


Stable geographically weighted Poisson regression for count data

Daisuke Murakami 

Institute of Statistical Mathematics, Japan
dmuraka@ism.ac.jp

Narumasa Tsutsumida 

Saitama University, Japan
narut@mail.saitama-u.ac.jp

Takahiro Yoshida 

The University of Tokyo, Japan
yoshida.takahiro@up.t.u-tokyo.ac.jp

Tomoki Nakaya 

Tohoku University, Japan
tomoki.nakaya.c8@tohoku.ac.jp

Binbin Lu 

Wuhan University, Japan
binbinlu@whu.edu.cn

Paul Harris 

Rothamsted Research, UK
paul.harris@rothamsted.ac.uk

Abstract

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 Poisson regression framework. The proposed approach approximates GWPR using the basic GWR modeling with transformed explained variables. Monte Carlo experiments show that the proposed GWPR outperforms the conventional GWPR in terms of both estimation accuracy and computationally efficiency. Finally, the proposed GWPR is applied to an analysis of coronavirus disease 2019 (COVID-19).

Acknowledgements This research was supported by JST-Mirai Program Grant Number JPMJMI20B2, Japan, and the Joint Support Center for Data Science Research at Research Organization of Information and Systems (ROIS-DS-JOINT) under Grant 003RP2020.

1 Introduction

Number of crimes, infected people, cars, and other counts have been monitored and opened to the public recently. Geographically weighted Poisson regression (GWPR) is a popular spatial regression approach to investigate spatially varying influencing factors on count outcome. For example, [7] applied GWPR to estimate spatially varying influence of the proportion of professional and technical workers, unemployment rate, and other covariates on working-age mortality counts. [5] used GWPR to analyze the number of vehicle collisions.

2 Stable geographically weighted Poisson regression

Still, as we will illustrate later, GWPR tends to be unstable. This is attributable to the following reasons. First, Poisson regression is identifiable only weakly or even unidentifiable depending on the data configuration [8]. For example, Poisson regression does not have the maximum likelihood solution if covariates are perfectly collinear for the sub-samples with positive observations. Second, the GWPR model, which is a local model, becomes unstable if there are many zeros nearby the regression point; if most observations nearby a site take zero values, the GWPR model at the site will be difficult to estimate. Because of these problems, it is not reasonable to rely on Poisson regression even if we want to estimate GWPR.

The objective of this study is to propose a stable version of GWPR. To achieve it, we apply a log-linear approximation of [6] estimating the conventional Poisson regression without annoying the identifiable problem, to GWPR.

2 Model

We approximate the following over-dispersed GWPR model:

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

where y_i is the explained count variable at i -th zone, z_i is the offset variable, $x_{i,k}$ is the k -th covariable, and $\beta_{i,k}$ is the spatially varying coefficient. $oPoisson(\mu_i, \sigma^2)$ is the over-dispersed Poisson distribution with mean μ_i and overdispersion parameter σ^2 . The count data $\{y_1, \dots, y_N\}$ is equi-dispersed if $\sigma^2 = 1$, which the usual GWPR assumes, over-dispersed if $\sigma^2 > 1$, and under-dispersed if $\sigma^2 < 1$.

To stably estimate the model, we replace the Poisson model estimation with a log-linear model estimation proposed by [6]. They showed that over-dispersed Poisson regression can be approximated by a log-linear regression model with explained variable $y_i^* = \log\left(\frac{y_i + 0.5}{z_i}\right) - \frac{1 + 0.5r}{y_i + 0.5}$ and the weight for i -th sample $w_i = y_i + 0.5$ where r is the ratio of zero counts. The log-linear model is estimated by the usual ordinary least squares fit. Thus, it is free from the identification problem in conventional Poisson model estimation. Despite the simplicity, the coefficient estimation accuracy is compatible to the usual (over-dispersed) Poisson regression for moderate to large samples while better for small samples owing to the stability.

This study applies their approach to GWPR. The resulting approximate GWPR model yields

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

Because the model is identical to the basic GWR model, the model is easily estimated, inferred, and extended in the same manner as the usual GWR model. It implies that the proposed model estimation is much faster than the conventional GWPR model estimation which iterates re-weighting samples and estimating the GWR model (i.e., iteratively re-weighted least squares estimation). If Eq. (2) achieves a reasonable estimation accuracy, it will be valuable as a simpler and faster alternative of the usual GWPR.

