ビッグデータのための空間加法混合モデリング：所得分析への応用

村上大輔・Daniel A. Griffith

Spatial Additive Mixed Modeling for Big Data: Application to Income Analysis

Daisuke MURAKAMI and Daniel A. GRIFFITH

Abstract: This study develops a spatial additive mixed model to estimate spatially varying effects, group effects, and other effects from large samples. This approach first eliminates matrices whose size depends on the sample size N before the model estimation. These procedures make our approach computationally efficient even if N is large (e.g., millions) and there are many spatial and non-spatial effects being estimated. The developed method is applied to a tract-level income analysis in USA.

Keywords: 空間統計（spatial statisics），ビッグデータ（big data），空間加法混合モデル（spatial additive mixed model），spmoran（spmoran）

1. はじめに

地理情報の多様化・大規模化が著しく，大規模な地理空間データを柔軟に解析することができるような方法が求められている．

回帰問題に対しては，空間相関をはじめとする地理的現象を明示的に捉えようという空間統計モデルが幅広く提案されてきた．しかしながら，標本数を N とすると，基本的な空間統計モデルの推定には共分散行列（$N \times N$）の逆行列計算が必要であり，その計算負荷は N^3 のオーダーで増加するため（$O(N^3)$；標本数が 2 倍になると計算負荷は 8 倍），基本的な空間統計モデルで扱うことができるのは，せいぜい標本数 1 万程度までである．

一方で，大規模データを十分に活用するにはモデルは十分に柔軟であることが望ましい．例えば空間相関，時系列相関，非線形効果，グループ効果といった多種多様な効果がありうることから，それらを大規模なデータからそれらを高速に推定するような方法が求められている．

村上大輔

統計数理研究所　データ科学研究系

dmuraka@ism.ac.jp

以上を踏まえ本研究では，大規模標本（例えば数百万標本）から複数の効果を柔軟に推定するような方法を新たに開発する．開発した手法の実装方法を紹介する．

2. 手法

2.1 モデル

本研究では以下のようなモデルを考える：

$$y_i = \sum_{k=1}^{K} f(x_{i,k}) + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2) \qquad (1)$$

y_iはi番目の被説明変数，ε_iは分散σ^2の誤差を表す．$f_k(x_{i,k})$は k 番目の説明変数$x_{i,k}$からの影響を表す関数であり，説明変数毎に設定される．この関数は線形関数（$f_k(x_{i,k}) = x_{i,k}\beta_k$；$\beta_k$は回帰係数）やスプライン関数など様々なものがありうる．また$x_{i,k}$をグループ id で定義した場合はグループ効果（例えば都道府県毎の異質性）を捉えることもできる．上式は加法混合モデル（additive mixed model）として知られており，近年幅広い応用がみられる．

本研究では，(1)式の一例として以下の空間加法混合モデルを考える：

$$y_i = b_{i,1} + \sum_{k=2}^{K-1} x_{i,k}\, b_{i,k} + b_{i\subseteq g_m} + \varepsilon_i \qquad (2)$$

ここで$b_{i,k}$は場所毎に変化する回帰係数（spatially varying coefficient: SVC）であり，以下のモデルを仮定する：

$$b_{i,k} = \beta_k + \beta_{i,k}^{SV} \quad \beta_{i,k}^{SV} \sim N(0, c(d_{i,j}; \boldsymbol{\theta}_k)) \qquad (3)$$

上式は回帰係数の平均値はβ_kであり，各地点における平均値からのズレを空間相関モデル$\beta_{i,k}^{SV} \sim N(0, c(d_{i,j}; \boldsymbol{\theta}_k))$で捉えようというものである．$c(d_{i,j}; \boldsymbol{\theta}_k)$は地点 *i*-*j* 間の距離$d_{i,j}$の減衰関数であり（$\boldsymbol{\theta}_k$は未知パラメータの集合），同関数で SVC の空間相関（距離が近いほど回帰係数の値は類似）を捉える．なお$b_{i,1}$（2 式参照）は誤差項の空間相関を捉えるものであり，以下の構造を仮定する：

$$b_{i,1} = \beta_{i,1}^{SV} \quad \beta_{i,1}^{SV} \sim N(0, c(d_{i,j}; \boldsymbol{\theta}_1)) \qquad (4)$$

