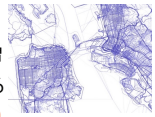
# 大規模データのための時空間回帰モデリング

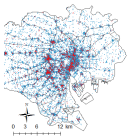
村上　大輔　　データ科学研究系　助教

## 【地理空間データのオープン化】



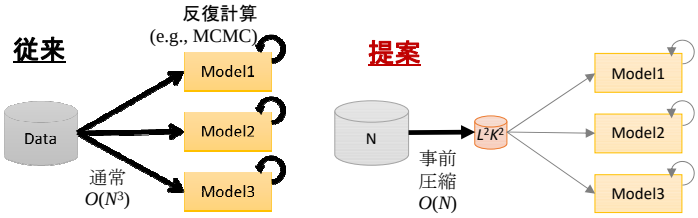
- Google Earth EngineやOpenSteetMapといったオープンデータが急増
- 大規模な地理空間データを効率よく解析する方法が求められている

**大規模な地理空間データの回帰・予測等を計算効率よく行うための実用手法の開発**



## 【方針：大規模データモデリングの冗長性排除】

- 従来：モデル推定毎に大規模データの反復処理が必要→重い。。
  → 大規模データを圧縮してからモデルを推定するようにすればよいのでは？




## 【提案手法】

- 線形混合効果モデルを仮定

$$\mathbf{y}=\sum_{k=1}^{K}\mathbf{x}_k{}^{\circ}\boldsymbol{\beta}_k+\mathbf{E}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}\qquad \boldsymbol{\gamma}\sim N(\mathbf{0},\tau^2\boldsymbol{\Lambda}^{\alpha})\qquad \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$$

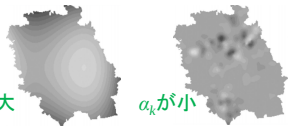
ガウス過程の低ランク近似　E: 空間近接行列の$L$個の固有ベクトル(の近似)　$\boldsymbol{\Lambda}$:$L$個の固有値からなる対角行列

**空間可変パラメータ（今回はこれ）**

場所毎に回帰係数を変える（場所以外も可）

$$\boldsymbol{\beta}_k=b_k\mathbf{1}+\mathbf{E}\boldsymbol{\gamma}_k\qquad \boldsymbol{\gamma}_k\sim N(\mathbf{0},\tau_k^2\boldsymbol{\Lambda}^{\alpha_k})$$

$\mathbf{E}\gamma_k$の空間分布



$\alpha_k$が大　　$\alpha_k$が小

**グループ効果**

小地域推定等にも使えそう

$$\boldsymbol{\beta}_k=b_k\mathbf{1}+\mathbf{G}\mathbf{v}_k$$
$$\mathbf{v}_k\sim N(\mathbf{0},\tau_k^2\mathbf{I})$$

**自己回帰、スプライン回帰**
等にも応用可能

- **$\boldsymbol{\Theta}\in\{\tau_1^2,\ldots\tau_K^2,\alpha_1,\ldots\alpha_K\}$の推定手順**

(1)事前圧縮→サイズが$N$の行列を消す（$K$個）

- $\mathbf{M}_{XX}=\mathbf{X}'\mathbf{X}$, $\mathbf{M}_{XE}=\mathbf{X}'[\mathbf{E},\ldots\mathbf{E}]$, $\mathbf{m}_{Xy}=\mathbf{X}'\mathbf{y}$, $\mathbf{m}_{Ey}=[\mathbf{E},\ldots\mathbf{E}]'\mathbf{y}$, $m_{yy}=\mathbf{y}'\mathbf{y}$

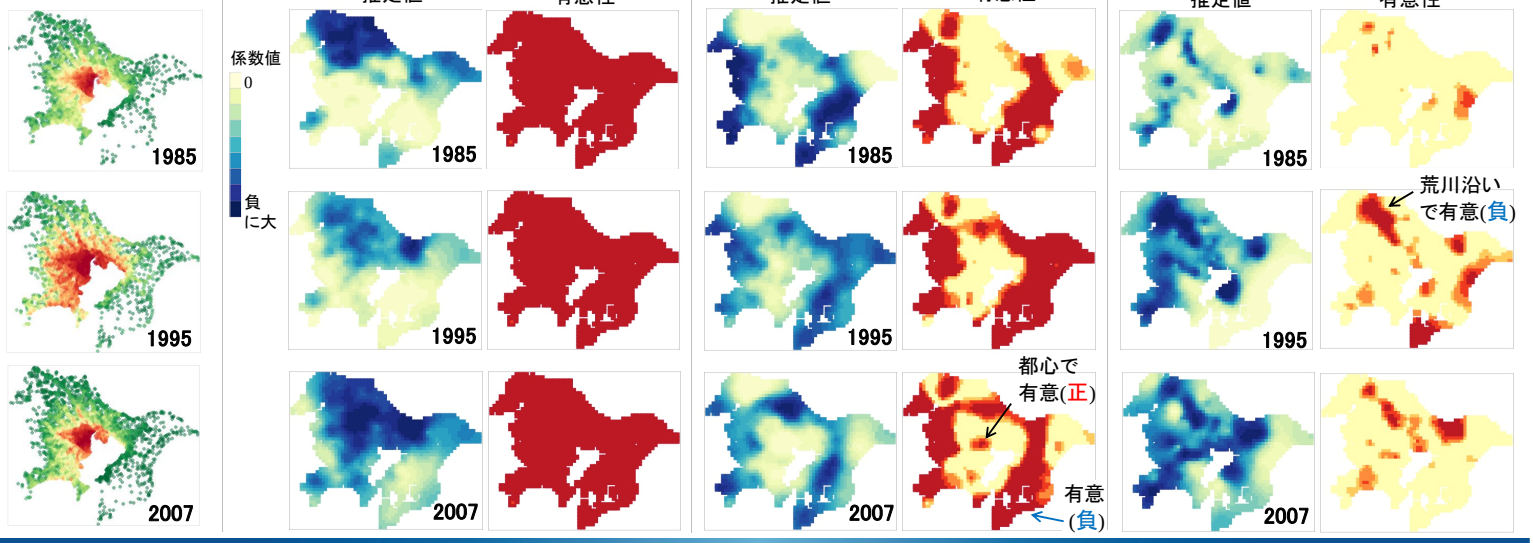
(2) 事後確率$p(\boldsymbol{\Theta}|\mathbf{y})$を以下のように書き換える→推定の計算量が$N$に依存しない

