複数階層の領域分割を考慮した Tree-guided Group Lasso による空間的異質性の抽出

竹本 一至*・井上 亮**

Extraction of multi-scale discrete spatial heterogeneity using Tree-guided Group Lasso

Kazushi TAKEMOTO*, Ryo INOUE**

Abstract: One approach to analyze discrete spatial heterogeneity is to set specific regional divisions in advance and estimate coefficients for each region. Sine there are multiple hierarchical levels of candidate regional divisions, such as prefecture, municipality, and city block, and the scale of spatial heterogeneity may vary by location, it is necessary to consider multiple candidate regional divisions simultaneously. However, efficient approaches to this problem have not been discussed. Tree-guided Group Lasso (TGL) is a type of sparse estimation method that defines a hierarchical group structure represented by a tree for explanatory variables and performs variable selection for each group. In this study, we propose to apply TGL to extract discrete spatial heterogeneity considering multiple hierarchical regional divisions. We tested the effectiveness of this method by applying it to real estate rental data for Setagaya, Tokyo.

Keywords: 空間的異質性 (spatial heterogeneity), 空間スケール (spatial scale), 階層性 (hierarchy), スパースモデリング (sparse modeling), Tree-guided group lasso (Tree-guided group lasso)

１．はじめに

近年，行政機関によるオープンガバメント政策や民間企業の積極的なデータの公開によって，幅広い種類のデータの入手が可能になり，蓄積されたビッグデータを活用した自然環境や社会経済現象の分析が盛んに行われている．位置に関する情報を持つ多様な空間データも流通するようになり，空間現象を対象とした分析の可能性も広がっている．

空間データ分析における関心事の一つは，空間現象の生成要因やその過程が場所によって異なる，空間的異質性の把握である．高い空間解像度を有するデータが入手可能になっており，これまでよりも詳細な分析ができる環境が整ってきている．空間的異質性を分析する手法の一つは，特定の領域境界で空間現象の生成過程が不連続に異なる空間的異質性の存在を仮定し，事前に分析者が設定した領域分割に基づき，各領域の係数を推定する方法である（例えば，Goodman and Thibodeau, 1998）．この分析にスパースモデリングを活用することが検討され，井上ら (2020) は fused lasso (Tibshirani et al., 2005)，Inoue et al. (2020) は fused-MCP (Jing et al., 2018) を導入した分析を行った．これらの分析では，分析の最小空間単位となる領域分割を事前に設定し，各領域に固有の係数を置いたモデルを立てた上で，隣接領域の係数間の差に対する正則化を行うことにより，係数が同じ値に推定される一連の領域を抽出できる．この正則化によって，事前設定の領域よりも大きな空間スケールで発生する空間的異質性を捉えられる．

しかし，fused lasso や fused-MCP による分析は，隣接領域を結合して抽出した領域が複雑な形状を示す場合があり，結果の解釈が難しい可能性がある．また，高い空間解像度の分析を目指して細かい領域分割を設定すると，隣接領域の組み合わせ数，すなわち正則化条件が増加し，計算量が大幅に増加するため，推定の実行可能性が乏しくなる．また，詳細な領域分割を基に，隣接領域を結合して広域で生じている空間的異質性を抽出することは難しい．

不連続的に起こる空間的異質性は，自治体毎の保育環境など住民サービス水準の違いや，小学校毎の教育サービス水準の違い，鉄道沿線毎の交通サービ

*　学生会員　東北大学 大学院情報科学研究科（Tohoku University）
〒980-8579　宮城県仙台市青葉区荒巻字青葉 6-6-06　E-mail：kazushi.takemoto.q7@dc.tohoku.ac.jp
**　正会員　東北大学 大学院情報科学研究科（Tohoku University）

ス水準の違い, 町名による地域イメージの違いなど, 多くの要素に影響を受け，市区町村，大字・字，街区などの行政界，小・中学校の学校区，駅勢圏，鉄道沿線など，複数の空間スケールで存在すると考えられる．そこで，空間現象の生成過程に関する空間的異質性を複数階層の領域分割に基づいて適切に表現できれば，詳細から広域なスケールまでの空間的異質性の分析が可能になると期待される．

そこで，本研究は，複数階層の領域分割を設定した上で，各場所で生じている空間的異質性を抽出する手法として，木構造で表せる階層的な係数構造を有するモデルに対するスパースモデリング手法である Tree-guided Group Lasso (TGL) に注目する．TGL の手法や他の分野での応用例を紹介した後，不動産賃料データに関する空間的異質性分析を通して，空間データ分析への適用可能性を検証する．

２．TGL を活用した空間的異質性分析手法の提案

2.1. TGL

TGL (Kim and Xing, 2010) はスパースモデリングの 1 種で，複数の変数をまとめて変数選択ができる Group Lasso (Yuan and Lin, 2006) の拡張手法である．