最後に$b_{i\subseteq g_m}$はグループ（$g_m | m \in 1 \cdots M$；例えば都道府県の場合 M=47）毎の異質性を捉える項であり期待値 0、分散σ_g^2の正規分布(5)に従うと仮定する：

$$b_{i\subseteq g_m} \sim N(0, \sigma_g^2) \qquad (5)$$

要約すると(2)式は$f_1(1) = b_{i,1}$，$f_k(x_{i,k}) = x_{i,k} b_{i,k}$，$f_K(g_l) = b_{i\subseteq g_l}$とした場合の加法混合モデルである．各項は誤差項の空間相関，各回帰係数の空間相関，グループ毎の異質性を捉える．

なお(2)式は基本的な空間統計モデルを包含する．例えば空間エラーモデルやクリギングは誤差項の空間相関（$b_{i,1}$）のみを考慮して，地理的加重回帰は（定式化の方法は異なるものの）回帰係数の空間相関（$b_{i,k}$）のみを考慮する．

2.2 推定

大規模データから(2)式を推定する方法を考える．ここでは経験ベイズ法による事後確率の最大化を想定する．残念ながら，以下の理由により近似なしでの推定は困難である：

(a) 各 SVC $\boldsymbol{\beta}_k^{SV} = [\beta_{1,k}^{SV}, \cdots, \beta_{N,k}^{SV}]'$の共分散行列を処理する必要があるため（「′」は転置）．具体的には(3)式内の$\beta_{i,k}^{SV} \sim N(0, c(d_{i,j}; \boldsymbol{\theta}_k))$が以下のように行列表記できることに注意する：

$$\boldsymbol{\beta}_k^{SV} \sim N(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_k)) \qquad (6)$$

$\mathbf{C}(\boldsymbol{\theta}_k)$は$c(d_{i,j}; \boldsymbol{\theta}_k)$を第$(i, j)$要素に持つ行列（$N$×$N$）である．その逆行列の計算量は $O(N^3)$であり大標本には適用できない．加えて，N が大きな場合，$\mathbf{C}(\boldsymbol{\theta}_k)$を保持すること自体がメモリ消費の観点で困難である．

(b) 多数のパラメータ$\{\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{K-1}, \sigma_g^2\}$を経験ベイズ推定するために，それらの値を変えながら事後確率を繰り返し求める必要があるため．

問題(a)に対処するために，本研究では$\boldsymbol{\beta}_k^{SV}$を主成分で近似することとする（低ランク近似）．具体的には$\mathbf{C}(\boldsymbol{\theta}_k) = \tau_k^2 \mathbf{C}^{\alpha_k}$という構造を仮定した上で$\mathbf{C}$を固有値分解する．ここで$\mathbf{C}$は既知の距離減衰関数$c(d_{i,j})$を要素に持つ行列，$\tau_k^2$は空間過程の分散の大きさを，$\alpha_k$は空間過程の空間スケール（$\alpha_k$が大きいほど大域的な空間パターン）を表すパラメータである．結果として得られる固有ベクトルは，$\mathbf{C}$で説明される空間相関成分を表し，固有値は各成分の強さを表す（Griffith, 2003）．ここでは，主成分分析と同様，固有値の大きな L（$\ll N$）個の固有ベクトル$\mathbf{E} = [\mathbf{e}_1 \cdots \mathbf{e}_L]$（つまり主成分）で$\mathbf{C}$を近似する．その際，高速化のために固有値と固有ベクトルは Nystom 法で近似する．結果として$\mathbf{C}$は$\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$で近似され(3)式は下式となる：

$$\boldsymbol{\beta}_k^{SV} \sim N(\mathbf{0}, \tau_k^2 \mathbf{E}\boldsymbol{\Lambda}^{\alpha_k}\mathbf{E}') \qquad (7)$$

