

Improved log-Gaussian approximation for over-dispersed Poisson regression: application to spatial analysis of COVID-19

Daisuke Murakami^{1*}, Tomoko Matsui²

¹Department of Statistical Data Science, Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

²Department of Statistical Modeling, Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

*Corresponding author

E-mail: dmuraka@ism.ac.jp

Abstract

In the era of open data, Poisson and other count regression models are increasingly important. Provided this, we develop a closed-form inference for an over-dispersed Poisson regression, especially for (over-dispersed) Bayesian Poisson wherein the exact inference is unobtainable. The approach is derived via mode-based log-Gaussian approximation. Unlike closed-form alternatives, it remains accurate even for zero-inflated count data. Besides, our approach has no arbitrary parameter that must be determined a priori. Monte Carlo experiments demonstrate that the estimation error of the proposed method is a considerably smaller estimation error than the closed-form alternatives and as small as the usual Poisson regressions. We obtained similar results in the case of Poisson additive mixed modeling considering spatial or group effects. The developed method was applied for analyzing COVID-19 data in Japan. This result suggests that influences of pedestrian density, age, and other factors on the number of cases change over periods.

Introduction

Currently, a wide variety of count data are corrected through sensors and used for smart urban and regional management (see Soomro et al., 2019). For example, in 2020–2021 when the coronavirus disease (COVID-19) spread globally, the daily number of people infected with coronavirus was monitored worldwide, and countermeasures were considered based on the observations (Viner et al., 2020).

Poisson and other regression models for count data have been used for analyzing the number of COVID-19 cases (e.g., Oztig and Askin, 2020; Vokó and Pitter, 2020) or other diseases (e.g., Wakefield, 2007; Lee and Neocleous, 2010). These regression models have also been used in ecology (e.g., Ver Hoef and Boveng, 2007; Lindén and Mäntyniemi, 2011), criminology (e.g., Osgood, 2000; Piza, 2012), and other fields.

Recently, Bayesian Poisson regression, which assumes Poisson distribution for the count data and Gaussian priors for latent variables describing spatial, group, and other effects, is widely used in applied studies. Owing to the lack of contiguity between the Poisson and Gaussian distributions, an approximate inference is necessary for the estimation. Unfortunately, the Markov Chain Monte Carlo method can be slow for large samples. Faster approximations have been developed for count data regression in a context of additive modeling (e.g., Wood, 2011; Rodríguez-Álvarez et al., 2015), mixed effects modeling (Pinheiro and Bates, 2011), and Gaussian process (e.g., Diggle et al., 1998; Rue et al., 2009). These approaches do not provide closed-form solutions.

Exceptionally, Chan and Dong (2011) and Chan and Vasconcelos (2011) proposed closed-form approximations for Poisson regression. Their approaches have the following advantages:

- (i) Easy to implement and extend. The Gaussian process model and other models for Gaussian data are readily transferred for count data modeling.
- (ii) Computationally efficient. Unlike alternatives, numerical optimization is not needed.
- (iii) Poisson regression estimates are unidentifiable or identifiable only weakly for certain data configuration (Silva and Tenreyro, 2010). As we will illustrate later, this property considerably worsen the accuracy of Poisson regression estimates especially for small samples with many zeros. The linear approximation is free from such difficulty and more stable.

Breslow (1984) and El-Sayyad (1973) proposed relevant closed-form approximations, too. Given the current situation wherein a wide range of researchers and practitioners use count data, these fast, stable, and practical approaches will become increasingly important. Unfortunately, these approximations have the following disadvantages:

- (iv) They have poor approximation accuracy for zero-inflated data (i.e., counts with many zeros) as we will demonstrate later. Zero-inflation is a special type of over-dispersion (Ver Hoef and Boveng, 2007). Thus, a closed-form approach accurately describing over-dispersion is needed.
- (v) An arbitrary parameter, which is used to avoid taking the logarithm of zero, must be determined a priori. The value is known to have substantial impact on the modeling result (Bellego and Pape, 2019). A closed-form approach without such an arbitrary parameter is needed.

Given that, we develop a closed-form approximation for the over-dispersed Poisson regression, especially (over-dispersed) Bayesian Poisson regression, that has (i)–(ii) and overcomes (iii)–(iv).

Methods

Improved log-Gaussian approximation

The mode of a log-Gaussian distribution grows slower than the mean whereas the mode and mean of a Poisson distribution grow at the same rate. Therefore, mean-based log-Gaussian approximation, such as the Taylor approximation use in Chan and Dong (2011), can have poor approximation accuracy around the mode, which is the distribution center. Considering the success of Laplace or other mode-based approximations in previous studies, it is reasonable to accurately approximate Poisson distribution around the mode. This study first introduces a mode-based approximation achieving it.

We first consider the following Poisson model for count variables $Y_i | i \in \{1, \dots, N\}$:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = z_i \exp(\mu_i), \quad (1)$$

where μ_i is represents the mean, which may be specified by a regression model with/without a Gaussian prior. z_i is a given offset variable. Y_i is known to have two modes $\{\lambda_i - 1, \lambda_i\}$ for integer-valued λ_i . Later, we will use the center of the two modes $\text{Mode}_c[Y_i] = \lambda_i - 0.5$.