Group Lasso は Lasso 推定で利用する L_1 ノルムの代わりに, L_1/L_2 ノルムというグループ毎の正則化を利用し，グループ単位の変数選択を可能にする．しかし，変数が複数のグループに重複して含まれる設定では推定ができない課題も有する．TGL は木構造に基づく罰則項を用いることにより，各変数が階層の異なる複数のグループに属していても，スパース推定を可能にする手法の 1 つとして提案された．

ここで，n件のデータをp種類の共変量で表す線形回帰モデルの推定を考える．$n \times 1$の被説明変数ベクトルを$\boldsymbol{y}$，$n \times p$行列で表される説明変数を$\mathbf{X}$，$p \times 1$のパラメータベクトルを$\boldsymbol{\beta}$とする．木構造の全体の深さを，根となるノードから葉となるノード群までのエッジ数dで表し，深さiでの共変量のグループ数をn_iとし，深さiでのj番目のグループをG_j^iで表す．G_j^iに対応するパラメータベクトルを$\boldsymbol{\beta}_{G_j^i}$, その正則化項の重みを$w_j^i$で表すと，TGL の推定は式(1)で書ける．

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2 + \lambda \sum_{i=0}^{d}\sum_{j=1}^{n_i} w_j^i \left\|\boldsymbol{\beta}_{G_j^i}\right\|_2 \qquad (1)$$

ただし，$\|\ \ \|_2$はユークリッドノルム (L_2 ノルム)である．λは正則化項に対する重みを表し，重みw_j^iと共にハイパーパラメータである．また，木構造のグループ設定は事前に設定する必要があり，同じ深さで 1 つの変数が複数のグループに重複して属することはできない．

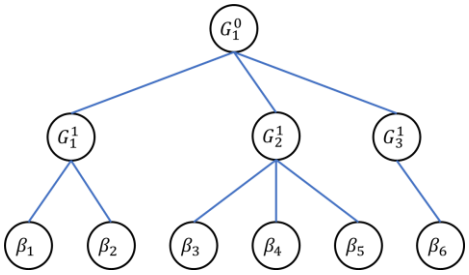


図 1　木構造の例

2.2. TGL の重み付けスキーム

TGL ではグループ全体と各係数にかかる罰則を調整する比を設定し，各グループと各係数に関する正則化項に対する重みを決定する．グループ全体にかかる罰則が強くなるように比を設定すれば上層のグループ単位で変数選択を行いやすくなり，各係数にかかる罰則を強くするように比を設定すれば Lasso に近づく．

TGL ではこの比を利用して，推定で生じうる偏りを防ぐ重み付けスキームが設定されている．重み付けスキームでは，親ノードと子ノードの重みの比を利用し，各グループに対する重みを根ノードから葉ノードに向かって課し，1 つの変数が複数階層のグループに属することで過剰に罰則がかけられることを防ぐ働きがある．

親ノードの重みの比率をg_i，子ノードへの重みの比率をs_iとする $(g_i + s_i = 1, (g_i, s_i > 0, g_i > s_i))$と，重み付けスキームは式(2)で書ける．

$$w^i = \begin{cases} g_i & if\ i = 0 \\ g_i \cdot \prod_{m\in(0,\dots,i-1)} s_m & if\ 0 < i < d \\ \prod_{m\in(0,\dots,i-1)} s_m & if\ i = d \end{cases} \qquad (2)$$

このスキームにより階層ごとに異なる重みが与えられ，同階層のグループの重みは等しくなる．

2.3. TGL の応用例と空間的異質性分析への適用

TGL の応用例には，説明変数が木構造となるマルチタスク回帰 (Kim and Xing, 2010) や，ある一定の領域が同時に反応することで疾患の原因となることが先験的に知られている DNA や塩基などのデータを用いた遺伝子変異の研究 (Hao, et al., 2018) などがある．

空間データ分析に近い TGL の応用例として，Liu ら (2012) の脳画像解析が挙げられる．1 人の被験者の複数の解像度の脳画像に対して，TGL を活用して脳疾患を分類した．この複数解像度の脳画像が階層的な木構造で表せるため，TGL が活用された．

空間データ分析において，市区町村界・町丁目界のように木構造で表現できる階層的な領域分割を対象領域に設定すれば，ある空間現象が有する複数の空間スケールからなる空間的異質性の抽出に，TGL を応用できる可能性があると考える．そこで，次章では TGL を空間的異質性抽出に活用し，その適用可能性を検証する．

３．TGL を活用した空間的異質性抽出の検証

不動産賃料データの空間性異質性分析を対象に，階層間の重みの比率を変えた複数の TGL のモデルを設定して推定し，その推定結果を基に TGL の適用可能性を検証する．

