

UC Santa Barbara

GIScience 2021 Short Paper Proceedings

Title

Stable geographically weighted Poisson regression for count data

Permalink

<https://escholarship.org/uc/item/8kg664zq>

Authors

Murakami, Daisuke
Tsutsumida, Narumasa
Yoshida, Takahiro
et al.

Publication Date

2021-09-01

DOI


10.25436/E2X59B

Peer reviewed

1 Stable geographically weighted Poisson regression 2 for count data

3 **Daisuke Murakami** 

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

6 **Narumasa Tsutsumida** 

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

9 **Takahiro Yoshida** 

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

12 **Tomoki Nakaya** 

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

15 **Binbin Lu** 

16 Wuhan University, Japan
17 binbinlu@whu.edu.cn

18 **Paul Harris** 

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

21 Abstract

22 Geographically weighted Poisson regression (GWPR) is widely used for spatial regression analysis
23 of count data. However, it tends to be unstable because of a fundamental drawback of Poisson
24 regression. To overcome the drawback, we introduce a log-linear approximation to estimate GWPR
25 without relying on Poisson regression framework. The proposed approach approximates GWPR
26 using the basic GWR modeling with transformed explained variables. Monte Carlo experiments
27 show that the proposed GWPR outperforms the conventional GWPR in terms of both estimation
28 accuracy and computationally efficiency. Finally, the proposed GWPR is applied to an analysis of
29 coronavirus disease 2019 (COVID-19).

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

37 **1** Introduction

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

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

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

55 2 Model

56 We approximate the following over-dispersed GWPR model:

$$57 \quad y_i \sim oPoisson(\mu_i, \sigma^2), \quad \mu_i = z_i \exp\left(\sum_{k=1}^K x_{i,k} \beta_{i,k}\right) \quad (1)$$

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

63 To stably estimate the model, we replace the Poisson model estimation with a log-linear
 64 model estimation proposed by [6]. They showed that over-dispersed Poisson regression can
 65 be approximated by a log-linear regression model with explained variable $y_i^* = \log\left(\frac{y_i + 0.5}{z_i}\right) -$
 66 $\frac{1 + 0.5r}{y_i + 0.5}$ and the weight for i -th sample $w_i = y_i + 0.5$ where r is the ratio of zero counts. The
 67 log-linear model is estimated by the usual ordinary least squares fit. Thus, it is free from the
 68 identification problem in conventional Poisson model estimation. Despite the simplicity, the
 69 coefficient estimation accuracy is compatible to the usual (over-dispersed) Poisson regression
 70 for moderate to large samples while better for small samples owing to the stability.

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

$$73 \quad y_i^* = \sum_{k=1}^K x_{i,k} \beta_{i,k} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \frac{\sigma^2}{w_i}\right) \quad (2)$$

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

80 3 Monte Carlo experiment

81 3.1 Outline

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

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

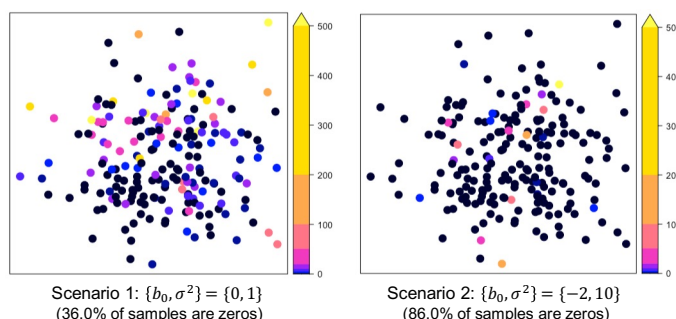
$$92 \quad y_i \sim oPoisson(\mu_i, \sigma^2), \quad \mu_i = \exp(\beta_{i,0} + x_{i,1}\beta_{i,1} + x_{i,2}\beta_{i,2}), \quad x_{i,k} \sim N(0, 1). \quad (3)$$

93 GW(P)R does not assume any process for the coefficients; for the simulation, we assume
 94 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
 95 sample size $N = 200$. b_k represents the mean of the k -th SVC. We assume $b_1 = 2.0$ and
 96 $b_2 = -0.5$. The spatial weight $c_{i,j}$ is given by the (i, j) -th element of a spatial proximity
 97 matrix whose (i, j) -th element equals $\exp(-(d_{i,j})^2)$ where $d_{i,j}$ is the Euclidean distance
 98 between the sample sites i and j . Following many data in regional science whose samples are
 99 concentrated in central urban areas while sparse in suburban areas, spatial coordinates for
 100 the samples are generated from two independent standard normal distributions. Estimation
 101 accuracy is compared in two scenarios (see Figure 1). The first assumes $\{b_0, \sigma^2\} = \{0, 1\}$
 102 whose samples are equi-dispersed and have a moderate number of zeros. The conventional
 103 GWPR is likely to work in this scenario. The second assumes $\{b_0, \sigma^2\} = \{-2, 10\}$ whose
 104 samples are over-dispersed and have many zeros. See Figure 1 for examples of samples in
 105 these scenarios.

