空間加法混合モデルを用いた犯罪の地理的要因の識別･選択

村上大輔*・梶田真実**

On Model Selection in Spatial Mixed Effects Modeling: Application to Crime Analysis

Daisuke Murakami *, Mami Kajita**

A rapid growth in spatial open datasets has led to a huge demand for regression approaches accommodating spatial and non-spatial effects in big data. Regression model selection is particularly important to stably estimate flexible regression models. However, conventional methods can be slow for large samples. Hence, we develop a fast and practical model-selection approach for spatial regression models, focusing on the selection of coefficient types that include constant, spatially varying, and non-spatially varying coefficients. Numerical experiments show that our approach selects the true model accurately and computationally efficiently, highlighting the importance of model selection in the spatial regression context. Then, the present approach is applied to open data to investigate local factors affecting crime in Tokyo, Japan. The results suggest that our approach is useful not only for extracting effective crime factors but also for predicting crime events. The developed model selection approach was implemented in the R package spmoran.

Keywords:犯罪分析（crime analysis），モデル選択（model selection），加法混合モデル（additive mixed model）, spmoran（spmoran）

１．はじめに

回帰は地理空間データの要因解析に幅広く用いられている．近年，幅広い効果を推定・識別することのできる回帰モデルの拡張として加法混合モデルが注目されている．例えば同モデルを犯罪分析に応用することで，場所毎・時間毎の犯罪リスクの違いや犯罪の繰り返されやすさの場所毎・時間毎の違いといった各種の犯罪発生要因を明らかとすることができる．

加法混合モデリングにおいて効果を精度良く推定するためにはモデル選択が必要となる．即ち，犯罪発生と無関係な要因は取り除き，犯罪発生に寄与する要因のみを選択することでモデルを特定化することが，犯罪要因を精度良くモデリングするうえで重要となる．残念ながら候補となるモデルの数は膨大となる場合が多い．例えば説明変数が 8 個あり，各説明変数からの効果（回帰係数のタイプ）の与え方が 4 種類存在する場合, この効果の与え方に応じて, 65,536（＝4^8）個の候補モデルが存在することとなる．大量の候補モデルの中から最も精度の良い加法混合モデルを選択するための計算効率の良いモデル選択手法が必要である．

以上を踏まえ，本研究では，最初に空間データのための加法混合モデル(以後, 空間加法混合モデル)のための高速なモデル選択法を開発する．次に同手法を，東京都を対象とした犯罪の地理的要因分析に応用する．本研究の開発手法はフリーの統計ソフトウェア R のパッケージ spmoran に実装した（５章参照）．

２．空間加法混合モデル

本研究では場所毎の効果の推定に特に着目した，以下の空間加法混合モデルを考える：

$$y_i = \sum_{p=1}^{P} x_{i,p}\beta_{i,p} + \sum_{q=1}^{Q} b_{i,q} + \varepsilon_i, \qquad (1)$$

y_iは i 番目の被説明変数（標本数：N），$x_{i,p}$は p 番目の説明変数，$\varepsilon_i \sim N(0, \sigma^2)$は誤差項である．

* 正会員　株式会社 Singular Perturbations（Singular Perturbations Inc.）
〒102-0074　東京都千代田区九段南 1-5-6 りそな九段ビル 5F KS フロア E-mail：dmuraka@ism.ac.jp
** 正会員　株式会社 Singular Perturbations（Singular Perturbations Inc.）

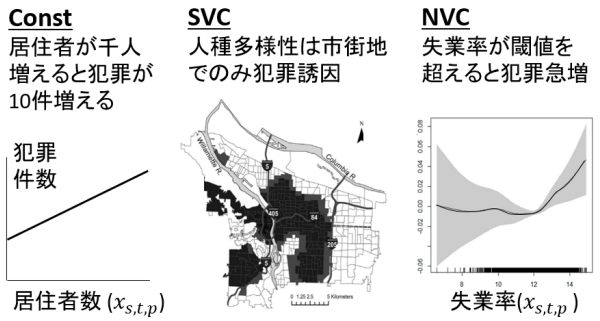

図１：各回帰係数のイメージ（SVC の図の出典：Cahill and Mulligan, 2007）

$\beta_{i,p}$は$x_{i,p}$からの影響を捉えるための回帰係数であり，以下のいずれかで与える（図１参照）：

(i-1)Const
- 一定値（通常の回帰モデルの回帰係数）

(i-2)Spatially varying coefficients (SVC)
- 回帰係数は場所毎に可変．地理的加重回帰と同様；例：東京駅に近いほど係数は大