3 Monte Carlo experiment

3.1 Outline

This section examines the estimation accuracy of the approximate GWPR with fixed kernel (Propose 1), the same with ridge regularization (Propose 2; see [9]), which imposes an ridge

prior, with the usual Poisson regression (GLM), GWPR with fixed kernel (GWPR), and GWPR with adaptive kernel (GWPRa). While geographically weighted models estimate spatially varying coefficients by locally weighting samples using a distance-decaying kernel, the fixed kernel means that the bandwidth, which is estimated from data, is the same across the study area. The adaptive bandwidth determines the the band to include a certain number of samples within the bandwidth distance (see, [2]). These bandwidths are optimized by leave-one-out cross-validation. The Gaussian kernel is used. These models are fitted to the synthetic count data generated from

$$y_i \sim \text{Poisson}(\mu_i, \sigma^2), \quad \mu_i = \exp(\beta_{i,0} + x_{i,1}\beta_{i,1} + x_{i,2}\beta_{i,2}), \quad x_{i,k} \sim N(0, 1). \quad (3)$$

GW(P)R does not assume any process for the coefficients; for the simulation, we assume the coefficients to obey a moving average process $\beta_{i,k} = b_k + \sum_{j=1}^N c_{i,j}u_j$, $u_j \sim N(0, 1)$, with sample size $N = 200$. b_k represents the mean of the k -th SVC. We assume $b_1 = 2.0$ and $b_2 = -0.5$. The spatial weight $c_{i,j}$ is given by the (i, j) -th element of a spatial proximity matrix whose (i, j) -th element equals $\exp(-(d_{i,j})^2)$ where $d_{i,j}$ is the Euclidean distance between the sample sites i and j . Following many data in regional science whose samples are concentrated in central urban areas while sparse in suburban areas, spatial coordinates for the samples are generated from two independent standard normal distributions. Estimation accuracy is compared in two scenarios (see Figure 1). The first assumes $\{b_0, \sigma^2\} = \{0, 1\}$ whose samples are equi-dispersed and have a moderate number of zeros. The conventional GWPR is likely to work in this scenario. The second assumes $\{b_0, \sigma^2\} = \{-2, 10\}$ whose samples are over-dispersed and have many zeros. See Figure 1 for examples of samples in these scenarios.

In the simulation, each model is estimated 200 times and the root mean squared error (RMSE) and the bias for the SVCs $\{\beta_{i,0}, \beta_{i,1}, \beta_{i,2}\}$ are evaluated. Note that, in the conference, we will also perform simulations assuming spatially dependent covariates.

3.2 Result

Figures 2 and 3 display the boxplots for the RMSE and bias for the estimated spatially varying coefficients under the two scenarios. The results in the two scenarios are similar. GLM, GWPR, and GWPRa tend to have large RMSE and bias values. The standard GLM-based approach including GWPR is found to be unstable. By contrast, our proposed models have considerably smaller RMSE and bias values than GLM, GWPR, and GWPRa across cases. For example, in case 1, the mean RMSEs for $\beta_{i,1}$ are 2.215 (GLM), 0.925 (Proposal 1), 0.958

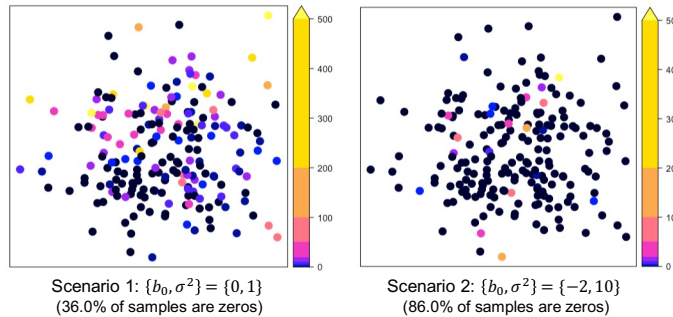


Figure 1 Examples of spatial plots for the count data generated under the scenarios A and B. Black dots represent zero values while lighter dots represent larger count values.

