空間加法モデルの高速なモデル選択法: 犯罪分析への応用

村上　大輔　　データ科学研究系　助教

【背景】

- **Additive mixed model (AMM)**は多様な効果(空間効果, 時間効果,…)を捉える回帰である
- **モデル選択**はAMMを安定的に推定する上で重要だが計算量が大きい(大標本の場合)

→ **そこで本研究では、以下の2つの研究に取り組む**

(1) AMMのための、高速なモデル選択アルゴリズムの開発

(2) 犯罪分析への応用

【Compositionally-warped additive mixed model: CAMM】

- 本研究では以下のモデルで町丁目s, 時期tの犯罪密度$y_{s,t}$をモデル化する。犯罪密度はガウス分布に従うとは限らため、合成変換関数$\varphi_{\boldsymbol{\theta}}(\cdot)$でできる限りガウス分布に近づける

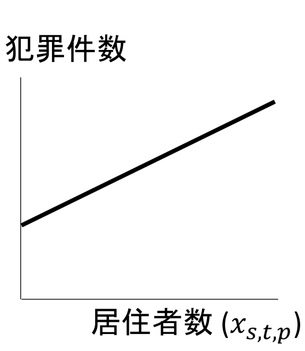
$$\varphi_{\boldsymbol{\theta}}(y_{s,t}) = \sum_{p=1}^{P} x_{s,t,p}\beta_{s,t,p} + \sum_{q=1}^{Q} b_{s,t,q} + \varepsilon_{s,t}, \qquad \varepsilon_{s,t} \sim N(0, \sigma^2),$$

$\beta_{s,t,p}$: p番目の共変量$x_{s,t,p}$に対する**Varying coefficient (VC)**
$b_{s,t,q}$: q番目の**Random effects** (varying intercept)

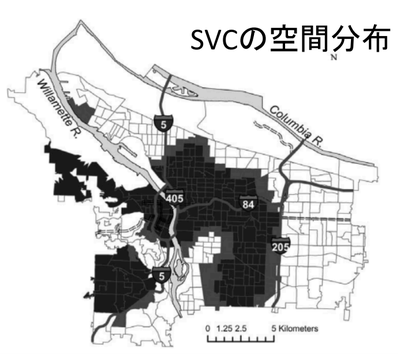
- Varying coefficient ($\beta_{s,t,p}$)は以下の4種類から選択:

Notation	Description
Constant	Constant coefficient
SVC (Spatially VC)	Coefficients varying over geographical space
NVC (Non-spatially VC)	Coefficients varying depending on $x_{s,t,p}$
SNVC	SVC + NVC

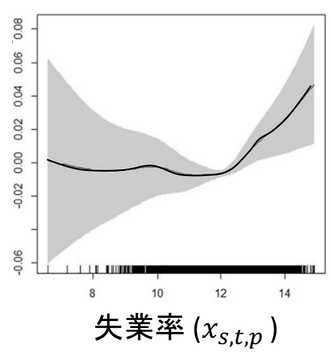
Constant
例: 居住者が千人増えると犯罪は10件増



SVC
例: 人種多様性は市街地で犯罪を増加させる



NVC
例: 失業率が一定以上の場合に犯罪増



- 各Random effects ($b_{s,t,q}$)は、With/withoutの2つの中から選択

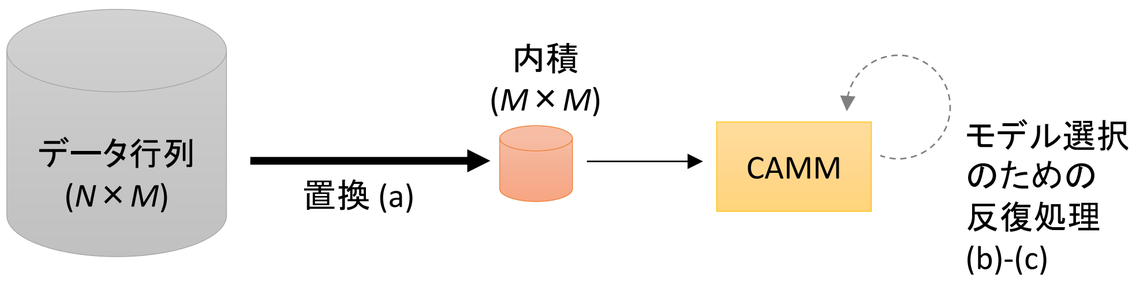
【モデル選択】

✓ Varying coefficientとRandom effectsの取り方に応じて4^P2^Q の候補モデルがある
- 例えば後述の犯罪分析(P=8, Q=1)の場合、131,072の候補モデル。高速化が必須

