疎なカウントデータのための地理的加重ポアソン回帰の安定化・高速化

村上大輔*・堤田成政**・吉田崇紘***・中谷友樹****

Stable geographically weighted Poisson regression for sparse count data

Daisuke Murakami*, Narumasa Tsutsumida**, Takahiro Yoshida***, Tomoki Nakaya****

Geographically weighted Poisson regression (GWPR) is widely used for spatial regression analysis of count data. However, it tends to be unstable because of a fundamental drawback of Poisson regression. To overcome the drawback, we introduce a log-linear approximation to estimate GWPR without relying on the conventional Poisson regression framework. The proposed approach approximates GWPR using the basic GWR modeling technique with transformed explained variables. Monte Carlo experiments demonstrate that the proposed GWPR outperforms the conventional GWPR in terms of estimation accuracy and computationally efficiency. The proposed GWPR is applied to a crime analysis, and demonstrated its usefulness empirically.

Keywords: 地理的加重ポアソン回帰, 線形近似, 犯罪, カウントデータ

１．はじめに

犯罪件数，陽性者数，滞留人口など，地理的に集計された数多くのカウントデータが収集・公開され実社会に役立てられている．地理的加重ポアソン回帰（GWPR: geographically weighted Poisson regression）はそういったカウントデータを対象とした局所回帰であり，地点（ゾーン毎）の回帰係数を推定するために幅広く用いられてきた．例えば Nakaya et al. (2005)は GWPR を用いて専門職・技術職従事者の割合や失業率が生産年齢死亡者数に及ぼす影響を分析した. また Hadayeghi (2010)は交通事故件数の地理的要因を GWPR を用いて分析した.

一方，GWPR の推定精度や安定性については必ずしも十分に精査されてこなかった．残念ながら GWPR の推定は不安定になりがちである. これは次の理由による。一点目はポアソン回帰が特定の条件下で解を持たない点である（Silva, 2010; Correira et al. 2019）．例えば，カウントデータ（被説明変数）がゼロ値ではない標本について，説明変数が線型独立ではない（完全な多重共線性がみられる）場合は解が識別できない．この特性により，特にゼロ値の多いカウントデータを用いたポアソン回帰は不安定になりがちである. 二点目は GWPR が周辺の標本のみを考慮する局所回帰である点である．周辺のカウントの値がゼロばかりでは一点目の影響を強く受けるためモデル推定はやはり不安定となる．以上の理由より, 特にゼロ値の多いカウントデータに GWPR を適用したい場合であっても，従来のポアソン回帰の枠組みに頼るのは危険である.

ポアソン回帰を近似するための線形近似はいくつか提案されており，それらを用いれば，ポアソン分布を仮定する必要がなくなり上記の２点の影響は緩和される．しかしながら，あとで示すように既存の線形近似は推定精度が低い．例外的に Murakami and Matsui (2021)で提案された線形近似は，ゼロ値が多い場合には従来のポアソン回帰を上回る精度で回帰係数が推定され，それ以外の場合にも従来のポアソン回帰とほぼ同等の精度となることを示している.

*　正会員　統計数理研究所 データ科学研究系（The Institute of Statistical Mathematics）
〒190-8562　東京都立川市緑町 10-3　E-mail：dmuraka@ism.ac.jp
**　正会員　埼玉大学 理工学研究科（Saitama University）
**　正会員　東京大学 大学院工学系研究科（The University of Tokyo）
**　正会員　東北大学 環境科学研究科（Tohoku University）

また同線形近似を用いれば，通常のポアソン回帰で必要な重み付き最小二乗推定の繰り返しが不要となり，大幅な計算時間の短縮が可能となる．

そこで本研究では Murakami and Matsui (2021)の線形近似を応用することで，GWPR の安定化を試みる．

２．ポアソン回帰の線形近似

2.1. ポアソン回帰

本研究では以下の擬似ポアソン回帰を考える：

$$y_i \sim Poisson(\lambda_i, \sigma^2),\ \lambda_i = \exp\left(\sum_{k=1}^{K} x_{i,k}\beta_k\right) \quad (1)$$

y_iはゾーン i の被説明変数（カウント），$x_{i,k}$は説明変数，β_kは回帰係数である．σ^2は過分散の度合いを決めるパラメータである．上式はカウントデータy_iを平均λ_i, 分散$\lambda_i\sigma^2$の(擬似)ポアソン分布で推定しようというモデルである．

