ を考えた場合，モデル M_k の事後確率は周辺尤度に比例して決まることになる．すなわち，BMA は周辺尤度の大きさ（モデルのエビデンスの強さ）に応じて決まるウエイトによって予測分布を重み付き平均している方法であると解釈することができる．BMA は候補モデル $M_1, \ldots, M_K$ のどれかが真のデータ生成過程に一致している（どれかはわからない）状況では妥当な方法として知られているが，そうでない状況では漸近的に Kullback-Leibler（KL）ダイバージェンスが最も小さくなるモデルに対するウエイトが 1 になる（それ以外のモデルは 0 になってしまう）問題点が知られている．

空間データ分析の文脈では，複数の空間自己回帰モデルを BMA によって統合するアプローチが Debarsy and LeSage（2022）や LeSage and Parent（2007）で検討されている．具体例として，空間誤差モデルにおいて異なる説明変数行列を用いた複数のモデルを BMA で統合するケースを考えてみる．y を n 次元の被説明変数ベクトル，X を $n \times p$ の説明変数行列，W を $n \times n$ の空間重み行列として，空間誤差モデルは

$$y = X\beta + e, \quad e = \rho We + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

と表せる．このモデルは行列 $Q(\rho) = (I_n - \rho W)^\top (I_n - \rho W)$ を用いて $y \sim N(X\beta, \sigma^2 Q(\rho)^{-1})$ と表すことができる．BMA を求めるためには周辺尤度 $p(y)$ を求める必要があるが，それは

$$p(y) = \int \phi_n(y; X\beta, \sigma^2 Q(\rho))\pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\rho)d\rho$$

で与えられる．ただし，$\pi(\beta|\sigma^2), \pi(\sigma^2), \pi(\rho)$ は β, σ^2, ρ の（条件付き）事前分布である．具体的には，β に対して g 事前分布 $\pi(\beta|\sigma^2) \sim N(0, \sigma^2\{gX^\top Q(\rho)X\}^{-1})$ および σ^2 に対して逆ガンマ事前分布 $\pi(\sigma^2) \sim \mathrm{IG}(\nu/2, \nu s^2/2)$ を用いる．すると周辺尤度は

$$p(y) = K\left(\frac{g}{1+g}\right)^{p/2}\int |Q(\rho)|^{1/2}\left[\nu s^2 + \frac{g}{1+g}\{y - X\hat{\beta}(\rho)\}Q(\rho)\{y - X\hat{\beta}(\rho)\}\right]^{-\frac{n+\nu-1}{2}}\pi(\rho)d\rho$$

となる．ただし，

$$K = \frac{\Gamma\left(\frac{n+\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}(\nu s^2)^{\nu/2}\pi^{-\frac{n-1}{2}}, \qquad \hat{\beta}(\rho) = \{X^\top Q(\rho)X\}^{-1}X^\top Q(\rho)y$$

である．$p(y)$ を求めるには $\pi(\rho)$ について積分をする必要があるが，一次元の有界区間 $(-1, 1)$ 上の積分のため比較的容易に数値積分を行うことができる．

2.2 スタッキング（Stacking）

スタッキング(Stacking)は複数モデルからの推定値を統合するアルゴリズムである（cf. Breiman, 1996）．$y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ をデータ y から y_i のみを除外したデータセットとし，$\hat{\mu}_k(y_{-i})$ をモデル M_k と y_{-i} に基づく未観測データ $\tilde{y}$ の leave-one-out（LOO）予測値とする．このとき，$\tilde{y}$ を $\hat{\mu}_k(y_{-i})$ の凸結合 $\sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})$（ただし，$w_k$ は $\sum_{k=1}^K w_k = 1$ を満たす)によって求めることを考え，最適なウエイト $w = (w_1, \ldots, w_K)$ を

$$\sum_{i=1}^n\left\{y_i - \sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})\right\}^2$$

を最小化するように決める．一般的に，二乗損失によって予測精度を比較した場合，BMA よりもスタッキングの方が予測精度が高くなることが知られている（Clarke, 2003）．

Clydec and Iversen（2013)は，スタッキングのアルゴリズムに対してベイズ的な解釈を与えている．$p_*(\tilde{y}|y)$ を真のデータ生成過程によって構成される予測分布(真の予測分布と呼ぶ)として，未観測データ $\tilde{y}$ を $\sum_{k=1}^K w_k\hat{\mu}_k(y)$ で予測することを考える．ここで，$\hat{\mu}_k(y)$ は予測分布の期待値

$$\int \tilde{y}p(\tilde{y}|\theta_k, M_k)p(\theta_k|y, M_k)d\theta_k d\tilde{y}$$

やパラメータの事後平均をプラグインした期待値

$$\int \tilde{y}p(\tilde{y}|\hat{\theta}_k, M_k)d\tilde{y}, \qquad \hat{\theta}_k = E[\theta_k|y, M_k]$$

などが考えられる．このとき，合成した推定量の期待二乗損失は

$$\int\left\{\tilde{y} - \sum_{k=1}^K w_k\hat{\mu}_k(y)\right\}^2 p_*(\tilde{y}|y)d\tilde{y} \tag{2.1}$$

と表せる．この量を最小にする w が最適な重みになるが，真の予測分布 $p_*(\tilde{y}|y)$ が未知のため上記の積分を評価することができない．その代わりに，スタッキングの考え方を使って積分の近似を構成することができる．最適なウエイトを LOO 二乗損失

$$\frac{1}{n}\sum_{i=1}^n\left\{y_i - \sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})\right\}^2 \tag{2.2}$$

を最小化するように決める．このとき，LOO 二乗損失(2.2)が漸近的に期待二乗損失(2.1)に一致することが Le and Clarke (2017)によって示されている．

従来のスタッキングは点予測値を合成する手法であるが，Yao et al. (2018)では，スタッキングを用いた予測分布の統合手法が提案されている．

$$p(\tilde{y}|y, M_k) = \int p(\tilde{y}|\theta_k, M_k)p(\theta_k|y, M_k)d\theta_k$$

をモデル M_k に基づく予測分布として，凸結合した予測分布 $\sum_{k=1}^{K} w_k p(\tilde{y}|y, M_k)$ を考える．また，LOO 予測分布を

$$p(y_i|y_{-i}, M_k) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$$

と定める．このとき，最適なウエイト $w_1, \ldots, w_K$ は以下の最適化問題によって求めることができる．

$$\max_{w} \frac{1}{n}\sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} w_k p(y_i|y_{-i}, M_k)\right\}, \quad \text{s.t.} \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1.$$

この方法を実行する場合の計算上の問題点として，LOO 予測分布 $p(y_i|y_{-i}, M_k)$ を n 回計算する必要がある点が挙げられる．最も単純に実行する場合は，各 y_{-i} ごとに θ_k の事後分布を(マルコフ連鎖モンテカルロ法などで)計算する必要がある．Yao et al. (2018)では，この問題に対して重点サンプリングによる効率的な近似計算手法を与えている．全データによる事後分布 $p(\theta_k|y, M_k)$ を用いて，LOO 予測分布は

$$p(y_i|y_{-i}, M_k) = \int p(y_i|\theta_k, M_k)\frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)}p(\theta_k|y, M_k)d\theta_k$$

と表すことができるため

$$\frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)} \propto \frac{1}{p(y_i|\theta_k, M_k)} \equiv w(\theta_k)$$

を重点ウエイトとした重点サンプリングによって計算することができる．$w(\theta_k)$ は密度の逆数で与えられるため，数値計算が不安定になってしまう問題があるが，Yao et al. (2018)ではパレート平滑化重点サンプリング (e.g. Vehtari et al., 2024)を用いることで安定的な計算方法を提案している．

近年では，スタッキングの考え方を用いて階層空間回帰モデルの推定自体に利用する方法も開発されている．X を $n \times p$ の計画行列として，以下のような階層空間回帰モデルを考える．

$$y|z \sim N(X\beta + u, \delta^2\tau^2 I_n), \quad u \sim N(0, \tau^2 C(\phi)).$$

ここで，u が潜在的な空間効果を表す n 次元ベクトルであり，$C(\phi)$ はパラメータ ϕ によって定まる相関行列である．Zhang et al. (2023)は，パラメータ (δ^2, ϕ) が所与のもとで事後分布が解析的に計算できる点に注目し，スタッキングを用いた高速な事後分布の近似方法を与えている．具体的には，(δ^2, ϕ) に対する候補値の集合を L セット用意しておき，それぞれを 1 つのモデル M_l $(l = 1, \ldots, L)$ とみなすことで，スタッキングのアルゴリズムを適用して最適なウエイト $\hat{w}_1, \ldots, \hat{w}_L$ を計算している．その結果，例えば β に対する合成された事後分布(stacked posterior)は $\sum_{l=1}^{L} \hat{w}_l p(\beta|y, M_g)$ の形($p(\beta|y, M_g)$ はモデル M_g のもとでの β の周辺事後分布)で与えられる．

2.3 古典的なアプローチの限界

これまでに紹介した手法は，点予測や予測分布を定数のウエイトによる凸結合の形で統合する方針を採用している．しかし，空間データを用いた予測タスクにおいてはモデルの重要度が地点によって異なることが想定され，既存のモデル統合のアプローチでは精度の高い空間予測を実現することは難しいと考えられる．このようなモデルウエイトの空間異質性を考慮したモ

デル統合を実現するため，Cabel et al. (2022)では，ベイズ的予測統合と呼ばれる枠組みを用いた空間モデルの統合手法を導入している．

3. ベイズ的空間予測統合(Bayesian spatial predictive synthesis)

3.1 ベイズ的予測統合

s を緯度・経度などの位置情報，$y(s)$ を地点 s における確率変数とする．n 個の地点 $s_1,\dots,s_n$ におけるデータ $y(s_1),\dots,y(s_n)$ を用いて，任意の地点 s における未観測な確率変数 $y(s)$ を予測する問題を考える．このデータに対して J 個の統計モデルをそれぞれ並列に適用することで，J 個の予測分布 $h_j(\cdot)$ $(j=1,\dots,J)$ および観測地点において予測分布に従う確率変数 $f_j(s_i)$ を得ることができる．$h_j(\cdot)$ は様々な分布があり得るが，典型的な例としては予測平均と分散を持つ正規分布である．J 個の予測分布が与える情報集合を $\mathcal{H}(s)=\{h_1(f_1(s)),\dots,h_J(f_J(s))\}$ と定義する．ここで, $f_j(s)$ は地点 s における確率変数である．このとき，情報集合 $\mathcal{H}(s)$ からどのように $y(s)$ の予測分布を構成するかが問題となる．

この問いに対する理論的な答えとして，ベイズ的予測統合(Bayesian predictive synthesis, BPS)の枠組み（e.g. Genest and Schervish, 1985; West and Crosse, 1992; West, 1992)において一般的かつコヒーレントな分布形が与えられている．BPS によって与えられる事後分布は以下のような形になる．

$$\Pi_{\mathrm{BPS}}\big(y(s)|\Psi(s),\mathcal{H}(s)\big)=\int \alpha\big(y(s)|f(s),\Psi(s)\big)\prod_{j=1}^{J}h_j\big(f_j(s)\big)df_j(s). \tag{3.1}$$

ここで，$\alpha\big(y(s)|f(s),\Psi(s)\big)$ は統合関数，$\Psi(s)$ は地点ごとのパラメータ，$f(s)=(f_1(s),\dots,f_J(s))$ は潜在変数のベクトルである．統合関数は複数のモデルをどのように統合するかを決定する関数であるが，一般的な理論では具体的な形までは指定されておらず，問題に応じて適切な統合関数を与える必要がある．例えば，統合関数として

$$\alpha\big(y(s)|f(s),\Psi(s)\big)=\sum_{j=1}^{J}w_j\delta_{f_j(s)}(y(s))$$

を考える．ただし，$\delta_a(y)$ は $y=a$ 上の Dirac 測度である．このとき，$\Pi_{\mathrm{BPS}}\big(y(s)|\Psi(s),\mathcal{H}(s)\big)=\sum_{j=1}^{J}w_jh_j(y(s))$ となり，2.2 節で扱ったような J 個の密度関数の重み付き平均の形になる．したがって，BPS による統合分布の形(3.1)は，既存のモデル平均化を含んだ一般的な枠組みになっていることがわかる．空間予測の問題で BPS を使う場合は，どのような形の統合関数を与えるかが肝となるが，Cabel et al. (2022)では

$$\alpha(y(s)|f(s),\Psi(s))=\phi\left(y(s);\beta_0(s)+\sum_{j=1}^{J}\beta_j(s)f_j(s),\sigma^2\right) \tag{3.2}$$

という形の統合関数を提案している．ここで，$\Psi(s)=\{\beta_0(s),\beta_1(s),\dots,\beta_J(s),\sigma^2\}$ である．この形の統合関数の妥当性として，Cabel et al. (2022)では，確率変数 $f(s)_1,\dots,f_J(s)$ が与えられたもとでの $y(s)$ の予測問題を考えたときに，ある条件のもとで最良近似を与える形であることが示されている．統合関数(3.2)を利用した予測統合手法はベイズ的空間予測統合(Bayesian spatial predictive synthesis, BSPS)と呼ぶ．

統合関数(3.2)の重要な性質として，$f_j(s)$ に対する重み $\beta_j(s)$ が地点に依存している点が挙げられる．これは，場所によって各モデルの重要度が異なる可能性があること(モデル統合における空間異質性)を考慮した形になっている．また，J 個のモデルが統合されているのに加

えて地点に依存する切片項 $\beta_0(s)$ が付随している．これは，もし統合された項 $\sum_{j=1}^{J}\beta_j(s)f_j(s)$ が $y(s)$ の変動を捉えるのに十分なほど柔軟ではなかった場合に，空間トレンド $\beta_0(s)$ として柔軟さを補う役割を果たしている．さらに，ウエイトに対して $\sum_{j=1}^{J}\beta_j(s)=1$ などの制約は与えておらず，地点によっては負の値になることも許した定式化になっている．このような性質は，2 節で紹介した古典的な方法論とは大きく異なる特徴である．

3.2 BSPS の推定アルゴリズム

BSPS を実行するためには，統合関数のパラメータ $\Psi(s)$ を推定する必要がある．$\beta_j(s)$ は位置情報 s の関数であるが，ガウス過程事前分布を用いることで推定が可能となる．$i=1,\ldots,n$ に対して，$y_i=y(s_i)$, $\beta_{ji}=\beta_j(s_i)$, $f_{ji}=f_j(s_i)$ とする．以下では簡単のため，各モデルの予測分布として予測値と予測分散をそれぞれ平均・分散に持つ正規分布を想定する．すなわち，$f_{ji}\sim N(a_{ji},b_{ji})$ を仮定する．このとき，BSPS によって統合事後分布を求めることは，以下の潜在変数モデルを推定することと同値である．

$$y_i=\beta_{0i}+\sum_{j=1}^{J}\beta_{ji}f_{ji}+\epsilon_i,\quad (\beta_{j1},\ldots,\beta_{jn})\sim N(m_j1_n,\tau_j^2C(\phi_j)),\quad \epsilon_i\sim N(0,\sigma^2). \tag{3.3}$$

ここで，$(\beta_{j1},\ldots,\beta_{jn})$ に対する同時分布はガウス過程に由来するものであり，m_j はその平均である．全てのモデルの単純平均を事前平均として採用する方法（$m_0=0$ かつ $m_j=1/J$ と設定するアプローチ）が 1 つの自然な方法であるが，m_j 自体もデータから推定することも可能である．$C(\phi_j)$ はカーネル関数によって決まる分散共分散行列であり，例えば指数型のカーネルを使ったモデル化（$C(\phi)$ の (i,j) 成分が $\exp(-\|s_i-s_j\|/\phi)$ によって与えられるモデル）が考えられる．

未知パラメータに対して事前分布を与えることで，潜在変数 β_{ji} および f_{ji} の事後推論が可能となる．その際には，マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo, MCMC）と呼ばれる計算アルゴリズムによって事後分布から乱数を生成する形で事後推測を行うことができる．今回のモデル (3.3) は潜在変数を用いた変化係数モデルと捉えることができるため，ギブスサンプラーによる MCMC によって乱数生成を行うことができる．ギブスサンプラーとは，パラメータ σ^2,τ_j^2,ϕ_j および潜在変数 β_{ji},f_{ji} それぞれの完全条件付き事後分布から繰り返し乱数を生成するアルゴリズムである．モデル (3.3) のもと，$\beta_j=(\beta_{j1},\ldots,\beta_{jn})$（地点 s_i におけるモデル j のウエイト）の完全条件付き分布は $N(A_j^\beta B_j^\beta,A_j^\beta)$

$$A_j^\beta=\left\{\frac{\Omega_j}{\sigma^2}+\frac{C(\phi_j)^{-1}}{\tau_j^2}\right\}^{-1},\quad B_j^\beta=\frac{1}{\sigma^2}f_j\circ\left(y-\beta_0-\sum_{k=1,k\neq j}^{J}f_k\circ\beta_k\right)+\frac{m_j}{\tau_j^2}C(\phi_j)^{-1}1_n$$

で与えられる．ここで，$\Omega_j=\mathrm{diag}(f_{j1}^2,\ldots,f_{jn}^2)$, $f_j=(f_{j1},\ldots,f_{jn})$, $\beta_j=(\beta_{j1},\ldots,\beta_{jn})$ であり，$\circ$ はアダマール積を表す．同様に，f_{ji} の完全条件付き分布は $N(A_{ji}^{(f)}B_{ji}^{(f)},A_{ji}^{(f)})$

$$A_{ji}^{(f)}=\left(\frac{\beta_{ji}^2}{\sigma^2}+\frac{1}{b_{ji}}\right)^{-1},\quad B_{ji}^{(f)}=\frac{\beta_{ji}}{\sigma^2}\left(y_i-\beta_{0i}-\sum_{k=1,k\neq j}^{J}\beta_{ki}f_{ki}\right)+\frac{a_{ji}}{b_{ji}}$$

で与えられる．σ^2,τ_j^2 については，事前分布として逆ガンマ分布を用いることで，完全条件付き分布も逆ガンマ分布となることが導出できる．ϕ_j の完全条件付き分布はよく知られた分布形にはならないが，ランダムウォーク型のメトロポリス・ヘイスティングスアルゴリズムを用いることでギブスサンプラーに組み込むことができる．このように，BSPS の MCMC アルゴリズムでは，f_{ji}（各モデルの予測値）の値に基づいて適切なウエイトを更新するステップ（β_{ji}

の生成ステップ)と β_{ji} (各モデルのウエイト)に基づいて予測値を更新するステップ (f_{ji} の生成ステップ)を繰り返し実行することで，モデルのウエイトや予測値に関する不確実性を事後分布を通して自然な形で得ることが可能となる．

MCMC によって事後サンプルが生成できると，任意の地点 s における統合予測を導出することができる．ガウス過程の定式化から，地点 s におけるモデルウエイト $\beta_j(s)$ の予測分布は $N(C_s(\phi_j)^\top C(\phi_j)^{-1}\beta_j, \{\tau_j^2 - \tau_j^2 C_s(\phi_j)^\top C(\phi_j)^{-1} C_s(\phi_j)\}^{-1})$ で与えられる(ただし，$C_s(\phi_j) = (C(\|s - s_1\|;\phi_j), \ldots, C(\|s - s_n\|;\phi_j))^\top$)ため，事後サンプルを用いて $\beta(s)$ の事後サンプルを生成することができる．さらに，地点 s における各モデルの予測分布が得られれば，BSPS のモデル式(3.2)を用いて $y(s)$ の乱数を生成することができる．この乱数の平均を計算することで点予測，分位点を計算することで区間予測を導出することができる．

3.3 最近傍ガウス過程を用いた高速実装

BSPS ではガウス過程を用いるため，β_j の完全条件付き分布から乱数を生成する(A_j^β を計算する)際には $n \times n$ 行列の逆行列を計算する必要があり，そのオペレーション回数は $O(n^3)$ である．したがって，地点数 n が大きい状況では計算コストが膨大なものになってしまい，現実的な計算時間で MCMC を実行することが困難になってしまう．このような問題を解決するための手法として，カーネル関数による重み付き平均，Karhunen-Loève 展開 (e.g. Banerjee et al., 2014, Section 12.3)や予測過程 (Banerjee et al., 2008)などの低ランク近似による方法が広く知られている．一般的に，低ランク近似による方法は，空間トレンドを過剰に平滑化してしまう(oversmoothing)問題が知られている (e.g. Datta et al., 2016)．近年では，低ランク近似によらないガウス過程の近似モデルが提案されているが，この節では特に Datta et al. (2016)で提案された最近傍ガウス過程(nearest-neighbor Gaussian process, NNGP)について解説する．Heaton et al. (2019)では，NNGP を含めた様々な高速化手法が紹介されており，手法間の包括的なレビューについてはこちらを参照されたい．

地点 $s_1, \ldots, s_n$ 上の確率変数をそれぞれ $x_i \equiv x(s_i)$ $(i = 1, \ldots, n)$ とする．通常の平均 0 のガウス過程では，$(x_1, \ldots, x_n) \sim N(0, C(\theta))$ となる．ここで，$C(\theta)$ はパラメータ θ を持つ $n \times n$ の共分散行列である．MCMC の各更新において，$C(\theta)$ の逆行列を扱う必要があり n が大きい状況では計算上のボトルネックとなる．Datta et al. (2016)では，n を固定したもとで，スパースな構造を持つ確率モデルによって元の多変量正規分布を近似する方法を与え，その後一般次元に対して確率分布を拡張することで確率過程として NNGP を構成している．一般に以下のような同時分布の分解公式が成り立つ．

$$p(x(s_1), \ldots, x(s_n)) = p(x(s_1)) \prod_{i=2}^{n} p(x(s_i)|x(s_1), \ldots, x(s_{i-1})).$$

添字 i が大きくなると，条件付ける変数の数が多くなるが，s_i から距離が離れている地点上の確率変数とは相関が小さく，そのような変数を無視したとしても大きな影響はないと考えられる．具体的に，$N(s_i)$ を $s_1, \ldots, s_{i-1}$ の中で地点 s_i から近い上位 m 地点を抽出した地点の集合として定義し，$x_{N(s_i)}$ を集合 $N(s_i)$ に含まれる各地点上の確率変数を並べた m 次元ベクトルとする．このとき，$p(x(s_i)|x(s_1), \ldots, x(s_{i-1}))$ を $p(x(s_i)|x_{N(s_i)})$ で置き換えることで同時分布の近似モデルを与えることができる．

$$\tilde{p}(x(s_1), \ldots, x(s_n)) = p(x(s_1)) \prod_{i=2}^{n} p(x(s_i)|x_{N(s_i)}).$$

$p(x(s_1), \ldots, x(s_n))$ が多変量正規分布の場合を考える．$C_{N(s_i)} \equiv \mathrm{Cov}(x_{N(s_i)})$ を $x_{N(s_i)}$ の分散

共分散行列とし，$C_{s_i,N(s_i)} \equiv \mathrm{Cov}(x(s_i), x_{N(s_i)})$ を x_i と $x_{N(s_i)}$ の共分散ベクトルとする．このとき

$$\begin{aligned}\tilde{p}(x(s_1),\ldots,x(s_n)) &= \phi(x(s_1);0,C(s_1,s_1))\prod_{i=2}^{n} p(x(s_i)|B_{s_i}x_{N(s_i)},F_{s_i}) \\ B_{s_i} &= C_{s_i,N(s_i)}C_{N(s_i)}^{-1}, \quad F_{s_i} = C_{s_i,s_i} - C_{s_i,N(s_i)}C_{N(s_i)}^{-1}C_{N(s_i),s_i}\end{aligned} \tag{3.4}$$

と表すことができる．上記の形で与えられる同時分布は多変量正規分布であり，元の分散共分散行列 $C(\theta)$ とは異なる共分散行列 $\tilde{C}(\theta)$ を持つ．また，$\tilde{C}(\theta)$ の重要な性質として，m が n に対して小さい場合，精度行列 $\tilde{C}(\theta)^{-1}$ がスパース行列(多くても $nm(m+1)/2$ 個の成分のみが非ゼロの行列)になる．一方で，低ランク近似の方法とは異なり，共分散行列 $\tilde{C}(\theta)$ は低ランク構造を持たない．これにより過剰な空間平滑化を防ぐことができる．近傍数 m の選択に関して，Datta et al. (2016) は n が大きいケースでも経験的に $m = 10, 15$ 程度で十分に正確な近似が得られると主張している．

この同時分布を元に，任意の地点集合 $(\tilde{s}_1,\ldots,\tilde{s}_k)$ における確率ベクトル $(x(\tilde{s}_1),\ldots,x(\tilde{s}_k))$ の同時分布を与えることができる．簡単のために，$(\tilde{s}_1,\ldots,\tilde{s}_k)$ と $(s_1,\ldots,s_n)$ は共通の地点を持たないと仮定する．このとき，$(x(\tilde{s}_1),\ldots,x(\tilde{s}_k))$ の同時分布は

$$p(x(\tilde{s}_1),\ldots,x(\tilde{s}_k)) = \int\prod_{l=1}^{k}\tilde{p}(x(\tilde{s}_l)|x(s_1),\ldots,x(s_n))\tilde{p}(x(s_1),\ldots,x(s_n))\prod_{i=1}^{n}dx(s_i)$$

で与えられる．ここで

$$\tilde{p}(x(\tilde{s}_l)|x(s_1),\ldots,x(s_n)) = \phi(x(\tilde{s}_l);B_{s_l}x_{N(x_l)},F_{s_l})$$

である．このように，任意の地点集合上で同時確率分布を定義することができ，確率過程として NNGP を構成することができる．

BSPS の推定を高速化するため，$\beta_j = (\beta_{j1},\ldots,\beta_{jn})$ の同時分布に対して以下のような NNGP を用いる．

$$p(\beta_{j1},\ldots,\beta_{jn}) = \prod_{i=1}^{n}\phi(\beta_{ji};B_{ji}\beta_j(N(s_i)),F_{ji}), \quad j=0,\ldots,J.$$

ただし

$$\begin{aligned}B_{ji} &= C_{s_i,N(s_i)}(\phi_j)C_{N(s_i)}(\phi_j)^{-1}, \\ F_{ji} &= \tau_j^2 - \tau_j^2 C_{s_i,N(s_i)}(\phi_j)C_{N(s_i)}(\phi_j)^{-1}C_{N(s_i),s_i}(\phi_j)\end{aligned}$$

であり，$N(s_i)$ は地点 s_i の m 近傍の地点の集合，$\beta_j(N_{s_i})$ は $s \in N(s_i)$ に対する $\beta_j(s)$ を並べた m 次元ベクトルを表す．また，$C_{s_i,N(s_i)}(\phi_j) = \mathrm{Cor}(\beta_j(s_i),\beta_j(N(s_i)))$ および $C_{N(s_i)}(\phi_j) = \mathrm{Cor}(\beta_j(N(s_i)),\beta_j(N(s_i)))$ である．このとき，$(\beta_{0i},\beta_{1i},\ldots,\beta_{Ji})$ の完全条件付き分布は $N(A_i^{(\beta)}B_i^{(\beta)}, A_i^{(\beta)})$

$$\begin{aligned}A_i^{(\beta)} &= \left\{\frac{f_i f_i^\top}{\sigma^2} + \mathrm{diag}(\gamma_{0i},\ldots,\gamma_{Ji})\right\}^{-1}, \quad \gamma_{ji} = \frac{1}{\tau_j F_{ji}} + \sum_{t;s_i\in N(s_t)}\frac{B_j(t;s_i)^2}{\tau_j F_{jt}}, \\ B_i^{(\beta)} &= \frac{f_i y_i}{\sigma^2} + (m_{0i},\ldots,m_{Ji})^\top, \\ m_{ji} &= \frac{B_{ji}^\top\beta_j(N(s_i))}{\tau_j F_{ji}} + \sum_{t;s_i\in N(s_t)}\frac{B_j(t;s_i)}{\tau_j F_{jt}}\left\{\beta_{jt} - \sum_{s\in N(s_t),s\neq s_i}B_j(t;s)\beta_j(s)\right\}\end{aligned}$$

で与えられる．ただし，$f_i = (1, f_{1i}, \ldots, f_{Ji})$ であり，$B_j(t;s)$ は係数ベクトル B_{jt} のうち $\beta_j(s)$ に対する係数を表す．f_{ji} の完全条件付き分布は 3.2 節と同様である．

3.4 他手法との比較

Cabel et al. (2022)では，数値実験やデータ解析例を通して BSPS と既存手法の予測精度を比較している．数値実験では，モデルの重要度が領域ごとに異なるシナリオにおいて予測精度の比較を行っており，(モデルの重要度の空間的異質性を考慮していない)既存手法に対する BSPS の優位性が数値的に示されている．また，データ解析においては，交差検証による精度比較を実施しており，その場面でも BSPS の予測精度が既存手法を上回ることが示されている．詳細については Cabel et al. (2022)を参照されたい．

4. まとめと議論

ベイズ的予測統合は空間予測の他にも時系列データ予測（e.g. McAlinn and West, 2019; Kobayashi et al., 2023)や因果推論（Sugasawa et al., 2023)などのタスクにおいて有効性が確認されている統合方法である．ベイズ的予測統合の実用上の限界点として，統合するモデルに対して何らかの不確実性が定量化されている必要がある点が挙げられる．機械学習的アプローチの中には点予測のみを与える手法も多く，それらの方法をベイズ的予測統合の枠組みで統合する場合は，ブートストラップ法などを使って予測の不確実性(例えば予測分散)を計算する必要があり，追加的な計算コストがかかってしまう点に注意が必要である．また，3 節の BSPS を実行するためには MCMC を用いる必要があり，(NNGP によって高速化をしているとはいえ)ある程度の計算時間がかかってしまうことが想定される．時系列予測の文脈では，時間ごとにモデルのウエイトが変化することを考慮しつつ，高速に逐次予測をするアルゴリズムなども開発されている（Bernaciak and Griffin, 2024)ため，空間データに対して同様なアプローチを開発していくことも今後の重要な課題になると考えられる．

参 考 文 献

Anselin, L. (2022). *Spatial Econometrics, Handbook of Spatial Analysis in the Social Sciences*, Edward Elgar Publishing, Glos, UK.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**, 825–848.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press, Boca Raton, Florida, USA.