4 Stable geographically weighted Poisson regression

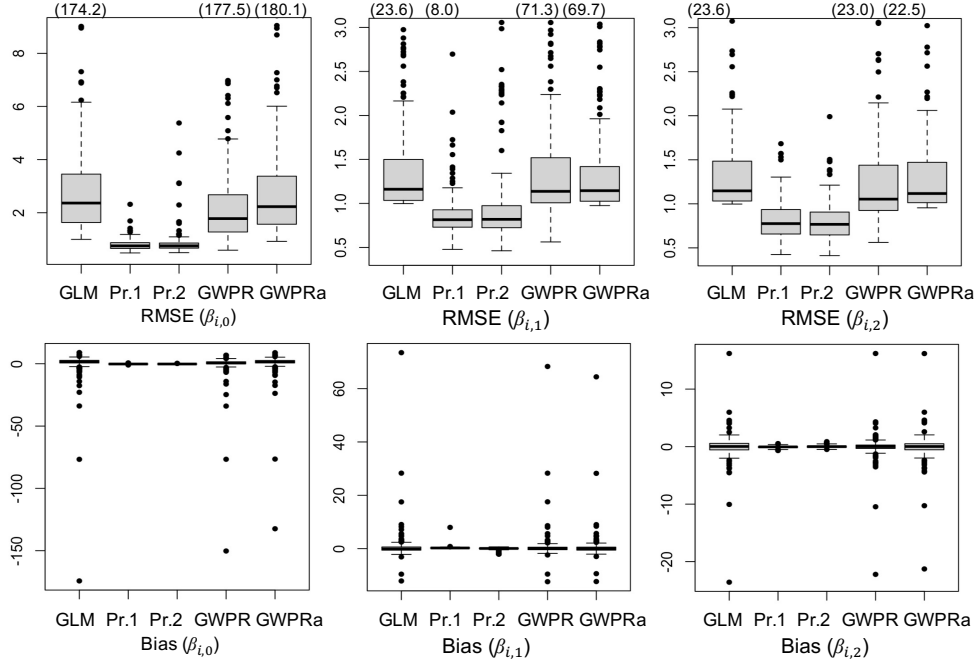


Figure 2 Boxplots of the RMSE and biases for the coefficients under the scenario A ($\sigma^2 = 1, b_0 = 0$). Pr.1 and Pr.2 represent Proposal 1 and Proposal 2. If the maximum RMSE value exceeds the displayed boundary in each panel, the maximum value is described above the panel.

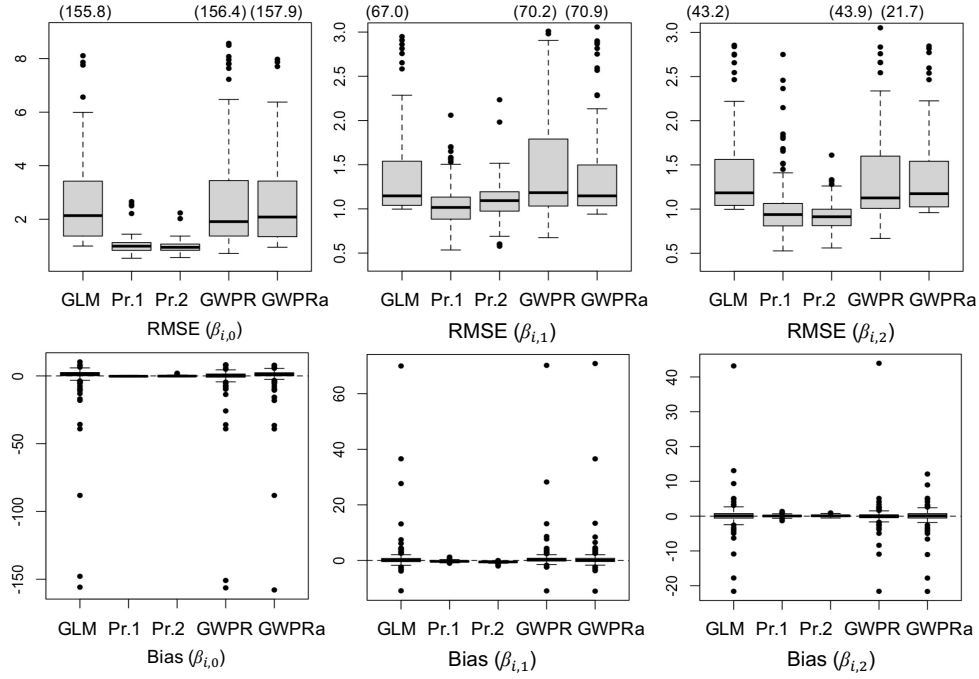
(Proposal 2), 2.126 (GWPR), and 2.133 (GWPRa). The result suggests that our approximate GWPR is more stable and accurate than the usual GLM-based approaches.

We also confirmed the computational efficiency of the proposed models. For example, for 2000 samples, GWPR took 620 seconds on average of five trials while Proposal 1 and 2 took only 9 and 72 seconds, respectively. While the accuracy of GWPR has been considered good enough, our study showed that GWPR can be unstable and it is better to employ the proposed approximation to stabilize it.