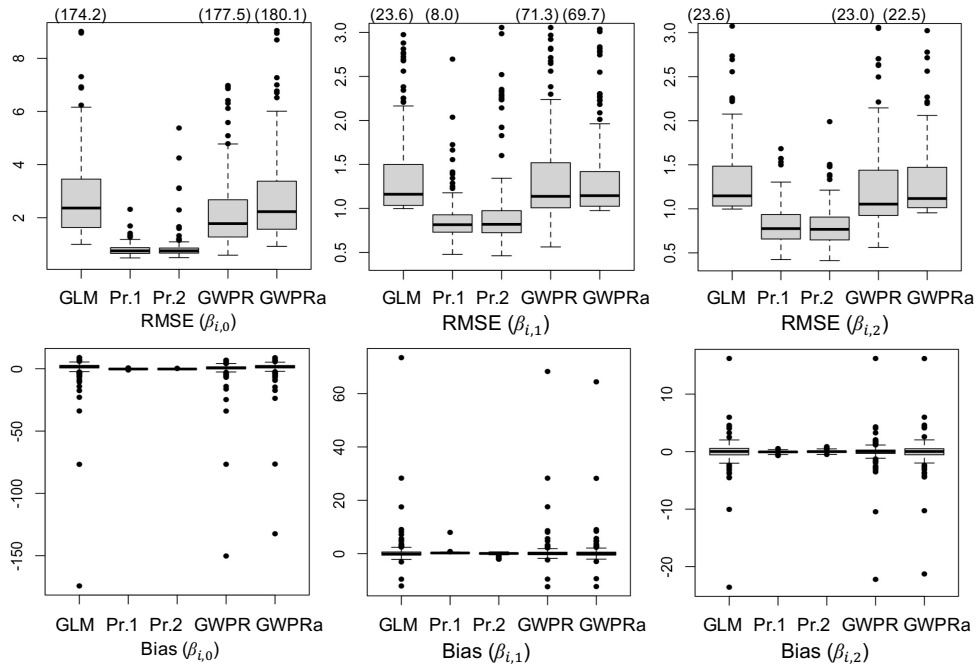
106 In the simulation, each model is estimated 200 times and the root mean squared error
 107 (RMSE) and the bias for the SVCs $\{\beta_{i,0}, \beta_{i,1}, \beta_{i,2}\}$ are evaluated. Note that, in the conference,
 108 we will also perform simulations assuming spatially dependent covariates.

109 3.2 Result

110 Figures 2 and 3 display the boxplots for the RMSE and bias for the estimated spatially varying
 111 coefficients under the two scenarios. The results in the two scenarios are similar. GLM,
 112 GWPR, and GWPRa tend to have large RMSE and bias values. The standard GLM-based
 113 approach including GWPR is found to be unstable. By contrast, our proposed models have
 114 considerably smaller RMSE and bias values than GLM, GWPR, and GWPRa across cases.
 115 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.



■ **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.

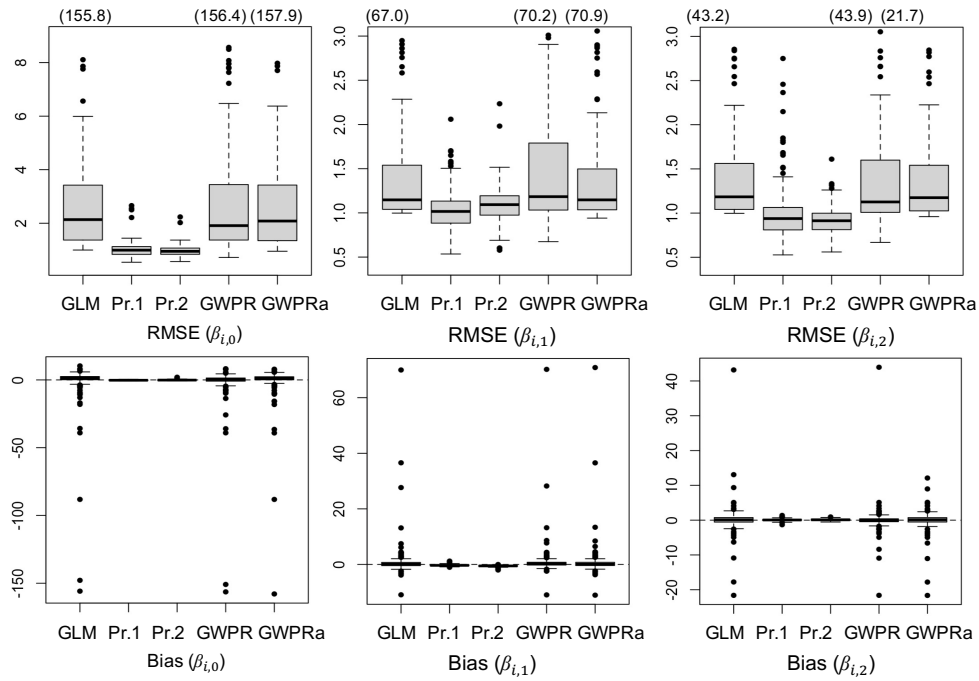
116 (Proposal 2), 2.126 (GWPR), and 2.133 (GWPRa). The result suggests that our approximate
 117 GWPR is more stable and accurate than the usual GLM-based approaches.