Bernaciak, D. and Griffin, J. E. (2024). A loss discounting framework for model averaging and selection in time series models, *International Journal of Forecasting*, **40**, 1721–1733.

Breiman, L. (1996). Stacked regressions, *Machine Learning*, **24**, 49–64.

Cabel, D., Sugasawa, S., Kato, M., Takanashi, K. and McAlinn, K. (2022). Bayesian spatial predictive synthesis, *arXiv preprint*, arXiv:2203.05197 (最終アクセス日 2024 年 10 月 23 日).

Clarke, B. (2003). Comparing Bayes model averaging and stacking when model approximation error cannot be ignored, *Journal of Machine Learning Research*, **4**, 683–712.

Clydec, M. and Iversen, E. S. (2013). Bayesian model averaging in the M-open framework, Bayesian Theory and Applications (eds. P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens), Oxford University Press, Oxford, UK.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian

process models for large geostatistical datasets, *Journal of the American Statistical Association*, **111**, 800–812.

Debarsy, N. and LeSage, J. P. (2022). Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices, *Journal of Business & Economic Statistics*, **40**, 547–558.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H. and Liu, W. (2020). Advances of four machine learning methods for spatial data handling: A review, *Journal of Geovisualization and Spatial Analysis*, **4**, 1–25.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression*: *The Analysis of Spatially Varying Relationships*, Wiley, New York.

Genest, C. and Schervish, M. J. (1985). Modelling expert judgements for Bayesian updating, *Annals of Statistics*, **13**, 1198–1212.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M. et al. (2019). A case study competition among methods for analyzing large spatial data, *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 398–425.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors), *Statistical Science*, **14**, 382–417.

Kobayashi, G., Sugasawa, S., Kawakubo, Y., Han, D. and Choi, T. (2023). Predicting COVID-19 hospitalisation using a mixture of Bayesian predictive syntheses, *arXiv preprint*, arXiv:2308.06134 (最終アクセス日 2024 年 10 月 23 日).

Le, T. and Clarke, B. (2017). A Bayes interpretation of stacking for M-complete and M-open settings, *Bayesian Analysis*, **12**, 807–829.

LeSage, J. P. and Parent, O. (2007). Bayesian model averaging for spatial econometric models, *Geographical Analysis*, **39**, 241–267.

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting, *Journal of Econometrics*, **210**, 155–169.

Oliver, M. A. and Webster, R. (1990). Kriging: A method of interpolation for geographical information systems, *International Journal of Geographical Information System*, **4**, 313–332.

Sugasawa, S., Takanashi, K., McAlinn, K. and Edoardo, A. (2023). Bayesian causal synthesis for meta-inference on heterogeneous treatment effects, *arXiv preprint*, arXiv:2304.07726 (最終アクセス日 2024 年 10 月 23 日).

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. and Gabry, J. (2024). Pareto smoothed importance sampling, *Journal of Machine Learning Research*, **25**, 1–57.

West, M. (1992). Modelling agent forecast distributions, *Journal of the Royal Statistical Society* (*Series B*: *Methodological*), **54**, 553–567.

West, M. and Crosse, J. (1992). Modelling of probabilistic agent opinion, *Journal of the Royal Statistical Society* (*Series B*: *Methodological*), **54**, 285–299.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion), *Bayesian Analysis*, **13**, 917–1003.

Zhang, L., Tang, W. and Banerjee, S. (2023). Exact Bayesian geostatistics using predictive stacking, *arXiv preprint*, arXiv:2304.12414 (最終アクセス日 2024 年 10 月 23 日).

Bayesian Model Synthesis for Spatial Prediction

Shonosuke Sugasawa

Faculty of Economics, Keio University

Estimating a model from spatial data observed at finite locations and predicting unobserved locations are the crucial tasks in spatial data analysis. Recently, various models ranging from classical geostatistics and spatial econometrics to machine learning approaches have become available. Therefore, selecting appropriate analysis methods for the data is a significant issue in spatial data analysis. This paper reviews the methodology of Bayesian model synthesis for spatial prediction in situations where multiple predictive models are obtained. Specifically, we discuss the differences between recently proposed Bayesian spatial predictive synthesis and classical synthesis methods, as well as provide explanations of concrete estimation algorithms.

Key words: Bayesian model averaging, stacking, Gaussian process, Bayesian predictive synthesis.

統計数理(2025)
第 73 巻 第 1 号 65–76
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
［原著論文］

深層学習モデルに拡張した非線形Cox回帰モデルと東京賃貸物件市場への応用

込山 湧士[†]・松田 安昌[†]

（受付 2024 年 7 月 1 日；改訂 9 月 4 日；採択 9 月 17 日）

要　　旨

本論文では，2019 年 3 月から 2021 年 3 月に収集した東京賃貸物件市場データに対して，賃貸物件の広告掲載期間を生存時間とみなして Cox 回帰モデルによる生存時間分析を行った．Cox 回帰モデルを深層学習モデルによって非線形に拡張し，賃貸物件のもつ流動性と価格弾力性の時空間特性を非線形に表現し，COVID-19 パンデミックが東京賃貸市場に及ぼした影響を評価した．流動性と価格弾力性のみをニューラルネットで表現し，解釈の可能性をなるべく残したところに本モデルの特長がある．本モデルによる分析の結果，パンデミック後には流動性の減少および価格弾力性の増加傾向が観察された．

キーワード：価格弾力性，Cox 回帰モデル，時空間モデル，ニューラルネットワーク，ハザード関数，流動性．

1. はじめに

生存時間分析は，もともと治療法や薬の有効性を患者の生存時間の結果から評価する方法である．Cox (1972)が提案した Cox 回帰が，最も標準的な生存時間分析モデルとして広く応用されている．生存関数 $S(t)$ を時間 t 以上生きている確率とする．ハザード関数 $h(t)$ は対数生存関数の微分にマイナスをつけたもの $(-d\log S(t)/dt)$ と定義され，時点 t における死亡リスクを意味する．Cox 回帰では，ある患者のハザード関数を，その患者の属性や特徴量をベクトル x とその線形結合 $x\beta$ を使って

$$h(t) = h_0(t)\exp(x\beta)$$

と表す．このモデルの特徴はハザード関数が時間の関数 $h_0(t)$(基準ハザードとよぶ)と特徴量の関数 $\exp(x\beta)$ の積で与えられることである．ここで $h_0(t)$ には特定のモデルを仮定しない．生存時間 (t) と特徴量 (x) のデータを収集することで，モデルのパラメータ β と基準ハザードを推定することが可能であり，推定したモデルによって患者の属性に応じた死亡リスクを評価することができる．データセットには打ち切りと呼ばれる生存時間が観測されないデータがしばしば存在するが，Cox 回帰では，部分尤度法によってパラメータ推定に使うことができる．打ち切りを死亡リスクの評価に生かすことができる点が Cox 回帰の重要な特徴の一つである．

本論文は，Cox 回帰を賃貸物件市場に応用して，賃貸物件市場の流動性および価格弾力性を分析することを動機とし，両特性の時空間分析を可能とする非線形モデルに拡張することを目

[†] 東北大学 大学院経済学研究科：〒 980–8576 宮城県仙台市青葉区川内 27–1

的とする．ある賃貸物件の情報が公開されてから契約が決まって掲載が終了するまでの広告掲載期間をその賃貸物件の生存時間とみなすことにより，生存時間モデルを賃貸物件市場の分析に応用することができる．特に，賃貸物件市場の流動性(物件の成約のしやすさの度合い)と価格弾力性(価格の変動によって物件の需要が変化する度合い)が物件の掲載時期と所在地によって変動することを許容するハザード関数を提案する．ハザード関数の時変性と空間異質性を深層学習モデルを用いて学習を行い，パラメータを推定する．なお，本モデルを推定するために，2019 年 3 月より 2021 年 3 月までの 2 年間にわたり，東京都 23 区内の広告記録を収集した．総計 1,762,243 件，うち 487,643 件が広告終了日が未記載の打ち切り物件であり，物件の所在地(緯度，経度)と様々な特徴量を含んでいる．

Cox 回帰を機械学習を使って拡張した研究は数多く行われている．例として Cox 回帰に Ridge 正則化を加えた Verweij and Van Houwelingen (1994)によるモデルや Lasso 正則化を加えた Tibshirani (1997)によるモデルがある．また，罰則を課しながらも特定の特徴量を高次元生存モデルに組み込む推定手法として Binder and Schumacher (2008)による CoxBoost がある．さらに，Faraggi and Simon (1995)は，従来のハザード関数をニューラルネットワークで非線形化する試みを行った．その後提案された Katzman et al. (2018)による DeepSurv によって従来の線形モデルを上回る成果が得られている．これを画像処理に応用した Zhu et al. (2016)による DeepConvSurv のように非構造データに対する応用も行われている．また，ニューラルネットによる非線形対応させた Cox 回帰を大規模データに適用するためミニバッチ処理での損失計算を提案した Kvamme et al. (2019)による Cox-Time がある．他に生存時間モデルに深層学習を応用した研究として，分布に仮定を置かず生存時間の分布を直接学習する Lee et al. (2018)による DeepHit, DeepHit に RNN と temporal attention mechanism を加えて長期にわたる，もしくは繰り返し行われた測定結果に対応できるようにした Lee et al. (2020)による Dynamic-DeepHit がある．しかし，これらのニューラルネットワークを用いたモデルは生存時間の予測精度向上を目指したものであり，特徴量の解釈性には乏しいという課題がある．本研究は DeepSurv と Cox-Time とに着想を得て，Cox 回帰の非線形化を行う．回帰関数の定数項と賃貸物件価格の偏回帰係数だけを緯度，経度，掲載時期に依存させ，ニューラルネットワークで非線形化する．それぞれを流動性と価格弾力性として解釈を与えられるという点に特徴がある．

賃貸物件市場に生存時間分析を適用した研究として，いくつかの重要な研究がある．Deng et al. (2003)は，1987 年から 1998 年の BLS-CPI 住宅サンプルからの独自データを用いて，賃貸住宅における居住期間の推定に Cox 回帰を使用した．また，Bhuiyan and Hasan (2016)は，観察期間内の過去の住宅販売情報をもとに，家の販売確率を予測するための生存分析に着想を得た監督付き回帰(Cox 回帰)モデルを提案した．さらに，Yilmaz et al. (2022)は，イギリスの賃貸市場における流動性が時間と場所による需要の変動にどのように反応するかを探求した．Minzat et al. (2018)は，リスティング日から販売契約までの期間を予測するために生存時間分析を適用した．特に，Bhuiyan and Hasan (2016)の研究では，物件の説明文に対して Topic Modeling や LDA，Doc2Vec を適用している．しかし，賃貸物件市場に生存時間分析を適用した研究の中で，推定モデルとして深層学習を適用した例は，私たちの調べた限り，みつかっていない．

本論文の構成を述べる．第 2 章で Cox 回帰を時空間に拡張するため深層学習モデルを組み込んだモデルを提案する．空間への拡張は物件所在地(緯度，経度)，時間への拡張は広告公開時期から構成する月ダミー変数を用いる．第 3 章では本モデルを東京都 23 区賃貸物件市場の分析に応用し，時空間分析を行う．加えて，データはコロナ前後の区間を含んでいるので，賃貸物件市場をコロナ禍との関連に注視して分析を行う．第 4 章にて結論を述べる．

2. Cox 回帰と深層学習

2.1 Cox 回帰の時空間拡張

賃貸物件の募集期間を生存時間と考えて，まず従来の Cox 回帰モデルをあてはめる．ここで，募集期間は掲載終了日から広告公開日の間の日数であり，公開日と終了日が同じ物件は生存時間が 0 とみなす．基準ハザード関数を $h_0(u)$ とおくとき，Cox 回帰では，物件 i のハザード関数 $h_i(u)$ は次の式で表現される．

$$\log h_i(u) = \log h_0(u) + x_i\beta, i = 1, \ldots, n, \tag{2.1}$$

ここで，x_i は物件 i の属性(説明変数あるいは特徴量ともよばれる)，β は偏回帰係数である．Cox 回帰では，基準ハザード，偏回帰係数いずれも物件 i の公開時期 t_i，所在地 s_i に依存せず，時間的，空間的双方に静的なハザード関数を仮定している．

賃貸市場において，募集期間を時間的，空間的に静的とするモデルは制約が極めて強い．そこで，ハザード関数が，物件の所在地 s と広告公開時点 t に依存することを許容するモデルを考案する．物件 i の広告公開時期を t_i，所在地を s_i とするとき，ハザード関数 $\tilde{h}_i(u)$ を，

$$\log \tilde{h}_i(u) = \log \tilde{h}_0(u) + f(x_i, s_i, t_i), i = 1, \ldots, n, \tag{2.2}$$

とする．ここで，非線形性関数 $f(x_i, s_i, t_i)$ のままでは表現すべき関数のクラスが広すぎるため，解釈不能になってしまう．

物件のもつ流動性，価格弾力性が時間的，地理的に変動するモデルを考える．特徴量 x_i の中から物件の 1 平米あたり単価の対数値を取り出し $\log P_i$ とおき，その他を z_i とおく．$f(x_i, s_i, t_i)$ を線形近似したとき，定数項と $\log P_i$ の係数だけを s_i, t_i に依存させたハザード

$$\log \tilde{h}_i(u) = \log \tilde{h}_0(u) + f_0(s_i, t_i) + f_1(s_i, t_i) \log P_i + z_i\beta, i = 1, \ldots, n, \tag{2.3}$$

を考える．$f_0(s, t)$ は，物件の条件をそろえた時の対数ハザード関数を表し，流動性の時空間変動を表すと解釈できる．$f_1(s, t)$ は，他の条件をそろえて価格を 1% あげたときのハザードの変化率を表し，価格弾力性の時空間変動を表すと解釈できる．なお，$f_1(s, t)$ は通常は負値をとるため，$-f_1(s, t)$ を価格弾力性と定義することに注意する．

$f_0(s, t), f_1(s, t)$ のニューラルネットによる非線形モデリングを考える．ここで時間 t を連続時間ではなく，離散時間であると仮定する．つまり，ハザード関数は t で連続的に変動するのではなく，月次や週次のように離散的に変動すると考える．全物件データを含む時間区間を $I = [0, T]$ とおき，K 個の背反な時間区間

$$I = I_1 \cup \cdots \cup I_K$$

に分割する．$I_j(t)$ を t が I_j に含まれていたら 1, 含まれなければ 0 をとるダミー変数とし，f_0, f_1 に次の制約

$$\begin{aligned} f_0(s, t) &= f_0(s, I_1(t), \ldots, I_K(t)), \\ f_1(s, t) &= f_1(s, I_1(t), \ldots, I_K(t)), \end{aligned} \tag{2.4}$$

を課す．f_0, f_1 に 4 層からなる深層学習モデルをあてはめる．緯度と経度の 2 次元からなる s と K 個のダミーからなる t を入力とし，出力を 1 次元とするニューラルネットである．f_0 と f_1 には同じ構造のニューラルネットを使用するが，ネットワークの重みは変わることに注意する．深層学習モデルの構造を図 1 に示す．実際のノード数はデータ分析のセクションで述べる．

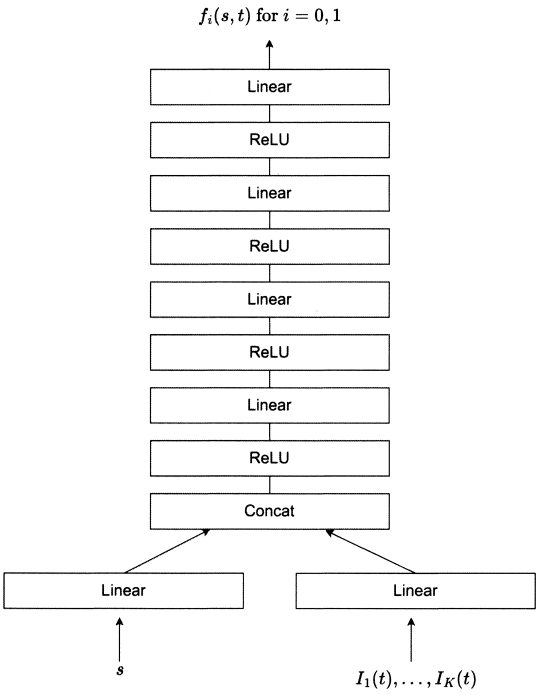


図 1. 深層学習モデルの構造.

2.2 推定法

前節で提案したハザード関数(2.3)では，物件の流動性と価格弾力性が広告公開時点 t と所在地 s に依存することを許し，t を離散化して深層学習モデルを適用した．本節では，深層学習モデルの推定法を紹介する．

本モデルでは，Cox (1972)による部分尤度法をそのまま応用することができる．物件 $i=1,\ldots,N$ の募集期間を u_i とおく．募集期間が u 以上の全物件の集合を $R(u)$ とおく．Cox (1972)は，ハザード関数が以下の線形関数

$$\log h_i(u) = \log h_0(u) + x_i\beta$$

で与えられるモデルに対して，基準ハザード関数 $h_0(u)$ を集約した対数尤度関数，つまり対数部分尤度，

$$Q(\beta) = \sum_{i=1}^{N} \left\{ x_i\beta - \log\left(\sum_{j\in R(u_i)} \exp x_i\beta \right) \right\}.$$

を構成し，対数部分尤度を最大化するパラメータ推定法を提案している．なお，広告期間が同じ日数となる物件が複数存在しても，物件 i ごとに定義通りに部分尤度を評価するものとする．これは，Breslow (1974)によるタイデータを含む場合の部分尤度近似法に対応する．

次に，Cox 回帰モデルの部分尤度を非線形ハザード関数に拡張する．ハザード関数が (s,t) に依存するモデル(2.3)に対して，ベースラインハザードを除いた部分を $g(s,t) = f_0(s,t) + f_1(s,t)\log P + z\beta$ とおく．このハザード関数の対数部分尤度に，Cox の対数部分尤度をそのまま拡張し，

$$Q(f) = \sum_{i=1}^{N} \left\{ f(s_i, t_i) - \log \left(\sum_{j \in R(u_i)} \exp f(s_j, t_j) \right) \right\}. \tag{2.5}$$

と与える．ここで $R(u_i)$ は広告期間が u_i 以上の打ち切りを含むすべての物件をあらわす．この対数尤度関数の負値を本モデルの深層学習における損失関数として，ニューラルネットワークの学習をおこなう．

2.3 C-index

モデルの評価指標として the concordance-index（C-index）を採用する．C-index は Harrell et al.（1982）によって提案され，生存時間解析において広く使用されている評価指標である．推定したモデルがどのくらい良く生存時間を予測できるかを示す．C-index の定義域は $[0, 1]$ で，1 に近いほど良いとされ．ランダムな予測では値が 0.5 となる．本論文で用いるデータにはタイデータを多く含むので，Kang et al.（2015）によるタイデータに対応した C-index を使って評価する．

我々の非線形 Cox 回帰モデルを，物件の特徴量を x として，

$$h(u) = h_0(u) \exp(f(x))$$

とおくとき，C-index は以下で定義される．物件 i の実測生存時間を T_i, 対数ハザードからベースラインを除いた $f(x_i)$ に対して，

$$C = \frac{1}{2} \left(\frac{\sum_{i \neq j} sgn(f(x_i), f(x_j)) csgn(T_i, \delta_i, T_j, \delta_j)}{\sum_{i \neq j} csgn(T_i, \delta_i, T_j, \delta_j)^2} + 1 \right), \tag{2.6}$$

ここで，δ_i は物件 i が打ち切りであれば 0 をとるダミー変数，I を指示関数として，

$$\begin{aligned} sgn(f(x_i), f(x_j)) &= I(f(x_i) \geq f(x_j)) - I(f(x_i) \leq f(x_j)), \\ csgn(T_i, \delta_i, T_j, \delta_j) &= \delta_j I(T_i \geq T_j) - \delta_i (T_i \leq T_j) \end{aligned}$$

である．タイデータを含む場合の定義なので意味が分かりにくいが，直観的には，物件 i, j の全ての組み合わせに対し，実測生存時間の大小と同じハザード関数の大小をもつ組み合わせの割合を意味している．

3. 東京賃貸物件市場の分析

3.1 データ

東京都 23 区内の賃貸物件を対象に，2019 年 3 月より 2021 年 3 月までの 2 年間の広告記録を収集した．広告記録には，物件名，住所，緯度，経度，公開日，広告終了日，賃料（円），面積（平米），最寄り駅からの徒歩距離（分），物件の位置する階数，物件の入る建物の階数，近辺の利便施設（例えばショッピングセンターや小学校）等の情報を含んでいる．本節では広告の公開日から掲載終了日までの日数をその物件の生存時間とみなし，賃貸物件に生存時間分析を応用して賃貸市場の地域特性とその時間変化を分析する．

本データには，総計 1,762,243 件の賃貸物件が収集され，そのうち 487,643 件が広告終了日が未記載の打ち切りデータとなっている．図 2 に，広告期間の分布を，図 3 には，23 区ごとの物件数を，打ち切りなしと打ち切りの各場合に集計したものを，図 4 には，物件数を広告公開月と掲載終了月の月別に集計したものを，それぞれ表している．広告期間の分布をみると，7 日，14 日，21 日,... の 1 週間ごとにジャンプがみられる．広告の掲載を 1 週間単位で契約している

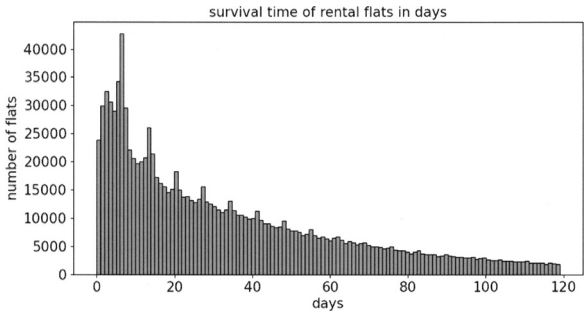


図 2. 東京 23 区賃貸物件の広告期間の分布.

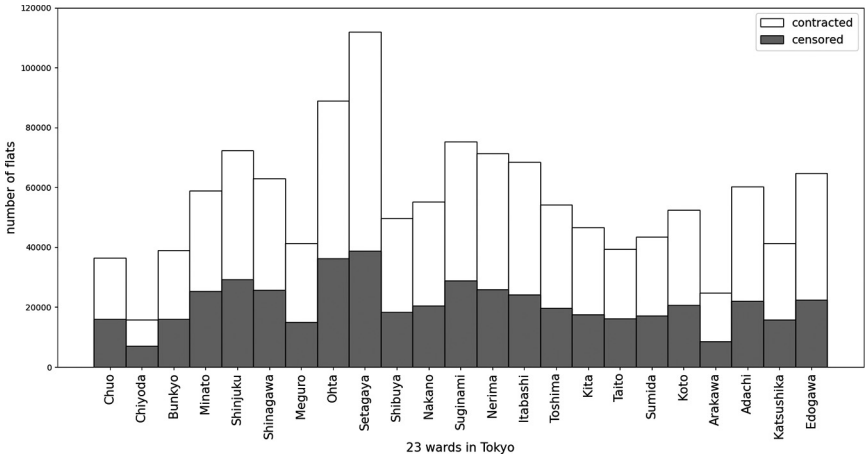


図 3. 東京 23 区賃貸物件数の分布. 区ごとの契約件数と打ち切り件数.

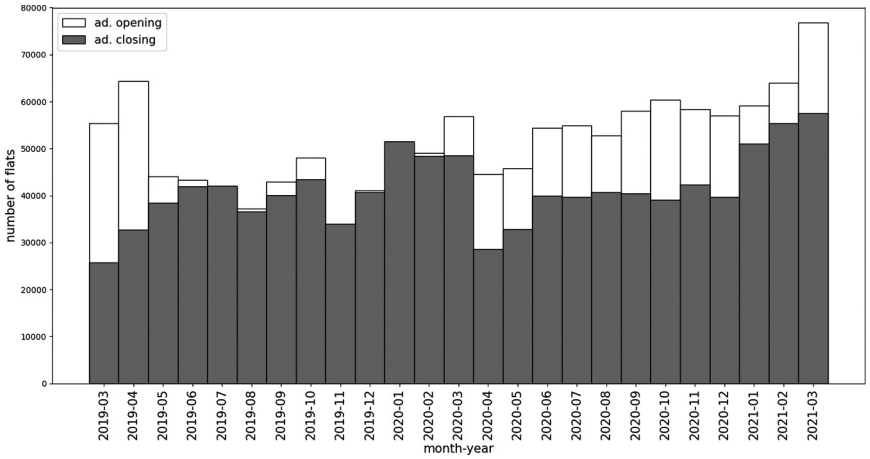


図 4. 東京 23 区賃貸物件数の分布. 月別の広告スタートと終了の賃貸物件数.

貸主がいるためであると予想される．広告期間が 0 日，つまり公開したその日に掲載が終了する人気物件も 20,000 件以上あることに注意を要する．23 区ごとの分布では，世田谷区が最大の物件数で 10 万件を超え，大田区，新宿区が続き，逆に千代田区が最小で荒川区が続き，ともに 2 万件未満の物件数しかない．最後に月別の分布では，1 月，2 月，3 月に公開，終了ともに物件数が増加する傾向にあり，年度前の 3 か月が賃貸物件取引が最も活発な時期であることを示している．

3.2 ハザード関数の時空間分析

2019 年 3 月から 2021 年 3 月まで全 25 か月間にわたり収集し，打ち切りを含め総計 1,762,243 件の賃貸物件の広告期間に対して，ハザード関数の流動性と価格弾力性が時空間的に変動するモデル(2.3)をあてはめ，流動性と価格弾力性のもつ地域特性の時間変化を分析する．ここで，z(対数家賃 $\log P$ 以外の説明変数)として，物件の築年数，最寄り駅からの徒歩距離(分単位)，1F ダミー，物件の入る建物の総階数を選択した．

流動性と弾力性にはそれぞれ 4 層からなるニューラルネット(図 1)を応用し，時間(月別ダミー)と所在地(緯度と経度)によるハザード関数の時空間変動を把握することを目指している．ここで，レイヤーごとにノード数を以下の式で設定した．$\mathrm{ReLU}(x)$ は活性化関数 ReLU による変換を表し，$\mathrm{Linear}_{a\to b}(x)$ は a 次元入力から b 次元へ出力する線形変換とし，2 次元の s, K 次元(本データでは 38 次元)からなる月ダミー変数 $I_1,\dots,I_K$ を入力とし，f_0 を出力とする関数を以下の 4 層からなるニューラルネットにあてはめた．

$$
\begin{aligned}
h_1 &= \mathrm{Linear}_{2\to 128}(s),\\
h_2 &= \mathrm{Linear}_{K\to 10}(I_1(t),\dots,I_K(t)),\\
h_3 &= \mathrm{Concatenate}(h_1,h_2),\\
h_4 &= \mathrm{ReLU}(\mathrm{Linear}_{128+K\to 256}(h_3)),\\
h_5 &= \mathrm{ReLU}(\mathrm{Linear}_{256\to 128}(h_4)),\\
h_6 &= \mathrm{ReLU}(\mathrm{Linear}_{128\to 64}(h_5)),\\
f_0 &= \mathrm{Linear}_{64\to 1}(h_6).
\end{aligned}
$$

f_1 を出力とするニューラルネットの構造も全く同じであるが，重みは異なることに注意する．

本モデルの 4 層ニューラルネットを記述するために必要な 154,896 個のパラメータの推定を行う．全物件 1,762,243 件に対し，本モデルの対数部分尤度(2.5)の負値を損失関数とし，損失関数を最小化するパラメータ推定を行う．

全サンプル 1,762,243 件の 60% を学習用データセット，20% を評価用データセット，残りの 20% をテスト用データセットにランダムに分割する．学習用データセットを用いて損失関数を最小化する深層学習モデルのパラメータを学習し，そのパラメータを使ってテスト用データセットの損失関数および C-index を評価し当てはまりの良さの基準とする．学習用データセットに対し確率的勾配降下法によって損失関数を最小化する推定を行った．評価用データセットは，学習時のエポック数やハイパーパラメータを決めるために用いられ，ここでは評価用データセットに対する C-index を最小化するように選択した．

従来の Cox 回帰モデル(2.1)を本モデル(2.3)のベンチマークとする．従来の Cox 回帰モデルは，本モデルにおいて流動性と弾力性をともに定数に仮定したときに一致する．本モデルと Cox 回帰モデルの損失関数(負の対数部分尤度)と C-index を評価し，表 1 に示す．本モデルは，Cox 回帰モデルに比べて，テストデータに対する損失関数，C-index とも改善している．損失

表 1. 損失関数(負の対数部分尤度)と C-index による Cox 回帰と深層学習モデルの比較.