$\boldsymbol{\Lambda}$は L 個の固有値$\{\lambda_1 \cdots \lambda_L\}$を要素に持つ対角行列（$L$×$L$）である．上式は回帰の形に書き直せる：

$$\boldsymbol{\beta}_k^{SV} = \mathbf{E}\boldsymbol{\gamma}_k \qquad \boldsymbol{\gamma}_k \sim N(\mathbf{0}, \tau_k^2 \boldsymbol{\Lambda}^{\alpha_k}) \qquad (8)$$

(5)式を用いた場合，$\boldsymbol{\beta}_k^{SV}$を求めるためには係数ベクトル$\boldsymbol{\gamma}_k$さえ求めればよく，$\mathbf{C}^{-1}$の計算は不要である．従って問題(a)は解消される．

次に問題(b)を考える．残念ながら，(a)を解消したとしても$\mathbf{E}$は N×L の行列であるため，例えば N=1, 000 万の場合は$\mathbf{E}$が巨大となりメモリを著し

く消費する．当然，パラメータ推定のための$\mathbf{E}$の反復処理もまた計算量の観点で非現実的である．そこで今回は，$\mathbf{E}$をH個のブロック毎に抽出・処理することでメモリ消費を$(N/H)\times L$に削減する．それにより巨大行列$\mathbf{E}$を明示的に保持することなく内積$\mathbf{E}'\mathbf{E}$（$L\times L$）が評価できる（Nystrom 近似で$\mathbf{E}$を近似したため$\mathbf{E}'\mathbf{E} \neq \mathbf{I}$である）．H を大きくすればするほど消費メモリは小さくなる．従って，.

説明変数やその他効果（今回の場合グループ効果）に関する内積計算も必要となるが，同様にブロック毎に処理可能である．以上によって得られた内積の集合を $\mathbf{M}$ とおくち， $\mathbf{M}$ さえあれば，サイズが N に依存する行列（説明変数行列や $\mathbf{E}$）を一切用いることなく事後確率が評価可能となる（Murakami and Griffith, 2019a）．結果として標本数に依存しない計算量でのパラメータ$\{\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{K-1}, \sigma_g^2\}$の経験ベイズ推定が可能となる．

最後に推定されたパラメータから$\boldsymbol{\beta}_k^{SV}$を推定する．その際にも$\mathbf{E}$が必要となるため，再びブロック毎の並列計算が必要となる．

以上の方法のイメージは図 1 に示した通りである．この方法を用いると，例えば N = 10,000,000, K = 7（全て SVC）の場合に Mac Pro (3.5 GHz, 12-コア; 64 GB メモリ)を用いた場合の計算時間は 4,224 秒となり，大規模データにも応用可能であることを確認した．なお同様の方法はより一般の加法混合モデル(1)式にも応用できる．詳しくは Murakami and Griffith (2019a,b)を参照されたい．