(i-3)Non-spatially varying coefficients (NVC)
- 回帰係数は説明変数の値毎に可変（例：説明変数が大きいほど回帰係数は大）

(i-4)SNVC（SVC＋NVC）.
- 回帰係数は場所毎にも説明変数の値毎にも可変.

ここで SNVC は下式で定義される：

$$\beta_{i,p} = \beta_p + \beta_p(lon_i, lat_i) + \beta_p(x_{i,p}), \qquad (2)$$

$\{lon_i, lat_i\}$は緯度経度である．β_pは Const，$\beta_p(lon_i, lat_i)$は SVC，$\beta_p(x_{i,p})$は NVC である．変数選択に際しては，それら３要素をモデルに入れるか否かを検証することでモデル選択が行える．

一方，$b_{i,q}$は地域毎・時点毎の違いを捉える Random effects 項である．本研究では，同項は以下のいずれかで与えることとする：

(ii-1) Const：一定値（通常の回帰モデルの定数項）

(ii-2) Random effects
- グループ毎の定数項．年次毎や市区町村毎の犯罪リスクの違いを捉える

同モデルは制限付き最尤法で推定する．詳しくは Murakami and Griffith (2019)を参照されたい.

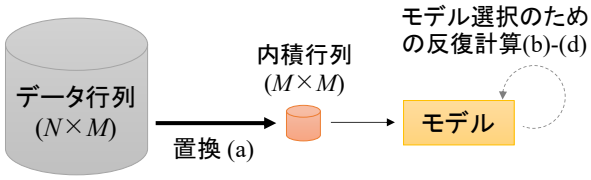

図２：提案するモデル選択手法のイメージ

３．モデル選択手法の開発

加法混合モデル(1)式は回帰係数$\beta_{i,p}$と定数項b_pのタイプの選択によって 4^P2^Q 通りモデルのとり方がある．従って，計算効率よくモデルを選択する手法がその実用に際して必要となる．

そこで本研究では，最良のモデルを高速に選択する以下のモデル選択アルゴリズムを提案する：

(a) 標本数 N にサイズが依存する行列（説明変数や被説明変数を含む $N \times M$ の行列）を内積計算によって $M \times M$ の行列に置き換える

(b) 以下を全ての$p \in \{1, \ldots, P\}$について順番に行うことで，各回帰係数$\beta_{i,p}$のタイプを選択する：
- (b-1) $\beta_p(lon_i, lat_i)$（SVC）ありのモデルを推定する
- (b-2) Bayesian information criterion（BIC; モデルの精度指標）を改善する場合は推定された SVC を採択. さもなくば$\beta_p(lon_i, lat_i) = 0$とする
- (b-3) $\beta_p(x_{i,p})$（NVC）ありのモデルを推定する
- (b-4) BIC を改善する場合は推定された NVC を採択．さもなくば$\beta_p(x_{i,p}) = 0$とする

(c) 同様の手順で定数項($b_{i,q}$)のタイプを選択

(d) BIC が収束した場合は，その時点のモデルが選択されたモデル．さもなくば手順(b)に戻る

手順(b)-(d)の反復計算より前の手順(a)で，標本数にサイズが依存する行列を小さな行列に置き換える．そのため，同手法のモデル選択ステップ（手順 b-d）の計算量ならびにメモリ消費量は標本数によらない．従って大規模標本に対しても高速にモデル選択が行えると考えられる（図２参照）.