モデル	損失関数（負の対数部分尤度）			C-index		
	学習	評価	テスト	学習	評価	テスト
Cox 回帰 (2.1)	5,232,433	1,590,255	1,588,879	0.547	0.547	0.546
深層学習 (2.3)	5,202,859	1,579,736	1,581,290	0.612	0.608	0.607

関数および C-index の基準からみて，本モデルは線形モデルを改善していることがわかる．ここで，本モデルの学習データとテストデータに対する損失関数，C-index にわずかな差しかみられないことに注意する．過学習することなくパラメータ推定が行われたことを示している．

式(2.4)で定義した流動性 $f_0(s,t)$ と価格弾力性 $f_1(s,t)$ の推定値を，特定の地域 s_0 に固定した場合と 23 区全体に分けて示す．なお，流動性は(2.3)式のベースラインハザードを除いた定数項であり，他の説明変数の単位によって変化する量であるから，流動性自体より流動性の差に意味があることに注意する．言い換えると，流動性は，同条件下での賃貸物件の契約の決まりやすさを表す．逆に，価格弾力性は，(2.3)式において，対数家賃 $\log P_i$ の偏回帰係数の負値で定義されるため，値自体に意味があり，他の条件をそろえて家賃が 1% 上昇したときのハザードの減少率を表すことに注意する．

まず，s_0 を杉並区荻窪 1 丁目に固定した場合の流動性 $f_0(s_0,t)$ と価格弾力性 $f_1(s_0,t)$ の月次時系列プロットを，図 5 と図 6 に示す．破線は，ベンチマークである Cox 回帰モデルによる流動性であり，時空間変化しない定数で評価される．なお，価格弾力性は $f_1(s,t)$ の負値であるから，値が小さい(高い)ほど弾力性が高い(低い)ことに注意する．図 5 より，荻窪の流動性は新型コロナウイルスが周知され始めた 2020 年 1 月以降に急減し，10 月以降に回復が始まっていることがわかる．また，通常の Cox 回帰による流動性とほぼ同じ水準である．荻窪の物件は，ほぼ平均的な流動性を持つことを示している．図 6 より，価格弾力性は 2020 年 2 月に 0.6 から 1.7 に急増した．コロナ禍による賃貸需要の急減のため，家賃に敏感に反応するようになったと考えられる．2021 年 1 月より回復傾向がみられるが，2019 年以前の値までには戻していない．また，Cox 回帰で評価された弾力性に比べて，荻窪の弾力性は高く評価されていることがわかる．荻窪では，平均からみて家賃の設定に強く反応する傾向があることを示している．

次に流動性 $f_0(s,t)$ と価格弾力性 $f_1(s,t)$ を，月別に東京都 23 区地図上に，それぞれ図 7 と図 8 に示す．なお，図は 2015 年度の国勢調査の境界データを基に加工して作成した．

図 7 の流動性の推移より，コロナ前(2019/9, 2019/12)とその後(2020/3, 2020/9, 2020/12, 2021/3)を比較する．全体的に 2019 年 3 月から 2020 年 9 月に大きく減少した流動性が 2020 年 12 月以降，徐々に回復していく様子がみられる．

図 8 の価格弾力性の推移において，コロナ禍の 2020 年 3 月以降，全体的に価格弾力性の上昇が観察される．ただし，都心のやや北に位置する浅草近辺がコロナ禍前後によらず高く評価されている．浅草エリアは都心でありながら比較的求めやすい家賃で賃貸物件が出されている地域であるといわれており，浅草では価格の設定に賃貸需要が影響されやすいことが考えられる．2020 年 9 月に江東区，大田区の湾岸地域の弾力性が広く上昇している．東京オリンピックの 2021 年 8 月の開催が危ぶまれた時期であり，湾岸地域の賃貸需要がオリンピック延期リスクに影響を受けたものと考えられる．

4. 終わりに

本論文では，大量の賃貸物件データを収集し，広告期間を生存時間とみなして Cox 回帰を応用した．古典的な Cox 回帰では，ハザード関数は物件特徴量の線形関数であらわされるため，

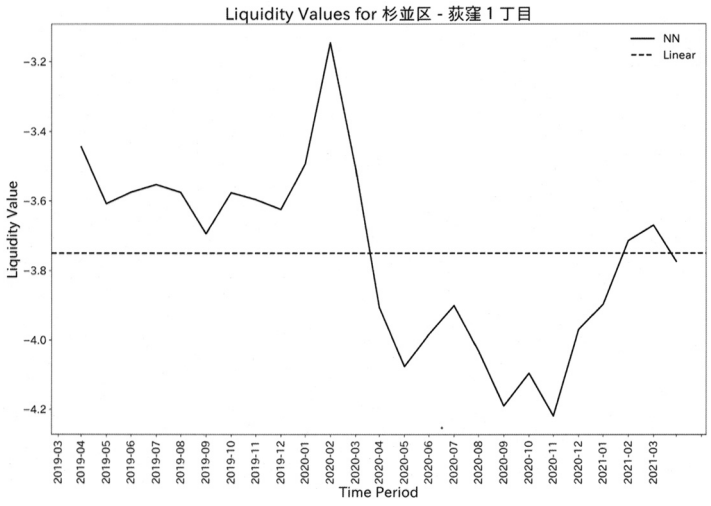


図 5．杉並区荻窪１丁目における流動性の時間変動．

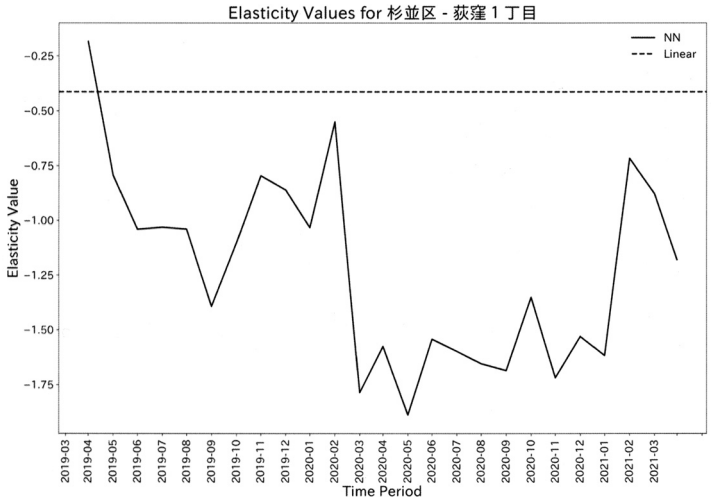
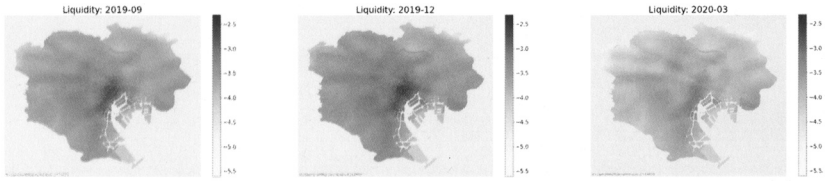
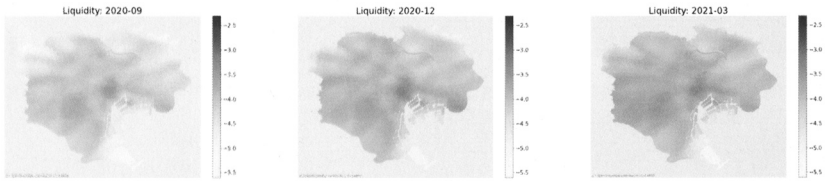


図 6．杉並区荻窪１丁目における価格弾力性の時間変動．

定数項にあたる流動性と対数家賃の偏回帰係数にあたる価格弾力性は，時間と所在地によらない．そこで，ハザード関数の流動性と価格弾力性が時空間変動することをゆるすハザード関数を考案した．ここで，時間を物件広告公開月の月ダミー変数，空間を物件所在地(緯度，経度)であらわし，流動性と価格弾力性を，時間と所在地を入力とするニューラルネットワークモデルで表現した．本モデルを学習させ，流動性および価格弾力性を評価し，東京都 23 区の賃貸物件市場の特徴をコロナ禍との関連を見ながら分析を行った．コロナ禍の発生後，都心部において賃貸物件市場は急速に流動性を下げ，価格弾力性を上げたが，徐々に回復する傾向がみられた．以上のように，広告掲載期間を生存時間とみなす生存時間分析により，賃貸物件市場を数理的な分析対象とすることができる．

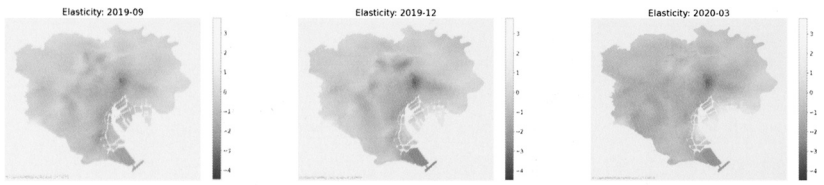


(a) 2019 年 9 月 (b) 2019 年 12 月 (c) 2020 年 3 月

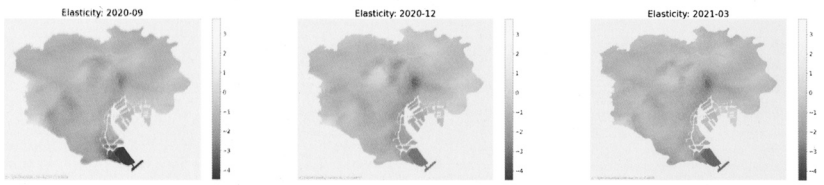


(d) 2020 年 9 月 (e) 2020 年 12 月 (f) 2021 年 3 月

図 7. 東京賃貸物件市場の流動性の推移.



(a) 2019 年 9 月 (b) 2019 年 12 月 (c) 2020 年 3 月



(d) 2020 年 9 月 (e) 2020 年 12 月 (f) 2021 年 3 月

図 8. 東京賃貸物件市場の価格弾力性の推移.

謝 辞

2 名の査読者よりいただいた貴重なコメントに感謝します.

参 考 文 献

Bhuiyan, M. and Hasan, M. A. (2016). Waiting to be sold: Prediction of time-dependent house selling probability, *2016 IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*), 468–477, https://doi.org/10.1109/DSAA.2016.58.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of

sparse high-dimensional survival models, *BMC Bioinformatics*, **9**(1), https://doi.org/10.1186/1471-2105-9-14.

Breslow, N. (1974). Covariance analysis of censored survival data, *Biometrics*, **30**(1), 89–99, https://doi.org/10.2307/2529620.

Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society*, *Series B* (*Methodological*), **34**(2), 187–220, https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Deng, Y., Gabriel, S. A. and Nothaft, F. E. (2003). Duration of residence in the rental housing market, *The Journal of Real Estate Finance and Economics*, **26**(1), 267–285, https://doi.org/10.1023/A:1022987010545.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data, *Statistics in Medicine*, **14**(1), 73–82, https://doi.org/10.1002/sim.4780140108.

Harrell, J., Frank, E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the yield of medical tests, *JAMA*, **247**(18), 2543–2546, https://doi.org/10.1001/jama.1982.03320430047030.

Kang, L., Chen, W., Petrick, N. A. and Gallas, B. D. (2015). Comparing two correlated C indices with right-censored survival outcome: A one-shot nonparametric approach, *Statistics in Medicine*, **34**(4), 685–703, https://doi.org/10.1002/sim.6370.

Katzman, J. L., Shaham, U., Cloninger, A. et al. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network, *BMC Medical Research Methodology*, **18**(1), 24, https://doi.org/10.1186/s12874-018-0482-1.

Kvamme, H., Borgan, Ø. and Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression, *Journal of Machine Learning Research*, **20**, 1–30.

Lee, C., Zame, W., Yoon, J. and van der Schaar, M. (2018). DeepHit: A Deep learning approach to survival analysis with competing risks, *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1), 2314–2321, https://doi.org/10.1609/aaai.v32i1.11842.

Lee, C., Yoon, J. and van der Schaar, M. (2020). Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data, *IEEE Transactions on Bio-medical Engineering*, **67**(1), 122–133, https://doi.org/10.1109/TBME.2019.2909027.

Minzat, D., Breaban, M. and Luchian, H. (2018). Modeling real estate dynamics using survival analysis, *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (*SYNASC*), 215–222, https://doi.org/10.1109/SYNASC.2018.00042.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**(4), 385–395, https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

Verweij, P. J. M. and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression, *Statistics in Medicine*, **13**(23-24), 2427–2436, https://doi.org/10.1002/sim.4780132307.

Yilmaz, O., Talavera, O. and Jia, J. Y. (2022). Rental market liquidity, seasonality, and distance to universities, *International Journal of the Economics of Business*, **29**(2), 223–239, https://doi.org/10.1080/13571516.2022.2033078.

Zhu, X., Yao, J. and Huang, J. (2016). Deep convolutional neural network for survival analysis with pathological images, *2016 IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), 544–547, https://doi.org/10.1109/BIBM.2016.7822579.

Deep Learning Extensions of Cox Regression and Their Applications to Rental Property Market in Tokyo

Yuji Komiyama and Yasumasa Matsuda

Graduate School of Economics and Management, Tohoku University

This paper applies Cox regression models to dataset in Tokyo rental property market collected from March, 2019 till March, 2021. We try a deep learning extension of Cox regression models to let liquidity and price elasticity of rental properties be dependent on space and time in a nonlinear way, and analyze the effects by the pandemic of COVID-19 to the rental market. We apply neural networks only to express liquidity and elasticity from which interpretability retains in our model. We find by the model applications that the pandemic makes the tendencies of decreasing liquidity increasing and elasticity all over the Tokyo 23 wards.

Key words: Cox regression, hazard function, liquidity, neural networks, price elasticity, spatio-temporal models.

統計数理(2025)
第 73 巻 第 1 号 77–99
©2025 統計数理研究所
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

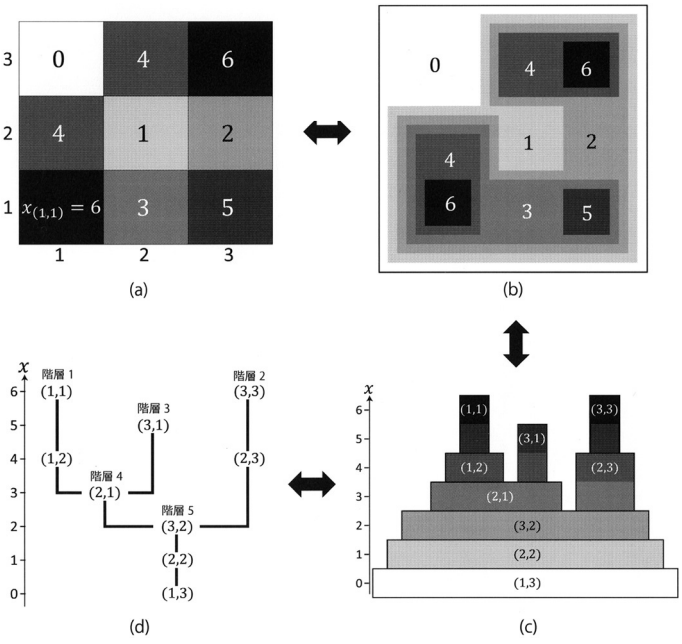


図 1. 3×3 の格子状のデータに対するエシェロン解析法の適用例．(a)3×3 の格子データ．各領域が 1 変量値 x を有している．ここに，$x_{(1,1)}$ は座標 $(1,1)$ の値を示している．(b)x の値に基づいて，データを同じ位相領域に分類する様子．(c)各位相領域の相対関係を，x を縦軸にとって表現したもの．ここに，$(1,1)$ は座標 $(1,1)$ の領域を指す．(d)デンドログラムの完成．

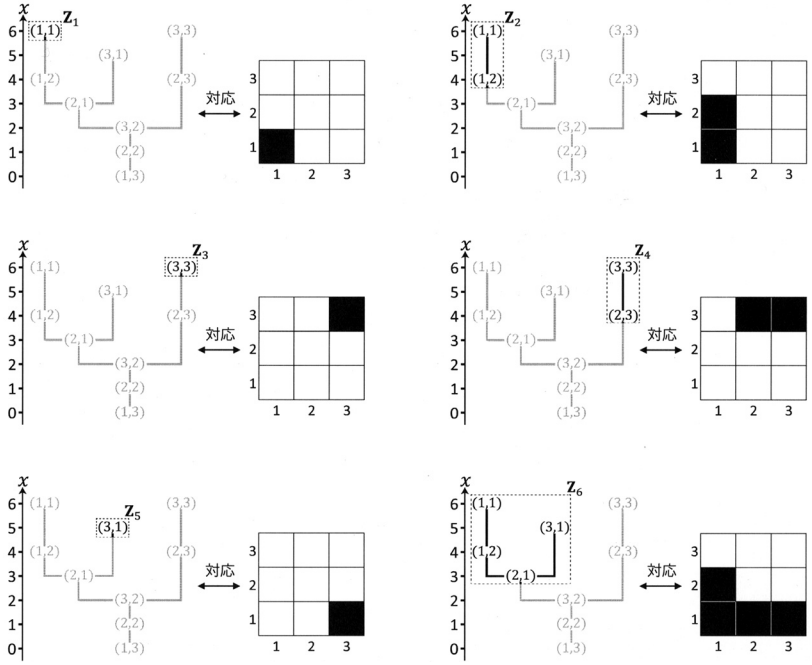


図 2. 3×3 の格子状のデータに対しエシェロンスキャン法を適用する様子．ここに，ウィンドウ内に許容する最大領域数は 4 としている．このとき，$\boldsymbol{Z}$ は全部で 6 つのパターンが存在する．

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．

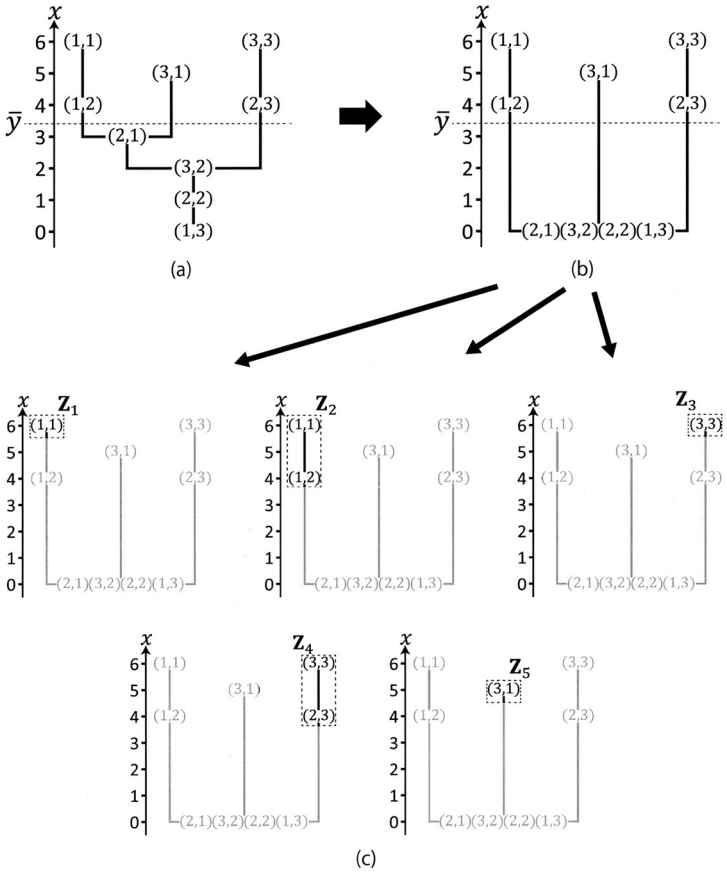


図 3. (a)図 1 の 3×3 の格子状のデータのデンドログラム．(b)$\bar{y}$ よりも低い値を持つ領域の値 x を y の最小値に変換した際に作成されるデンドログラム．(c)変換後のデンドログラムに対するスキャン．ここに，ウィンドウ内に許容する最大領域数は 4 としている．

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

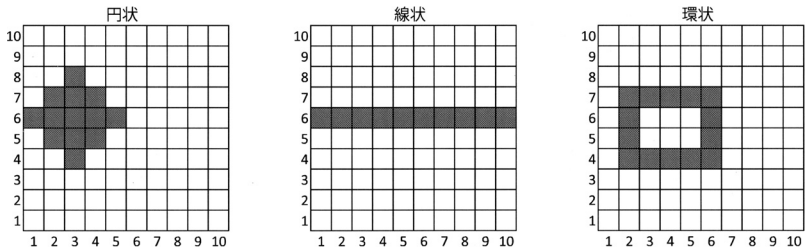


図 4．想定した真のホットスポット $\boldsymbol{Z}^+$．

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

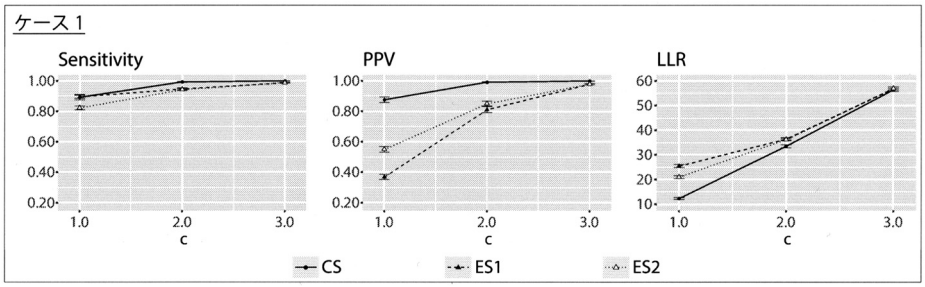


図 5．従来法(CS)と提案法(ES1，ES2)に対する *Sensitivity*，*PPV*，*LLR* の結果(ケース 1)．

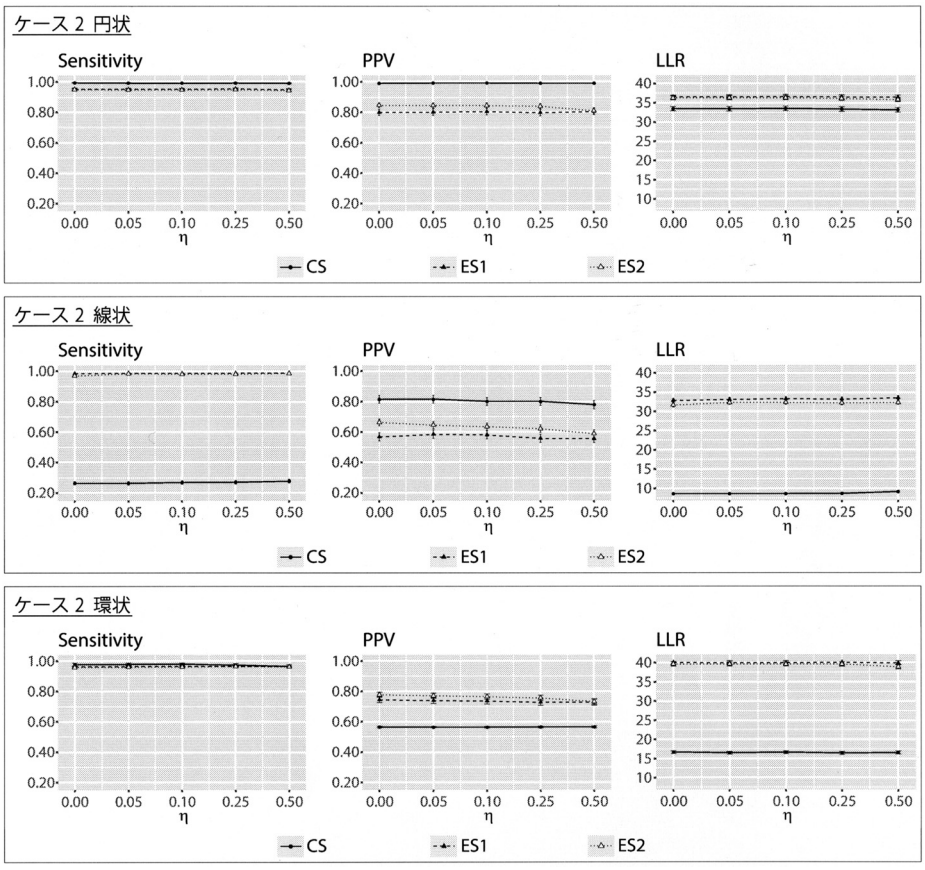


図 6．従来法(CS)と提案法(ES1，ES2)に対する *Sensitivity*，*PPV*，*LLR* の結果(ケース 2)．

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．

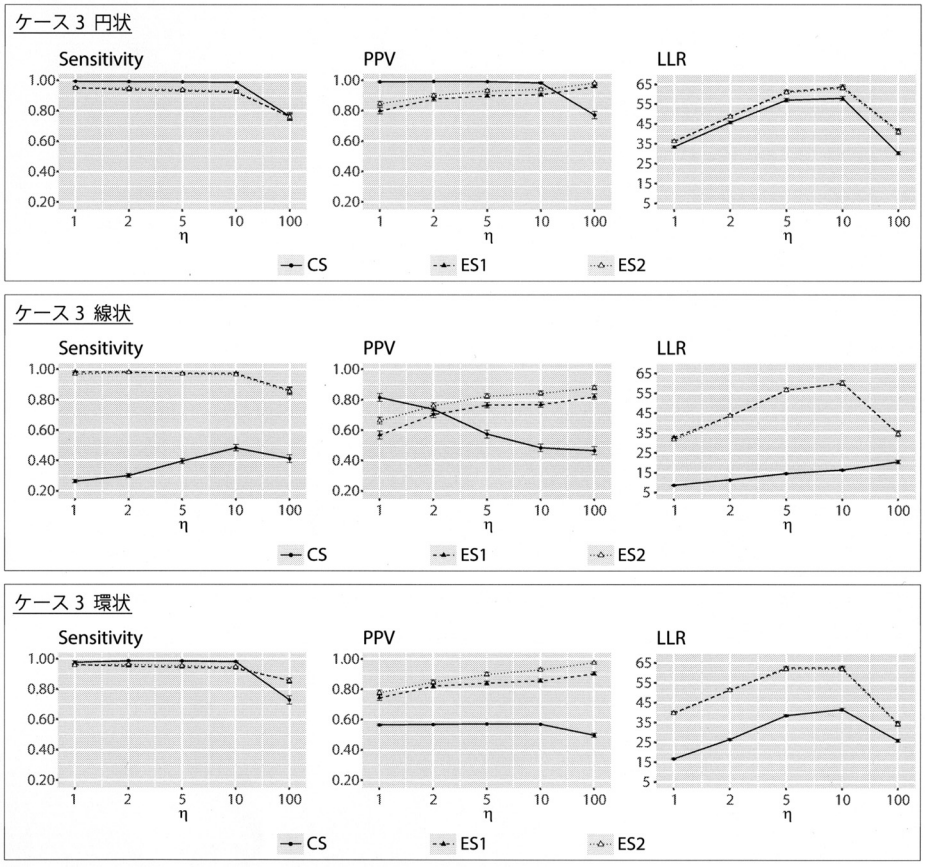


図 7. 従来法(CS)と提案法(ES1, ES2)に対する *Sensitivity*, *PPV*, *LLR* の結果(ケース 3).

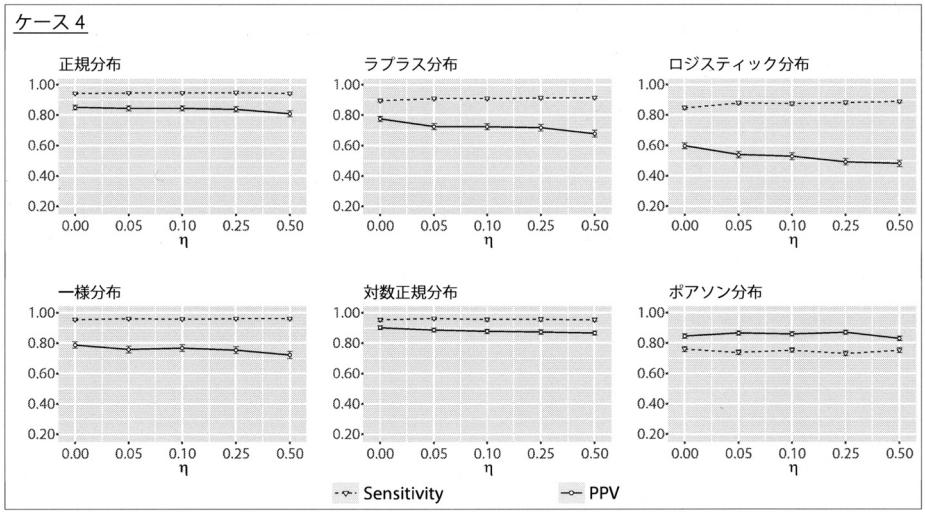


図 8. 提案法(ES2)に対する *Sensitivity*, *PPV* の結果(ケース 4).