2.2. 線形近似

残念ながら(1)式はゼロ値が多い場合に不安定となる．従って，本研究では以下の線形近似を考える：

$$y_i^* = \sum_{k=1}^{K} x_{i,k}\beta_k + \varepsilon_i,\ \varepsilon_i \sim N\left(0, \frac{\sigma^2}{y_i + 0.5}\right) \quad (2)$$

$y_i^* = \log(y_i + 0.5) - \frac{1+0.5r}{y_i+0.5}$であり，$r$は被説明変数$y_i$に締めるゼロ値の割合である．Murakami and Matsui (2021)の重み付き最小二乗推定で得られる回帰係数の推定精度は通常のポアソン回帰と同等である．またゼロ値が多い場合は通常のポアソン回帰を上回る精度となったことから，GWPR に対してもこの近似は有効な可能性がある．従って次章では GWPR に同近似を導入する．

２．GWPR の線形近似

2.1. GWPR

GWPR は下式で定義される：

$$y_i \sim Poisson(\lambda_i, \sigma^2),\ \lambda_i = \exp\left(\sum_{k=1}^{K} x_{i,k}\beta_{i,k}\right) \quad (3)$$

(1)式との唯一の違いは回帰係数$\beta_{i,k}$が場所毎に与えられている点である．ゾーン i の係数$\beta_{i,k}$は，同ゾーンの重心点からの距離$d_{i,j}$に応じて減衰するカーネルw_j（本研究では$w_j = \exp\left(-(d_{i,j}/b)^2\right)$）で各標本を重みづけた上で，減衰の速さを決めるバンド幅パラメータ b を交差検証などで最適化する。次に b を所与として反復重み付き最小 2 乗法などによって各回帰係数を推定する．

2.2. 線形近似（提案手法）

(2)式と同様の近似を GWPR に適用すると以下となる：

$$y_i^* = \sum_{k=1}^{K} x_{i,k}\beta_{k,i} + \varepsilon_i,\ \varepsilon_i \sim N\left(0, \frac{\sigma^2}{y_i + 0.5}\right) \quad (4)$$

(4)式は誤差項の分散が$\frac{\sigma^2}{y_i+0.5}$となっている以外は線形の地理的加重回帰（GWR; geographically weighted regression）と同じであり，GWR を推定する際に用いられる推定法が直ちに応用できる．具体的には，上記と同様にカーネルw_jで各標本を重みづけた上で，バンド幅 b を交差検証などで最適化する．次に b を所与とした重み付き最小二乗推定により回帰係数を推定$\beta_{k,i}$する．

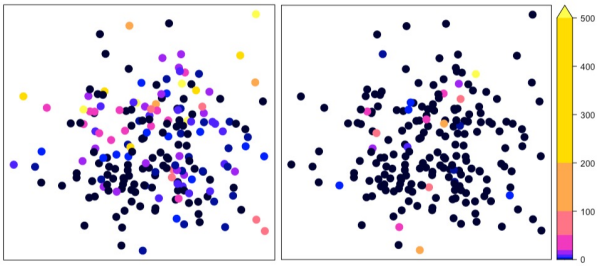
３．シミュレーション実験

3.1. 概要

(4)式は実用的ではあるが GWPR の精度良い近似となっているのだろうか．本章では，ポアソンモデル(5)式から生成したカウントデータ(標本数: 200)に対するあてはめを繰り返すことで，提案モデルの近似精度を検証する：

$$\begin{aligned} &y_i \sim Poisson(\lambda_i, \sigma^2), \\ &\lambda_i = \exp\left(\beta_{i,0} + x_{i,1}\beta_{i,1} + x_{i,2}\beta_{i,2}\right) \end{aligned} \quad (5)$$

説明変数は標準正規分布$x_{i,k} \sim N(0,1)$から生成する．データの得られる XY 座標もまた標準正規分布で与える．その上で，地点毎の回帰係数は空間移動平均過程$\beta_{i,k} = b_k + \sum_{k=1}^{K} c_{i,j}u_j$, $u_j \sim N(0,1)$で与える．ただし$b_1 = 2, b_2 = -0.5$である．$c_{i,j}$は$\exp\left(-(d_{i,j}/b)^2\right)$を$(i, j)$要素に持つ行列を行基準化したもので与える．



シナリオ1 (ゼロ比率: 0.36)　　シナリオ2 (ゼロ比率: 0.86)

図 1: シナリオ毎に得られたカウントデータy_iの例

以上の設定の下，2 つのシナリオ下で GWPR と提案モデルを比較する. シナリオ 1 では$\{b_0, \sigma^2\} = \{0,1\}$を仮定する．このシナリオ下では被説明変数のゼロ値は比較的少ない．もう一つは$\{b_0, \sigma^2\} = \{-2,10\}$を仮定するゼロ値が多いシナリオである（図 1）.