パラメータ推定には，Matlab の SLEP パッケージ (Liu et al, 2009) の TGL の関数を使用する．

3.1. 使用データ

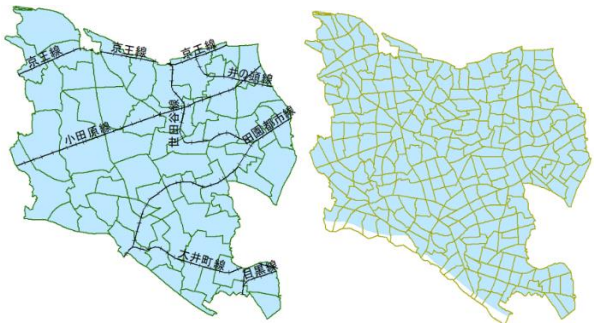
賃料データは 2019 年にアットホーム株式会社が収集した東京 23 区のマンションの募集賃料データの中から世田谷区のデータを使用する．築年数が 35 年以下の物件を対象とし，異常値・欠損値を取り除いた 279,302 件のデータを分析対象とする．

3.2. モデル設定

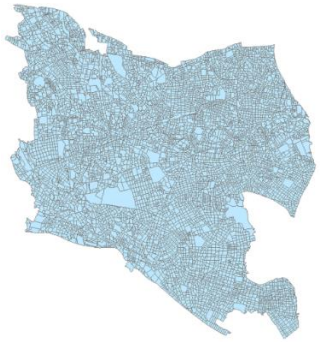
被説明変数は単位面積当たりの賃料の対数値とし，説明変数は，専有面積，物件の所在階層，築年数，最寄り駅までの徒歩時間，最寄り駅から山手線の駅 (新宿，代々木，原宿，渋谷，恵比寿，目黒) までの最短所要時間のそれぞれの対数値と，街区ごとに設定したダミー変数とする．最寄り駅から山手線の駅までの所要時間は Yahoo!路線情報の乗換案内で出発を 2022/08/01/12:00 として検索し，表示されたものを利用した．

グループ設定は 3 階層の構造を設定する．上層は町（大字），中間層は丁目，下層は分析の最小空間分割単位として街区を設定した．世田谷区では，59 町，271 丁目，5,388 街区それぞれに少なくとも 1 件の賃料データが存在した．正則化項の数となる総グループ設定数は 5,718 である．

本研究では，木構造の階層間の重みの比を $(g_i, s_i) = (0.8, 0.2), (0.7, 0.3), (0.6, 0.4)$ に設定した 3 つのモデルを検討する．各モデルで正則化項に対するハイパーパラメータ λ を変化させて探索し，BIC 最小となる設定の結果を採用する．以降は，$g_i = 0.6$ のモデルをモデル 06, $g_i = 0.7$ のモデルをモデル 07, $g_i = 0.8$ のモデルをモデル 08 と表記する．



(a) 町(大字)界　　(b) 丁目界



(c) 街区界

図 2　階層的なグループ設定

3.3. 分析結果

表 1 に，各モデルについて，BIC が最小となったハイパーパラメータ λ 設定の推定結果について，決定係数，非 0 となった係数の数，および，各層への重みを示し，図 3 にモデル 06 とモデル 08 の街区の係数の推定結果を示す．

以後，推定結果を表す図では，係数が正に推定された街区を赤，負に推定された街区を青で表し，色の濃さが絶対値の大きさを表す．また，賃料データが一件もなく係数を設定していない街区を灰色で表す．緑色と黄色の太枠はそれぞれ町(大字)界，丁目界の分割領域を表す．また，図 4 は丁目の分割で，図 5 は町(大字)の分割ですべての街区が 0 に推定された領域を表す．

モデル 06 では，丁目単位で 0 に推定された領域は 54 領域で全体の約 20% (= 54/271)であるのに対し，モデル 08 は 36 領域が 0 に推定され，全体の約 13%となった (図 4)．また，モデル 06 はモデル 08 と比べ，街区単位で 0 に推定された領域が多い (図 3)．モデル 06 は，モデル 08 に比べて親ノードへの重みの比率が低く，より上側の階層に対する正則化項の働きが抑えられるため，より大きな空間スケールでの空間的異質性が抽出されたと考えられる．また，表 1 から 3 つのモデルの決定係数は大きく変わらないことが確認できる．