４．モデル選択手法の性能検証

4.1. 回帰係数の推定精度の比較

本章では，定数項ならびに 9 個の回帰係数からなるモデルを用いたシミュレーション実験を行う．ここでは，最初 3 つの回帰係数が Const（b_p），次の 3 つが SVC（$\beta_q(lon_i, lat_i)$），残りの 3 つが NVC（$\beta_r(x_{i,r})$）の(3)式から生成される擬似データを用いて提案手法の推定精度を評価する：

$$y_i = b_0 + \sum_{p=1}^{3} x_{i,p} b_p + \sum_{q=1}^{3} x_{i,q} \beta_q(lon_i, lat_i) + \sum_{r=1}^{3} x_{i,r}\, \beta_r(x_{i,r}) + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \tag{3}$$

各説明変数$N(0,1)$は正規分布から生成する．$\beta_q(lon_i, lat_i)$の真値は空間移動平均過程から，$\beta_r(x_{i,r})$はスプライン基底を用いた確率過程から，それぞれ生成する．

(3)式から生成されたデータに対するモデルのあてはめを 200 回繰り返すことで，モデルの回帰係数の推定精度を比較する．比較対象とするモデルは次の通りである：

(I) 線形回帰モデル
- 全係数が Const（b_p）と仮定

(II) 正解モデル
- 各係数のタイプを(3)式に従って与える

(III) SVC モデル
- 全係数が SVC（$b_p + \beta_q(lon_i, lat_i)$）と仮定（地理的加重回帰はこのモデルと同種）

(IV) SNVC モデル
- 全係数が SNVC（2 式）と仮定

(V) 提案モデル
- 回帰係数のタイプを(i-1)〜(i-4)から選択

(VI) 提案モデル(ver2)
- 提案モデルと同様だが，局所解に陥るのを避けるためにモデル選択手順(b)における変数選択の順番$p \in \{1, \ldots, P\}$を無作為に変えながら 30 回繰り返し，そこから得られた結果のうち BIC が最良のものを用いる

なお正解モデルが最良となるはずだが，真のデータ生成過程(3)式は通常は未知であり推定できない．ここでの目的は，提案モデルがどれだけ精度良く正解モデルを近似できるかを検証することである．

回帰係数$\beta_{i,p}$（Const，SVC，NVC のそれぞれ）の推定精度を Root Mean Squared Error（RMSE）で評価する：

$$RMSE[\hat{\beta}_{i,p}] = \sqrt{\frac{1}{200}\sum_{s=1}^{200}(\hat{\beta}_{i,p}^{(s)} - \beta_{i,p}^{(s)})^2}, \tag{4}$$

$\hat{\beta}_{i,p}^{(s)}$は推定値，$\beta_{i,p}$は確率過程から生成した真値である．$RMSE[\hat{\beta}_{i,p}]$が小さいことは推定誤差が小さいことを意味する．

有意性検定に用いられる回帰係数の標準誤差$SE(\hat{\beta}_{i,p})$のバイアスを下式で評価して比較する：

$$Bias[SE(\hat{\beta}_{i,p})] = \frac{1}{N}\sum_{s=1}^{N}\{SE(\hat{\beta}_{i,p}) - SE(\beta_{i,p})\}, \tag{5}$$

$Bias[SE(\hat{\beta}_{i,p})]$が正であることは統計的有意性の過小評価を招き，負であることは過大評価を招く．

図３に推定結果を要約した．RMSE の結果から，線形回帰モデルの精度が低いこと，空間統計分野で広く用いられる SVC モデルもまた，真の回帰係数が NVC の場合に著しく精度が低下することが確認できる．対照的に，提案モデル（SNVC モデル＋変数選択）は全体として精度が良く，その精度はモデル選択なしの SNVC モデルを上回る傾向が確認できる．この傾向は特に，Const が真値である場合，ならびに標本数が少ない場合に顕著である．以上より，回帰係数の推定精度の観点からの提案手法の有効性が確認された．

次に標準誤差のバイアスの評価結果を図３にまとめた．この結果から，提案手法のバイアスはモデル選択なしの他手法に比べて著しく小さく，提案した変数選択法が統計的な有意性を適切に評価する上で極めて重要であるとの示唆を得た．