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

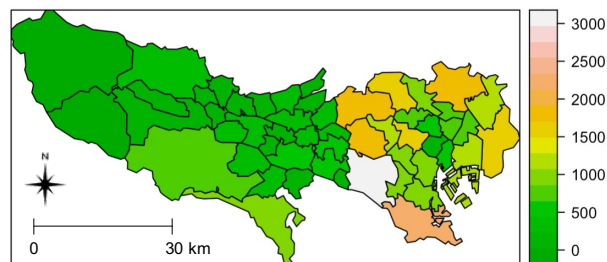
123 4 Application to COVID-19 data

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

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

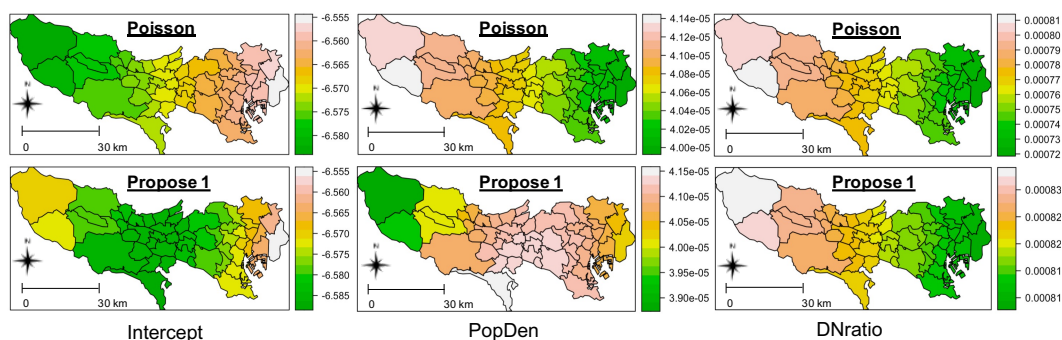


■ **Figure 3** Boxplots of the RMSE and biases for the coefficients under the scenario B ($\sigma^2 = 10, \beta_0 = -2$). Pr.1 and Pr.2 mean Proposal 1 and Proposal 2. If the maximum RMSE value exceeds the displayed boundary in each panel, the maximum value is described above the panel.



■ **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)

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

141 5 Summary

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

148 — References —

- 149 1 Peter J Diggle, Jonathan A Tawn, and Rana A Moyeed. Model-based geostatistics. *Journal*
150 *of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.
- 151 2 A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted*
152 *regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.
- 153 3 A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Geographical and temporal weighted
154 regression (gtwr). *Geographical Analysis*, 47(4):431–452, 2015.
- 155 4 A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted
156 regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265,
157 2017.
- 158 5 Alireza Hadayeghi, Amer S Shalaby, and Bhagwant N Persaud. Development of planning
159 level transportation safety tools using geographically weighted poisson regression. *Accident*
160 *Analysis & Prevention*, 42(2):676–688, 2010.
- 161 6 Daisuke Murakami and Tomoko Matsui. Improved log-gaussian approximation for over-
162 dispersed poisson regression: application to spatial analysis of covid-19. *arXiv preprint*
163 *arXiv:2104.13588*, 2021.
- 164 7 Tomoki Nakaya, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. Geo-
165 graphically weighted poisson regression for disease association mapping. *Statistics in medicine*,
166 24(17):2695–2717, 2005.
- 167 8 JMC Santos Silva and Silvana Tenreiro. On the existence of the maximum likelihood estimates
168 in poisson regression. *Economics Letters*, 107(2):310–312, 2010.
- 169 9 David C Wheeler. Diagnostic tools and a remedial method for collinearity in geographically
170 weighted regression. *Environment and Planning A*, 39(10):2464–2481, 2007.