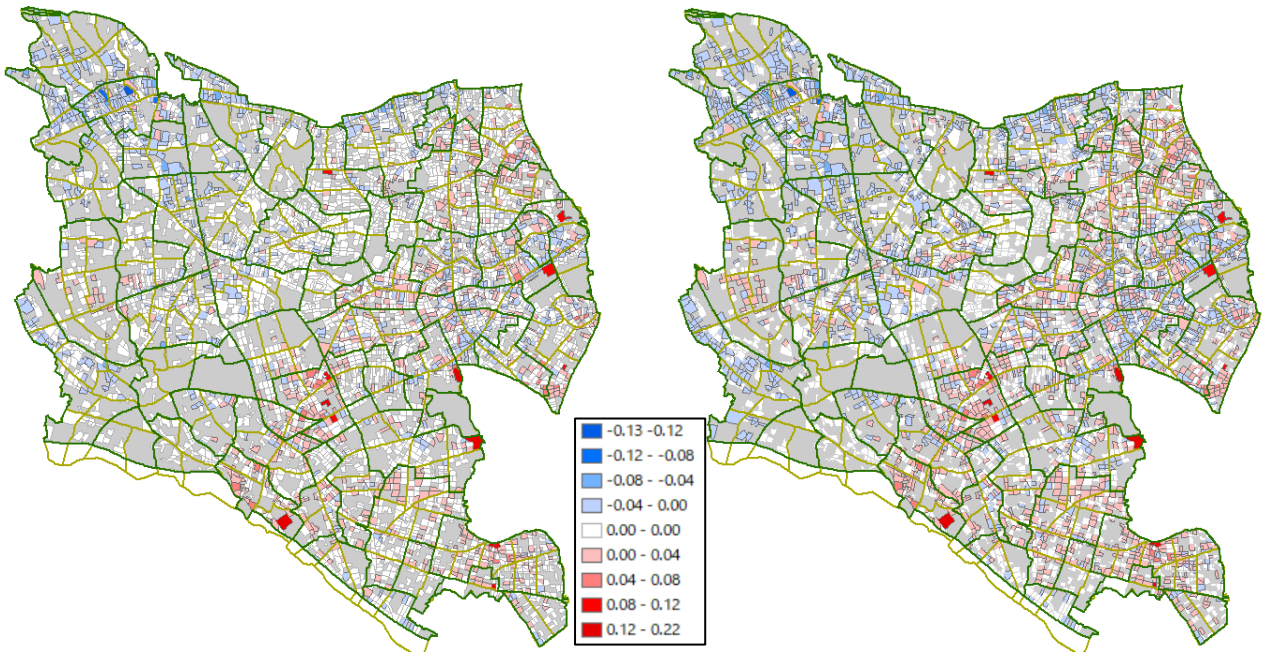
これらの結果から，TGL で重みの比を変えると，各層へ課せられる罰則の大きさが調整されるため，結果として，モデルが抽出する空間的異質性の空間スケールが変化することを確認した．説明力が高い結果を出す重みの比を選択することで，データが有する空間的異質性のスケールを把握できる．

今回の世田谷区のデータではモデル 06 が BIC 最小となるため，適した空間スケールで空間的異質性を抽出しているモデルであると考えられる．

表 1　各モデルの統計量と 3 つの階層に対する重み

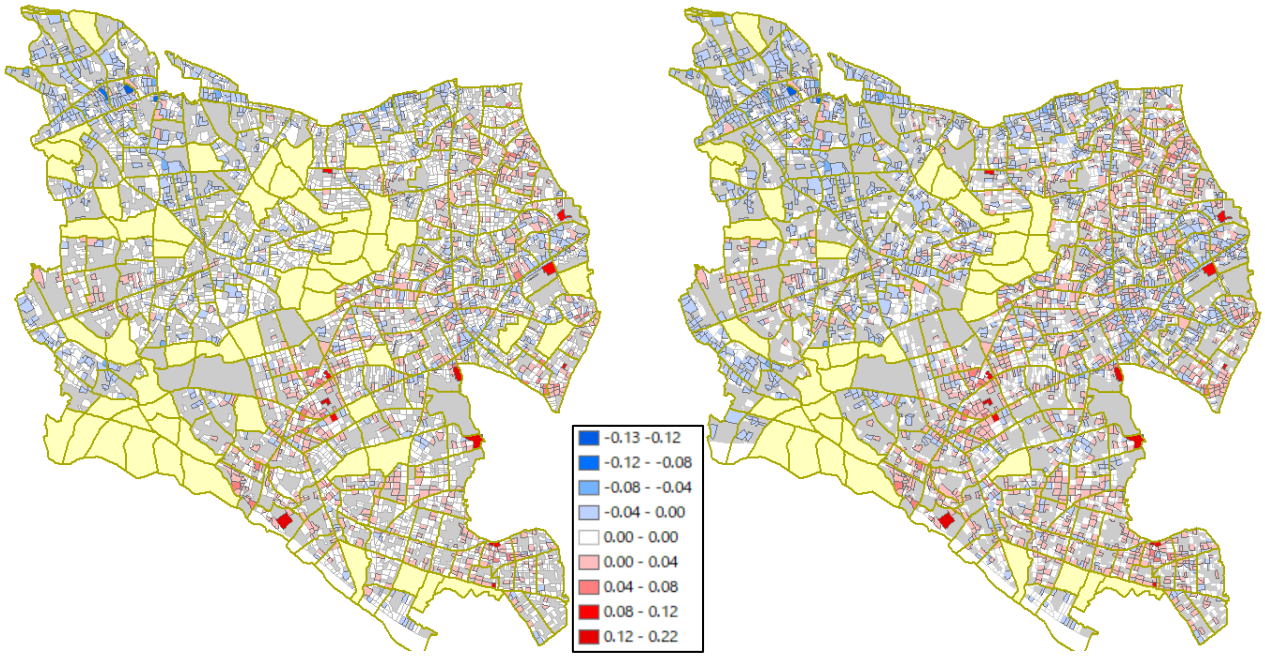
重み比率(g_l)	BIC	決定係数	自由度調整済みR_2	非0の係数の数
0.6	-407585	0.679	0.677	1785
0.7	-401230	0.679	0.676	2283
0.8	-391512	0.681	0.677	3209