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，

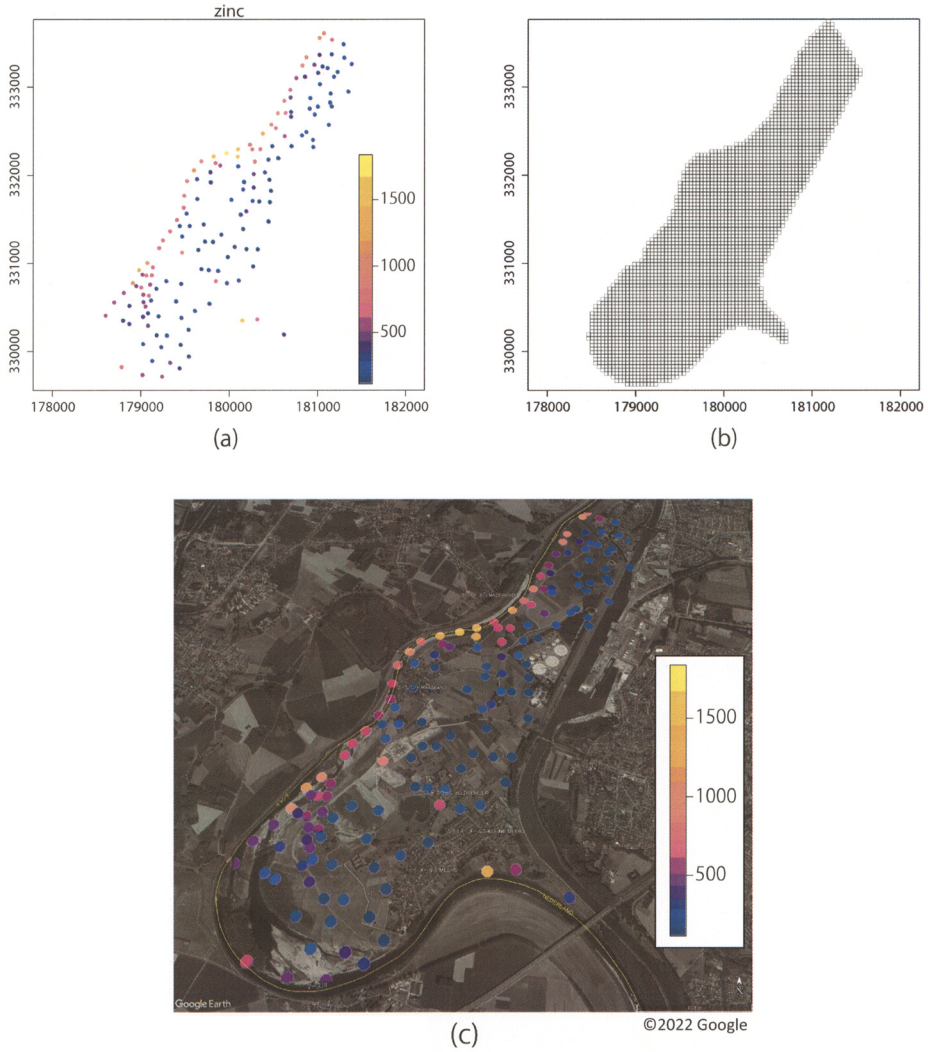


図 9. (a)元データにおける 155 の観測地点とその亜鉛濃度．(b)推定する地点を表した格子．(c)観測地点を Google Earth 上に重ねた図．マース川北部で亜鉛濃度が高くなっている様子が見て取れる．

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

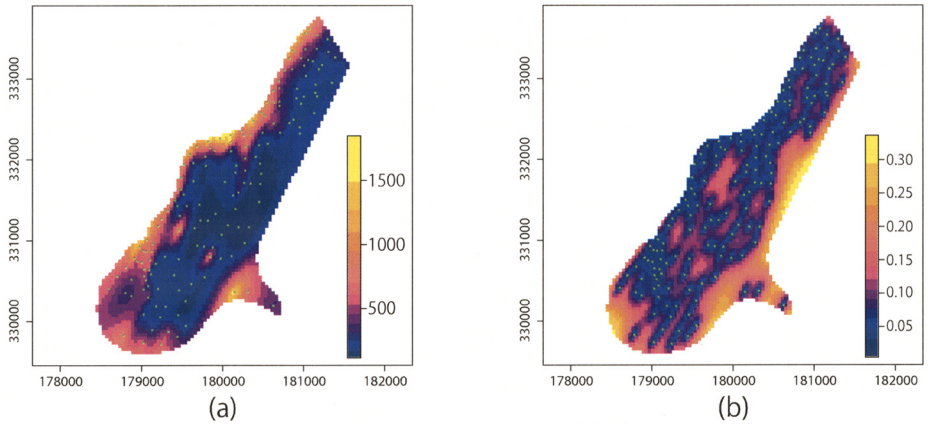


図 10. (a)クリギングによる亜鉛濃度の推定結果の可視化．なお，可視化に際し対数値を元の単位に戻している．(b)クリギング分散の可視化．図内の緑点は元データの観測地点を示す．

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

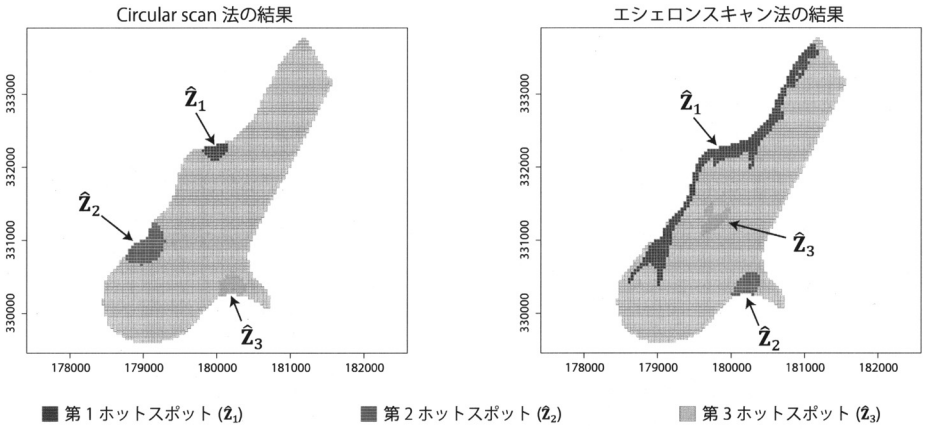


図 11. 亜鉛濃度の予測値に対するホットスポットの検出結果．

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

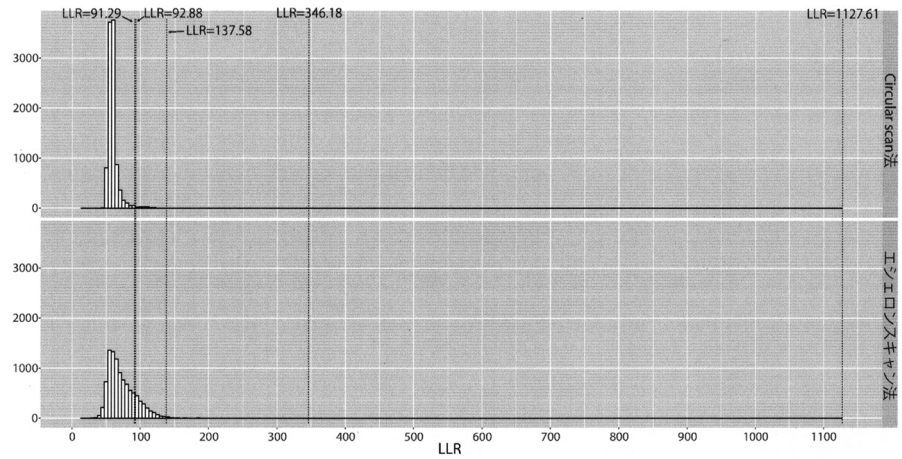


図 12. 2 つのスキャン方式による H_0 の下での最大対数尤度比の分布，ならびに今回検出された各ウィンドウにおける対数尤度比の位置.

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

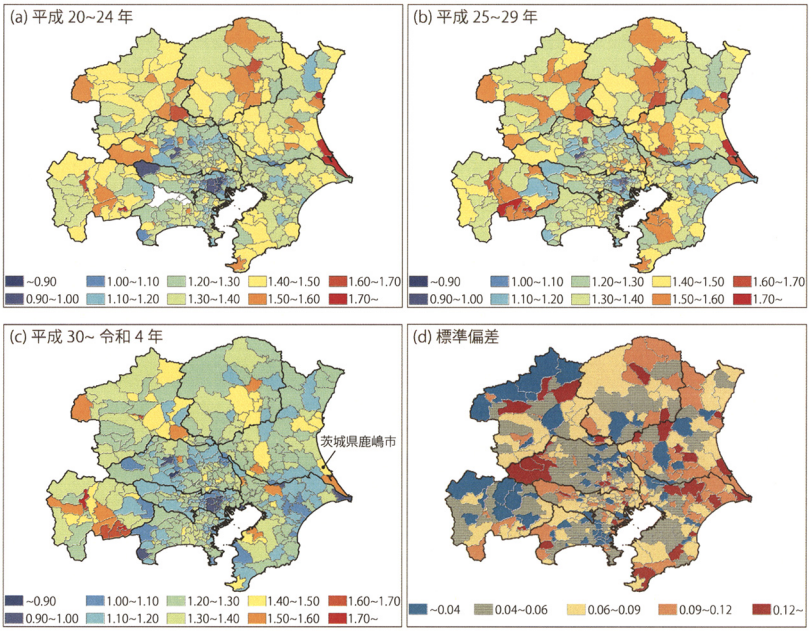


図 13. 厚生労働省公表による首都圏市区町村別の合計特殊出生率のベイズ推定値とその標準偏差を地図上に可視化した結果．(a) 平成 20〜24 年．ここに，神奈川県相模原市の緑区・中央区・南区については欠損値(白色)となっている．(b) 平成 25〜29 年．(c) 平成 30 年〜令和 4 年．(d) 3 期間での標準偏差．例えば，茨城県鹿嶋市の合計特殊出生率のベイズ推定値は，平成 20〜24 年：1.77，平成 25〜29 年：1.79，平成 30〜令和 3 年：1.49 であり，その標準偏差は 0.169 となる．

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

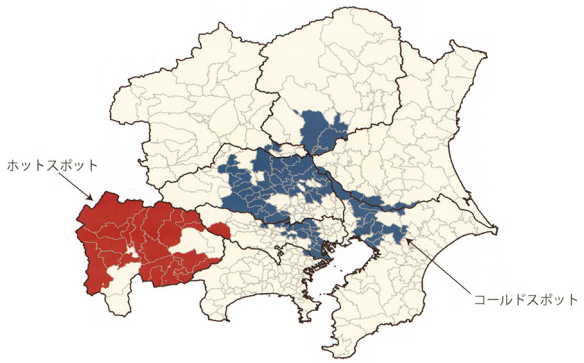


図 14. 有意な空間集積性が認められた地域群．それぞれ赤色がホットスポット，青色がコールドスポットを示す.

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.

統計数理(2025)
第 73 巻 第 1 号 101–116
©2025 統計数理研究所
特集「空間統計モデリング：理論と応用」
［原著論文］

位置登録情報を利用した長崎の観光地間の相互影響力評価

一藤 裕[1]・村上 大輔[2]

（受付 2024 年 6 月 30 日；改訂 11 月 11 日；採択 11 月 11 日）

要　　旨

観光は，日本の主要産業の一つである．特に地方都市では，地域活性化の役割を担っている．新型コロナウィルスにより大きなダメージを受けたが，コロナ前の状況，もしくはそれ以上の観光客数の増加により観光が活性化している．今後は，観光客をさらに誘致する方策や，オーバーツーリズムとなっている地域では観光客の抑制や制御する方策が必要となっている．しかし，個人旅行が増加しているため，全体的な動きや観光地間の移動に関する分析は難しい．そこで本研究では，通信キャリアの位置登録情報を利用して，観光地間の関係性を評価し解釈する方法を提案し，定量的に評価することで観光客誘致のための基礎的なデータを創出することを目的とする．本提案手法は，各観光地のユニークユーザ数と観光地間の移動数，ユーザの居住地情報を利用して，観光地間の相互影響力を評価するものであり，各地における来訪者数の増加の影響が，位置登録情報から推定された移動者数に比例的に拡散していくことを仮定する多変量時系列モデルがベースとなっている．この手法を長崎県の観光地に適用し，観光地間の関係性を評価した結果，居住地・年代ごとに行動が異なることが示され，居住地・年代ごとに広告宣伝方法を変えることで効率的に観光客の誘致ができる可能性が示された．

キーワード：観光，位置登録情報，個人属性情報，相互相関，ブースティング．

1. はじめに

観光は日本の主要産業の一つとなっている．東京や京都といった主要都市だけでなく，特に地方都市にとっては地域活性化を実現するための大きな柱の一つと言える．観光客誘致のために，地方自治体は様々な活動を積極的に行っている．例えば，観光地ごとに観光地域づくり法人(DMO)が立ち上がり，観光庁 (2024) が登録を管理し，令和 6 年 4 月時点で全国 301 団体が登録されている．DMO は，観光庁によると「地域の多様な関係者を巻き込みつつ，科学的なアプローチを取り入れた観光地域づくりの司令塔となる法人」とされている．よって，観光客誘致のための科学的根拠となるデータを準備することが求められている．

従来の観光は団体旅行が主流だったため，観光客の行動を把握することが容易であった．しかし，現在の日本国内の観光旅行は個々人の嗜好が多様化したため個人旅行が主流となっており，画一的な行動分析では観光客誘致の根拠として十分な対応が出来なくなっている．よって，個々人の行動を把握し，観光客が訪れる要因や行動分析に対応することが必要となる．個々人

[1] 長崎大学 大学院総合生産科学研究科：〒 852–8521 長崎県長崎市文教町 1–14
[2] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

の観光客の行動を把握する手段として，スマートフォンに搭載される Wi-Fi 機能や GPS，携帯基地局のログデータが挙げられる．

Wi-Fi 機能に注目した行動分析結果を利用したサービスとして，GMO（2024）や Cinarra Systems（2024）が提供する来店分析が挙げられる．これは，Wi-Fi の特性を利用したものである．店にアクセスポイントを置いておくと，Wi-Fi の機能を ON にしている端末を持った利用者が入店し通信可能範囲に入ると，通信のためのやり取りを行い通信した端末 ID がログに残る．通信した端末 ID と通信契約時に同意した個人属性情報を紐づけることで，個人の属性別に来店傾向を分析でき，集客につなげるための根拠を提示する．このサービスは，店のアプリと連携することでポイント付与などのインセンティブを用意できるため，利用者の行動分析には向いている．しかし，観光地のような広大な場所に設置することは難しく，また，観光地は定期的に訪れる場所ではないため，アプリと紐づけるといったことが難しく，データ数が十分に得られない問題がある．また，Wi-Fi パケットセンサーを使ってデータを収集し，来場者の推定や Origin-Destination（OD）票の作成，観光地間の移動パターンを推定する関連研究が多くされている（Gao and Schmocker, 2022; 中西 他, 2018; 寺部 他, 2019; 神谷, 2018; 遠藤 他, 2019）．これらの研究は，観光地の限定的なエリアに Wi-Fi パケットセンサーを設置し，端末の MAC アドレスを利用して観光客数や OD データの生成，得られたデータから移動パターンの分析を行っている．観光客の細かな誘導・行動分析に関しては有益であるが，Wi-Fi パケットセンサーが捉えられるデータ範囲が理想値で半径 100m であり，対象観光スポットが広域の場合，設置場所や数を考慮しなければならない．また，MAC アドレスを接続先などでランダム化する機能が搭載されたスマートフォンの出現や，観光客がどこから来たのかなど観光客誘致に必要な個人の属性情報がこのデータには紐づかないため，観光客誘致の根拠として利用するには十分ではない．

次に，GPS のログデータを利用した行動分析サービスとして，Agoop（2024）の流動人口データが挙げられる．これは，スマホアプリから取得した GPS などの位置情報を秘匿化・統計加工した情報を使い，エリア毎の時間経過による来訪・滞在人口の推移や，OD などの人流を可視化したものである．マーケティング，需要予測，観光調査など様々な分野で活用できる．しかし，GPS データはアプリで取得することが主で，そのアプリをインストールされていない観光客の情報は取得できない．特に地方で分析を行う場合，数が十分にないと統計化された時点で具体的な数値が分からず分析できない可能性がある．また，GPS データを使った観光客の行動分析研究も多く行われている．例えば，Liu et al.（2022）は，登山に関する GPS データを用いた行動パターンのクラスター分析を行った．この GPS データはオープン化されているものであり，また，登山道にも限りがあるため，データ量が少なくてもデータの補完が可能でクラスター分析には耐えられる．しかし，県全体の観光地を想定した場合，十分なデータが集まらないという問題がある．D'Angelo et al.（2023）は，イタリアとクロアチアでクルーズ客に GPS 追跡デバイスを配布し，データ収集およびアンケート調査を実施し，観光客の移動パターンを収集した．観光地とそのルートを直線で結び空間点過程モデルを用いて分析し，観光客の滞留パターンを定量的に評価した．その結果，観光地からの距離やクルーズ客同士の相互作用が最も行動に影響力があることを明らかにした．個人を特定できる範囲で同じ属性を持った集団を対象とする分析は，観光行動の要因を明らかにするために非常に重要であるが，本研究のように広域かつ不特定多数を長期間対象として分析を行うにはあまり適さない．特に GPS と個人属性情報が紐づいたデータセットを日々収集することは難しいからである．

最後に，位置登録情報を利用した行動分析として，NTT docomo（2024）のモバイル空間統計を筆頭に，各社，携帯電話端末が事前準備として，携帯電話網に接続した際に登録される基地局単位の位置情報を用いた統計サービスが挙げられる．これらは，システム上ほぼリアルタ

イムに様々な人口統計を作成可能となっており，社会基盤の一つとしてデータ利活用が期待されている．例えば，大井（2022）は，KDDI が公表している全国主要観光地の人流データを用いて 23 の観光地の日時データを作成した．そのデータを用いて観光に対し新型コロナウィルス感染症の影響が縮小している傾向を示した．また，観光需要の変動要因として時間軸に着目し，その要因として 1 週間内の変動が最大の影響を与えていることを明らかにしている．また，Yamaguchi and Nakayama（2023）は，モバイル空間統計を使って，新型コロナ感染症の感染拡大時期における長距離移動パターンの変化を分析し，観光目的の移動が大幅に減少したがビジネス目的の移動はある程度維持されていることを明らかにしている．Nakanishi et al.（2018）もモバイル空間統計データを利用し，日本全国を 9 つの地域に分割し，その地域間の旅行者数の変化の要因を分析した．その結果，季節や週ごと，経済状況および気候が要因であることを明らかにしている．大井（2020）はモバイル空間統計データの訪日外国人データを用いて，大阪市と和歌山市のインバウンドの観光の季節変動を分析した．その結果，大阪市では観光需要の安定化が進んでいるが，地域格差が存在していることを示した．その一方で和歌山市はデータ数が限られていたため，特定のエリアでのみ分析できなかったことを述べている．Qian et al.（2021）は，観光地，休憩場所，交通ハブの空間的な相関を明らかにし，観光施設の相互関係を評価する枠組み，特に観光地間の結びつきを強化するためのバスの路線の追加について検討した．このように観光客の行動を分析するために，携帯電話の位置情報を利用することは非常に重要であり，観光客の滞在数や移動に着目しそのパターン化や要因を分析する研究が活発に行われている．また，観光地間のつながりについても移動手段や OD データによって，隣接する観光地や交通のハブなどの結びつきを評価し，ネットワークのハブとして整備する都市計画の効率化に主眼が置かれている研究が多く，特定の観光地の観光客数を観測し，その結果から他の観光地の観光客数を推定するような研究は行われていない．加えて，Aasa et al.（2021）が携帯電話の位置情報を利用した統計情報の有用性について報告しているが，その一方で，個人属性情報が位置登録情報に完全に紐づくため，その扱いに関する議論が継続的に行われており，データ利活用の障壁が高いことも併せて報告している．また，詳細に広く追跡できてしまうため，可視化した後に自治体側がどう活用してよいか分からないという問題もあり，位置登録情報の社会での利活用方法の実装は議論の余地が多くある．

そこで本研究では，位置登録情報の観光客誘致への利用に向けた検討として，位置登録情報から得られる観光地間のつながりが，各観光地の来訪者数に及ぼす影響を分析する．具体的には，観光地別・週別来訪者数の時間変化傾向を解析し，週を追う毎に変化する観光地間の結びつきが，翌週以降の来訪者数に及ぼす影響を評価することを目的とする．もし観光地間の結びつきの強さが来訪者数に影響しているのであれば，個別地域ごとの誘致だけでなく，結びつきの強い観光地も含めた一体的な誘致も有効である可能性がある．また，他の観光地に強く影響するような観光地が特定できれば，同観光地への誘致を積極的に行うことで，周囲の観光地への波及効果も期待できる．反対に，観光地間の結びつきの強さが来訪者数に影響していないのであれば，個別観光地ごとに誘致を行えば十分ということになる．以上のように，各観光地間の相互影響関係とその影響を明らかとすることは，誘致を効果的に行う上で重要である．

しかしながら，相互影響関係を評価する方法は必ずしも明らかではない．観光地別・週別の時系列データが利用できる場合であれば Vector autoregressive（VAR）モデルなどの多変量時系列モデルが応用できる．しかし，本研究のように数十の観光地が分析対象の場合，仮に全ての時点で観光地間の結びつきの強さは同じと仮定したとしても，観光地間のつながりを表すパラメータの数は（観光地数）^{2} と多くなり効果の識別は困難である．さらに，結びつきの強さが時点によって変わるというより現実的な仮定をおいた場合，同パラメータ数は（地域数）$^2 \times$ 時点数 となり，効果の識別はより困難となる．時点毎にパラメータを推定するために，パラメータ

を時間の関数で与える VAR モデルの拡張（Christopoulos and Leon-Ledesma, 2008; Bringmann et al., 2018）や，直近の標本により大きな重みを与えることで局所推定を行おうという VAR モデルの拡張（Casas and Fernandez-Casal, 2019; 村上・松井, 2022）もあるが，それらは必ずしも推定結果の解釈性を担保するものではない．

以上を踏まえ，本研究の章立ては次の通りとした．2 章では位置登録情報および個人属性情報を利用して，長崎県内の観光地毎の属性別観光客の来訪者数データを作成する．また，観光地間の結びつきの強さを表すデータとして OD データも併せて作成する．3 章では，同 OD データで表される観光地間のつながりが各地の来訪者数に及ぼす影響を評価するための新たな手法を提案する．4 章では，同手法を用いて観光地間の相互影響力を評価し，ある観光地に人が増加した場合，他の観光地にどのような影響を及ぼすか，居住地によって行動がどう異なるかを定量評価する．最後に 5 章で結果を要約し今後の展望について論じる．なお，周遊行動をはじめとした個別観光客の時間帯毎の移動パターンなども考慮したより詳細な解析については今後の課題としたい．

2. 利用データとプライバシー保護の取組み

2.1 位置登録情報の詳細

本稿では，大手携帯電話事業者の一つである株式会社ソフトバンクの位置登録情報を利用する．今回利用した位置登録情報には，居住地（関東エリア，福岡県，長崎県），毎週日曜日のユニーク検知数，年代（5 歳刻み．紙面の制限のため，今回は 30–34 歳，60–64 歳のみを比較）の属性情報が付随する．位置情報が記録されるタイミングは，基地局の通信カバー範囲に入ったタイミングであり，同じ範囲に居続ける限り情報は更新されない．通信カバー範囲と観光地の対応付けはされているため，観光地ごとの検知数を算出できる．また，居住地の関東エリアは，東京・千葉・神奈川・埼玉・茨城・栃木・茨城・長野・群馬の羽田空港を利用して長崎を訪れる可能性の高い地域をまとめたものである．これらのデータからユニークユーザ数を集計した来訪者数データと OD データの 2 種類を利用する．

来訪者数データは，観光地毎に居住エリア別，日別かつ年代別で集計したものである．また，OD データは，検知された観光地に来る前に，最後にどこで検知されたかの情報を利用して，2 地点間の移動検知数を集計したものである．ただし，直前の位置情報が検知された日時については考慮しないものとする．例えば，A 地点で 6 月 1 日 8 時〜10 時に検知され，直前に B 地点にいた場合を考える．B 地点で検知された日時は考慮しないため，B 地点で 6 月 1 日 6 時〜8 時に検知されていても，もしくは，それ以前に検知されていた場合でも同様に「B → A」の検知数に集計される．対象期間は，2022 年 8 月 1 日から 2023 年 7 月 31 日までとした．また，今回は観光客を分析対象とするため，毎週日曜日データのみを用いることとした．

2.2 位置登録情報のプライバシー保護の取組みについて

位置登録情報には，通信の秘密には該当しないものの，プライバシーの観点で強い保護が求められることから，十分な匿名化の実施が必要となる．「十分な匿名化」には，位置情報と付帯情報とを結合して作成したデータを用いることができる．ただし，結合することのできる付帯情報（電気通信事業者が保有する契約者に係る情報）は，性別，年齢，市区町村までの住所や利用者の趣味趣向情報である．また，「十分な匿名化」により加工した位置情報について，次のすべての条件を満たせば，事前の包括的同意に基づいて活用をすることができるとしている．

- オプトアウトが適切に提供されている．
- 加工及び運用管理体制が適切である．

• プライバシー影響評価(PIA)が適切に運用されている．

株式会社ソフトバンクでは，位置登録情報を含む運用データの取得・保有・利用に関する取扱いを Web 上に掲載している（SoftBank, 2024）．これに基づき，個人を特定可能な項目を符号や番号等へ置き換える仮名化と，個人を特定可能な項目を要約等する一般化の方法を用いて匿名加工処理を施す．その後，匿名加工情報を，性別，年代，居住エリア毎に集計し，かつ対象となるデータ内に同じ属性を持つ数値データが k 件以上存在するようにデータを変換する k-匿名化処理(k=10)も併せて実施し，十分に匿名化した上で，統計データを作成する．このように処理されたデータを用いて観光地間の相互影響力を評価する．

2.3 観光地の選定と観光客の傾向

本稿で選定した長崎県のエリア(観光地や交通ハブ)とその特徴について述べる．長崎県内の観光地は県内に散らばっており，すべての観光地を 1 日で回ることは難しい．そこで，図 1 に示す通りに観光エリアを選定した．具体的には，長崎市内(長崎駅，浜町，平和公園，グラバー園，出島，ペンギン水族館)，時津町，佐世保市内(佐世保駅，ハウステンボス)，諫早市，西海市，平戸市，波佐見町，松浦市，島原市，雲仙市，九十九島エリア，大村空港，口之津の計 20 エリアである．この中で，長崎駅，佐世保駅，大村空港，口之津は交通のハブである．また，雲仙市は小浜温泉エリアと雲仙温泉，島原市は世界遺産群の一部，九十九島エリアはリゾート地，波佐見町は焼き物，諫早市は V・ファーレン長崎のホームスタジアム，その他のエリアは歴史的遺産があるという特徴がある．

上記エリアの来訪者数は，エリアと紐づく基地局のログから算出される．そのログから 2 時間ごとに集計しているため，2 時間ごとの集計では同一端末は重複してカウントされないが，1 日単位では重複してカウントされる場合もある．例えば，別のエリアに移動し，その後，元のエリアに戻ってきた場合は，2 時間を超えていれば再度カウントされることになる．この集計ルールに基づいてエリア別の来訪者数を居住地(長崎県，福岡県，関東エリア)と年齢(30–34 歳，60–64 歳)について図 2 にプロットした．図 2 から，来訪者数の週毎の変動は比較的少ないことや，エリア毎の差は大きいことが確認できる．長崎県居住者は各エリアを比較的満遍なく訪問している．一方で，福岡県民と関東エリアの住民の訪問先は，長崎駅，浜町，ハウステンボス，佐世保駅といった中心エリアに集中していることがわかる．また，福岡県や関東エリ

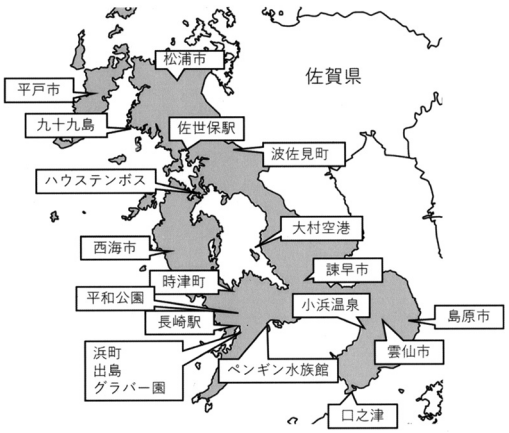


図 1．長崎県内の観光地および交通ハブの位置関係．

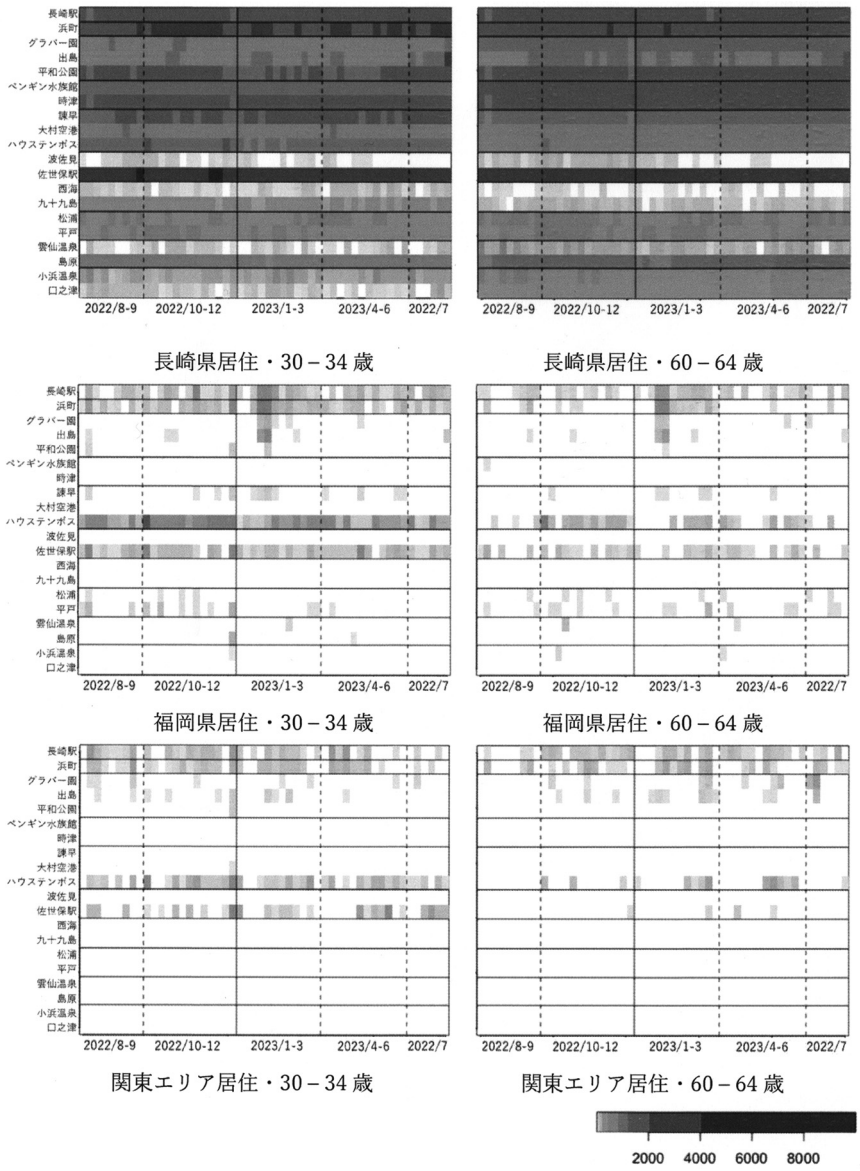


図 2. 居住エリア・年齢別の滞在者数.