→ そこで、以下の高速なモデル選択法を提案する

(a) Replace all the data matrix ($N \times M$) with the inner products ($M \times M$)
(b) Select specification of varying coefficients ($\beta_{s,t,p}$) for $p \in \{1, \ldots, P\}$:
  (b-1) Estimate p-th SVC
  (b-2) Select the SVC if it improves Bayesian information criterion (BIC).
  (b-3) Estimate p-th NVC.
  (b-4) Select the NVC if it improves the BIC.
(c) Select specification of random effects ($b_{s,t,q}$) in a similar way.
(d) End if BIC converges. Otherwise, go back to (b).

事前にデータ行列を内積行列に置換
→ 大規模データに対しても高速なモデル選択が可能



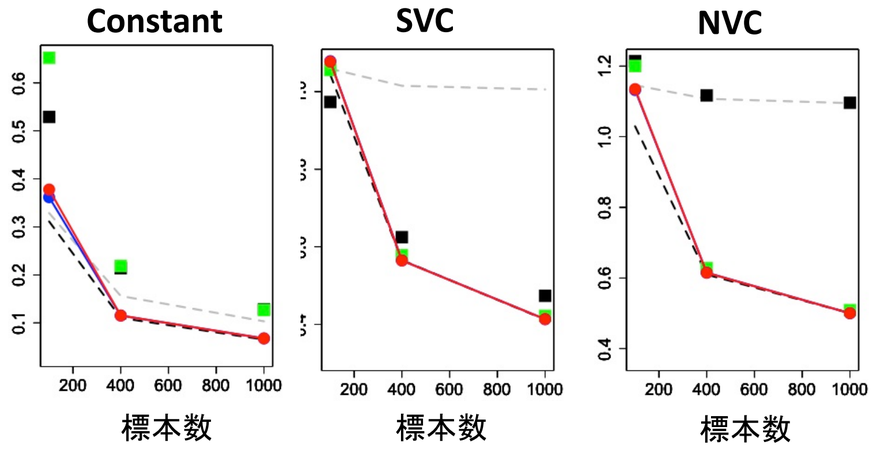
【モンテカルロ実験】

- Const3つ, SVC3つ, NVC3つを仮定した説明変数9つの回帰モデルから生成したデータに対する推定精度を検証することで、モデル選択の有効性を検証

結果:精度

- 提案したモデル選択法による推定精度の改善を確認

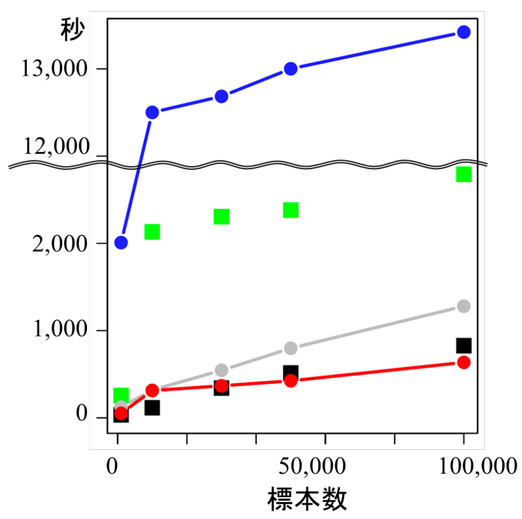
回帰係数のRMSE



結果:計算効率

- 既存の高速なモデル選択法(Rパッケージmgcv内の手法)よりも高速であることを確認

計算時間(秒)