Our objective is to approximate Eq. (1) by using the following log-Gaussian model for variables $y_i | i \in \{1, \dots, N\}$:

$$y_i + c \sim \log N\left(\mu_{i(G)}, \frac{1}{y_i + c}\right) \quad (2)$$

where $\mu_{i(G)}$ represents the mean, and c is a constant required to avoid taking the logarithm of zero. $\frac{1}{y_i + c}$ is an approximate variance for a log-transformed Poisson random deviate (see El-Sayyad, 1973).

We approximate the Poisson model (Eq. (1)) using the log-Gaussian model (Eq. (2)) so that the mode $Mode[y_i]$ under Eq. (2) equals the mode center $Mode_c[Y_i]$ of the Poisson model. The following condition is obtained from the mode matching $Mode_c[Y_i] = Mode[y_i]$:

$$z_i \exp(\mu_i) - 0.5 = \exp\left(\mu_{i(G)} - \frac{1}{y_i + c}\right) - c. \quad (3)$$

Eq. (3) suggests that μ_i and $\mu_{i(G)}$ do not generally have a linear relationship. Exceptionally, they have the following linear relationship if $c = 0.5$:

$$\mu_{i(G)} = \log(z_i) + \mu_i + \frac{1}{y_i + 0.5}. \quad (4)$$

While existing studies have determined c somewhat arbitrary, $c = 0.5$ is found to be a necessary for applying the linear approximation under our assumption.

Let us substitute $c = 0.5$ and Eq. (4) into Eq. (2). Then, we obtain the following log-Gaussian model approximating the Poisson model:

$$y_i + 0.5 \sim \text{LogN}\left(\log(z_i) + \mu_i + \frac{1}{y_i + 0.5}, \frac{1}{y_i + 0.5}\right). \quad (5)$$

By organizing Eq. (5), we have the following model:

$$\log(y_i^*) \sim N\left(\mu_i, \frac{1}{y_i + 0.5}\right), \quad (6)$$

where $y_i^* = \frac{y_i + 0.5}{z_i} \exp\left(-\frac{1}{y_i + 0.5}\right)$. In short, the log-Gaussian distribution approximates the Poisson distribution around the mode center.

Unfortunately, the mode-based approximation is available if only $\lambda_i = E[Y_i] \geq 0.5$, which assures the non-negativity of $Mode_c[Y_i]$. If $\lambda_i < 0.5$, the mode of the Poisson and log-Gaussian distributions behave somewhat differently: the Poisson mode is always zero while the mode of the log-Gaussian distribution gradually converges to zero as λ_i (or μ_i) declines. Mode-based approximation is not suitable in this case. By contrast, the mean of the two distributions both converge to zero as λ_i (or μ_i) approaches zero.

Thus, for $E[Y_i] < 0.5$, we rely on a mean-based approximation based on the following relationship, which is obtained from $E[y_i^*] = \exp\left(\mu_i + \frac{0.5}{y_i + 0.5}\right)$ (see Eq. 6):

$$E[y_i^*] \exp\left(-\frac{0.5}{y_i + 0.5}\right) = E[Y_i]. \quad (7)$$

Eq. (7) implies that, when approximating the Poisson mean using y_i^* , it should be rescaled by multiplying $\exp\left(-\frac{0.5}{y_i+0.5}\right)$. By applying the rescaling for Eq. (6), we have the following mean-based approximation:

$$\log(y_i^{**}) \sim N\left(\mu_i, \frac{1}{y_i + 0.5}\right), \quad (8)$$

where $y_i^{**} = \frac{y_i+0.5}{z_i} \exp\left(-\frac{1+0.5}{y_i+0.5}\right)$. Because of the disadvantage of the mean-based approximation explained in the beginning of this section, we assume Eq. (6) as long as $E[Y_i] \geq 0.5$ (i.e., the mode-center is available) while Eq. (8) otherwise.

Still, $E[Y_i] = \lambda_i$ is unknown a priori. Considering the property that $P(Y_i < 0.5) = P(Y_i = 0)$, we approximate $P(E[Y_i] < 0.5)$ using the ratio r of zero counts in $\{Y_1, \dots, Y_N\}$. Given the approximation, Eqs. (6) and (8) are applied with probabilities r and $1 - r$, respectively. By combining these equations using r , our proposed approximation is formulated as

$$\log(y_i^+) \sim N\left(\mu_i, \frac{1}{y_i + 0.5}\right), \quad (9)$$

where $y_i^+ = \frac{y_i+0.5}{z_i} \exp\left(-\frac{1+0.5r}{y_i+0.5}\right)$, which yields Eq. (6) if $r = 0$ and Eq. (8) if $r = 1$. If all counts are non-zero, the mode-based approximation is applied for all the samples. As the share of zero counts increases, the mean-based approximation is emphasized.