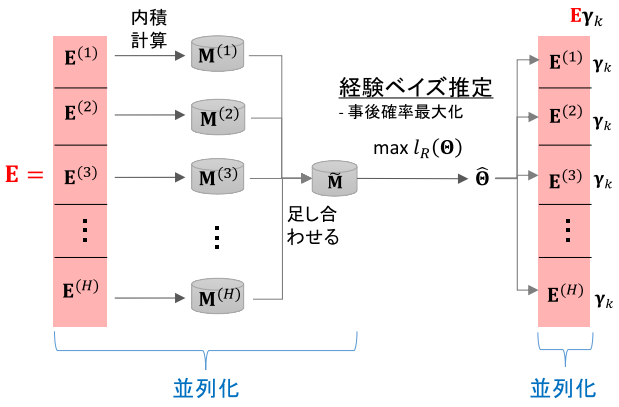

図-1 推定手順。赤はサイズが N に依存する行列

3. 所得分析への応用

3.1 概要

空間加法混合モデル(2)式をアメリカ合衆国の街区(tract)レベルの世帯所得(2015 年の中央値; 1,000 USD)に適用する（図 2）．ここでは学位取得者の割合（大卒率），英語が主要言語である居住者の割合（英語率），および平均年齢（5 歳階級別人口から推定）が世帯所得に及ぼす影響を，各説明変数に対する SVC（$\mathbf{b}_{大卒}$，$\mathbf{b}_{英語}$，$\mathbf{b}_{年齢}$）で，誤差項の空間相関を(4)式で，州毎の異質性をグループ効果(5)式でそれぞれ推定する．標本数は 72,160 である．

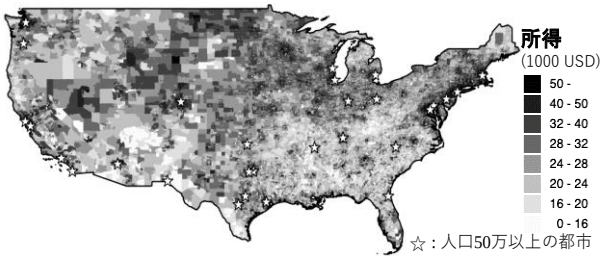

図-2 街区別の世帯所得(2015; N=72, 160)

3.2 結果

並列化なしでモデル推定を行なった結果 103. 7 秒を要した（固有値分解 3. 62 秒 + パラメータ推定 100. 08 秒）．

推定された各 SVC を図 3 にプロットした．$\mathbf{b}_{大卒}$の結果を見ると，サンフランシスコ，ロサンゼルス，ワシントン DC，ニューヨーク等の主要都市で特に大きな正の値を示しており，その他人口 50 万人以上の都市の周辺でも正の値が大きくなる傾向が確認できる．学位取得者は主要都市付近で高い所得を得やすいとの示唆を得た．$\mathbf{b}_{英語}$の推定結果に基づけば，北部では英語が喋れることによる所得上昇が示唆されるが，南部ではそのような効果は見られない．これは比較的所得水準の低い南部では英語の使用可否によらず低所得となるための可能性ある．$\mathbf{b}_{年齢}$の推定結果からは，北東部では年齢が所得増加に特に強く寄与することが確認された．ただし，主要都市周辺ではその効果

が弱まる傾向があり，年齢の効果は非都市部で強くなるの結果が得られた．

州毎のグループ効果の推定結果を図5に示したものの統計的に有意な効果が得られた州は存在しなかった．州毎ではなく，空間的に滑らかに所得構造が変化しているとの示唆を得た．