両シナリオ下でモデルを 200 回ずつ推定して，回帰係数の推定値の Root mean squared error（RMSE）とバイアスを評価する．比較するモデルは次の通りである：通常のポアソン回帰（Basic），提案した GWPR の近似 (p-GWPR), p-GWPR の縮小推定版 (p-GWPRs)，GWPR，適応型カーネルを用いた GWPR（GWPRa）．ここで適応型カーネルとは，標本の密度に応じてカーネルのバンド幅を調整する方法であり，安定性が期待できることから比較対象とした.

3.2. 結果

図 2 と図 3 は，2 つのシナリオ下で推定された回帰係数$\beta_{i,k}$の RMSE とバイアスの箱ひげ図である. Basic の RMSE やバイアスは両シナリオで大きくなった．残念ながら GWPR と GWPRa の RMSE とバイアスもまた Basic と同等であり，回帰係数を場所毎に推定することによる精度改善はみられなかった. また 3 手法いずれも RMSE とバイアスが極めて大きな値となる場合があることが確認され，それら手法が不安定であることを確認した．対象的に p-GWPR や p-GWPRs は RMSE とバイアスが小さくなった。例えば、ケース 1 では、$\beta_{i,1}$の平均 RMSE は、2.215（Basic）、0.925（p-GWPR）、0.958（p-GWPRs）、2.126（GWPR）、2.133（GWPRa）となった．この結果から、我々が近似した近似は通常の GWPR よりも安定しておりむしろ高精度であるという期待通りの結果が得られた.

次に, 提案手法の計算時間を通常の GWPR と比較した．ここでは標本数を変えながら各 5 回ずつモデル推定を行なって平均計算時間を評価した（図 4）. 図 4 から，例えば標本数 2000 の場合， GWPR の平均計算時間は 620 秒であったのに対し，p-GWPR と p-GWPRs はそれぞれ 9 秒，72 秒であった.

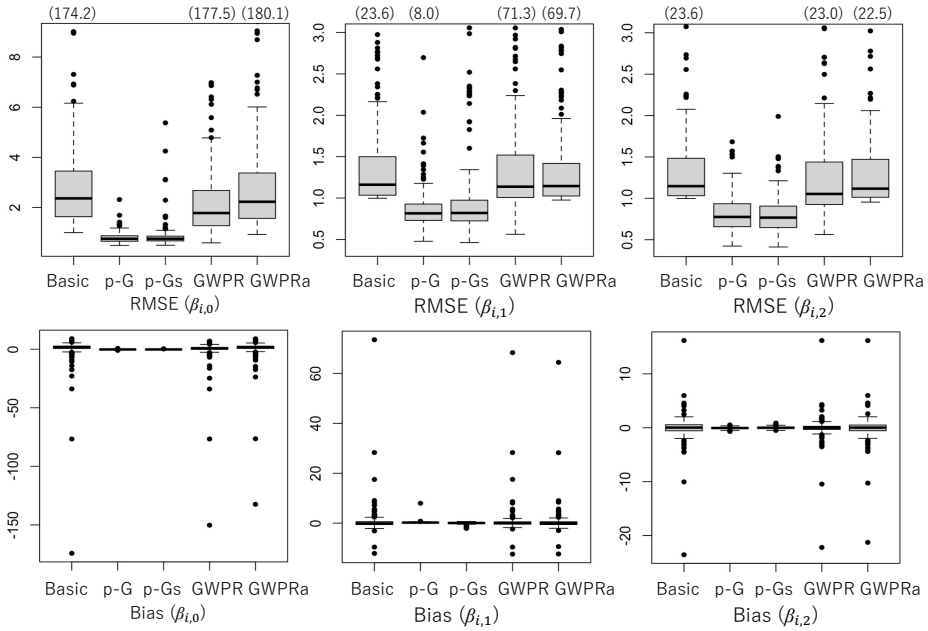


図 2: シナリオ 1 での RMSE（上）とバイアス（下）の評価結果（左: $\beta_{i,0}$，中: $\beta_{i,1}$，右: $\beta_{i,2}$）．p-G は p-GWPR，p-Gs は p-GWPRs を表す．最大値が枠外の場合は，最大値のみパネル上部に示した.