重み比率(g_l)	λ	街区層への重み	丁目層への重み	町(大字)への重み
0.6	17.44	2.79	4.19	6.28
0.7	20.1	1.81	4.22	9.85
0.8	23.25	0.93	3.72	14.88



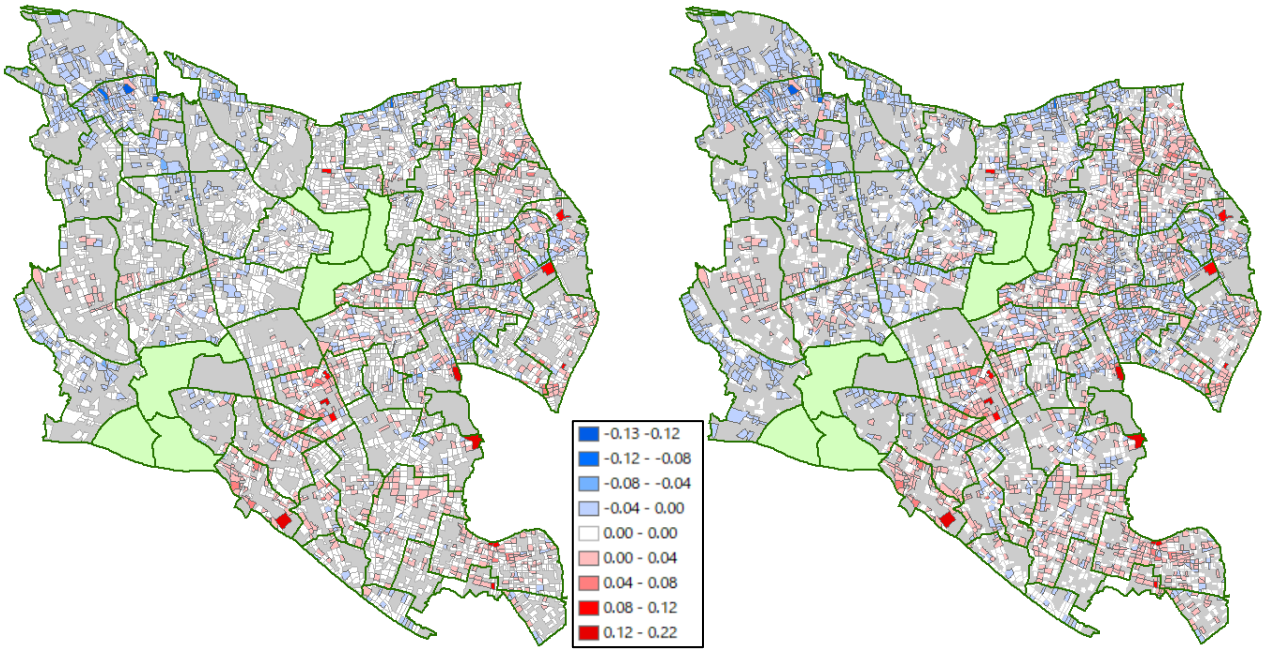
(a) モデル 06　　(b) モデル 08

図 3　モデル 06 とモデル 08 の分析結果



(a) モデル 06　　(b) モデル 08

図 4　丁目界に着目した分析結果



(a) モデル 06　　(b) モデル 08

図 5　町(大字)界に着目した分析結果

3.4. 世田谷区の分析結果の考察

TGL に基づく分析結果から，世田谷区の賃料データが有する空間的異質性に関して考察する．

まず，山手線など主要路線が走る東側と南側が全体的に正に推定され，逆に西側と北側が負に推定され，世田谷区全体の傾向を捉えられている．

多摩川付近で内水氾濫での浸水が 2~3m となる喜多見，烏山川の流域で浸水の恐れがある烏山地域は町単位で負に推定されている領域が多い (図 6)．逆に，ブランドエリアとして知られる二子玉川や成城，