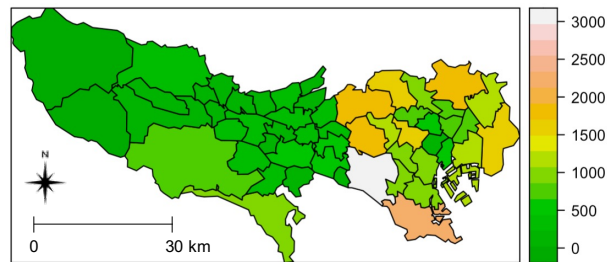
4 Application to COVID-19 data

This section applies the proposed approach (Proposal 1) and GWPR to an analysis of coronavirus disease 2019 in the Tokyo metropolis, Japan. The explained variable is the number of reported cases by municipality during January 2021 (see Figure 5). The covariates are nighttime population density (PopDen) and day-night population ratio (DNrat) (source: National census 2015). For offset variable, we use nighttime population. Thus, we estimate spatially varying influence of PopDen and DNrat on the number cases standardized by the population.

The optimized bandwidth values are 88.4 km for GWPR and 55.3 km for Propose 1. The estimated coefficients are plotted in Figure 6. This figure suggests that the proposed and usual GWPR estimates often have considerably different map patterns. Based on the simulation result, ours is more reliable. The intercept estimated from Proposal 1 suggests higher infection risk in the eastern part of the study area including the center of Tokyo. The estimated coefficients on PopDen increases in the residential area in the center of this figure. These results are intuitively reasonable. The estimated coefficient on DNrat demonstrates

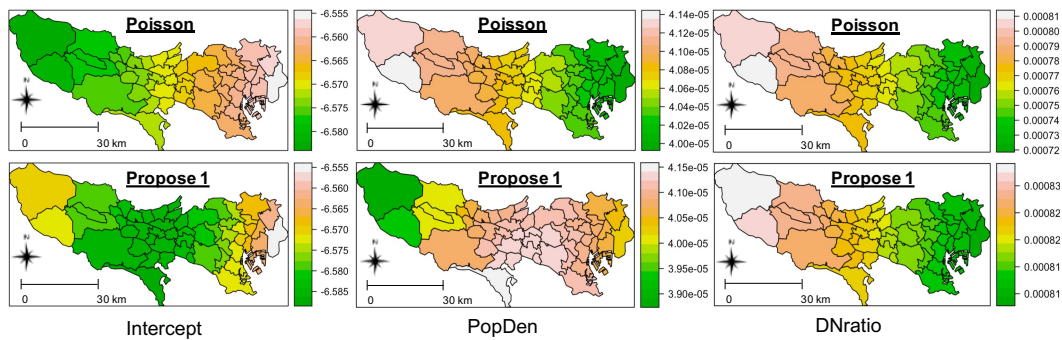


■ **Figure 3** Boxplots of the RMSE and biases for the coefficients under the scenario B ($\sigma^2 = 10, \beta_0 = -2$). Pr.1 and Pr.2 mean Proposal 1 and Proposal 2. If the maximum RMSE value exceeds the displayed boundary in each panel, the maximum value is described above the panel.



■ **Figure 4** Number of cases by municipality in January 2021.

6 Stable geographically weighted Poisson regression



■ **Figure 5** Estimated spatially varying coefficients (Top: GWPR; Bottom: Propose 1)

that, in the western suburban area, infection risk tends to increase in municipalities with population concentration during daytime. These findings will be useful to consider measures against COVID-19.

5 Summary

We demonstrate that GWPR can be estimated accurately and computationally efficiently through the basic GWR procedure if only a simple transformation is applied to the explained variables. The proposed approach will enable us applying multiscale GWR ([4]), geographically and temporally weighted regression ([3]), and other extended GWR, which was developed for Gaussian data, to count data. Based on studies in geostatistics for non-Gaussian data (e.g., [1]), it is also important to consider residual spatial dependence.

References

- 1 Peter J Diggle, Jonathan A Tawn, and Rana A Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.
- 2 A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.
- 3 A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Geographical and temporal weighted regression (gtwr). *Geographical Analysis*, 47(4):431–452, 2015.
- 4 A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265, 2017.
- 5 Alireza Hadayeghi, Amer S Shalaby, and Bhagwant N Persaud. Development of planning level transportation safety tools using geographically weighted poisson regression. *Accident Analysis & Prevention*, 42(2):676–688, 2010.
- 6 Daisuke Murakami and Tomoko Matsui. Improved log-gaussian approximation for over-dispersed poisson regression: application to spatial analysis of covid-19. *arXiv preprint arXiv:2104.13588*, 2021.
- 7 Tomoki Nakaya, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. Geographically weighted poisson regression for disease association mapping. *Statistics in medicine*, 24(17):2695–2717, 2005.
- 8 JMC Santos Silva and Silvana Tenreiro. On the existence of the maximum likelihood estimates in poisson regression. *Economics Letters*, 107(2):310–312, 2010.
- 9 David C Wheeler. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, 39(10):2464–2481, 2007.