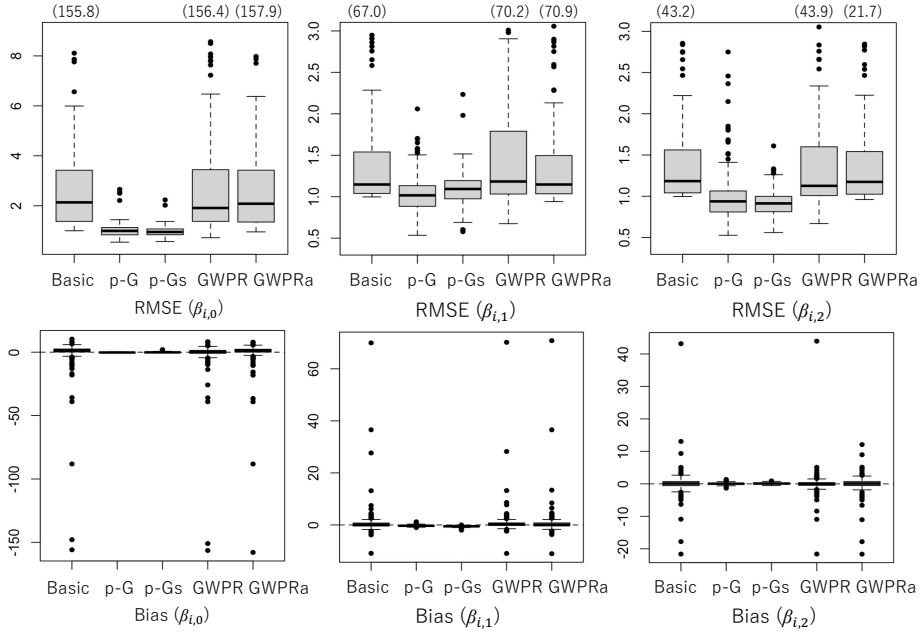


図 3: シナリオ 2 での RMSE（上）とバイアス（下）の評価結果（左: $\beta_{i,0}$，中: $\beta_{i,1}$，右: $\beta_{i,2}$）．

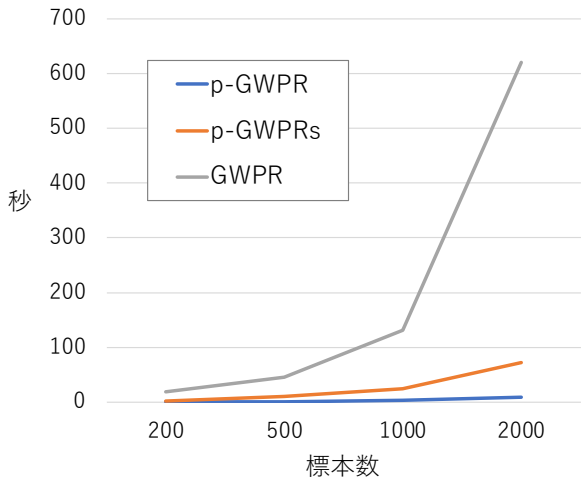


図 4：計算時間の比較

以上より，これまで幅広く使われてきた GWPR の回帰係数の推定に際しては，従来の推定法を使うのは必ずしも推奨はされず，むしろ今回提案した線形近似を使う方が，より高精度で高速に推定できることを確認した．

4．犯罪件数データへの適用

4.1．概要

本章では，提案した p-GWPR や p-GWPRs を東京都内の犯罪分析に応用する．被説明変数は自動車盗難件数（2019 年 4 月，町丁目別）（出典：大東京防犯ネットワーク https://www.bouhan.metro.tokyo.lg.jp/）とする．町丁目 1,529 の 45.8%にあたる 700 の町丁目で盗難件数は 0 件であった．説明変数は一月前の自動車盗難件数（一月前件数），夜間人口密度，大学卒業者が人口に占める割合（大卒率）とした．またオフセット変数として人口を用いた．