中心街へのアクセス，公園，お店の多さなど総合的な観点から住みやすい街とされている用賀は町単位で正に推定がされている領域が多い (図 7). 一方で，下北沢や三軒茶屋などでは，1 つのエリアに正と負の推定結果を持つ領域が混在する地域では局所的な空間的異質性を抽出する結果となっている (図 8). これらのエリアは木賃ベルトと呼ばれる，木造住宅が密集し，火災に対する脆弱性を持つ反面，商店街や活発的なコミュニティが形成されている東京近郊の住宅地域に属している．

分析結果から世田谷区では物件の賃料が形成される要因が複数の空間スケールで存在する傾向を持ち，TGL を活用した手法は階層構造に基づき解釈のしやすい結果を導くと考えられる．

3.5. TGL のモデルと Lasso のモデルの比較

街区単位に係数を設定した Lasso のモデルで分析し，TGL のモデル 06 の結果と比較する．Lasso のモデルは，TGL のモデルから上層・中間層単位の正則化条件を外したものに概ね対応する．

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1 \qquad (2)$$

式(2)は Lasso のモデルで，$\|\ \ \|_1$は L_1 ノルムを，λは正則化項に対する重みを表わす．分析では正則化パラメータを変化させて BIC 最小のものを選択する．

図 9 に Lasso の析結果を示す．TGL の分析と同様に，図 10 は丁目の分割で，図 10 は町(大字)の分割ですべての街区が 0 に推定された領域を表す．表 2 は TGL のモデル 06 と Lasso の統計量を表す．

Lasso のモデルは街区単位では TGL のモデルより 0 に推定された領域が多い (図 9)．また，丁目単位で 0 に推定された領域は 32 領域，町単位で 0 に推定された領域は 1 領域で，丁目単位や町単位ではモデル 06 のほうが 0 に推定された領域が多い結果となった (図 10，11).

表 2　TGL のモデルと Lasso のモデルの統計量

モデル	BIC	決定係数	自由度調整済みR_2	非0の係数の数	街区への重み
TGL	-407585	0.679	0.677	1785	2.79
Lasso	-419138	0.686	0.685	1370	4.75