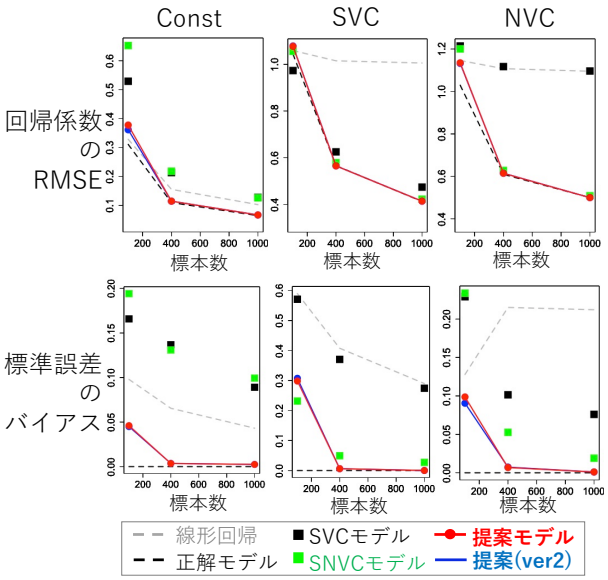

図３：回帰係数とその推定精度の評価経過

4.2. 計算時間の比較

次に，提案手法の計算効率を評価するために，(3)式を再びデータ生成過程とした上で，モデル推定と選択に要する計算時間を評価した．ここでは加法混合モデルのための高速なモデル選択法が実装されている R パッケージ mgcv の bam 関数との比較を行った．同関数はモデル推定を fast REML (Wood, 2017), モデル選択を double-penalty 法(Marra and Wood, 2010)で行うものである．

計算時間の比較結果を図４に整理した．この図より，当然ではあるが提案モデルの推定を 30 回繰り返す提案モデル(ver2)は極めて遅いことが確認でき，実用的でないことを確認した．一方，提案モデルの結果からは，(a)mgcv を上回る計算効率であること，(b)モデル選択を行ったほうがかえって計算時間が短くなることが確認された．(b)は，モデル選択で SVC と NVC が 0 に置き換えられるたびに，手順(b)の反復計算での処理される行列のサイズ（M）が小さくなるためと考えられる（図２参照）．

以上より，提案するモデル選択法の計算効率の良さが確認できた．なお，提案手法の回帰係数の推定精度が SVC と NVC に対しては mgcv と同等，Const に対しては mgcv を上回ることもまた確認している（Murakami et al., 2020）．

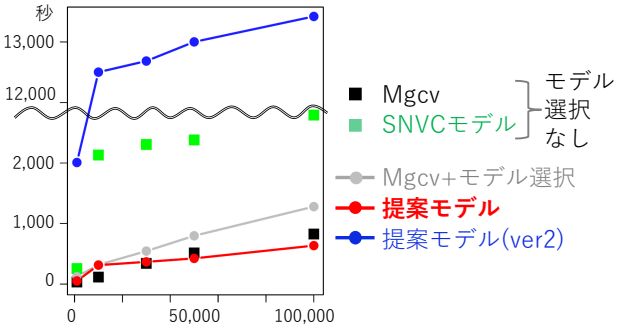

図４：計算時間の比較結果

３．犯罪の要因分析への応用

3.1. 概要

東京都提供のウェブサイトである大東京防犯ネットワーク（https://www.bouhan.metro.tokyo.lg.jp/）で公開されている犯罪件数データに上記の空間加法混合モデル(1 式＋提案したモデル選択法)を適用する．今回は，非侵入窃盗に分類される犯罪のうち，東京都で最も頻発する自転車盗と万引きを分析対象とする．両方について，東京都の町丁目別・四半期別（2017 年～2018 年）の面積あたり犯罪件数を被説明変数とする．標本数は 12,232 である．図 5 からもわかる通り、自転車盗は広域で満遍なく発生し，万引きは主要駅周辺に集中する傾向がある．