アからの来訪者数は，30–34 歳の方が 60–64 歳よりも多いことなども確認できる．

3. 観光地の相互影響力評価モデルの提案

本研究では，あるエリアでの来訪者の増加が，同エリアのその後の来訪者数や，周辺エリアのその後の来訪者数に及ぼす影響を推定する．それにより，エリア別来訪者数の相互依存関係を明らかとする．

まず，エリア $s \in \{1, \ldots, S\}$ で週 $t \in \{1, \ldots, T\}$ に来訪者が 1 単位増えることを，第 s 要素が

1 でその他が 0 の one-hot ベクトル $\delta(s,t)(S\times 1)$ を表す．本研究では，この 1 単位増加に伴う L^* 週先のエリア別来訪者数の増分は下式に従うと仮定する：

$$z(s,t+L^*,L)=\prod_{l=1}^{L^*}\rho_{(L)}^{l}W_{t+l}\delta(s,t). \tag{3.1}$$

W_t は第 t 週におけるエリア間の移動者数の行列 $(S\times S)$ であり，その要素は OD データから与える．L を来訪者数増加の影響が残る期間とすると，パラメータ $\rho_{(L)}$ は期間内での時間減衰を表現するために $\rho_{(L)}^{L}=0.1$ を満たすように与える $(\rho_{(L)}=0.1^{1/L})$．ただし L は未知とする．上式はエリア s，時点 t における来訪者増加の影響が，移動者数に従う形で他地域に拡散していき，その影響の 90% が L 期先までに消失することを表している．例えば $\rho_{(L)}^{L}=0.01$ とすれば L 期までの影響の消失率が 99% となるなど，$\rho_{(L)}$ の設定には恣意性が残るが，影響の及ぶ期間は L で調整されるため，L との識別の難しさも勘案して消失率は 90% で固定することとした．

本研究では，エリア s，週 t，期間 L を変えながら来訪者数増加の影響を足し合わせていくことで，単純集計データから得られるエリア別，週別の来訪者数行列 Y $(S\times T)$ に対する予測行列 $\hat{Y}$ を最適化する．(3.1)式で記述される $t+1$ 週から $t+L$ 週までの来訪者数の増分を対象期間全体で並べた行列を $Z(s,t,L)$ $(S\times T)$ と表すと，その推定手順は以下のとおりである：

(1) 予測行列の初期値 $\hat{Y}_0$ を設定．$i=0$.
(2) 以下の手順で予測行列を更新：
  (a) 週 t_i をランダムに選択
  (b) エリア $s\in\{1,\dots,S\}$ と拡大期間 $L\in\{1,\dots,L_{max}\}$ の全組み合わせについて，以下の手順で目的関数を評価（グリッドサーチ）：
    i. 週 t_i，エリア s で 1 単位増加した影響が L 期先までの続いた場合の，来客数の増分 $Z(s,t_i,L)$ を(3.1)式を用いて評価．
    ii. $vec(Z(s,t_i,L))$ を残差 $vec(Y-\hat{Y}_i)$ に回帰して回帰係数 $\hat{\beta}_i$ を推定．新たな予測行列 $\hat{Y}_i+\hat{\beta}_iZ(s,t_i,L)$ を得る．ただし $vec(\cdot)$ は行列をベクトル化する演算子である．効果の識別を容易にするために，ここでは $\hat{\beta}_i\geq 0$ を満たすように制約を課して最小 2 乗推定している（Lee and Seung, 2000）．
    iii. 同予測値の目的関数を評価．
  (c) 手順(2) (b)で目的関数を最良にしたエリア s_i と拡大期間 L_i を用いて，予測行列の候補 $\hat{Y}^*_{i+1}=\hat{Y}_i+\hat{\beta}_iZ(s_i,t_i,L_i)$ を与える．
  (d) $\hat{Y}_i$ よりも目的関数が改善する場合は $\hat{Y}^*_{i+1}$ を採択 $(\hat{Y}_{i+1}=\hat{Y}^*_{i+1})$，さもなくば $\hat{Y}_{i+1}=\hat{Y}_i$ のままとする．
(3) i を $i+1$ に更新．
(4) 手順(2)–(3)を，目的関数が改善しなくなるまで繰り返す．

同手法は誤差を減らすように弱学習器を合成していくブースティングの一種である（Freund and Schapire, 1996）．なお，週 t_i をランダムに与えたのは，週数が大きくグリッドサーチにした場合に計算コストの増大が見込まれた点，並びに局所解への収束を避けるためである．目的関数として本研究では Bayesian Information Criterion(BIC)を用いた．以上で得られる時点 T の予測来訪者数の予測値ベクトル $(S\times 1)$ は下式となる：

$$\hat{Y}_T=\sum_{i=1}^{l}W^*_{T,i}\delta(s_i,t_i) \tag{3.2}$$

ただし $W^*_{T,i}=\hat{\beta}_i\delta(t_i+1\leq T\leq t_i+L_i)\prod_{l_i=1}^{T-t_i}\hat{\rho}_{(L_i)}^{l_i}W_{t_i+l_i}^{l_i}$ である．

$W^*_{T,i}$ 同行列を全反復について足し合わせた $W^*_T = \sum_{i=1}^{l} W^*_{T,i}$ は時点 T におけるエリア間の相互相関関係を表す行列になっている．その第 (s, s') 要素は，エリア s' の来訪者数が増えた場合に，翌週以降にエリア s の来訪者数が増える傾向がある場合には正，さもなくば 0 となる．従って，W^*_T の列和が大きいエリアは各地の来訪者数に影響するエリア(以後，主要エリアと呼称)，W^*_T の行和が大きいエリアは主要エリアに依存して来訪者数が上下する従属エリアと考えることができる．以上のような解釈をもとに，4 章では W^*_T の推定結果をもとに，エリア間の主従関係をみていく．ここで，本研究における主要エリアとは，あくまで他地域への強い影響が推定された地域のことであり，観光客の多くが同エリアを経由して周遊していることを意味しているわけではない点には注意されたい．

なお，一連の推定ではパラメータ β_i の符号を非負に制約することで識別の問題を緩和している．具体的には，極端に大きな正の推定値が，極端に小さな別パラメータの負の推定値で相殺されることにより生じる特異推定を回避している (村上・松井, 2022)．また，予測式は観光地における来訪者増加の影響の拡散を明示的にモデル化した(3.1)式の線形和に制限されているため特異になりにくく，かつ解釈しやすい．以上より，提案手法は推定の安定性や解釈性を重視して観光地間の相互影響関係を評価しようという手法となっている．

4. 提案手法を利用した長崎県における観光地間の相互影響力評価

本章では，毎週日曜日のエリア別来訪者数(30–34 歳と 60–64 歳)を提案モデル((3.2)式)に適用することで，観光地間間の相互影響力を評価し，関係性を明らかとする．

4.1 年間の観光地間の相互影響力評価結果

提案モデルから推定されたエリア間の相互相関行列を見ていく．まずは，全期間の平均値を図 3 に示す．行列の第 (i, j) 要素の色が濃いことは，エリア i の傾向が，エリア j の翌週以降の傾向をよりよく説明することを意味する．図 3 上段から，長崎県民にとっての主要エリアは，長崎駅，浜町，平和公園，時津，諫早，佐世保駅などであり，長崎県内の各エリアの集客状況は，それら主要エリアの集客状況で説明されるとの示唆を得た．

福岡県民にとっての主要エリアは，長崎駅，浜町，ハウステンボスエリアなどであり，それらのエリアの傾向から，翌週以降の長崎市内(長崎駅〜平和公園)の傾向が説明されること，ならびに，ハウステンボスの傾向から，翌週以降の佐世保周辺エリア(ハウステンボス含む)の傾向が説明されるとの推定結果となった．関東エリアからの来訪者にも類似した傾向がみられたが，長崎駅・浜町エリアの来訪者数は翌週のより広域の来訪者数を説明する傾向がみられた．

以上のように，各地の来訪者数は比較的少数の主要エリアの傾向で説明されることや，主要エリアの数は，県外からの来訪の場合に少なくなる傾向がみられた．これは，長崎県内居住者の方が，比較的マイナーな観光地を訪れることや，県外からの来訪者が主要な観光地を選好するための可能性がある．

参考までに，時系列相互作用を評価する標準的手法である Vector autoregressive(VAR)モデルを用いて地域間の相互相関を推定した．ここでは，安定した解を得るために SparseLag 損失 (Nicholson et al., 2017)を導入した．VAR モデルから得られた相互相関行列を図 4 に示す．図 4 から，提案手法とは異なり長崎駅や浜町がハブとなっておらず，また例えば雲仙温泉が出島エリアとだけ強い相互相関を持つなど解釈に窮する箇所が散見される．提案手法を用いることで，エリア間の主従／相互関係が，より解釈しやすい形で推定されることが確認された．

解釈しやすい結果が得られた理由には，ある観光地における来訪者増加の影響の拡散を明示的にモデル化した点や，同影響の符号を非負に限定することで効果の識別を容易にした点

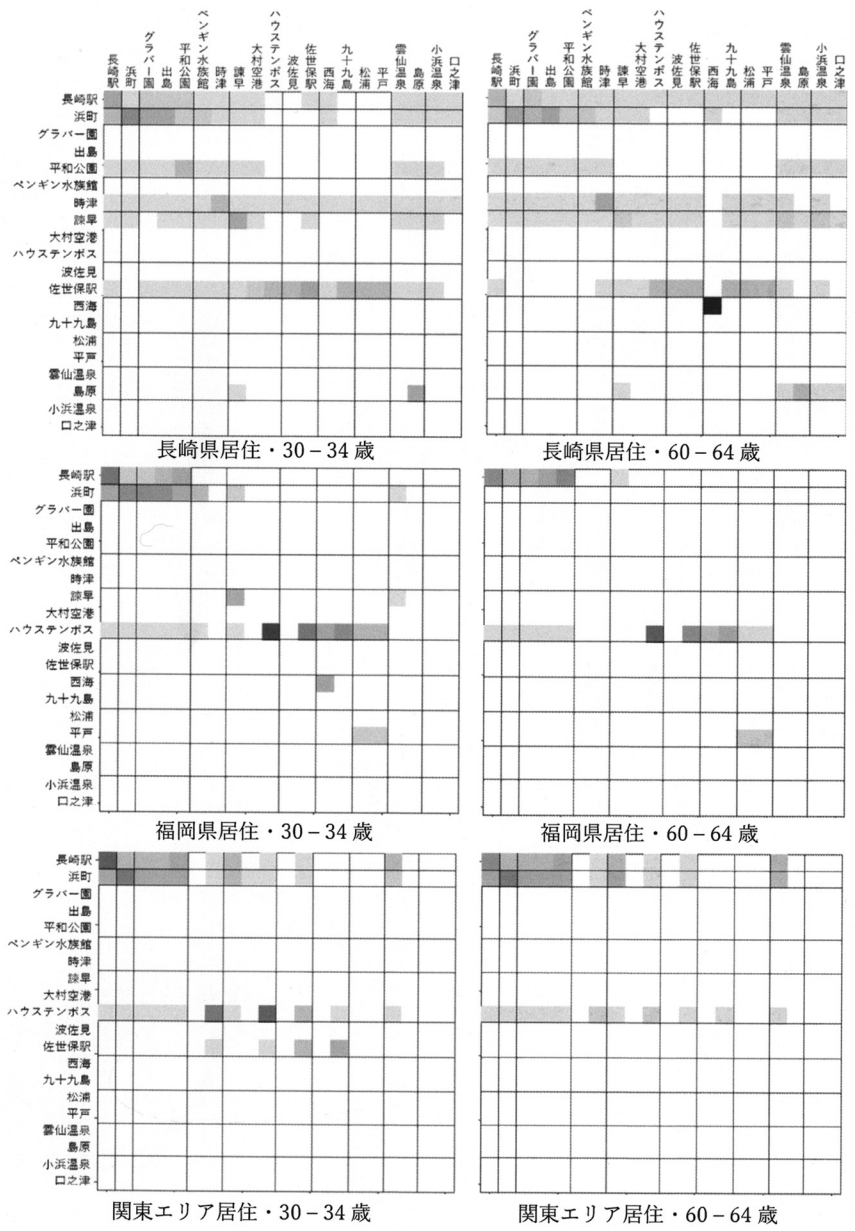


図 3. エリア間の相互作用行列の推定結果(全期間の平均).

($\hat{\beta}_i \geq 0$) などが考えられる．推定したパラメータの数が極めて多いことを踏まえると(1 章参照)，妥当と思われる以上の仮定を導入することで地域間・週別の相互影響関係が解釈しやすい形で評価できるという知見は興味深い．言い換えると，おおむね予想されていた傾向が実データ分析を通して確認された.

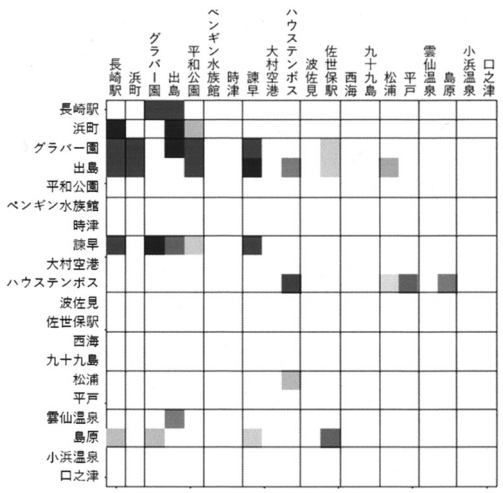


図 4. VAR モデルから推定された相互相関行列(福岡県居住・30–34 歳).

4.2 週毎の観光地間の相互影響力評価結果

週毎に推定された相互相関行列 W_T^* をみていく．紙面の関係から，ここでは同行列の対角要素，列和，行和についてそれぞれ週別に見ていく．

W_T^* の対角要素の大小を濃淡で表したものを図 5 に示す．図 5 で色が濃いことは，同一エリア内の時系列相関で説明される来訪者数が多く，先週までと似た来訪者数となりやすいことを意味する($\hat{\beta}_i \geq 0$ を仮定したため)．長崎県内からの来訪者数(30–34 歳)に関しては，半数程度のエリアで系列相関がみられた．特に長崎駅，浜町，諫早市，佐世保駅，島原市などの主要エリアでは，期間を通して系列相関が強く，直近と同数程度のまとまった来訪があるとの示唆を得た．同じ長崎県民でも，60–64 歳に関しては，2022 年末頃までは系列相関がみられないエリアも多く，通年で系列相関がみられたのは長崎駅・浜町エリアのみであった．これはコロナウィルスの感染を回避するための自粛をより長い期間行ったためである可能性がある．

福岡県からの来訪者数に関しては，長崎駅，浜町，ハウステンボス，佐世保駅，平戸市で系列相関がみられ，関東からの来訪者に関しては，より少数の主要エリアでのみ系列相関が確認された．以上より，遠方であるほど，複数週にわたる(系列相関で説明されるような)まとまった来訪者があるエリアは限られるとの示唆を得た．なお，長崎県民からハウステンボスへの来訪者に関しては系列相関がみられず週毎に傾向が異なると推定された一方で，福岡県または関東エリアからハウステンボスの来訪者に関しては強い系列相関がみられた点は興味深い．

次に，週毎の W_T^* の列和を並べて図 6 に示す．色が濃いことは，翌週以降の他エリアの傾向をより良く説明する主要エリアであることを意味している．図 6 から，長崎県民の各地の来訪者数は，長崎駅，浜町，佐世保エリアにおける傾向で説明されると推定された．それらエリアが人口密集地であり多くのトリップが発生していることを踏まえると，この結果は自然である．福岡県民と関東エリアの住民に関しては，上記 3 エリアに加えてハウステンボスでも色が濃くなっている．県外からの来訪者にとっては，ハウステンボスもまた他エリアの傾向を説明する主要エリアとなっているとの示唆を得た．

最後に，週毎の W_T^* の行和を並べて図 7 に示す．色が濃いことは，主要エリア(長崎駅，浜町，佐世保，ハウステンボスエリア等)でより多くの傾向が説明される従属エリアであることを意味する(3 章参照)．図 7 から，長崎県民に関しては，各地の来訪者数(30–34 歳)が主要エ

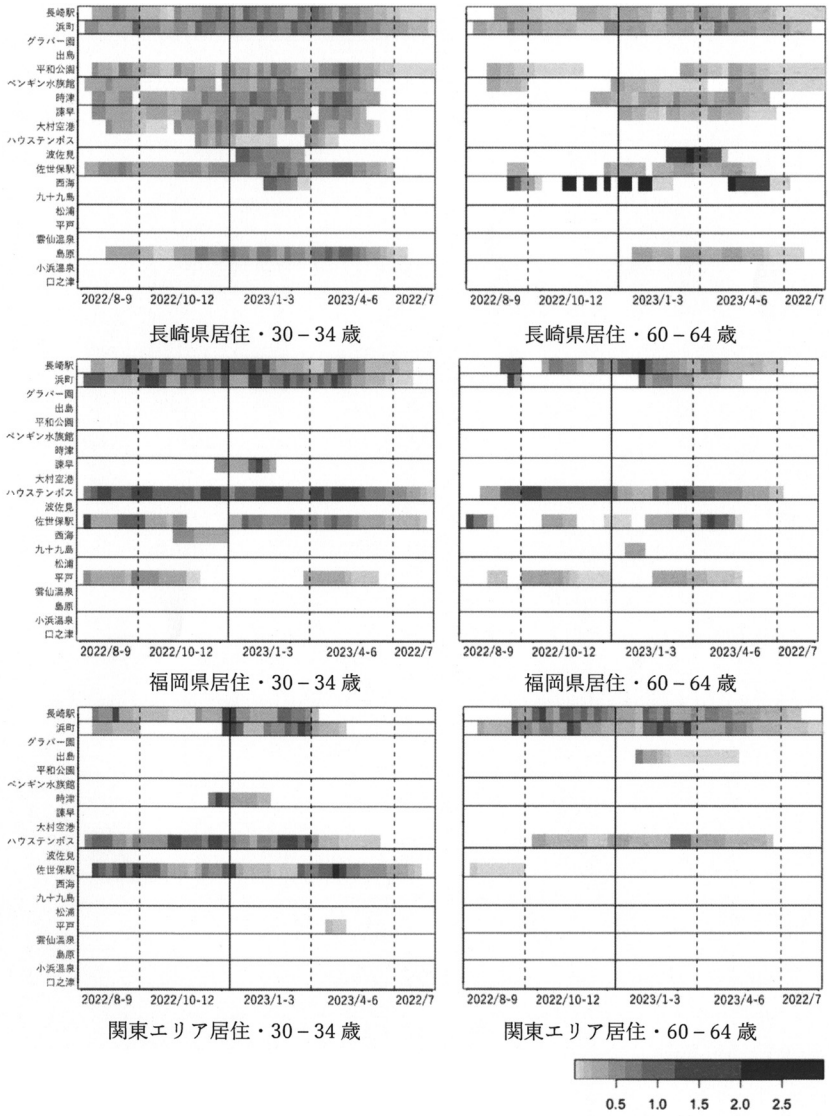


図 5. 同一エリア内の系列相関で説明される来訪者数の推定結果(相対値).

リアの傾向で説明される従属エリアとなっていることが確認できる．60–64 歳については 2022 年末までは相互相関がみられない地域もあり，コロナ渦で年齢による来訪パターンに違いがあった可能性がある．福岡県民に関してはペンギン水族館，佐世保駅，九十九島エリアなどの傾向が長崎駅エリアの傾向で説明されると推定された．関東エリアの居住者に関しても，長崎駅と平和公園の間で顕著な相互相関がみられたものの，福岡県民とは異なり，ペンギン水族館や九十九島エリアとは相互相関がみられず，また全体として相互相関がみられたエリアは少なかった．以上に加え，居住地によらず長崎市内の観光地間で比較的強い相互相関がみられた点や，福岡県民と関東エリアの住民で差がみられた点は興味深い．

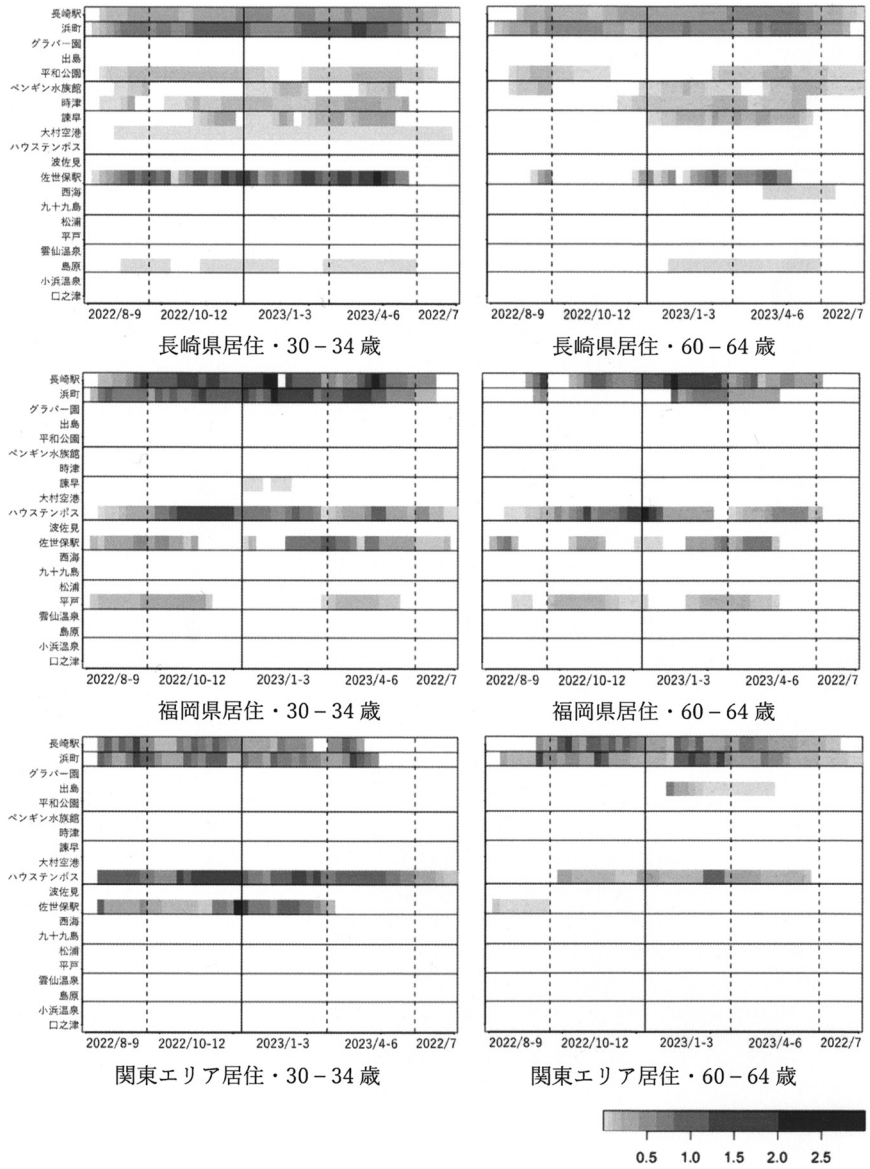


図 6. 他エリアへの影響力の推定結果.

5. まとめ

本稿は，主要産業の一つである観光を活性化するために，観光客誘致の根拠となるデータの創出を目的とした．そのため，携帯電話の位置登録情報を利用して，あるエリアでの来訪者数の増加が，同エリアおよび他のエリアのその後の来訪者数に及ぼす影響を推定する手法を提案し，それを基に，30–34 歳と 60–64 歳の年齢別かつ居住地別でエリア別来訪者数の相互依存関係を明らかにし観光地間の相互影響力について考察した．その結果，長崎県民は年齢別でみると，コロナの影響による自粛が 60–64 歳のグループでは見られたが，30–34 歳グループには見

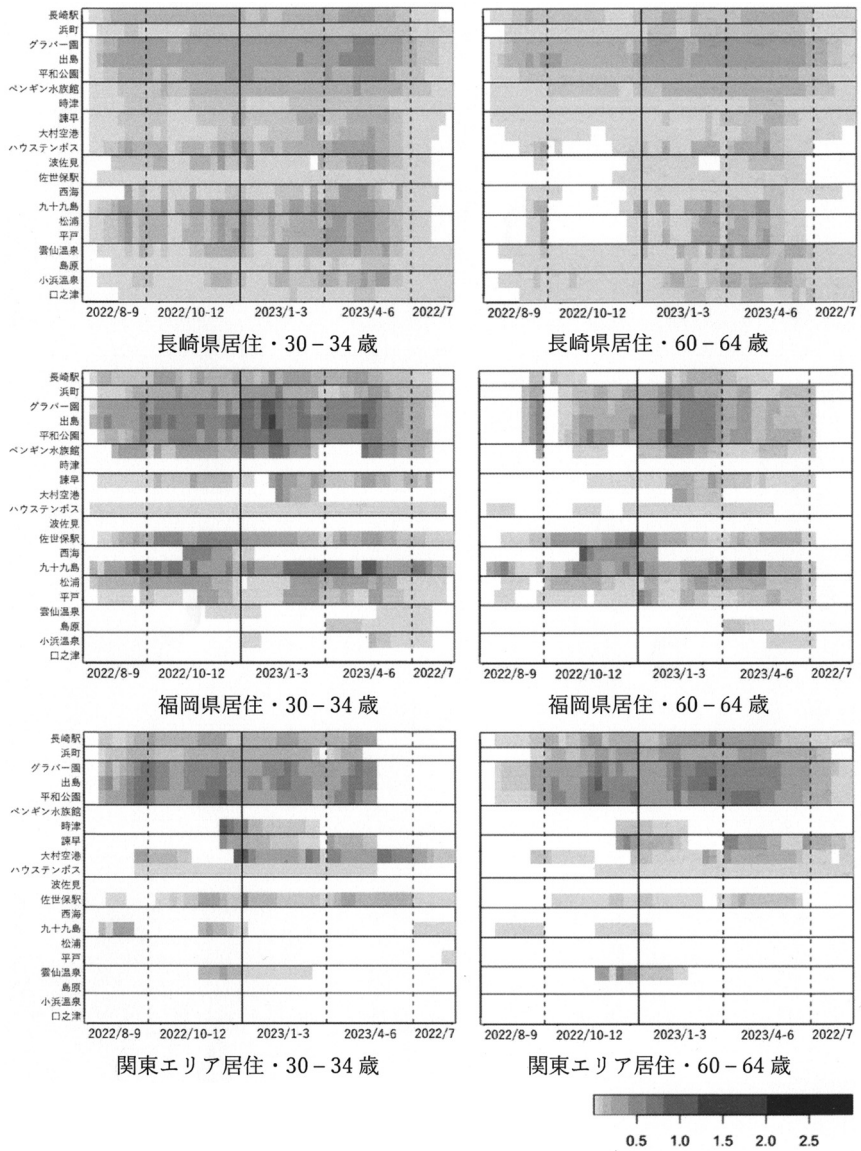


図 7. 他エリアから受ける影響力の推定結果.