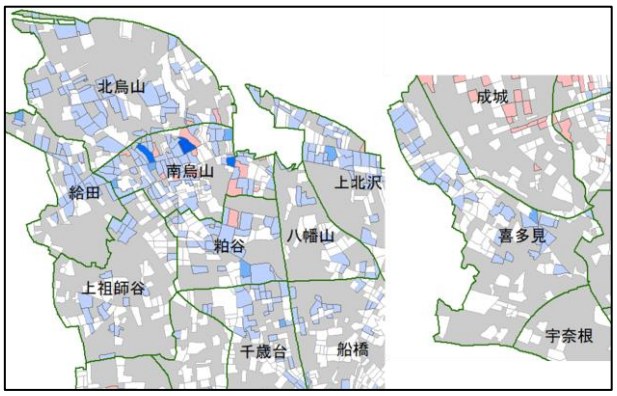


図 6　喜多見・烏山地域

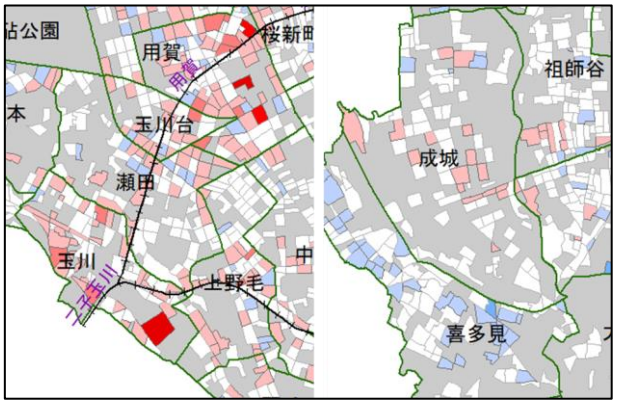


図 7　二子玉川・用賀・成城地域

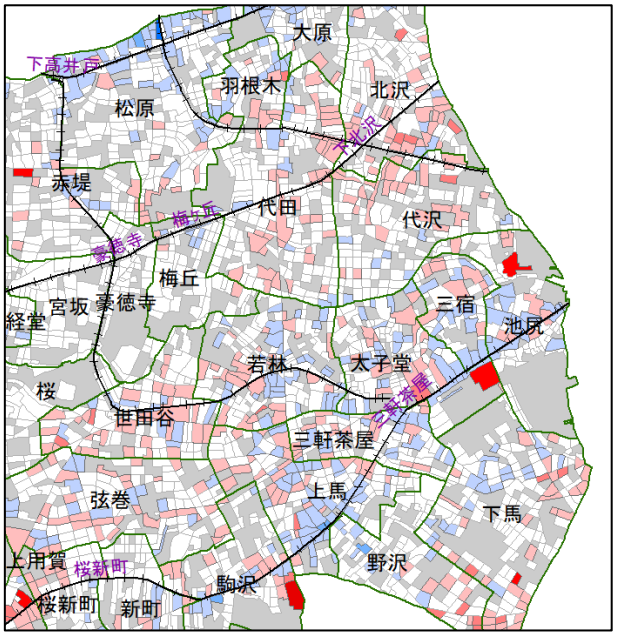


図 8　木賃ベルト地帯 (下北沢・三軒茶屋)

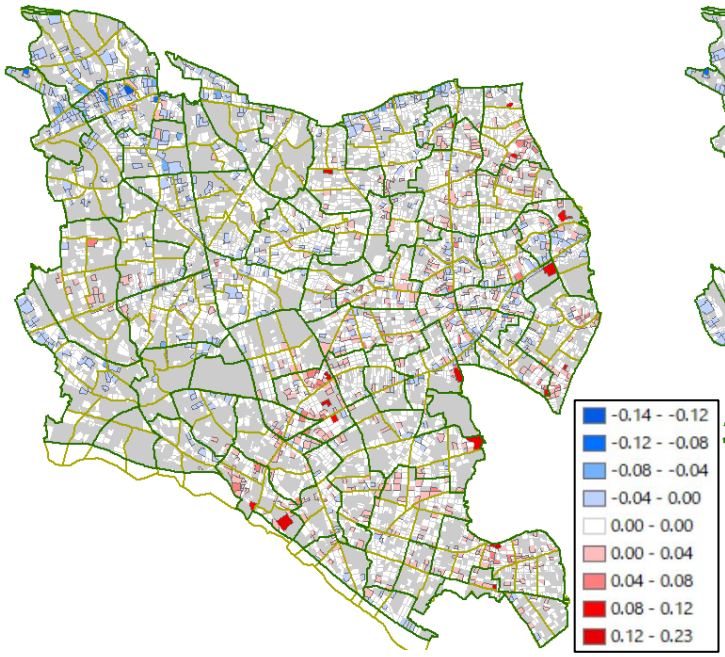


図 9　Lasso モデルの分析結果

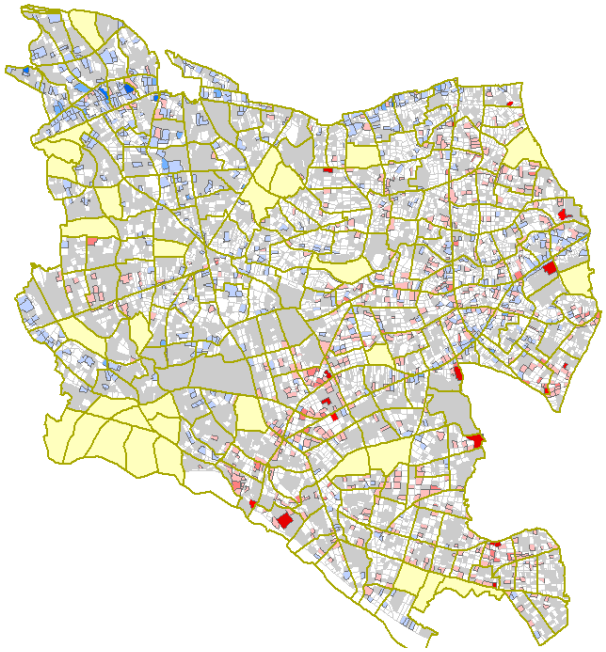


図 10　丁目界に着目した Lasso モデルの分析結果

一方，表 2 から Lasso の方が決定係数・BIC 共に優れた結果を示したことが確認できる．TGL は，上位 2 階層に対する制約を加えているため，データの表現は Lasso より劣る結果となった．