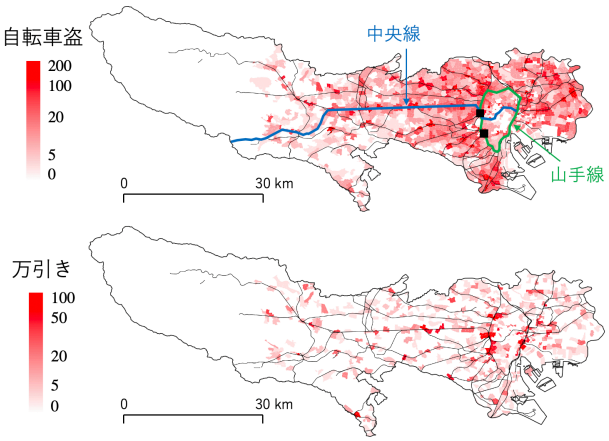

図５：自転車盗と万引きの面積（1km^2）あたり件数（2017 年第 1 四半期）

本実証分析では，(1)式の一種である以下のモデルで面積あたりの自転車盗件数・万引き件数を各々モデリングする：

$$y_{i,t}^{c} = \beta_{i,0} + \beta_{i,1} y_{i,t-1}^{c} + \beta_{i,2} Y_{i,t-1}^{-c} + \sum_{k=3}^{6} \beta_{i,k} x_{i,k} + g_{d(i)}^{(s)} + g_{q(i)}^{(t)} + \varepsilon_{i}, \quad (6)$$

$y_{i,t}^{c}$は町丁目 i の時期 t における面積あたりの犯罪 c の件数である（c は自転車盗または万引き）．犯罪は同じ地区で繰り返される傾向があることから（近接反復被害），１期前の面積あたり犯罪件数$y_{i,t-1}^{c}$（Repeat）を説明変数に加えた．また他罪種の件数が増えた場合にも犯罪リスクが上昇する可能性があることから，非侵入窃盗（罪種 c 以外）の面積あたり犯罪件数$Y_{i,t-1}^{-c}$（RepOther）も説明変数とした．その他の説明変数$\{x_{i,3}, x_{i,4}, x_{i,5}, x_{i,6}\}$は次の通りである：夜間人口密度 (Popden)；昼間人口密度（Dpopden）；外国人居住者比率 (Fpopden), 業率 (UnEmp)；大学卒業者比率(Univ)．各変数は 2015 年国勢調査から収集した．

前章と同様，説明変数の回帰係数$\{\beta_{i,1}, \dots, \beta_{i,6}\}$ のタイプは{Const, SVC, NVC, SNVC}の 4 種類から選択する．ただし$\beta_{i,0}$については，残差の空間相関を除外するために，SVC（正確には spatially varying intercept）で与える．

一方，$g_{d(i)}^{(s)}$と$g_{q(i)}^{(t)}$は街区毎の異質性をとらえるための Random effects 項である．$g_{d(i)}^{(s)} \sim N(0, \tau_{(s)}^{2})$は町丁目毎（添字 : $d(i)$ ）の犯罪リスク水準を表す項であり，$g_{q(i)}^{(t)} \sim N(0, \tau_{(t)}^{2})$は四半期毎の犯罪リスク水準を表す項である．$\tau_{(s)}^{2}$と$\tau_{(t)}^{2}$は両効果の分散を表す．両効果の有無もまた提案したモデル選択法で同時推定する．

3.2. 要因分析への応用結果

選択された回帰係数のタイプを表１に，回帰係数に占める統計的な有意性の割合を表２に，それぞれ要約した．

自転車盗は Repeat, RepOther, Popden, Dpopden のみが統計的に有意となり，夜間・昼間の人口密度が多い場所で繰り返されやすいという直感に整合する結果が得られた．有意な係数のうち，Repeat の係数は SNVC と判定された．同係数をプロットした図６からは都心からやや離れた中央線沿線や西武線沿線で自転車盗が繰り返されるという結果が得られた．一方, RepOther, Popden, Dpopden の回帰係数は NVC と判定された. Popden と Dpopden の回帰係数をプロットした図７からは，基本的には人口が増えるほど自転車盗が増えるものの（係数の符号が一貫して正のため），人口の影響力は，人口増加に伴って低減していくという傾向が確認された．

万引きに関しては，Repeat，RepOther，Dpopden のみが統計的に有意となり，昼間人口密度が多い場所で万引き繰り返されやすいという結果が得られた．有意な係数のうち，Repeat の係数は SNVC と推定された. 同係数をプロットした図６からは新宿, 渋谷, 池袋，吉祥寺といった主要商業地で万引きが繰り返されるという結果が得られた．一方，Dpopden の係数は NVC と判定されており，自転車盗の場合と同様，人口が増えるほど万引きが増えるが，人口が増加するほどその影響は弱まるという傾向が確認された．