られないなど年代でも行動が大きく変わることが示された．また，エリア別で比較すると，長崎県民と他県の住民の観光地の選択が異なっており，長崎県民は主要な観光地をはずし，県外からの観光客は主要観光地を目指して訪れる傾向があることが示された．さらに，他県からの観光客が訪れる先が主要観光地のため，特定の観光地に注目することで他のエリアの観光客の傾向を十分に推定できることが明らかとなった．これらを踏まえ，例えば，他県からの来訪者に対しては，マイナーな観光地の誘導が十分ではないため，情報発信を積極的にするなどし，主要観光地以外へ誘導するといった観光客誘致方法が効果的であるなどの複数の可能性が示された．

その一方で，今回の分析に利用したデータは，観光客とビジネス客を区別していない．よって，今後は観光政策の立案の根拠として役立てるために，観光客のみを対象とした集計データを使って観光地間の結びつきを追加評価することも検討したい．これに加え，日曜日だけでなく平日や時間帯別での評価方法の確立や，他県の観光地における観光客の行動との比較を行い各県の特徴を明らかにすること，観光客誘致のための応用方法についての検討も行いたい．今回は実態把握のために移動データのみで観光地間の相互相関関係を評価したが，より解釈性の高い分析を行って観光政策に役立てるためには観光地間の因果関係や観光客の移動理由を考慮してモデル化することが望ましい．時空間データに特化した因果推論法（Reich et al., 2021）や，Directed acyclic graph を用いた柔軟な因果推論法（Textor et al., 2016）なども提案されており，それらを応用した因果関係の特定も今後取り組みたい．また，分析期間を伸ばすことも重要な課題である．今回は 1 年分しかデータが得られなかったが，複数年のデータが得られれば，季節変動，過去から未来への長期的変動，パンデミックに伴う比較的短期の変動などが識別できる可能性がある．それらの識別も今後検討したい．

謝　辞

本研究は JSPS 科研費 JP23K21816 の助成を受けたものです．

参 考 文 献

Aasa, A., Kamenjuk, P., Saluveer, E., Šimbera, J. and Raun, J. (2021). Spatial interpolation of mobile positioning data for population statistics, *Journal of Location based Services*, **15**(4), 239–260, https://doi.org/10.1080/17489725.2021.1917710.

Agoop (2024). マチレポ, https://agoop.co.jp/（最終アクセス日 2024 年 6 月 1 日）.

Bringmann, L.F., Ferrer, E., Hamaker, E.L., Borsboom, D. and Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model, *Multivariate Behavioral Research*, **53**(3), 293–314.

Casas, I. and Fernandez-Casal, R. (2019). tvReg: Time-varying coefficient linear regression for single and multi-equations in R, SSRN, http://dx.doi.org/10.2139/ssrn.3363526.

Christopoulos, D.K. and León-Ledesma, M.A. (2008). Testing for Granger (non-)causality in a time-varying coefficient VAR model, *Journal of Forecasting*, **27**(4), 293–303.

Cinarra Systems (2024), 位置情報マーケティング, https://cinarra.co.jp/（最終アクセス日 2024 年 6 月 1 日）.

D'Angelo, N., Abbruzzo, A., Ferrante, M., Adelfio, G. and Chiodi, M. (2023). GPS data on tourists: A spatial analysis on road networks, *AStA Advances in Statistical Analysis*, **108**, 477–499, https://doi.org/10.1007/s10182-023-00484-w.

遠藤幹大, 高橋央亘, 浅田拓海, 有村幹治 (2019). Wi-Fi パケットセンシングによる道央広域エリアの時空間周遊パターン分析, 土木学会論文集 D3（土木計画学）, **75**(5), 475–483.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of International Conference on Machine Learning*, **96**, 148–156.

Gao Y. and Schmocker, J-D. (2022). Distinguishing different types of city tourists through clustering and recursive logit models applied to Wi-Fi data, *Asian Transport Studies*, **8**, https://doi.org/10.1016/j.eastsj.2021.100044.

GMO (2024). Town Wi-Fi, https://townwifi.jp/（最終アクセス日 2024 年 6 月 1 日）.

神谷達夫 (2018). 位置情報データを活用した観光地指標―海の京都観光圏 Wi-Fi パケットセンサーの情報量解析から―, 日本観光学会誌, **59**, 41–48.

観光庁 (2024). 観光地域づくり法人（DMO）とは, https://www.mlit.go.jp/kankocho/seisaku_seido/dmo/dmotoha.html（最終アクセス日 2024 年 6 月 1 日）.

Lee, D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems*, **13**, 535–541.

Liu, W., Wang, B., Yang Y., Mou, N., Zheng, Y., Zhang, L. and Yang, T. (2022). Cluster analysis of microscopic spatio-temporal patterns of tourists' movement behaviors in mountainous scenic areas using open GPS-trajectory data, *Tourism Management*, **93**, https://doi.org/10.1016/j.tourman.2022.104614.

村上大輔, 松井知子 (2022). 世代間・地域間の時系列相互相関に着目した COVID-19 の分析, 統計数理, **70**(1), 27–39.

中西航, 小林巴奈, 都留崇弘, 松本拓郎, 田中謙大, 菅芳樹, 神谷大介, 福田大輔 (2018). Wi-Fi パケットセンサーによる観光周遊パターンの把握可能性：沖縄・本部半島における検討, 土木学会論文集 D3（土木計画学）, **74**(5), 787–797.

Nakanishi, W., Yamaguchi, H. and Fukuda, D. (2018). Feature extraction of inter-region travel pattern using random matrix theory and mobile phone location data, *Transportation Research Procedia*, **34**, 115–122.

Nicholson, W., Matteson, D. and Bien, J. (2017). Bigvar: Tools for modeling sparse high-dimensional multivariate time series, ArXiv, https://doi.org/10.48550/arXiv.1702.07094.

NTT docomo (2024). モバイル空間統計, https://mobaku.jp/（最終アクセス日 2024 年 6 月 1 日）.

大井達夫 (2020). 小地域統計を利用したインバウンド観光の季節変動分析, 観光学, **22**, 13–24.

大井達夫 (2022). タイル指数による観光地の人流データの変動要因分析—新型コロナウイルス感染症の初期段階と現在を比較して—, 観光振興研究, **2**(4), 29–41.

Qian, C., Li, W., Duan, Z., Yang, D. and Ran, B. (2021). Using mobile phone data to determine spatial correlations between tourism facilities, *Journal of Transport Geography*, **92**, https://doi.org/10.1016/j.jtrangeo.2021.103018.

Reich, B. J., Yang, S., Guan, Y., Giffin, A. B., Miller, M. J. and Rappold, A. (2021). A review of spatial causal inference methods for environmental and epidemiological applications, *International Statistical Review*, **89**(3), 605–634.

SoftBank (2024). プライバシーセンター, https://www.softbank.jp/privacy/（最終アクセス日 2024 年 6 月 1 日）.

寺部慎太郎, 一井啓介, 柳沼秀樹, 小野瑞樹, 田中皓介, 康楠 (2019). Wi-Fi パケットセンサーを用いた歩行者行動・観光客周遊行動研究の包括的レビューとそれを踏まえた分析例示, 土木学会論文集 D3（土木計画学）, **75**(5), 669–679.

Textor, J., Van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M. and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package 'dagitty', *International Journal of Epidemiology*, **45**(6), 1887–1894.

Yamaguchi, H. and Nakayama, S. (2023). Pattern analysis of Japanese long-distance travel change under the COVID-19 pandemic, *Transportation Research Part A*: *Policy and Practice*, **176**, https://doi.org/10.1016/j.tra.2023.103805.

Evaluation of Interactions among Tourist Spots in Nagasaki Using Location Data

Yu Ichifuji[1] and Daisuke Murakami[2]

[1]Graduate School of Integrated Science and Technology, Nagasaki University
[2]The Institute of Statistical Mathematics

Tourism is a major industry in Japan, playing a crucial role in regional revitalization, especially in local cities. Despite being heavily impacted by the novel coronavirus, tourism is starting to recover. The number of tourists is returning to pre-pandemic levels or even surpassing them. To further attract tourists and manage overtourism, it will be necessary to implement strategies and measures. However, the increase in individual travel makes it challenging to analyze overall trends and movements between tourist destinations. To address this, the study aims to propose a method for evaluating relationships between tourist destinations using location registration data from mobile carriers. Through a quantitative assessment, the study seeks to create fundamental data to support tourist attraction efforts. Our proposed method evaluates the influence between tourist destinations using the number of unique users at each destination, the number of movements between destinations, and users' residential information. It is based on a time series model assuming that the influence of an increase in the number of visitors at each location diffuses proportionally to the number of movements estimated from the location registration data. When applied to tourist destinations in Nagasaki Prefecture, the evaluation of relationships between destinations revealed differences in behavior according to residence and age group. The findings suggest that modifying advertising methods based on residence and age group could efficiently attract tourists.

Key words: Tourism, location data, personal attribute information, cross-correlation, boosting.

統計数理(2025)
第 73 巻 第 1 号 117–139
©2025 統計数理研究所
［原著論文］

無回答誤差と調査票の返送時期の関係

―「高槻市と関西大学による高槻市民郵送調査」の調査不能と項目無回答―

松本 渉†

（受付 2023 年 10 月 17 日；改訂 2024 年 3 月 7 日；採択 11 月 19 日）

要　　旨

無回答には調査不能と項目無回答がある．無回答誤差の解消は重要であるが，無理に調査不能を減らそうとしても項目無回答を高め，全体としての誤差の減少にはつながらないという議論がある．本研究では，「高槻市と関西大学による高槻市民郵送調査」の 12 年分の結果を用いて，返送時期が調査不能と項目無回答とどのように関わっているかについて明らかにすることにした．

その結果，集計レベルの分析では，平均年齢に関する無回答誤差と回収率との間に負の相関関係がみられ，調査不能者(回答を得られなかった者)の割合が増加すると年齢に関する無回答誤差が増加するという影響が確認された．一方，個票レベルの分析では，項目無回答の場合は，回答済みの場合よりも平均返送日数が長くなる傾向がうかがえた．そこで，項目無回答と返送日数の関係性を調べると，項目無回答の個数と返送日数の間には双方が影響を及ぼしあう関係が成り立つ可能性がでてきた．その一方で，回収率が高い場合に項目無回答の個数が増加する傾向もうかがえた．郵送調査では，導入部分が良質であれば早く返送されて回収率も上昇するが，消極的回答者を含むために，項目無回答が増える可能性が考えられた．項目無回答の増加の極限には調査不能であるが，返送時期が長引くことの延長上に調査不能があるという関係と連動して成立する可能性が確認できた．

キーワード：郵送調査，無回答誤差，調査不能，項目無回答，返送日数，回収率．

1. はじめに

無回答誤差(nonresponse error)とは，回答を得られた調査のデータだけから算出した統計量と調査対象全体のデータから算出した統計量とのずれである（Groves et al., 2004a）．無回答(nonresponse)は，調査対象全体の測定ができずに全ての質問の回答が無効となる調査不能(unit nonresponse)と質問の一部で回答を得られなかった項目無回答(item nonresponse)に分かれる．慣習的・実務的には，項目無回答を単に無回答と呼ぶことが多いが，本研究における無回答誤差(nonresponse error)とは，調査不能と項目無回答の両方を原因として生じる誤差を意味する．どちらの意味の無回答も，現実には一定の方向性を持ちながら系統的に真の値とのずれを生じさせることが多く，無回答の偏り(nonresponse bias)となる．この偏りが深刻になった場合には調査結果が不適切なものになる．そのため，調査不能と項目無回答は，調査方法論上重要な課題として議論されてきた．

† 関西大学 総合情報学部：〒 569–1095 大阪府高槻市霊仙寺町 2–1–1

ただし，多くの調査研究において，調査不能と項目無回答は異なる問題として別々に議論されてきた．回答をするかどうかの決定過程を議論する場合であっても，項目無回答だけを扱うもの（例えば，Beatty and Herrmann, 2002; 阪口, 2023）がある一方で，調査不能（あるいは回収率）だけを扱うもの（例えば，Groves et al., 1992, 2000, 2004b; 林 他, 2003）があるという具合である．確かに調査主体の権威や信頼性，協力依頼状の表現の違いのように，調査不能には影響するが項目無回答には影響しそうにない事項などもあることを考えれば，項目無回答と調査不能のそれぞれに特化して研究することは，基本的には有益なアプローチであるといえる．

しかし一方で，調査不能と項目無回答との関連性に注目する議論もある．例えば，吉野（1994, 2001）は，情報回収量（information collection rate）という概念を用いて，調査不能が減少した場合に項目無回答が増加する問題を指摘している．ここでいう情報回収量とは，「わからない」「その他」以外の明確な選択肢を選んだ率（明確回答率，definite response rate）を各質問について計算してその平均値に回収率を乗じた値である（吉野, 1994）．吉野（1994）は，三つの異なる機関で同じ質問文を用いて実施された同時期の調査について情報回収率を計算したところ，それらがほぼ一定であることから，無理に回収率を上げる努力をしても明確な回答を得るのが難しいとしたのである．土屋（2005）も同様の立場を示し，「日本人の国民性第 11 次全国調査」への協力理由に関する事後調査における未返送者の特性から，一般的な調査不能者の特性を推察し，回収率の強引な引き上げが必ずしも非標本誤差の低減につながらない可能性を指摘している．Yan and Curtin（2010）も，米国の消費者調査の 20 年分のデータを用いて，調査不能率の上下変動と項目無回答率の上下変動の関連性を指摘し，項目無回答率の減少は調査不能率の増加の結果によって生じていることを指摘している．

つまり，一連の調査不能と項目無回答の関係に注目する研究は，調査不能を減らし，回収率の上昇によって無回答誤差を小さくしようとしても，項目無回答が増加するために，全体の無回答誤差が減少しないという問題点を示唆したのである．ここで調査不能が減少する際に項目無回答が増加すると考えられる理由としては，それほど調査に協力的でない人に無理に回答を求めても，結局明確な回答は得られないためであると推察されている（吉野, 1994, 2001）．このようなメカニズムが想定されているのは，主に面接調査の場合であり，吉野（1994, 2001）では面接調査の場合を前提に議論されている．この考え方を敷衍し，面接調査に限らず，郵送調査など他のデータ収集法についてもあてはまるような表現でその本質だけを述べるとすれば，「消極的な回答者がしぶしぶ回答をしたとしても，その回答内容は，積極的な回答者の回答ほどには有益な情報にはならない」ということになる．

回答者の積極性・消極性を推察する方法は，調査モードによって異なる．例えば，調査員が介在する面接調査や電話調査においては，調査員の感想を得ることによって判断するといった方法がありえる．他方，郵送調査の場合は，そのような方法はできないが，回答者の返送時期から推察することができる．なぜなら，郵送調査においては，調査票を発送するタイミングを統一することによって，調査対象者全員へのアプローチの開始をほぼ同時にすることができるという特徴があり（松本, 2023），そのために，返送の時期が早ければ早いほど，回答者として積極的である可能性が高いと予想されるからである．なお，このように予想することの前提として，調査においては，調査のタイトルや実施主体から調査の重要性を理解した時，さらには調査票の導入部分や第 1 問目を見たときの印象で，回答をするかどうかの意欲が異なる可能性があることも付言しておきたい．対象者は，自分にとってあまりに無縁な内容であればその調査に回答する気になれない一方で，自分にかかわりの深いもので負担感が低そうな調査であれば回答しようとすると考えられるからである．つまり，ここで積極性・消極性と呼んでいるものは，このように調査対象者が調査そのもの（調査内容，実施主体，調査票の出来栄え等）に刺激を受けて反応する程度を総合的に要約した概念である．実際には，調査票を受け取った際に

多忙かどうかといった個々の調査対象者側の事情によって回答する時期は前後するし，投函に至るタイミングも一定ではないので，返送時期から積極性・消極性を推し量るには，ある程度の誤差をおりこむ必要はある．

しかしいずれにせよそういった事情から返送時期の違いに目をむけた郵送調査研究は珍しいものではない．特に回答者の特性との関係を扱った返送時期の研究がいくつも見られる．具体的には，返送時期と調査回答者の基本的属性との関係に注目した研究（Siemiatycki and Campbell, 1984; 小島, 1993; 林・村田, 1996; 前田, 2005; 松田, 2017），返送時期と調査対象者の調査テーマへの関心度合いに目を向けた研究（Donald, 1960; 渡邊, 2007），返送時期と基本属性・内容面への関心の強さの統合的な関係に目をむけた研究（松本, 2023）等がある．このうち，調査回答者の基本的な属性について注目した研究には，返送時期と回答者基本属性との関連に否定的なもの（Siemiatycki and Campbell, 1984; 小島, 1993; 林・村田, 1996）がある一方で，基本属性における関連性を指摘するものもあり（前田, 2005; 松田, 2017），郵送調査の返送の時期によって，回収層の基本属性の異同がどのように関わっているのかについては議論が続いている．しかし，調査対象者の調査テーマへの関心度合いとの関連性については，内容面への関心の強さが返信速度を早めるという見解で一致している（Donald, 1960; 渡邊, 2007; 松本, 2023）．返送の時期が早ければ早いほど，回答者として調査の内容に関心が高いという関係が安定的に成立していることを踏まえれば，返送時期に回答者の積極性・消極性の違いが反映されているという予想も，それほど不自然な帰結ではないはずである．

さらに，回答者の積極性・消極性の違いが返送時期に表れているという考えを拡張すると，遅い時期の返送者（消極的回答者）の延長上に返送に至らなかった調査不能者がいると位置づけることができる．Zimmer（1956）は，調査対象者を早期回答者・督促後の回答者・調査不能者の 3 グループに分類し，それぞれを返送確率の高・中・低の三段階と位置づけて三者の違いを分析しているが[1]，このような位置づけが見られる研究の一つである．田中（1962）においても同様の考えが見られ，調査不能によって生じた回答の偏りを補正する方法を考案するに際して[2]，対象者を返送確率の連続体の上に位置づけるとしている．

Zimmer（1956）や田中（1962）に見られるような調査不能者を後期返送者の延長上に位置づける考え方は，郵送調査に限定された表現に聞こえるが，調査研究一般の議論においては，回答が得られなかった調査不能者と回答を得るまでに時間がかかるなどの困難がある回答者との間には抵抗についての連続性（continuum of resistance）があるという表現が用いられている（Filion, 1976; Fitzgerald and Fuller, 1982）．Lin and Schaeffer（1995）は，このような考えに立脚するモデルを抵抗についての連続性モデル（continuum of resistance model）とよび，調査不能層と回収層を区別する分類モデル（classes model）と明確に区別している．このような連続性の仮定は，面接調査・留置調査・電話調査といった調査モードの場合，その妥当性に疑問が生じるかもしれない．調査を実施する側がどこかの時点で調査対象者への再接触を中止し，たまたま会えなかった回答意欲の高い調査対象者を調査不能者としている可能性があるため，調査不能者の中に遅い時期の回答者よりも回答意欲の高い人がいないとは限らないからである．しかし，調査実施側の締め切りとは無関係に調査対象者の側が返送できる郵送調査の場合は，締め切り後の回答を実際に有効とするかどうかは別として，ある種の飽和状態に到達するまで回答を得ることができる．このような状態が生じるのは，受け取って（あるいは締め切り日）から相当日数が経過したことにより，仮に当初非協力的であった調査対象者において偶発的に回答しようという意欲が生じたとしても調査に応じることのハードルが高くなっているために調査対象者は返送しなくなるからである．もちろん住所不明で戻ってきたケースはそのようなモデルの例外となるが，多くの調査不能者がそのように構成されると考えないと，調査によって回収率が上下することの説明がつかない．そのため，調査不能者と遅い時期の回答者との間には断

絶が少ないと予想できる（松本, 2023）．

以上の議論から，本研究では，郵送調査における調査不能と項目無回答といった二種類の無回答が，返送日数と何らかの関連があると予想する．回答者の積極性・消極性が返送日数に反映されるとすれば，調査不能者は返送日数の短い積極的回答者の対極，かつ返送日数の長い消極的回答者の延長上に位置づけられるし[3)]，回答者の積極性・消極性が項目無回答の発生頻度に現われるのであれば，返送日数と項目無回答にも関連が生じるはずだからである．

なお，郵送調査における返送時期と調査不能との関係性については，Zimmer（1956），田中（1962），土屋（2005）が取り扱っているが，Zimmer（1956）の郵送調査は米国空軍関係者に対して実施されたもので標本が特殊で 220 とサイズも小さい．田中（1962）や土屋（2005）の分析は，一般的な市民を対象とする調査であるが，どちらも先行する面接調査の事後に実施された郵送調査を利用したものである．事後的な郵送調査の未返送者は，面接調査段階での項目無回答が多いことなど一定の知見が明らかになっている面もあるが（土屋, 2005），逆に言えば一度面接調査に応じた人々に対する郵送調査という特殊性も否めない．郵送調査における返送時期と調査不能・項目無回答の関係性についての研究に蓄積はまだ十分とは言えない．

そこで，本研究では，高槻市民に対して実施した郵送調査の返送日数を用いて，郵送調査の返送時期が調査不能と項目無回答という無回答誤差要因にどのように関わっていくのかを明らかにする．地域限定の調査結果について分析することになるが，一般市民を対象とする多くの郵送調査にとって有益な知見を見出だそうとするものである．

2. データの概要と分析の手順

本研究では，高槻市民に対する意識調査「高槻市と関西大学による高槻市民郵送調査」（2011〜2022 年度）のデータを活用する．「高槻市と関西大学による高槻市民郵送調査」は，高槻市と関西大学の連携によって 2011 年度から実施されている郵送調査であり，高槻市が施策に利用するための市民意識調査であると同時に，関西大学における教育・研究での活用を意図した調査でもある．高槻市側の質問項目は，年度ごとに各部署の希望に応じて提案され，関西大学側の質問項目は，社会調査実習の授業の中で学生から提案されたものから構成されるが，属性項目さらには調整に必要と考えられた項目も追加される．そのため調査票の質問数は，年度によって変動するが，調査票全体は 8 ページ以内に統一して実施される．

またこの調査は，層化無作為抽出により選出された高槻市民が調査対象者となっている標本調査である．標本抽出に際しては，2011〜2016 年度の 6 年間は，20 歳以上 85 歳未満の男女を 20 代，30 代，40 代，50 代，60 代，70 代以上の 6 つに分類したため，12 層からなる男女別の年齢区分で層化している．2017 年度以降は，これに 18・19 歳の男女の 2 層が加わり，合計 14 層が層化において利用されている．

調査票には，対象者名簿との照合が可能になるような識別子がつけられていない．そのため，性・年齢のような基本情報についても対象者名簿との照合ができず，これらについての項目無回答に由来するバイアスは本データでは検討できないことが，分析の前提となる．その一方，返信される封筒に押されている消印の日付についても調査開始以来毎年継続して記録している．そこで本研究では，消印の日付を個々の調査が完了した返送日とみなし，送付の日からのその日付までに経過した日数を返送日数として計算している．なお，この調査においては，同封のボールペンをそのまま使用して良いとしているものの，事前・事後ともに換金性の高い謝礼は行っていない（松本・李, 2015）．したがって，謝礼による返送日数の影響については考慮しなくても良い．

調査の実施概要と返送日数の概要を示すと，表 1 の通りである．

表 1．調査の実施概要と返送日数の概要．

調査年度	対象年齢	回収標本	回収率	返送日数の中央値	返送日数の平均	返送日数の標準偏差
2011	20-85 歳	1225	61.3%	5	6.8	5.0
2012	20-85 歳	1230	61.5%	5	6.8	5.5
2013	20-85 歳	1233	61.7%	5	7.3	5.3
2014	20-85 歳	1198	59.9%	6	8.2	6.6
2015	20-85 歳	1224	61.2%	6	7.4	5.7
2016	20-85 歳	1232	61.6%	5	6.7	5.1
2017	18-85 歳	1196	59.8%	5	6.9	5.2
2018	18-85 歳	1233	61.7%	5	6.4	4.9
2019	18-85 歳	1172	58.6%	5	6.8	6.1
2020	18-85 歳	1227	61.4%	5	6.7	5.2
2021	18-85 歳	1211	60.6%	5	6.5	5.3
2022	18-85 歳	1214	60.7%	7	8.7	4.9
全体		14595	60.7%	5	7.1	5.5

回収率が概ね 6 割前後で推移しているため[4)]，回収標本サイズも概ね 1200 前後で安定している．返送日数の平均も 7 日程度，中央値は平均より短めの値で安定しているが，2014 年度と 2022 年度がやや日数が長くなっている．2014 年度は手違いで予告はがきが早めに発送されたことで締め切りまでの期間が長くなり，返送が間延びした，2022 年度は郵便サービスの変更(土曜日配達中止および配達日繰り下げ)に対応していなかったといった事情による（松本, 2023）．なお 6 割前後の回収率を実現しているものの督促を行っていない（松本・李, 2015）．そのため督促の前後によって返送の性質が変わるということは生じていないため，返送過程の連続性を仮定しやすい．そこで，本研究においては，返送時期を期間で区切るのではなく返送日数で理解する．

本研究では，この郵送調査の返送時期が，調査不能と項目無回答という無回答誤差要因にどのように関わるのかを明らかにするため，返送日数・調査不能・項目無回答の関係を調べる．まず 3 章で，全体像を把握するため調査年度別の回収標本と枠母集団の状況の比較から無回答誤差の状況を確認し，回収率や返送日数との集計レベルでの関連性を把握する．次に個票レベルでの関連性を分析する．4 章では，無回答誤差を生じる要因のうち項目無回答が返送日数とどのように関わるか属性別の集計結果の比較に基づいて考察する．5 章では，個票レベルの詳細な分析として，項目無回答と返送日数の関係性のうち項目無回答が返送日数に影響を与えると想定した場合と返送日数が項目無回答の発生に影響を与えると考えた場合と両方の場合を検討する．特に後者では，回収率と項目無回答の発生との関わりの有無も検証する．最後に，返送時期と，調査不能・項目無回答ひいては無回答誤差との関係を総括する．

3. 回収標本の分布と枠母集団の分布の比較

無回答誤差を確認するにあたって，刊行済みの 12 点の社会調査実習報告書（関西大学総合情報学部, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023)に基づき確認できる属性について 12 年分の枠母集団と標本の人口分布について整理する．ただし，本調査では，無記名で回収する郵送調査においては，性別や年齢も質問によってしか確認できない．性別や年齢といった質問も項目無回答があるため，両者の乖離は，調査不能と項目無回答の両方に由来することに注意がいる．

表 2. 調査年度別の人口性比のずれ.

調査年度	枠母集団	計画標本	回収標本	標本誤差	乖離	無回答誤差
2011	92.8	93.1	74.4	0.2	-18.5	-18.7
2012	92.6	92.7	70.8	0.1	-21.8	-21.9
2013	92.3	92.3	78.4	0.0	-13.9	-13.9
2014	92.0	91.9	75.3	0.0	-16.7	-16.6
2015	91.6	91.8	75.9	0.1	-15.7	-15.8
2016	91.5	91.6	81.5	0.1	-10.0	-10.1
2017	91.8	91.8	71.3	-0.1	-20.6	-20.5
2018	91.9	91.9	73.7	0.0	-18.3	-18.3
2019	91.9	91.8	72.3	-0.1	-19.5	-19.4
2020	91.9	91.9	76.1	0.0	-15.8	-15.8
2021	91.8	91.8	78.2	-0.1	-13.7	-13.6
2022	91.6	91.8	69.3	0.1	-22.4	-22.5
	(1)	(2)	(3)	(2) - (1)	(3) - (1)	(2) - (3)

3.1 男女別の人口分布

男女別の人口分布に注目し，枠母集団(当該年度 6 月末の高槻市の住民基本台帳人口)と計画標本，回収標本(最終的に回収できた調査票全体)における人口性比の推移を示した(表 2)．なお人口性比とは，統計局の調査などでよく利用されるもので，女性 100 人に対する男性の人数を表した値である．回収標本における人口性比が枠母集団の人口性比を 12 年間継続して下回っている．女性よりも男性の回収率が低いまま推移したことが反映されている．

計画標本における人口性比から標本誤差(計画標本と枠母集団のずれ)を計算し，枠母集団と回収標本との乖離とともに表示すると，枠母集団と回収標本との乖離に寄与する割合はわずかであることが確認できる(時には正負逆向きのずれもある)．「高槻市と関西大学による高槻市民郵送調査」の計画標本の決定においては，枠母集団における性・年齢別の人口分布を用いて層化抽出をしているので，標本誤差は実質的に丸め誤差であるため当然の帰結である．なお，回収標本と計画標本のそれぞれから算出される統計量のずれは，枠母集団と回収標本との乖離から標本誤差を除いたものに一致する．これは調査不能と項目無回答の両方に由来する無回答誤差であるが，性別においては，後述するように項目無回答が少ないため，実質的には調査不能の影響が大きいと予想される．そしてその値は，絶対値でみると，最小の時に 10.1(2016 年)，最大の時に 22.5(2022 年)となり，年度によって変動がある．最小の場合も無視できないずれが生じていると言えよう．

3.2 年齢別の人口分布

年齢別の人口分布に着目し，同様の乖離の傾向を確認するために，2 章で述べたようなカテゴリーの形で取得されている年齢についてのデータを，各年齢カテゴリーの階級値に変換して，調査年度別の年齢の平均値を算出し，平均年齢とすることとした．表 2 と同様に，枠母集団(当該年度 6 月末の高槻市の住民基本台帳人口)，計画標本，回収標本について平均年齢を算出して，それぞれの差も整理したものが表 3 である．

2 章ですでに述べたように 2011〜2016 年度の枠母集団は 20 歳以上 85 歳未満であるが，2017 年度以降 18 歳以上 85 歳未満になっている．枠母集団の平均年齢が基本的には上昇傾向にあり，高槻市における高齢化を反映しているにもかかわらず，2016 年度から 2017 年度にかけて

表 3. 調査年度別の平均年齢のずれ.