4.2．結果

提案手法を用いた回帰係数の推定結果を図 5 に整理した．ここでは有意水準 5%で統計的に有意な効果の見られた町丁目のみを着色している．定数項の推定結果は，都心では人口あたり盗難件数が少ない傾向を反映している（オフセット変数として人口を用いた点に注意）．これは同地区には公共交通が従実しておりそもそも自転車利用が少ないための可能性がある．人口密度に対する回帰係数は都心で負，郊外で正になった．この結果は，都心では人の目がない場所で自転車が盗まれやすい一方で郊外では人の

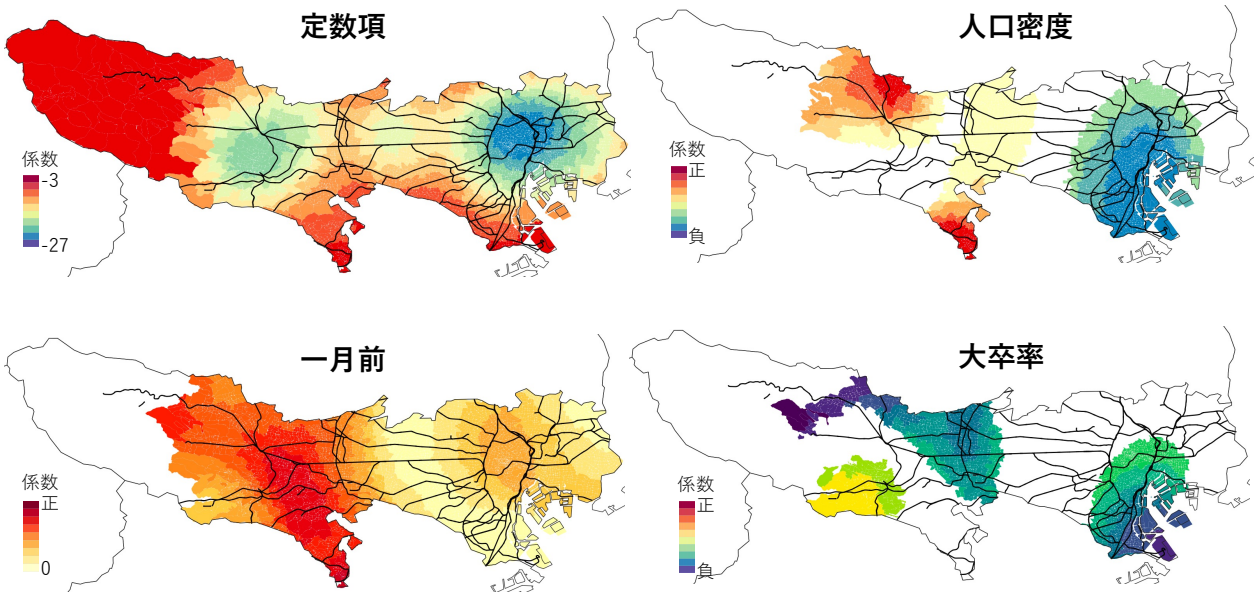


図 5：推定された回帰係数の空間分布

多い繁華街等で盗まれやすいことを意味する可能性がある．この知見は防犯対策を検討する上で有用な可能性がある．一月前件数の係数は一貫して正であり基本的に同じ町丁目で自転車が盗まれ続ける傾向が確認された．一方，係数値は郊外で大きく，同傾向は郊外部で特に顕著との示唆を得た．最後に大卒率の係数は一貫して負であり，例えば都心沿岸部などでは大卒者が多い地域ほど自転車が盗まれにくいという結果となった．

謝辞

本研究は JSPS 科研費 17H02046，20K13261，ならびにデータサイエンス共同利用基盤施設の公募型共同研究 DS-ROIS-JOINT（課題番号 005RP2021）の助成を受けたものです．

参考文献

Carrel, M., Escamilla, V., Messina, J., Giebultowicz, S., Winston, J., Yunus, M., ... & Emch, M. (2011). Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. *International journal of health geographics*, *10* (1), 1-9.

Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis & Prevention*, *42* (2), 676-688.

Murakami, D., & Matsui, T. (2021). Improved log-Gaussian approximation for over-dispersed Poisson regression: application to spatial analysis of COVID-19. *ArXiv*, 2104.13588.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, *24* (17), 2695-2717.

Silva, J. S., & Tenreyro, S. (2010). On the existence of the maximum likelihood estimates in Poisson regression. *Economics Letters*, *107* (2), 310-312.