Random effects 項（$g_{d(i)}^{(s)}$，$g_{q(i)}^{(t)}$）に関しては，$g_{q(i)}^{(t)}$が自転車盗に対して選択されたのみで，他は全て効果なしと判定された．その推定結果から，第 1 四半期（１～３月）に自転車盗が増加して第 2 四半期（４～６月）に減少するという季節性が確認された（推定値は省略）．

表１：選択された回帰係数のタイプ

	自動車盗	万引き
Intercept ($\beta_{i,0}$)	SVC	SVC
Repeat	SNVC	SNVC
RepOther	NVC	NVC
Popden	NVC	Const
Dpopden	NVC	NVC
Fpopden	Const	Const
UnEmp	Const	Const
Univ	Const	Const

表２：推定された回帰係数に占める統計的な有意性の割合

有意水準	自転車盗						
	Repeat	RepOther	Popden	Dpopden	Fpopden	UnEmp	Univ
10%	0.000	0.000	0.000	0.000			
5%	0.000	0.000	0.000	0.001	0.000	0.000	0.000
1%	1.000	1.000	1.000	0.999			
有意水準	万引き						
	Repeat	RepOther	Popden	Dpopden	Fpopden	UnEmp	Univ
10%	0.000	0.000		0.000			
5%	0.000	0.000	0.000	1.000	0.000	0.000	0.000
1%	1.000	1.000		0.000			

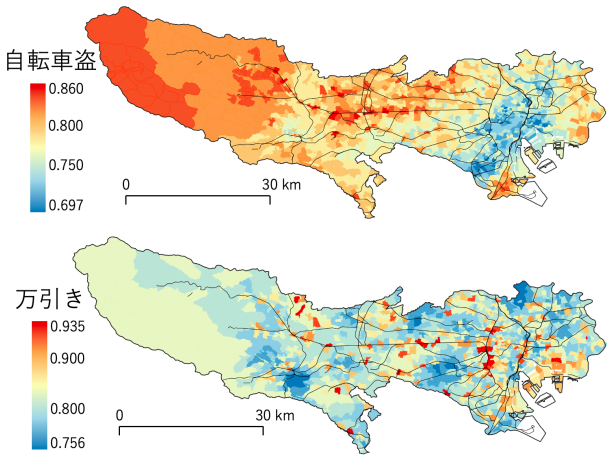

図６：推定された Repeat の回帰係数．赤いほど犯罪が繰り返されやすい

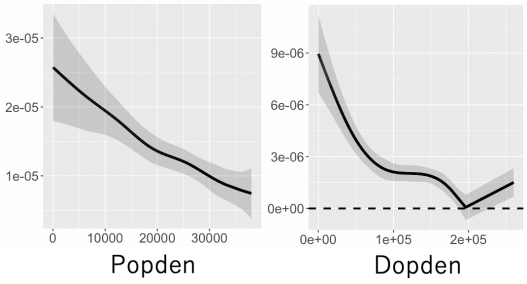

図７： Popden と Dpopden に対する回帰係数(NVC)の推定結果（自転車盗）．横軸が人口密度，縦軸が計数の値である．灰色の領域は 95%信頼区間．

3.3. 犯罪予測への応用結果

最後に，2017～2018 年のデータから推定された上記モデルを 2019 年第 1 四半期の面積あたり犯罪件数の予測に応用した．ここでは提案手法ならびにカーネル密度推定法（Kernel density estimation; KDE）を比較した．KDE は「近隣での犯罪件数が多いほど犯罪リスクが高い」という仮定の下で，犯罪予測を行う手法であり，その予測結果は空間的に滑らかな分布を持つ．

自動車盗と万引きに対する両結果を図８，９にそれぞれ示す．両図から，両ケースとも，提案手法は KDE を大きく上回る精度で犯罪が予測できていることが確認できる．このことより，例えば「KDE を使う」のようにモデルを決め打ちするのでなく，提案手法などを用いてモデルを選択することで，モデルの予測精度が大きく改善することを確認した。