$$p(\boldsymbol{\Theta}|\mathbf{y})=-\frac{1}{2}ln\left|\begin{bmatrix}\mathbf{M}_{XX} & \mathbf{M}_{XE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})\\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}'_{XE} & \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}_{EE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})+\mathbf{I}\end{bmatrix}\right|-\frac{N-K}{2}\left(1+ln\left(\frac{2\pi d(\boldsymbol{\Theta})}{N-K}\right)\right)$$

(3) $p(\boldsymbol{\Theta}|\mathbf{y})$を$k$毎に逐次最大化 (省略)→$K$が大きくても高速推定可

## 【実データへの適用】

- 住宅地公示地価の回帰に適用した後、係数を補間




## 【シミュレーション実験】

- $N\in\{6{,}000, 9{,}000, 12{,}000\}$、空間可変パラメータ数$K\in\{2,4,6,8\}$ を想定
- 緯度経度は標準正規分布からそれぞれ生成
- 以下から生成されるデータへのあてはめを200回繰り返す

※$K$=8の場合

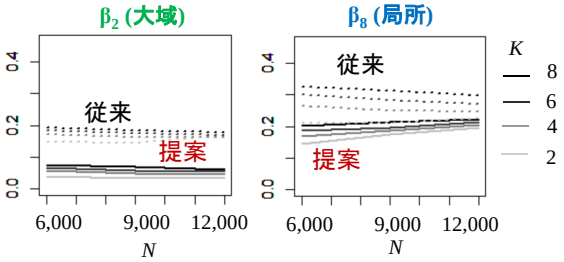
$$\mathbf{y}=\sum_{k=1}^{4}\mathbf{x}_k{}^{\circ}\boldsymbol{\beta}_k+\sum_{k=5}^{8}\mathbf{x}_k{}^{\circ}\boldsymbol{\beta}_k+\boldsymbol{\varepsilon}\qquad \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$$

大域的 ($\alpha$=3)　　局所的 ($\alpha$=0.5)

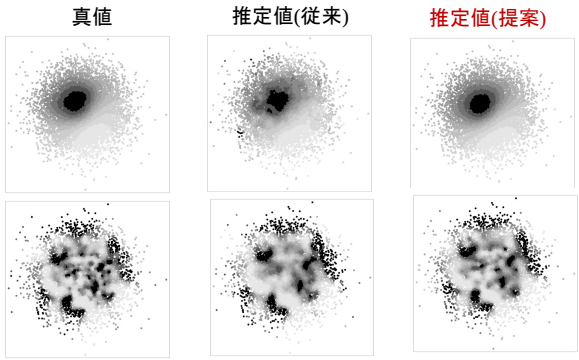
$R^2$=0.7になるよう設定

**推定精度**

- 大域変動の推定精度が特に良い

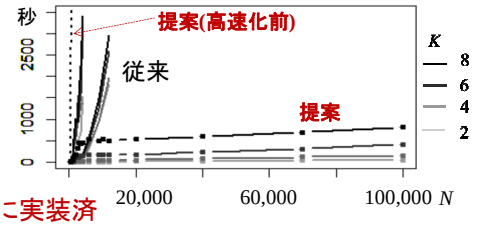



-結果の例




**計算時間の比較**

- 計算時間を大幅に短縮
- 事前圧縮の計算量のみ$N$に応じて増加




→Rパッケージspmoranに実装済

$$d(\boldsymbol{\Theta})=m_{yy}-2[\hat{\mathbf{b}}',\hat{\mathbf{u}}']\begin{bmatrix}\mathbf{m}_{Xy}\\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{m}_{Ey}\end{bmatrix}+[\hat{\mathbf{b}}',\hat{\mathbf{u}}']\begin{bmatrix}\mathbf{M}_{XX} & \mathbf{M}_{XE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})\\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}'_{XE} & \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}_{EE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}+\|\hat{\mathbf{u}}\|^2$$

誤差分散　　回帰係数の分散

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{M}_{XX} & \mathbf{M}_{XE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})\\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}'_{XE} & \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{M}_{EE}\tilde{\mathbf{V}}(\boldsymbol{\Theta})+\mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{m}_{Xy}\\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\mathbf{m}_{Ey}\end{bmatrix}\qquad \tilde{\mathbf{V}}(\boldsymbol{\Theta})=\frac{1}{\sigma}\begin{bmatrix}\tau_1\boldsymbol{\Lambda}^{\alpha_1} & & \\ & \ddots & \\ & & \tau_K\boldsymbol{\Lambda}^{\alpha_K}\end{bmatrix}\qquad \tilde{\mathbf{u}}=\begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_K\end{bmatrix}$$

$\mathbf{u}_k\sim N(0,\sigma^2\mathbf{I})$