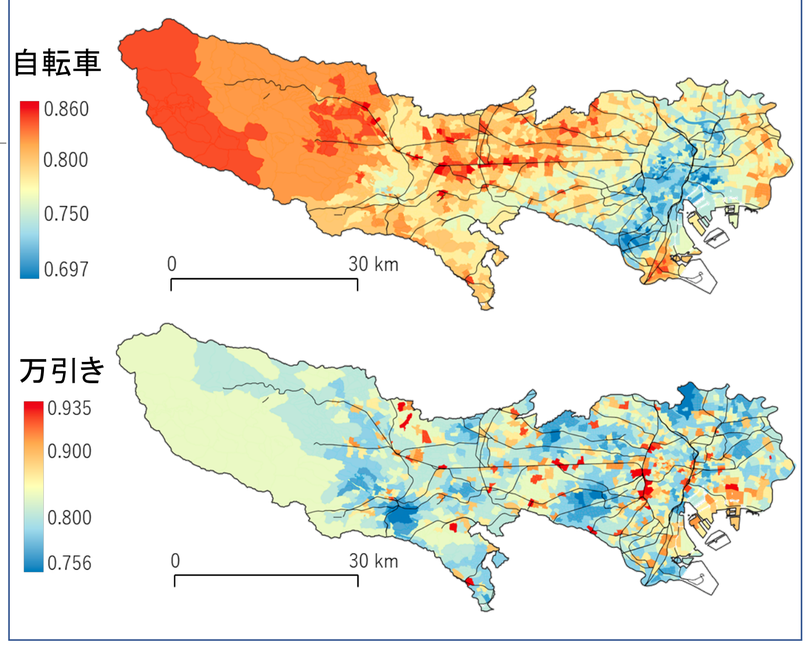
【犯罪の地理的要因分析への応用】

- 提案手法(CAMM + モデル選択)を自転車盗と万引きの密度 (件数/km2;町丁目別; 四半期別)の要因分析に応用(標本数: 12,232)

選択された効果
- 赤:正の効果, 青: 負の効果

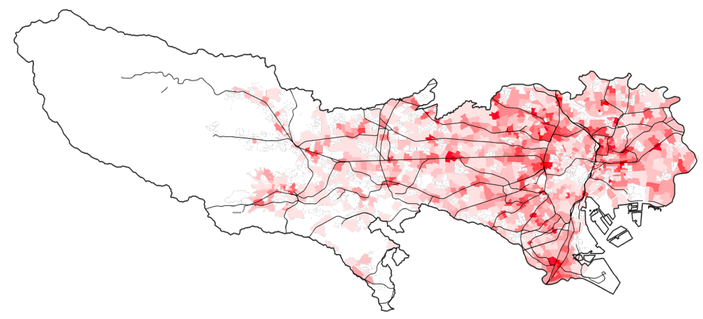
	自転車	万引き
空間相関	SVC	SVC
時間相関	SNVC	SNVC
他罪種との共起	NVC	NVC
夜間人口	NVC	Const.
昼間人口	NVC	NVC
外国人比	Const.	Const.
失業率	Const.	Const.
大卒率	Const.	Const.

推定された時間相関の強さ
-赤 = 強い時間相関 = 犯罪の反復傾向あり

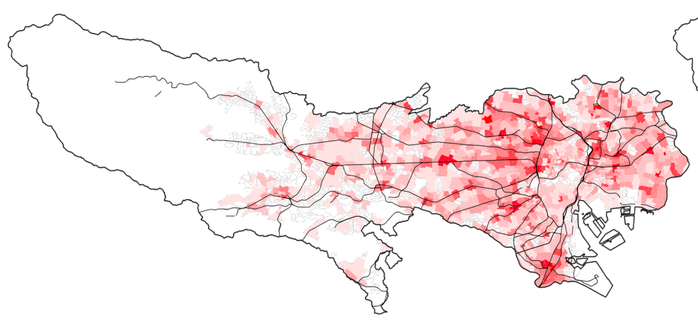


犯罪予測への応用

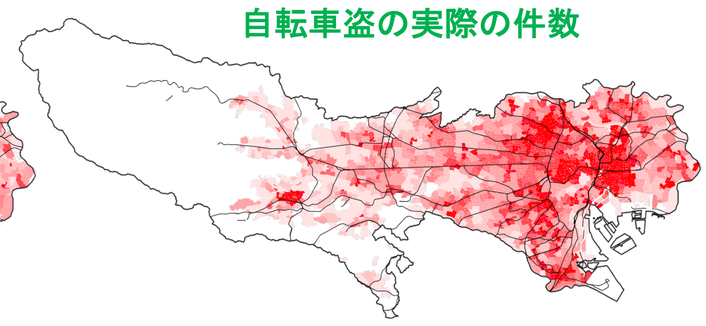
- 2019年第一四半期の自転車盗難件数を2017-2018年データから予測
- 提案手法の精度がカーネル密度推定法を上回ることを確認



自転車盗の実際の件数



提案法



カーネル密度推定法