調査年度	枠母集団	計画標本	回収標本	標本誤差	乖離	無回答誤差
2011	50.9	50.9	54.6	0.0	3.6	3.6
2012	51.3	51.3	54.9	0.0	3.6	3.6
2013	51.6	51.6	55.5	0.0	3.9	3.8
2014	51.9	51.9	55.9	0.0	4.0	4.0
2015	52.2	51.9	55.8	-0.3	3.6	3.9
2016	52.3	51.9	55.7	-0.4	3.4	3.9
2017	51.7	51.7	55.8	0.0	4.1	4.1
2018	51.9	51.9	55.3	0.0	3.4	3.4
2019	52.1	52.1	56.8	0.0	4.6	4.7
2020	52.2	52.2	55.6	0.0	3.4	3.4
2021	52.3	52.3	56.4	0.0	4.1	4.1
2022	52.4	52.4	56.4	0.0	4.0	4.0
		(1)	(2)	(3)	(2) - (1)	(3) - (1)

平均年齢が下がったのは 18–19 歳の若年層が追加されたためである．なお回収標本においては，2016 年度から 2017 年度にかけて平均年齢が上昇しているのは，18–19 歳の若年層の回収率が 31.3% と低いため，全体的な高齢化傾向を打ち消すほど若年層の回収標本が追加されなかったためと考えられる．

枠母集団と各標本に関しては，人口性比の場合と同様の理由から，標本誤差はほとんどないため，枠母集団と回収標本との間の乖離と無回答誤差は，ほぼ一致している．なお回収標本と計画標本の平均年齢は年度によって変動があるが，回収標本の方が計画標本よりも毎年 4 歳程度高齢となっている．高齢層の方が若年層よりも一般的に回収率が高いことが影響している．

3.3 その他の属性の人口分布

回収標本と母集団を厳密に比較できる変数は限られている．婚姻状態の分布については，国勢調査で集計されているので，2015 年度と 2020 年度に実施された国勢調査の結果から高槻市の調査対象者年齢に限定して集計することができた．これと「高槻市と関西大学による高槻市民郵送調査」の回収標本との分布を比較することとし，両者を併記して，表 4 を作成した．

まずこの解釈に先立ち，2020 年度の調査では，12 年間の平均が 4% となる無回答率が高く（阪口, 2023），かつ既婚者（配偶者あり）の回答割合 59% が，例年（平均 68%）よりも低いことに注目したい．当該年度の調査票を確認すれば，この年度に限り，婚姻の質問が最終頁の前に配置され，かつ直後の分岐により既婚者のみ 4 問追加で回答する必要があることが一見して分かる状態であり，2020 年度に無回答率が突出して上昇した分は，追加質問を回避した既婚者によるものと考えるのが自然である．そう考えると，①既婚（離婚・死別）の割合は，国勢調査より 2% ほど高い，②未婚の割合は，国勢調査より 8〜9% ほど低い，といった 2 回分の比較で共通する傾向に加え，③2015 年度の比較に見るように既婚（配偶者あり）の場合は，国勢調査よりも高い割合になることの 3 点が意味のある比較として理解できる．残念ながら調査不能と項目無回答の影響を分離できないため，厳密なことは言えないが，この既婚者割合が高めに出て未婚者が低めにでる傾向は，郵送調査によっても未婚者を補足しづらい実情を踏まえれば，調査不能を主な要因として生じていると考えられる．

表 4. 婚姻状況の比較(2015, 2020 年度). (出典)国勢調査における婚姻状況については,「平成 27 年国勢調査結果」(総務省統計局)(https://www.stat.go.jp/data/kokusei/2015/kekka.html)ならびに「令和 2 年国勢調査結果」(総務省統計局)(https://www.stat.go.jp/data/kokusei/2020/kekka.html)から高槻市の該当年齢(20 歳以上または 18 歳以上 85 歳未満)の婚姻状況のデータを集計して作成.

婚姻状況 (人, %)	既婚 (配偶者あり)	既婚 (離婚・死別)	未婚	無回答	合計
国勢調査	175,978	30,669	61,204	7,557	275,408
2015 年度	64%	11%	22%	3%	100%
郵送調査	868	157	172	27	1224
2015 年度	71%	13%	14%	2%	100%
国勢調査	169,291	28,928	68,135	11,387	277,741
2020 年度	61%	10%	25%	4%	100%
郵送調査	729	144	196	158	1,227
2020 年度	59%	12%	16%	13%	100%

表 5. 返送日数・回収率・無回答誤差の推移.

		返送日数			無回答誤差	
調査年度	度数	平均値	中央値	回収率	人口性比	平均年齢
2011	1225	6.8	5	61.3%	-18.7	3.6
2012	1230	6.8	5	61.5%	-21.9	3.6
2013	1233	7.3	5	61.7%	-13.9	3.8
2014	1198	8.2	6	59.9%	-16.6	4.0
2015	1218	7.4	6	61.2%	-15.8	3.9
2016	1222	6.7	5	61.6%	-10.1	3.9
2017	1194	6.9	5	59.8%	-20.5	4.1
2018	1233	6.4	5	61.7%	-18.3	3.4
2019	1172	6.8	5	58.6%	-19.4	4.7
2020	1227	6.7	5	61.4%	-15.8	3.4
2021	1211	6.5	5	60.6%	-13.6	4.1
2022	1214	8.7	7	60.7%	-22.5	4.0

3.4 集計レベルの関連性

無回答誤差, 返送日数[5]及び回収率の関連を検討するため, これらの 12 年間の推移を整理し(表 5), これらについての相関係数を確認した(表 6).

その結果, 回収率と無回答誤差(平均年齢)との間に強い負の相関関係が見られた[6]. 回収率が減少した年度において, 平均年齢についての無回答誤差が増加する傾向にあるためと予想される. 回収率が低い場合, 年齢層によって回収率の下がり方が異なるため, 枠母集団と回収標本との間で年齢層別の人口分布の違いが大きくなることを示唆している.

そもそも回収率は, 100% から調査不能率を除した値であり, 調査不能と表裏一体の関係にある. 無回答誤差(平均年齢)と回収率の相関が高いことは, 年齢についての無回答誤差は, 無回答よりも調査不能に起因する割合が大きいと考えられる. 一方で, 人口性比についての無回答誤差は平均年齢についての無回答誤差ほどには回収率の影響が見られず, 男性の回収率低下による男性の人口割合の低下よりも, 若年層の人口割合の低下の方が回収率低下(調査不能の

表 6．返送日数・回収率・無回答誤差の相関行列．

		返送日数			無回答誤差	
		平均値	中央値	回収率	人口性比	平均年齢
返送日数	平均値	1.00				
	中央値	**0.92**	1.00			
回収率		-0.20	-0.12	1.00		
無回答誤差	人口性比	-0.31	-0.32	0.32	1.00	
	平均年齢	0.22	0.16	**-0.86**	-0.06	1.00

太字：$p<0.01$

増大）に関わっていることを示唆している．

また，返送日数については，回収率や無回答誤差と明確な関連性は見いだせなかった．年度別平均返送日数については，送付手順の違いや郵便事情の変化などにより長引いた年度もある（松本, 2023）．そういった要因は，年度別の平均日数に影響しても，必ずしも調査不能や項目無回答に影響するものではない．返送日数と諸要因の関連性については，年度ごとの集計レベルでの分析では限界がある．次章以降では，個票レベルでの分析を行う．

4. 返送日数・項目無回答・回答者の状況の関係

本研究では，無回答誤差と返送時期の関係に関心がある．この章では，無回答誤差を生じる要因のうち，項目無回答が返送日数とどのように関わるかについて属性別集計から確認する．なお郵送調査における項目無回答とは，選択式回答においては事前に用意されたどの選択肢も選ばれなかったケース，または自由記述式回答の場合は，何も記入されなかったケースを意味する．本研究では，この項目無回答が回収標本に占める割合を単に無回答率と呼んでいる．

「高槻市と関西大学による高槻市民郵送調査」（2011–2022 年度）において継続的に扱われている質問の無回答率の単純集計の変遷についての検討は，阪口（2023）でなされており，特に調査票構成の変更を原因とする無回答率の変動が確認されている．本研究では，同じ記述を繰り返すのではなく，返送時期を扱うことを考慮し，この章以降の分析では，性別・年齢といった基本的属性に加え，職業・市内居住年数の項目に注目することとした．職業に注目するのは，日常の多忙さが異なる就業状態が返送時期と関わりがあると予想されると考えられたため統制変数として重要であると考えられたからである．これに加え，市内居住年数という質問に注目するのは，単に属性変数として統制するためというよりも，高槻市という限定された地域での調査において市内居住年数は間接的に調査への理解や関心を高める強い要因と考えられるので，返送速度に関わりがあると考えられたからである（松本, 2023）．つまり，居住年数は，地域関心度の代理変数であり，調査の内容に対する関心を表すものとして扱っている．

そこでまず回収率に占める性別・年齢・職業・市内居住年数の無回答率を調査年度別に整理した（表 7）．なお回収標本サイズも確認しやすいように一番右の列に追記した．

表 7 をみると，大部分が 1〜3% 程度を安定的に推移しており，それほど大きい割合ではない．なお 2018 年度と 2020 年度の年齢の質問で無回答率が不自然に高いのは，どちらも当該質問で分岐がなされており，2020 年度には直前の質問でも分岐があるという事情による（阪口, 2023）．

次に性別・年齢・職業・居住年数の順に平均返送日数との関係について検討する．

性別・年齢・職業・市内居住年数に関する質問についても回答・無回答別に平均返送日数を示すと，順に図 1・図 2・図 3・図 4 のようになる．どの質問とも長期的には回答群よりも（項目）無回答群の方が平均返送日数が長いように見えるが，性別や年齢の質問に関しては，回答

表 7. 回収標本に占める無回答率.

無回答				市内	
%	性別	年齢	職業	居住年数	回収標本
2011	0.7%	0.9%	2.0%	1.2%	1225
2012	0.7%	0.7%	0.7%	0.5%	1230
2013	1.6%	2.1%	2.7%	1.5%	1233
2014	1.1%	1.6%	2.1%	1.3%	1198
2015	1.4%	1.7%	2.9%	2.0%	1224
2016	1.1%	1.6%	2.0%	1.9%	1232
2017	1.3%	1.3%	2.8%	1.1%	1196
2018	1.0%	6.2%	1.5%	1.1%	1233
2019	1.6%	2.1%	3.1%	2.2%	1172
2020	1.4%	10.4%	1.8%	0.7%	1227
2021	1.0%	1.3%	2.6%	1.2%	1211
2022	2.0%	1.6%	2.6%	1.2%	1214
全体	1.2%	2.7%	2.2%	1.3%	14595

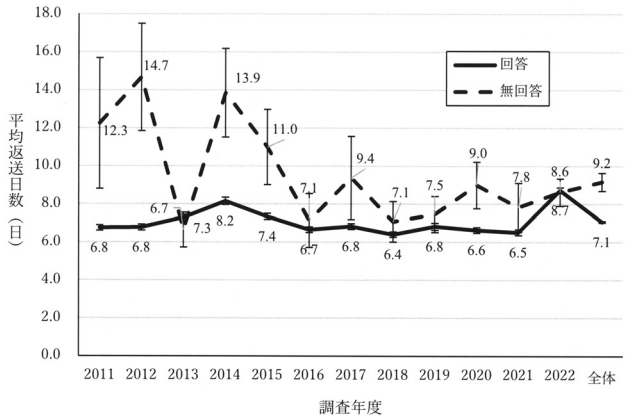


図 1. 性別の回答・無回答による平均日数. 注：エラーバーは，標準誤差による.

群の方が無回答群よりも長い年度がある．性別については 2013 年と 2022 年に，年齢については 2018 年と 2020 年に逆転している（ただし誤差を考慮すると，どれも再逆転がありえる）．どちらも無回答割合自体が例年より高かった年度である．性別に関しての原因ははっきりしないが，年齢に関しては当該年度だけ分岐があるという調査票上のデザインの違いがあった（阪口, 2023）．そのため，たまたま返送日数が短めの人も無回答に転じてしまったと考えれば，もともとの無回答群は回答群よりも平均返送日数は長かった可能性がある．仮に，平均返送日数を従属変数とし，質問項目の回答・無回答と調査年度を固定因子とする二元配置の分散分析を，性別・年齢・職業・市内居住年数のそれぞれについて行った場合，性別（$p < .001$），年齢（$p < .001$），職業（$p < .01$），市内居住年数（$p < .001$）ともに有意となり，固定因子としての効果が認められる[7)]．

このことから返送日数が遅い人ほど項目無回答が生じやすいとか，項目無回答がある人ほど返送するのが遅いといった関係性が疑われるのであるが，その因果関係は少し難しい．例え

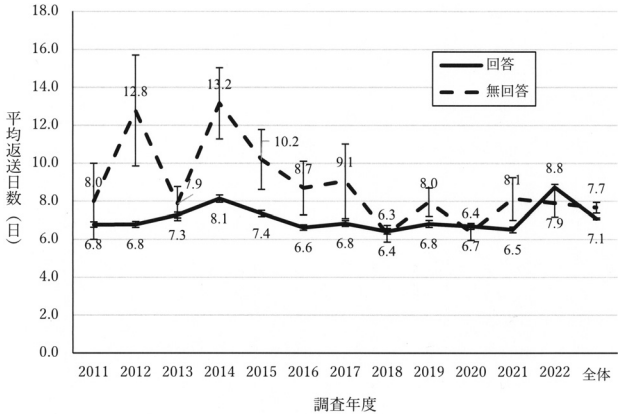


図 2．年齢の回答・無回答による平均日数．注：エラーバーは，標準誤差による．

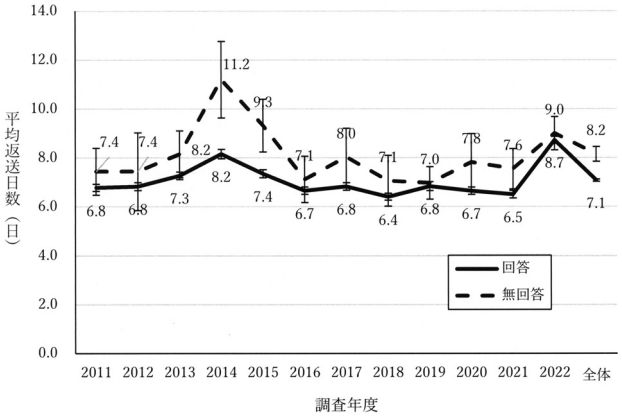


図 3．職業の回答・無回答別の平均日数．注：エラーバーは，標準誤差による．

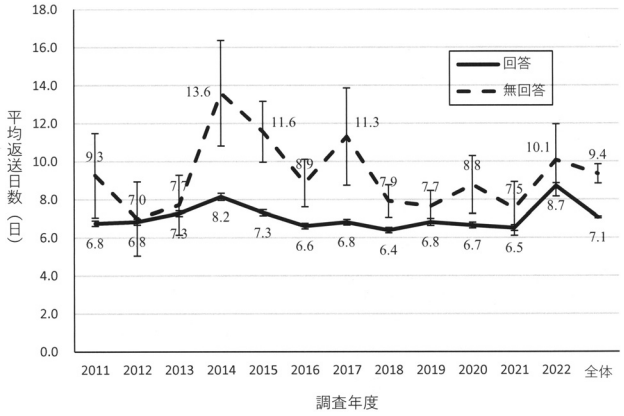


図 4．市内居住年数の回答・無回答別の平均日数．注：エラーバーは，標準誤差による．

表 8. 職業別の返送日数の統計量(12 年間全体).

職業	度数	中央値	平均値	標準偏差
学生	329	6	8.6	7.1
無回答	323	6	8.2	5.4
常時雇用の勤め人	4148	6	7.8	6.1
その他	271	6	7.8	6.2
臨時雇用、パート、アルバイト	2417	6	7.3	5.6
自営業主	569	5	6.9	5.7
経営者、役員	307	5	6.9	5.8
家事専業	2318	5	6.8	4.7
自営業の家族従業者	270	5	6.6	5.0
無職	3625	5	6.2	4.5
合計	14577	5	7.1	5.5

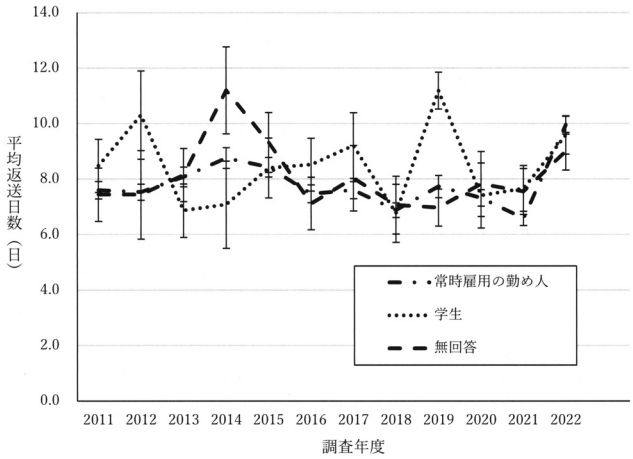


図 5. 職業別(常時雇用の勤め人・学生・無回答)の平均日数. 注：エラーバーは，標準誤差による.

ば，項目無回答は調査票記入時点に発生し，その後調査票は返送されている．そのため，調査票記入後に日数が経過するのであれば，項目無回答は(回答に悩んだ結果，日数が経過するなどで)返送日数の原因になると考えられる．しかし，もし調査票記入時点ですでに日数が経過していた場合は，長くかかった日数の方が項目無回答の原因となりえる．例えば，早く出そうと焦って無回答が生じるような場合などである．そうすると，項目無回答と返送日数は，相互に影響を及ぼしあう可能性も考えられる．

また仮に，平均返送日数を従属変数と考えられると仮定した場合も，女性単独であれば無回答群よりも平均返送日数が長い年度が存在するというように，回答別に比較すると，無回答群より長い回答群が存在するという場合がある（松本, 2023)．無回答群を含めた職業別の返送日数(12 年間全体)を要約し，平均返送日数の長い順に並べると(表 8)，学生は無回答群よりも平均返送日数が上回っていることもわかる．学生や常時雇用の勤め人といった不在がちな回答群と無回答群に絞って 12 年間の平均返送日数の推移を示すと図 5 のようになる．学生・常時雇用の勤め人・無回答の平均返送日数についての上下関係は，調査年度によって変動していることが確認できる．

表 9．返送日数の対数変換．

2011-2022 年度	度数	最小値	中央値	最大値	平均値	標準偏差
返送日数	14577	1.0	5.0	82.0	7.1	5.5
返送日数の対数	14577	0.0	1.6	4.4	1.7	0.7

5. 返送日数と項目無回答の詳細分析

返送日数と項目無回答の間の因果関係については，第 4 章で論じたように，両者がともに原因と結果になる可能性がある．そのため，返送日数を従属変数とする場合と，項目無回答の発生の状況を従属変数とする両方の回帰モデルを検討し，その総合的な結果から考察することとする．

返送日数のデータは，右に裾が長い分布をしていることから(松本, 2023)，回帰モデルを適用するにあたって，モデルのあてはまりを考慮し，本章以降では，返送日数の自然対数に変換して分析を行う．なお，対数変換前後の統計量を示すと，表 9 のようになる．

5.1 返送日数の対数を従属変数とする場合

返送日数の対数を従属変数とする分析にあたっては，性別，年齢，職業，市内居住年数の 4 つの質問項目を主要な独立変数とする回帰モデルを出発点として考える．1 章で議論したように多くの既存研究において，どのような要因が返送日数に影響を与えているかは関心のある事項であり，特に本研究で扱っている 4 項目は返送速度との関わりがある程度想定されるからである．ただし，4 章で述べたように，性別，年齢，職業の 3 項目は単なる属性としての統制変数であるが，市内居住年数は地域関心度の代理変数であり，調査の内容に対する関心を反映する項目として想定されている．

ただし，本研究では，項目無回答の影響の有無を確認することを重視している．そのため，項目無回答の影響を見るために，このモデルに回答・無回答についての変数を追加する必要がある．その際，各項目についての独立変数が無回答を欠測データとして除外すると計算ができないため，4 項目の各変数については，以下のようなダミー変数を用意する．

(a) 性別：男性を基準カテゴリーとして，女性 = 1，女性以外=0 とする女性ダミー変数を用いる．無回答については，無回答=1, 回答(男性・女性)=0 とするダミー変数となる．
(b) 年齢：10 代，20 代，30 代，40 代，50 代，60 代，無回答それぞれについてのダミー変数を作成する．どれも該当するカテゴリー=1，それ以外=0 となる．70 代以上(70 歳以上 85 歳未満)が基準のカテゴリーとなっている．
(c) 職業：常勤者ダミー(常勤者 = 1, 常勤者以外 = 0)，学生ダミー(学生 = 1, 学生以外 = 0)，無回答ダミー(無回答=1，回答=0)を作成する．常勤者とは，表 8 に示す「常時雇用の勤め人」(カテゴリー)のことである．
(d) 市内居住年数：1 年未満，1 年以上 3 年未満，3 年以上 5 年未満，5 年以上 10 年未満，10 年以上 20 年未満，20 年以上 30 年未満，30 年以上 40 年未満，40 年以上 50 年未満，無回答それぞれのダミー変数を作成する．どれも該当するカテゴリー=1，それ以外=0 となる．50 年以上の市内居住年数が基準のカテゴリーとなっている．

ここで，それぞれの項目無回答について識別するダミー変数(無回答 = 1，回答=0)を用意したが，返送日数に対する項目無回答の影響を分析する上では，それぞれの事情によって生じる個別の無回答のダミー変数よりも，ある程度の項目の無回答の個数(ダミー変数の合計値)を用いる方が好ましいと考えられる．そこで，性別，年齢，職業，市内居住年数の 4 項目に，最終

表 10. 無回答の個数の年度別基本情報.

調査年度	平均	標準偏差	0 個	(%)	1 個	(%)	2-10 個	(%)	11 個	(%)	回収標本
2011	0.55	1.05	781	(64)	331	(27)	112	(9)	1	(0)	1225
2012	0.36	0.98	952	(77)	211	(17)	65	(5)	2	(0)	1230
2013	0.56	1.38	886	(72)	217	(18)	126	(10)	4	(0)	1233
2014	0.36	1.07	947	(79)	188	(16)	61	(5)	2	(0)	1198
2015	0.43	1.31	961	(79)	178	(15)	83	(7)	2	(0)	1224
2016	0.42	1.25	959	(78)	200	(16)	73	(6)	0	(0)	1232
2017	0.38	1.14	946	(79)	181	(15)	65	(5)	4	(0)	1196
2018	0.37	1.07	954	(77)	212	(17)	64	(5)	3	(0)	1233
2019	0.49	1.48	925	(79)	154	(13)	83	(7)	10	(1)	1172
2020	0.69	1.04	685	(56)	353	(29)	189	(15)	0	(0)	1227
2021	0.36	1.14	980	(81)	163	(13)	65	(5)	3	(0)	1211
2022	0.43	1.28	945	(78)	187	(15)	76	(6)	6	(0)	1214
全体	0.45	1.19	10921	(75)	2575	(18)	1062	(7)	37	(0)	14595

学歴，居住地域，住居，居住形態，婚姻状態，世帯人数，世帯収入の 7 項目を加えた計 11 項目に関する無回答の個数を合計した変数を用意した．この 11 項目を用いたのは，これらが 12 年間共通して使用されているためである．なお，子供の有無については，12 年間を通じて集計することが，質問の内容(「子供の有無」「子供の人数」)だけでなく，既婚者に尋ねている場合と全員に尋ねている場合があって前提とする条件も年度によって異なっているため，利用していない．

この 11 項目の無回答の個数の統計量等を年度別に整理したのが表 10 である．年齢や婚姻状態の質問の位置や形式が例年と異なっている 2020 年度の平均個数 0.69 が最も高い(ただし 11 個全てのケースは 1 件もない)．この年度を例外とすれば，どの年度も 0 個の割合が 6 割以上となっており，1 件の発生確率が低いカウントデータの分布であることが確認できる．

性別，年齢，職業，居住年数の 4 項目に関して作成した無回答以外の全てのダミー変数と無回答項目数を用いて，2011–2022 年までの調査年度ごとおよび全体データに重回帰分析を実施すると，表 11 の結果が得られた．なお，VIF の値から，両方の分析のどの独立変数間でも多重共線性が起きていないことは確認ずみである．

12 年全体の分析では，多くの項目で符号が正かつ有意となっていて，男性よりも女性の方が，70 代以上よりも若い年齢層の方が，職業の中では常勤者が，(地域に関心が強いと想定される)居住年数 50 年以上の人々よりも 40 年未満の人々の方が大体において返送日数が長くなるという傾向を示すと同時に，無回答項目数が多いほど返送日数も長くなる傾向を示唆する結果となっている．

調査年度別にみても，一部の年度で例外はあるが，多くの年度で無回答項目数について有意となる結果が見られ，返送日数(の対数へ)の影響がうかがえる．どの年度においても係数は正の符号となっており，無回答項目の発生が高い場合は，返送日数が長くなると考えられる可能性がうかがえた．つまり，調査票記入時点に無回答が発生するほど回答しづらい事情があると，調査票の返送にあたって日数を要するようになるという関係性がある程度成立すると考えられる．

参考までに，表 11 の下部に，無回答項目数による単回帰分析の結果も示した．重回帰分析

表 11．返送日数の対数を従属変数とする回帰分析の結果（調査年度別，全体）．

	年度	2011		2012		2013		2014		2015		2016		2017		2018		2019		2020		2021		2022		全体	
	独立変数	β		β		β		β		β		β		β		β		β		β		β		β		β	
重回帰分析	(定数)		***		***		***		***		***		***		***		***		***		***		***		***		***
	性別																										
	女性	0.018		0.057	†	0.073	*	0.040		0.024		0.057	†	0.054	†	0.022		0.032		0.056	†	-0.008		0.035		0.038	***
	年齢																										
	10 代	-		-		-		-		-		-		0.063		0.116	**	-0.051		-0.037		0.084	*	0.018		0.025	**
	20 代	0.039		0.031		0.044		0.088	*	0.069	†	0.044		0.093	*	0.042		0.069	†	0.001		0.044		0.079	*	0.052	***
	30 代	0.050		0.036		0.110	**	0.006		0.112	**	0.043		0.097	*	0.006		0.057		0.066	†	0.036		0.071	†	0.054	***
	40 代	0.090	*	0.093	*	0.134	**	0.088	*	0.100	**	0.065	†	0.129	**	0.030		0.040		0.065	†	0.020		0.105	**	0.078	***
	50 代	0.024		0.024		0.106	**	0.069	†	0.096	**	0.037		0.043		-0.005		0.038		0.047		0.027		0.059	†	0.048	***
	60 代	-0.019		0.015		0.021		-0.030		0.082	*	-0.016		0.034		-0.073	*	-0.027		-0.022		-0.011		0.000		-0.004	
	職業																										
	常勤者	0.052		0.099	**	0.023		0.003		0.060	†	0.091	**	0.046		0.035		0.054		0.048		0.008		0.116	**	0.051	***
	学生	0.047		0.046		-0.018		-0.037		0.009		0.058	†	0.028		-0.078	*	0.100	*	0.052		-0.022		0.008		0.018	†
	居住年数																										
	1 年	0.046		0.025		0.018		0.009		0.009		0.012		-0.026		0.005		0.036		-0.047		-0.015		0.017		0.004	
	1-3 年	0.059	†	-0.004		0.043		0.010		-0.009		0.005		0.031		0.114	**	0.031		0.089	**	0.041		-0.036		0.030	**
	3-5 年	0.017		0.037		0.021		-0.014		0.024		0.015		0.030		0.072	*	0.020		0.047		-0.040		-0.017		0.019	*
	5-10 年	0.017		0.034		0.012		0.019		0.037		0.063	†	-0.022		0.026		0.031		-0.005		-0.016		0.011		0.018	†
	10-20 年	0.036		-0.013		0.046		0.007		0.022		0.011		0.007		0.050		0.085	*	0.095	*	0.008		0.034		0.025	*
	20-30 年	0.057		0.021		0.063		0.040		0.023		0.038		0.002		0.057		0.072	†	0.055		0.033		0.062		0.039	**
	30-40 年	0.084	†	0.010		0.012		0.026		0.014		0.035		0.031		0.069	†	0.055		0.025		-0.031		0.017		0.025	*
	40-50 年	0.046		-0.043		0.050		-0.035		0.015		0.071	†	-0.004		0.046		0.044		0.010		-0.009		0.001		0.012	
	無回答項目数	0.048		0.109	***	0.055	†	0.098	**	0.119	***	0.094	**	0.094	**	0.044		0.107	***	0.091	**	0.091	**	0.082	**	0.082	***
	R	0.166		0.206		0.193		0.187		0.193		0.191		0.216		0.199		0.220		0.214		0.156		0.233		0.162	
	R^2	0.028		0.042		0.037		0.035		0.037		0.037		0.046		0.040		0.049		0.046		0.024		0.054		0.026	
	Adj R	0.014		0.029		0.024		0.021		0.024		0.023		0.032		0.025		0.034		0.032		0.009		0.040		0.025	
	F	2.014	†	3.153	***	2.762	***	2.502	***	2.727	***	2.689	***	3.183	***	2.778	***	3.267	***	3.226	***	1.642	*	3.802	***	21.680	***
単回帰分析	(定数)		***		***		***		***		***		***		***		***		***		***		***		***		***
	無回答項目数	0.021		0.090	**	0.009		0.088	**	0.090	**	0.069	*	0.089	**	0.035		0.081	**	0.057	*	0.094	**	0.059	*	0.062	***
	R	0.021		0.090		0.009		0.088		0.090		0.069		0.089		0.035		0.081		0.057		0.094		0.059		0.062	
	R^2	0.000		0.008		0.000		0.008		0.008		0.005		0.008		0.001		0.006		0.003		0.009		0.003		0.004	
	Adj R	0.000		0.007		-0.001		0.007		0.007		0.004		0.007		0.000		0.006		0.002		0.008		0.003		0.004	
	F	0.523		9.964	**	0.096		9.440	**	9.873	**	5.829	**	9.447	**	1.551		7.636	**	3.949	*	10.779	**	4.166	*	56.643	***