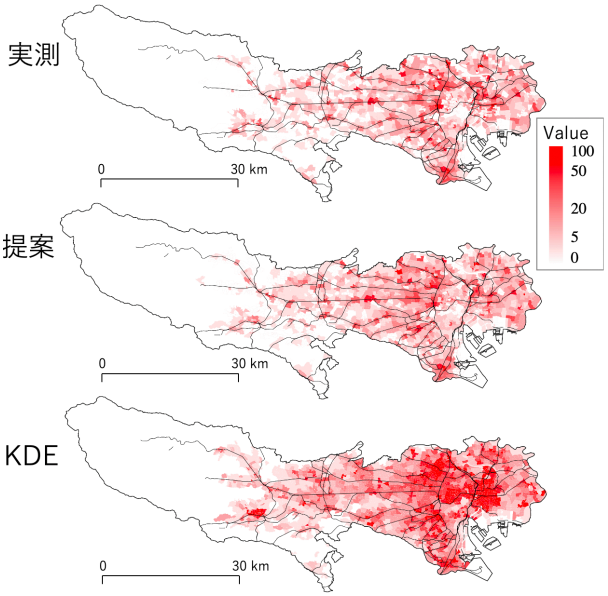

図８：自転車盗の実測値と予測結果

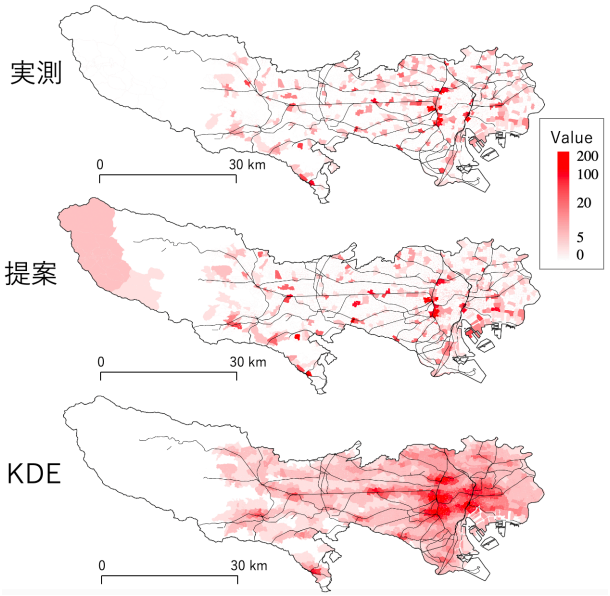

図９：万引きの実測値と予測結果

４．R パッケージ spmoran への実装

２章で開発したモデル選択法をフリーの統計ソフトウェア R のパッケージ spmoran 内の resf_vc 関数に実装した．例えば本日 3.2 節で紹介したモデルを実装するサンプルコードは以下である.なお同関数では default でモデル選択を行うようになっている：

```
meig <- meigen_f(coords=coords)
res      <- resf_vc(y=y, x=x, xgroup=xgroup,
                    meig=meig, x_nvc =TRUE)
```

入力する変数は以下の通りである：

- coords：緯度経度の行列
- y　　：被説明変数のベクトル
- x　　：説明変数の行列
- xgroup: Random effects のグループを決める ID
- x_nvc：TRUE の場合 SVC と NVC の両方を考慮（Const, SVC, NVC, SNVC の中からモデル選択）

spmoran パッケージについて詳しくはマニュアル（https://github.com/dmuraka/spmoran）を参照されたい.

謝辞

本研究は，情報通信研究機構の委託研究「犯罪オープンデータを活用したデータ駆動型犯罪予測手法の開発と市民・自治体向けの犯罪予測アプリケーションの構築」の成果の一部である.

参考文献

Cahill, M.・Mulligan, G. (2007) Using geographically weighted regression to explore local crime patterns.「Social Science Computer Review」, **25** (2), 174-193.

Marra, G.・Wood, S. N. (2011) Practical variable selection for generalized additive models.「Computational Statistics & Data Analysis」, **55** (7), 2372-2387.

Murakami, D.・Kajita, M.・Kajita, S. (2020) Scalable model selection for spatial additive mixed modeling: application to crime analysis.「ArXiv」, 2008.03551.

Wood, S. N.・Li, Z.・Shaddick, G.・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.