しかし，TGL は，上層・中間層毎に下層に設定した係数がすべて 0 となる推定がされるため，複数階層から領域ごとに適切な空間スケールを用いて結果

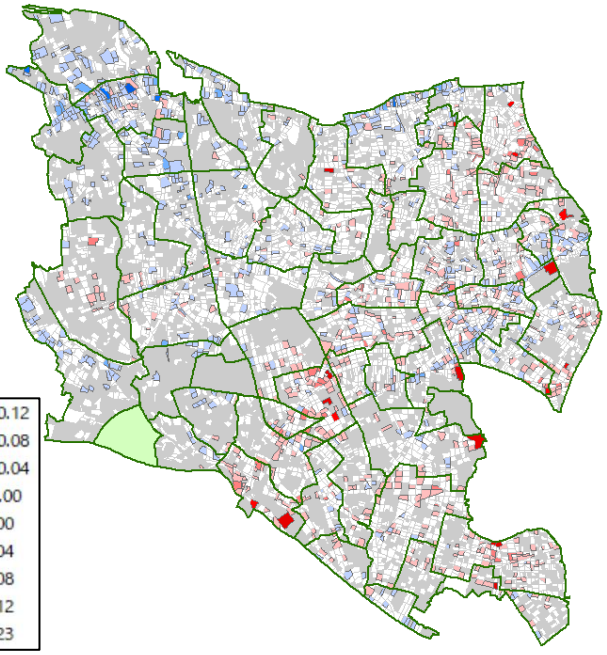


図 11　町(大字)界に着目した Lasso モデルの分析結果

を解釈することができる手法である．また，Lasso に比べて計算時間の増大は見られなかったことから，計算コストの増大を抑えながら階層性を考慮した分析を可能にする手法だと評価できる．

4．おわりに

本研究では，空間現象が有する空間的異質性の空間スケールを把握する分析方法として，木構造に基づき階層的なグループを設定する TGL を用いて係数を推定するモデルを検討した．重みの比率を変えたモデルの検証から，モデルごとに異なるスケールの空間的異質性を抽出できる可能性を確認した．また，Lasso と比較すると少し精度は劣るが，計算コストを抑えながら複数階層構造を考慮した空間的異質性の分析を可能にした手法であることを確認した．

本稿では世田谷区の賃料データと階層構造に行政界を活用した分析を行ったが，他の領域のデータや，学校区や駅勢圏などの異なる領域分割設定を用いた分析の検証を進める必要がある．計算コストも異なる空間スケールを扱う手法と比べて大きくないので，広範囲の対象領域から複数階層の分割単位で空間的異質性を抽出するモデルの検討を今後行う．

また，本研究の分析では，説明変数の領域ダミー変数のみに対して木構造に基づく罰則項を与えたが，今後，複数の説明変数に対する係数に対して罰則を与えた分析の検討も進めたい．さらに，領域分割が木構造に基づく階層的な構造ではなく，下層領域が複数の上層領域と重なる場合にも分析可能なより柔軟な手法の検討も必要である．

謝辞

本研究は, JSPS 科研費 21H01447 の助成を受けた．また，本研究は，東京大学 CSIS 共同研究（No. 815）による成果である（利用データ: 不動産データライブラリー 戸データ 全国 2013-2017 データセット（アットホーム株式会社提供））．

参考文献

Goodman, A. C., and Thibodeau, T. G. (1998) Housing market segmentation. *Journal of Housing Economics*, **7** (2): 121–143.

Inoue, R., Ishiyama, R., and Sugiura, A. (2020) Identifying local differences with fused-MCP: An apartment rental market case study on geographical segmentation detection. *Japanese Journal of Statistics and Data Science*, **3**: 183–214.

Jing, B., Yang, G., Yu, X., and Zhang, C. (2018) Fused-MCP with application to signal processing. *Journal of Computational and Graphical Statistics*, **27** (4): 872–886.

Kim, S. and Xing, E. P. (2010) Tree-guided group lasso for multi-task regression with structured sparsity. *27th International Conference on Machine Learning*, 1–14.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67** (1): 91–108.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58** (1): 267–288.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68** (1):49-67

Liu, J., Ji, S., and Ye, J. (2009) SLEP: Sparse Learning with Efficient Projections. Arizona State University.

Hao, X., Yao, X., Risacher, S.L., Saykin, A.J., Yu, J., Wang, H., Tan, L., Shen, L., Zhang, D., (2018). Identifying candidate genetic associations with MRI-derived AD-related ROI via tree-guided sparse learning. *IEEE/ACM Transactions on Computational. Biology and Bioinformatics.,* **16**(6): 1986-1996.

Liu, M., Zhang, D., Yap, P.T., Shen, D. (2012). Tree-Guided Sparse Coding for Brain Disease Classification. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, *Part III, LNCS* 7512, 239-247

井上 亮, 石山 里穂子, 杉浦 綾子 (2020) 東京都区部の賃貸マンション市場の地理的分割の実態把握－スパースモデリングによるアプローチー. 土木学会論文集 D3 (土木計画学), **76**(3): 251–263.