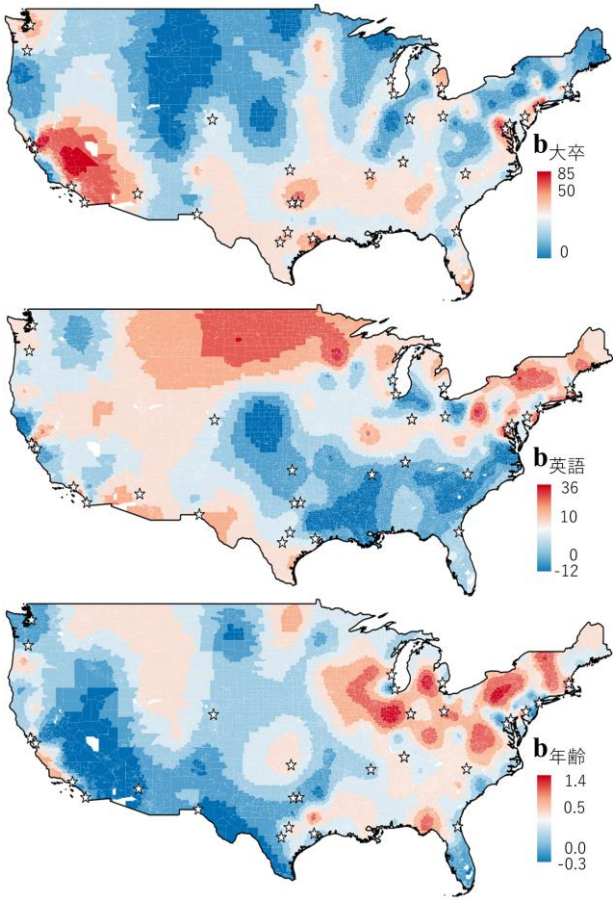

図-3 場所毎の効果（SVC）の推定結果

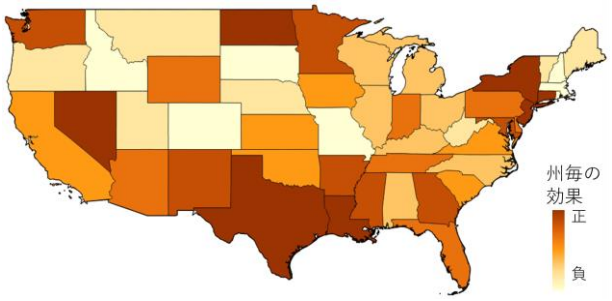

図-4 グループ効果の推定結果

4. 実装方法

今回の方法は統計ソフト R のパッケージ spmoran の resf_vc 関数に実装した．spdep パッケージ内のサンプルデータを対象とする場合，今回のモデル実装の R コードは以下の通りである．

```
Install.packages(“spmoran”)#パッケージのインストール
Install.packages(“spdep”)  #spddep の boston データの場合
library(spmoran); library(spdep)
data(boston)              #データの読み込み
y       <- boston.c[, “CMEDV”]
x       <- boston.c[,c(“ZN”, “LSTAT”)]
xconst <- boston.c[,c(“CRIM”, “NOX”, “AGE”, “DIS”,
                  ”RAD”, ”TAX”, ”PTRATIO”, ”B”, ”RM”)]
xgroup <- boston.c[,”TOWN”]
coords <- boston.c[,c(“LAT”,”LON”)]

meig   <- meigen_f(coords=coords)          #固有値分解
res    <- resf_vc(y=y, x=x, xconst=xconst, #モデル推定
                  xgroup=xgroup, meig=meig)
res$e                    #誤差統計量(擬似 R2, BIC 等)
res$b_vc[ 1:10, ]        #推定された SVC (最初の 10 行)
res$p_vc[ 1:10, ]        #SVC の有意性 (最初の 10 行)
res$b_g                  #推定されたグループ効果
```

各変数の定義は以下の通りである：

y	：被説明変数
x	：回帰係数を場所毎に変える説明変数
xconst	：（3 章では用いなかったが）回帰係数を一定（通常の回帰と同様）とする説明変数
xgroup	：グループの id または名称
coords	：緯度経度

なお標本数が数十万以上の場合は besf_vc 関数で並列計算可能である．詳しくはマニュアル（Murakami, 2017）を参照されたい。

謝辞

本研究は科研費（18H03628，17H02046，17K12974）の助成を受けたものである．

参考文献

Murakami, D. 2017. spmoran: An R package for Moran's eigenvector-based spatial regression analysis. *ArXiv*, 1703.04467.

Murakami, D. and Griffith, D. A. 2019a. Spatially varying coefficient modeling for large datasets: Eliminating N from spatial regressions. *Spatial Statistics*, **30**, 39-64.

Murakami, D., and Griffith, D. A. 2019b. A memory-free spatial additive mixed modeling for big spatial data. *ArXiv*, 1907.11369.