R：重相関係数，R^2：決定係数，Adj R^2：自由度修正済み係数，F：F 値（分散比），†:p<0.10, *:p<0.05, **:p<0.01, ***:p<0.001, β:標準化回帰係数

の結果と比較すると，全ての年度において自由度修正済み係数が改善され，かつ標準化回帰係数の値が大きくなっている．単純に無回答項目数で返送日数を説明するよりも，4 項目に関する変数も含めた重回帰分析の方が関係性を明確に示せることが確認できる．

5.2 項目無回答の発生度合いを従属変数とする場合

本研究では，返送日数が長くかかった場合に項目無回答が発生しやすい可能性にも注目している．そのため，項目無回答の発生度合いを従属変数として，独立変数として返送日数の対数を用いる．なお，属性などの使用は論理的におかしいので当然用いない．

次に，調査不能発生の影響も調べるために，調査年度ごとの回収率を各ケースに割当て，独立変数として用意した．ここで回収率は，調査不能の発生を通じて調査票の総合的な質を反映した変数の代理的な役割を果たすことを予定している．ページ数を 8 頁に統一しており，その制限に応じて質問数も一定数以内におおむね統一されているため，調査票の質自体を指標化できないためである．回収率も項目無回答の多さも共通原因としての「調査票の質」の結果ではないかという異論もありえるが，回収率は，調査のタイトルや実施主体から調査の重要性を理解したかどうか，さらには調査票の導入部分や第 1 問目の印象などの影響が大きいと考えられるのに対し，項目無回答の個数は，具体的に調査に回答してからの回答の難度を反映すると予想され，やや異なった質を反映すると考えられる．さらに，返送日数はある程度回収率と裏返しの関係にあるはずだが，それとは別に個人の事情も反映されている．

その一方，調査年度ごとの影響という点で，2 章で述べたように 2014 年と 2022 年の 2 回の調査は特殊な理由で返送日数が長引いている．そのため，返送日数の影響を見るにあたって統制するために，2014 年度と 2022 年度に該当するか否かのダミー変数も用意した．

ここまで用意した変数を用いて，無回答項目数の回帰モデルとして，1)返送日数の対数，2)返送日数の対数，回収率，3)返送日数の対数，回収率，返送日数の対数 × 回収率(交互作用項)，4)返送日数の対数，回収率，5)返送日数の対数，回収率，返送日数の対数 × 回収率(交互作用項)，2014・2022 年度ダミーの 5 通りを用意した．最後に，これらのモデルを分析するが，無回答項目数がカウントデータであることを考慮し，ポアソン回帰と負の二項回帰による分析を行った．その結果，表 12 の結果が得られた．

まずポアソン回帰と負の二項回帰では，全体的な傾向はそれほど変わらなかったが，モデルの評価としては，負の二項回帰の適合度の方が良い．またポアソン回帰モデルでも負の二項回帰モデルモデルでも 5 番目のモデルが他よりも適合度が良いことが分かる．

そこでモデル 2–5 に注目して解釈を試みると，①返送日数の対数，②回収率が正の符号で，③返送日数の対数 × 回収率，④2014・2022 年度ダミーが負の符号で無回答項目数に影響があることがうかがえる．

1 点目の返送日数が長いほど項目無回答数が増加する傾向については，当初の想定通りで，回答への消極性が反映しているともいえるし，遅い時期の返送では回答が雑になってしまうともいえる．なお，2014・2022 年度ダミー変数があることによって当該年度で特別な事情で返送日数が長くなったことによる影響は調整されている．

2 点目の回収率については，回収率が上昇しても項目無回答の発生が増加することを示唆している．面接調査において指摘されてきた見かけ上の向上はかえって項目無回答を生じさせるという指摘（吉野, 1994)と同様の結果がうかがえる．ただし，郵送調査の場合は，面接調査の場合のように調査員が何度も調査対象者を訪問するといった努力ができず，しぶしぶ回答させているわけではない．せいぜい督促状を送付するぐらいであるが，この調査ではそれも実施されていない．そうすると回収率の上下変動に影響する要因は，事前に準備した郵送物の見栄えや調査票の良し悪し，日程上の都合などに限られる．特にこの調査の場合は，調査主体につい

表 12. 項目無回答の発生度合いについての回帰分析.

従属変数:										
無回答項目数(全体)	モデル 1-1		モデル 1-2		モデル 1-3		モデル 1-4		モデル 1-5	
ポアソン回帰	*B*		*B*		*B*		*B*		*B*	
(切片)	-1.213	***	-4.451	***	-11.993	***	-3.876	***	-11.022	***
日数の対数	0.234	***	0.235	***	4.383	***			4.619	***
回収率			0.053	***	0.177	***	0.051	***	0.161	***
日数の対数×回収率					-0.068	***			-0.072	***
2014・2022 年度ダミー									-0.207	***
LL	-15260.207		-15252.809		-15246.395		-15383.092		-15229.810	
AIC	30524.414		30511.617		30500.789		30770.184		30469.619	
AICC	30524.415		30511.619		30500.792		30770.185		30469.623	
BIC	30539.588		30534.379		30531.138		30785.361		30507.555	
CAIC	30541.588		30537.379		30535.138		30787.361		30512.555	
LR X^2(df)	177.823(1)	***	192.619(2)	***	205.447(3)	***	13.5598(1)	***	238.617(4)	***

従属変数:										
無回答項目数(12 年)	モデル 2-1		モデル 2-2		モデル 2-3		モデル 2-4		モデル 2-5	
負の二項回帰	*B*		*B*		*B*		*B*		*B*	
(切片)	-1.207	***	-4.519	***	-11.960	***	-3.778	***	-11.179	***
日数の対数	0.231	***	0.233	***	4.421	**			4.741	***
回収率			0.054	***	0.177	***	0.049	**	0.164	***
日数の対数×回収率					-0.069	**			-0.074	**
2014・2022 年度ダミー									-0.210	***
LL	-13032.723		-13027.251		-13022.742		-13088.628		-13010.965	
AIC	26069.447		26060.502		26053.484		26181.256		26031.929	
AICC	26069.448		26060.504		26053.487		26181.257		26031.933	
BIC	26084.621		26083.264		26083.833		26196.430		26069.865	
CAIC	26086.621		26086.264		26087.833		26198.430		26074.865	
LR X^2(df)	120.792(1)	***	131.736(2)	***	140.754(3)	***	8.983(1)	**	164.31(4)	***

LL：対数尤度，*AIC*：赤池情報量規準，*AICC*：有限サンプル相関 AIC，*BIC*：ベイズ情報量規準，*CAIC*：一致 AIC，*LR* X^2(df) ：尤度比カイ 2 乗(自由度)，†:p<0.10, *:p<0.05, **:p<0.01, ***:p<0.001, *B*:回帰係数

ても高槻市と関西大学という大きな枠組みは 12 年間変わっていないので，12 年間の結果からは調査主体に起因する効果も変わっていないとは考えられる．郵送物の内容や調査日程もほとんど変わっていないことから（松本, 2023），回収率の変動は，導入部分や第 1 問目の印象を中心とする調査票の質を反映している．一方，属性を中心とする項目無回答の個数は，回答をほぼ終えた終盤部分での回答状況であり，回収率の上昇する状況では消極的回答者がより多く含まれることから，調査票の後半では，無回答へのハードルが下がると項目無回答も増加しやすいのである．現実に毎年の調査票の最終ページ冒頭では，答えたくない場合や答えにくい場合は答えなくてもよい旨が明記されている．

3 点目の交互作用項が負の符号を示している点については，回収率が高い状況では返送日数が長い人ほど，前出の 2 つの主効果がやや弱まることを示している．これはまったく不自然な

話ではない．回収率が上昇した場合は，調査票の冒頭のできが良いわけだが，その場合は全体としても回答しやすいため，後期の回答者が消極的回答者だったとしても，別途項目無回答の発生率が下がる効果も働くと考えられるからである．

4 点目のダミー変数が負の符号を示している点は，当該年度の該当者であることは，他の年度の該当者である場合より，項目無回答が少なくなることを示している．すでに表 10 から単変量の効果という点でも確認できることであるが，他の変数の条件が一定の下でも個数が下がる傾向が示唆される．

6. まとめ

本研究では，郵送調査の返送時期が，調査不能と項目無回答という無回答誤差要因にどのように関わっているかを明らかにするため，「高槻市と関西大学による高槻市民郵送調査」の 2011 年度から 2022 年度までの 12 年分の調査結果を分析した．

まず 12 年間の調査年度別の回収標本と枠母集団の比較から無回答誤差の発生状況を確認した上で，無回答誤差と回収率・返送日数との集計レベルでの関連性を把握した．その結果，平均年齢に関する無回答誤差と回収率との間に負の相関関係が見られた．一方で返送日数については明確な関連性が見いだせなかった．ただしこれは，年度別に調査事情が異なることを考慮すると妥当な結果と考えられる．

次に個票レベルでの関連性を分析した．まず，性別・年齢・職業・居住年数の 4 項目について，無回答誤差を生じる要因のうち項目無回答が返送日数とどのように関わるか確認した．その結果，例外はあるものの，項目無回答の場合の方が回答している場合よりも平均返送日数が長くなることが多い傾向が読み取れた．

項目無回答と返送日数の間に関わりがあるとしても，両者の間の因果関係ははっきりと言えないため，項目無回答と返送日数の関係性のうち項目無回答が返送日数に影響を与えると想定した場合と返送日数が項目無回答の発生に影響を与えると想定した場合の両方の可能性を検討した．

前者の可能性については，多くの調査年度で，項目無回答の発生が返送日数に影響を与えていることが確認された．特に単回帰モデルの場合よりも，性・年齢・職業(常勤者・学生)，居住年数といった項目についてのダミー変数も投入した重回帰分析の場合の方がモデルとしても適合しており，標準偏回帰係数も強くでることから，無回答の発生に伴って，すなわち回答に悩んだ結果返送に日数を要するようになるという構造は，(具体的な構造はともかく)性別，年齢，職業，居住年数といった項目が返送日数に影響を与える構造と矛盾せず両立するものと考えられる．

後者の可能性についても，返送日数が長い場合ほど項目無回答がより多く発生する可能性が確認された．それだけでなく，調査の冒頭部分の質を代弁すると考えられる回収率も高い場合は，項目無回答の発生を高めていることが確認できた．回収率が高い場合は，消極的回答者が多く含まれるため，最終頁の無回答を許容する表現に影響され，項目無回答が増加しやすくなる可能性も確認できた．いずれにせよ返送日数の長期化が，回答への消極性を反映しているためであるか，遅い時期の返送であるがゆえに回答が雑になったかなどの事情により，項目無回答を多く発生させている関係性が確認できたことになる．

以上から，郵送調査において項目無回答が増加すればするほど返送日数が長くかかることと，郵送調査の返送時期が遅くなればなるほど項目無回答が増加するという双方向の因果関係が成り立っている可能性が示された．項目無回答とは，調査票の一部の項目において無回答が生じる現象であるが，これが増加し，全項目が無回答(白紙回答)となると，それは定義上も調

査不能とされる．これは，返送日数の長期化の延長上に調査不能があるという考え方とも整合性が取れる．論理関係を簡潔に表現すると次のようになる．

①（調査票に項目無回答の占める割合 $n\% \to 100\%$）$\Rightarrow$ 調査不能
②（返送日数 $\to \infty$）$\Rightarrow$ 調査不能
③①と②より，（調査票に項目無回答の占める割合 $n\% \to 100\%$）$\propto$（返送日数 $\to \infty$）

項目無回答の増加の極限には調査不能であるが，返送時期が長引くことの延長上に調査不能があるという関係と連動して成立する可能性が示唆される．

一方で，本研究の限界もある．1 つには，高槻市という地域限定の調査事例であるという点である．高槻市は，都市化が進んだ商業地域もある一方で，農業地域や山間部もあり，自治体自身が「とかいなか」を標榜するほどで，多様な側面を有している．いわゆる市部であり，人口も 35 万人程度で，近い規模の自治体も複数存在する．そのような観点からいっても，参考事例としてある程度有用性はあると考えられる．しかし，全国の調査や他の地域の調査においては，新しい課題が生じる可能性も否定できない．そういった点では，異なるタイプの調査でも類似の取り組みを行って知見を積み重ねることも重要である．

もう 1 つは分析面での限界である．返送日数を説明する回帰モデルにおいて対象者のもつ特性の影響を考慮する際にダミー変数を使って処理している．これは間違いとは考えていないが，年齢や居住年数の場合は，線形性や単調増加を仮定したほうが明確になる場合もあったかもしれない．しかしながら連続変数となる項目と無回答を表現する変数は重回帰モデルにおいて両立しない．これに限らず，モデルの改善の余地は残っているであろう．例えば，項目無回答数の分析においては，調査時点が 12 と比較的多く取れるので，年度をレベル 2，回収個票をレベル 1 としたマルチレベル分析的なアプローチも考えられる．今後の課題としたい．

ところで，本研究では扱わなかったが，返送時期が長引くと，変化しやすい意識項目などでは，調査開始に近い初期の回答と最後に回収された回答を同列に論じて良いのかという問題もある．無回答誤差とは別の測定誤差の問題に広がることにもなるが，政治意識や新しい社会課題などでは，当初は理解しづらいために項目無回答を生じやすいが，時間の経過につれて回答がはっきりする場合など考えられ，あながち無回答誤差の問題と無関係とも言い切れない．本研究ではそのような質問の項目無回答を扱っていないが，調査の種類によっては重要な課題となりえるであろう．

注．

[1] Zimmer（1956）は，早期回答者，督促後回答者，調査不能者の順に，年齢が下がり，教育年数が短くなり，軍の階級が低くなる傾向を示している．

[2] 田中（1962）は，調査不能で生じた回答の偏りを補正する方法として 4 つ提示しているが，締切期限後や後期の返送者を調査不能者の回答に近いとみなす方法（遅延標本法）と，時間的推移に基づいて調査不能者の回答を推測する方法（時系列法）の 2 つが有力であるとしている．

[3] この傾向は，年齢において反映されており，若い人ほど回答時期が遅く，また調査不能分に多く含まれるという傾向が確認されている（前田, 2005；松本, 2023）．

[4] 実質白紙で返送したと判断できる場合は不能票としているが，属性項目で主に構成されている 8 頁（最終頁）では，ページ冒頭で性別の質問があるため，答えたくない場合は答えなくてよい旨を明記している．このような事情から，属性項目のみ白紙の場合も有効回答（中途返送者）に含めることとしている．属性項目=最終頁ではないが，12 年間を通じて

用いられた 11 の属性項目全てが無回答な場合は，合計 37 件であった(表 10 参照)．

[5] 返送日数の度数の合計が，回収標本の合計より 18 件少ない 14577 になるのは，返送日数についての欠測があるためである．

[6] 返送日数の平均値と中央値の相関が強いのは当然なので議論しない．

[7] 従属変数に 5 章で後述する返送日数を対数変換した値を用いて，同様の 4 つの二元配置の分散分析を行った場合も，それぞれ性別($p < .001$)，年齢($p < .001$)，職業($p < .001$)，市内居住年数($p < .001$)ともに有意となる．なお調査年度 × 性別，調査年度 × 年齢の交互作用項については，従属変数に日数を用いた場合(調査年度 × 性別：$p < .001$，調査年度 × 年齢：$p < .001$)，対数変換後の日数を用いた場合(調査年度 × 性別：$p < .01$，調査年度 × 年齢：$p < .01$)ともに有意な結果を示す．これは，一部の年度において，回答者の方が無回答者よりも平均日数が長いという逆転現象を反映していると考えられる．

謝　辞

「高槻市と関西大学による高槻市民郵送調査」(2011〜2022 年度)は，高槻市と関西大学が共同で高槻市民を対象に実施した調査です．当該調査は，関西大学総合情報学部の「社会調査実習」の一環として行われています．調査に関わった全ての皆様に謝意を表します．さらに，本研究の成果をまとめるにあたっては，2023 年度関西大学学術研究員制度の御支援もいただきました．ここに記して感謝申し上げます．

参 考 文 献

Beatty, P. and Herrmann, D. (2002). To answer or not to answer: Decision process related to survey item nonresponse, S*urvey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), 71–85, John Wiley & Sons, New York.

Donald, M. N. (1960). Implications of nonresponse for the interpretation of mail questionnaire data, *Public Opinion Quarterly*, **24**(1), 99–114.

Filion, F. L. (1976). Exploring and correcting for nonresponse bias using follow-ups of nonrespondents, *Pacific Sociological Review*, **19**(3), 401–408.

Fitzgerald, R. and Fuller L. (1982). I hear you knocking but you can't come in: The effects of reluctant respondents and refusers on sample survey estimates, *Sociological Methods and Research*, **11**(1), 3–32.

Groves, R. M., Cialdini, R. and Couper, M. (1992). Understanding the decision to participate in a survey, *Public Opinion Quarterly*, **56**(4), 474–495.

Groves, R. M., Singer, E. and Corning, A. (2000). Leverage-salience theory of survey participation: Description and illustration, *Public Opinion Quarterly*, **64**(3), 299–308.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2004a), *Survey Methodology*, Wiley-Interscience, Hoboken, New Jersey (大隅昇 監訳, 氏家豊, 大隅昇, 松本渉, 村田磨理子, 鳰真紀子 訳 (2011).『調査法ハンドブック』, 朝倉書店, 東京.)

Groves, R. M., Presser, S. and Dipko, S. (2004b). The role of topic interest in survey participation decisions, *Public Opinion Quarterly*, **68**(1), 2–31.

林英夫, 村田晴路 (1996). 郵送調査における応答誤差：応答の正確度および安定度ならびに返信時期による応答の差異, 関西大学社会学部紀要, **28**(1), 171–189.

林英夫, 土田昭司, 林直保子, 松本敦, 箱井英寿, 矢島誠人, 池上和之, 小城英子, 吉川聡一 (2003). 郵送調査における返送率を左右する効果要因：質問紙のサイズおよび枚数ならびに協力依頼状の要請

表現の効果, 関西大学社会学部紀要, **34**(3), 259–278.
関西大学総合情報学部編 (2012).『2011 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2013).『2012 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2014).『2013 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2015).『2014 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2016).『2015 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2017).『2016 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2018).『2017 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2019).『2018 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2020).『2019 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2021).『2020 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2022).『2021 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
関西大学総合情報学部編 (2023).『2022 年度社会調査実習報告書―高槻市と関西大学による高槻市民郵送調査―』, 関西大学総合情報学部, 大阪.
小島秀夫 (1993). TDM による郵送調査の実践, 茨城大学教育学部紀要, **42**, 185–194.
Lin, I. and Schaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation, *Public Opinion Quarterly*, **59**(2), 236–258.
前田忠彦 (2005). 郵送調査法の特徴に関する一研究―面接調査法との比較を中心として―, 統計数理, **53**(1), 57–81.
松田映二 (2017). 新たな時代への地域づくり：標本調査を用いて人口減少への対応を考察, 政策と調査, **12**(2), 1–196.
松本渉 (2023). 郵送調査における返送日数に関する分析―「高槻市と関西大学による高槻市民郵送調査」の消印日付の活用, 情報研究, **57**, 1–20.
松本渉, 李容玲 (2015). 高品質な郵送調査の実践をめざして：高槻市と連携した関西大学総合情報学部の社会調査実習, 社会と調査, **15**, 107–111.
阪口祐介 (2023). 質問形式による無回答率の差と生活満足度の変化：高槻市民郵送調査の累積データの基礎分析, 情報研究, **57**, 21–44.
Siemiatycki, J. and Campbell, S. (1984). Nonresponse bias and early versus all responders in mail and telephone surveys, *American Journal of Epidemiology*, **120**(2), 291–301.
田中富士夫 (1962). 郵送調査における返送バイアス補正の試みとその妥当性, 金沢大学法文学部論集(哲学史学篇), **9**, 100–120.
土屋隆裕 (2005). 調査不能者の特性に関する一考察―「日本人の国民性 第 11 次全国調査」への協力理由に関する事後調査から―, 統計数理, **53**(1), 35–56.
渡邊勉 (2007). 郵送調査における早期回答者, 後期回答者, 非回答者の特徴, 信州大学人文学部人文科学論集人間情報学科編, **41**, 66–77.
Yan, T. and Curtin, R. T. (2010). The relation between unit nonresponse and item nonresponse:

A response continuum perspective, *International Journal of Public Opinion Research*, **22**(4), 535–551.

吉野諒三 (1994). 国民性意識の国際比較調査研究—統計数理研究所による社会調査研究の時間・空間的拡大—, 統計数理, **42**(2), 259–276.

吉野諒三 (2001).『心を測る—個と集団の意識の科学—』, 朝倉書店, 東京.

Zimmer, H. (1956). Validity of extrapolating nonresponse bias from mail questionnaire follow-ups, *Journal of Applied Psychology*, **40**(2), 117–121.

Relation between Nonresponse Error and the Return Timing of Questionnaires: Unit Nonresponse and Item Nonresponse in the Takatsuki Citizen Mail Survey by Takatsuki City and Kansai University

Wataru Matsumoto

Faculty of Informatics, Kansai University

Nonresponse includes unit nonresponse and item nonresponse. Although eliminating nonresponse errors is important, forcibly reducing nonresponse by increasing item nonresponse does has been suggested to not lead to an overall reduction in errors. This study examined the relation between the return timing of surveys and unit or item nonresponse using the "Takatsuki Citizen Mail Survey by Takatsuki City and Kansai University 2011–2022." The results at the aggregate level show a negative correlation between the nonresponse error for the average age and the response rate and confirm that increase in unit nonresponse influences the nonresponse error. An analysis of the raw data showed that the average number of return days tended to be longer in the item nonresponse than in the completed response. Therefore, we examined the relation between item nonresponse and the number of return days, and found the possibility of a reciprocal relationship between them. On the other hand, the number of item nonresponses tended to increase when the response rate was high. In a mail survey, if the introduction is of high quality, the questionnaires are returned quickly and the response rate increases; however, the number of item nonresponses may increase because of reluctant respondents. The results confirmed the possibility that unit nonresponse represents the extreme limit of the increase in item nonresponse. Further, considering the reciprocal relationship between item nonresponse and number of return days, unit nonresponse can be viewed as an extension of a prolonged return period.

Key words: Mail survey, nonresponse error, unit nonresponse, item nonresponse, return days, response rate.

2025 年 4 月改訂

「統計数理」投稿規程

1. 「統計数理」は，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指すものである．投稿原稿は，統計科学に関連した内容を持つもので，和文の原稿に限る．
2. 投稿原稿は次の 6 種とする．
   a. 原著論文 (Paper)
      統計科学の発展に貢献すると考えられる研究結果.
   b. 総合報告 (Review Article)
      特定の主題に関する一連の研究およびその周辺領域の発展を著者の見解に従って総括的，かつ体系的に報告したもの.
   c. 研究ノート (Letter)
      研究速報，新しい発想，提言，問題提起，事例報告など研究上，記録にとどめておく価値があると認められるものや，既発表の論文等に対するコメントで，研究上，記録にとどめておく価値があると認められるもの.
   d. 研究詳解 (Research Review)
      特定の研究領域における理論的あるいは応用的成果を，最近の結果や知見を加えてわかりやすく説明したもの.
   e. 統計ソフトウェア (Statistical Software)
      有用な計算法や解析法に関する短いプログラムおよびサブルーチンのリスト，利用手引き，実行例など.
   f. 研究資料 (Research Archives)
      歴史的なデータ，入手困難なデータや統計的手法の比較検討のために有用なデータ，あるいは，歴史的文献の翻訳や解説など.
   いずれも原則として，未発表のものに限る.
3. 投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定する.
4. 投稿原稿は電子投稿査読システム https://www.editorialmanager.com/toukei/ より投稿するものとする．原稿は pdf ファイルとし，必要なフォントはすべて埋め込み，原稿全体を一つのファイルにまとめることとする．論文が採択になった場合，著者は最終稿のソースファイルとハードコピーを提出するものとする.
5. 著作権
   (1) 掲載される論文等の著作権はその採択をもって統計数理研究所に帰属するものとする．統計数理研究所は，紙媒体の「統計数理」のほか電子媒体などを通じて論文等を公表することができる．特別な事情がある場合は，著者と本編集委員会との間で協議の上措置する.
   (2) 投稿原稿の中で引用する文章や図表の著作権に関する問題は，著者の責任において処理する.
   (3) 著者が自分の論文等を複製，転載，翻訳，翻案等の形で利用するのは自由である．この場合，著者は掲載先に出典を明記する.
6. 原稿は次の執筆要項に従って作成する.

「統計数理」執筆要項

1. 原稿は A4 用紙に 1 行 36 字から 40 字で 1 行おき，1 頁あたり 22 行程度とする．原稿の長さは原則として表・図を含めて 30 頁相当以内とし，各ページにページ番号を付す．図表は別紙にまとめ，本文中には挿入箇所のみを指定する．LaTeX で原稿を作成する場合は，「統計数理」スタイルファイルの使用を推奨する.
   https://www.ism.ac.jp/editsec/toukei/
2. 文体は「である体」とし，句読点は「, 」「. 」を用いる.
3. 原稿は以下の順に書くものとする.
   ［第 1 頁］ 標題，著者名，所属名および所在地，メールアドレス，和文要旨 (500 字程度，文献の引用および数式は原則として避ける)，和文キーワード (6 語以内).
   ［第 2 頁］ 英語による標題，著者名，所属名，Abstract (450 ワード程度)，Key words (6 words and phrases 以内)．Abstract は，問題の所在と得られた結果等がそれだけで理解できるようなものとする.

［第 3 頁以降］

① 本文：章，節の番号は，第 1 章にあたるものは，“1.”, 第 1 章 第 1 節にあたるものは，“1.1” というようにつける．また，式の番号は，章ごとに (2.1), (2.2) のようにし，式の左側に配置する．

② 数式：数式は簡明さを心がけ，添字にさらに添字をつけるのはなるべく避ける．

③ 参考文献：書き方は本要項 第 4 項を参照.

④ 表：一枚の用紙に一つの表を書く．表の番号は論文中に現れる順に従って，表 1，表 2,... または，Table 1, Table 2,... のようにする．

⑤ 図：一枚の用紙に一つの図を描く．図はそのまま写真製版できる鮮明なものを用意する．大きさは印刷出来上がりの 1〜2 倍とし，トレースが必要な場合は原則として著者が行うものとする．図の番号は論文中に現れる順に従って，図 1，図 2,... または，Fig. 1, Fig. 2,... のようにする．

⑥ 注：本文中の注釈は極力避ける．やむを得ず注釈をつける場合は脚注とせず，論文末尾に後注とする．後注は，順番に “1, 2,...” の番号を付け，本文中では上付きで示す.

4. 本文中での参考文献の引用は，著者名 (出版年) とする．たとえば，Efron (1982), 清水・湯浅 (1984), Cox and Snell (1981), 坂元 他 (2004), Nakano et al. (2000).

5. 参考文献の書き方

［雑誌の場合］いずれの場合も雑誌名は省略しないものとする．

①ページ番号がある文献

著者名 (出版年)．標題，雑誌名，巻，ページ〔始–終〕.

【例】Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

【例】Bligh, E. G. and Dyer, W. J. (1959). A rapid method of total lipid extraction and purification, *Canadian Journal of Biochemistry and Physiology*, **37**, 911–917, https://doi.org/10.1139/o59-099.

②ページ番号がない文献

著者名 (出版年)．標題，雑誌名，巻に続けて，DOIがあるものはDOIを入れ，DOIがない場合はURLと最終アクセス日を記載する.

【例】Lien, G. Y., Kalnay, E. and Miyoshi, T. (2013). Effective assimilation of global precipitation: Simulation experiments, *Tellus A,* **65**, https://doi.org/10.1175/WAF-D-13-00032.1.

【例】Kiraly, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data, *Journal of Machine Learning Research*, **20**(31), http://jmlr.org/papers/v20/16-314.html（最終アクセス日 2023 年 10 月 1 日）.

［叢書の中の一巻の場合］

著者名 (出版年)．書名 (編集者名)，叢書名，発行所名，発行地名.

【例】Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, Mathematics and Its Applications, Reidel, Dordrecht.

［単行本等の場合］

著者名 (出版年)．書名，発行所名，発行地名.

【例】Cressie, Noel (1993). *Statistics for Spatial Data*, Wiley, New York.

［編集書の中の一部の場合］

著者名 (出版年)．標題，編集書名 (編集者名)，巻，ページ，発行所名，発行地名.

【例】Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

なお，同じ著者によるものが同一年に複数個現れる場合には，(1980a), (1980b) などとして区別する．文献は，日本人も含め，著者名のアルファベット順に並べる.

6. 著者校正は原則として一回とする．その際，印刷上の誤り以外の字句や図版の訂正，挿入，削除等は原則として認めない.