Property of the proposed approximation

Table 1 summarizes closed-form approximations for the Poisson regression models. These methods perform approximations through the estimation of a linear regression model using the log-transformed explained variables and the inverse weight based on y_i . These practical methods will be useful for not only researchers but also practitioners. However, existing methods are accurate only for a moderate to large μ_i (Chan and Vasconcelos, 2011). In other words, they should not be used for counts with many zeros. Besides, the c parameter, which has a considerable impact on analysis result, must be determined a priori (see the Introduction section). These drawbacks inhibit the practical use of these approximations.

Table 1: Closed-form approximations for the Poisson regression model Eq. (1). c is a tuning parameter that must be determined a priori. For the offset variable, $z_i = 1$ is assumed.

Method	Explained variables (pseudo-data)	Weight	Value of the tuning parameter c used
Posterior approx. (EL-Sayyad, 1973)	$\log(y_i + c)$	$\frac{1}{y_i + c}$	0.0
Taylor approx. (Chan and Dong, 2011) Log-Gamma approx. (Chan and Vasconcelos, 2011)	$\log(y_i + c) - \frac{c}{y_i + c}$	$\frac{1}{y_i + c}$	1.0
Our approximation	$\log(y_i + 0.5) - \frac{1 + 0.5r}{y_i + 0.5}$	$\frac{1}{y_i + 0.5}$	

In contrast, our method does not have any unknown tuning parameters. Because of the mode matching, the proposed method accurately approximates the mode of the Poisson distribution irrespective of μ_i . As we will show later, this property dramatically improves the approximation accuracy for zero-inflated count data.

Note that our mode-matching method is akin to the Laplace approximation, which is based on the mode-matching of a Gaussian distribution and the target distribution. Considering studies demonstrating the accuracy of the Laplace approximation, our mode-based approach is expected to be accurate as well.

Results: Monte Carlo experiments

Case 1: Basic over-dispersed Poisson regression model

This section compares the estimation accuracy of the proposed approximation (Proposed) with standard Poisson regression (Poisson), over-dispersed Poisson regression (odPoisson), and negative binomial regression alternatives (NB). We also compare ours with the posterior approximation of EL-Sayyad (1973) (Posterior) and the Taylor approximation of Chan and Vasconcelos (2011) (Taylor).

The simulated count data y_i is generated from the over-dispersed Poisson regression with mean λ_i and the overdispersion parameter σ^2 :

$$y_i \sim \text{odPoisson}(\lambda_i, \sigma^2), \quad \lambda_i = \exp(\beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2), \quad (10)$$

where $x_{i,1}$ and $x_{i,2}$ are generated from standard normal distributions $N(0, 1)$, and $\{\beta_1, \beta_2\} = \{2.0, 0.5\}$. We refer to β_1 as a strong and β_2 as a weak coefficient. $\sigma^2 = 1$ implies the standard Poisson regression without over-dispersion while $\sigma^2 > 1$ means over-dispersion. The β_0 parameter implicitly controls the ratio of zero counts; a smaller β_0 value yields more zero counts.

The coefficient estimation accuracy is compared across models while varying $\beta_0 \in \{-2, -1, 0, 1, 1\}$, $\sigma^2 \in \{1, 5\}$, and $N \in \{50, 200\}$. In each case, the simulations were iterated 500 times and the root mean squared error (RMSE) and the mean bias are evaluated:

$$RMSE[\beta_k] = \sqrt{\frac{1}{N} \sum_{iter=1}^{500} (\hat{\beta}_k^{(iter)} - \beta_k)^2}, \quad Bias[\beta_k] = \frac{1}{N} \sum_{iter=1}^{500} (\hat{\beta}_k^{(iter)} - \beta_k) \quad (11)$$

where $\hat{\beta}_k^{(iter)}$ is the estimated β_k in the $iter$ -th iteration.

The evaluated RMSE and bias values are plotted in Figs 1 and 2 in a case without overdispersion $\sigma^2 = 1.0$ whereas Figs 3 and 4 in cases with overdispersion $\sigma^2 = 5.0$. Posterior and Taylor tend to have large RMSEs and biases across cases, and the errors inflate if y_i has many zero values (i.e., small β_0). These approximations do not work for zero-inflated count data. In contrast, the RMSE values for the proposed method are as small as the exact Poisson and odPoisson specifications across cases. Poisson, odPoisson, and NB have very large RMSE values if the counts are over-dispersed ($\sigma^2 = 5.0$) and have many zero values (small β_0); this is attributable to the identification problem explained in the ‘‘Introduction’’ section. Proposed, which does not suffer from this problem, advantageous in terms of stability. Although the bias of the proposed method tends to be larger than that of Poisson and odPoisson, the value is still considerably smaller than that of Posterior and Taylor. It is suggested that the proposed method estimates regression coefficients accurately, especially for zero-inflated small data. Fig 5 shows the coefficient standard error (SE) estimates. If y_i has less zero values (i.e., large β_0), the SEs estimated from the proposed method are similar to those of odPoisson, especially in the over-dispersion case. In contrast, the proposed method indicated smaller SEs if y_i had many zero values (i.e., small β_0). Considering the better estimation accuracy of the proposed method for small β_0 , the smaller SEs might suggest the stability of the proposed method.

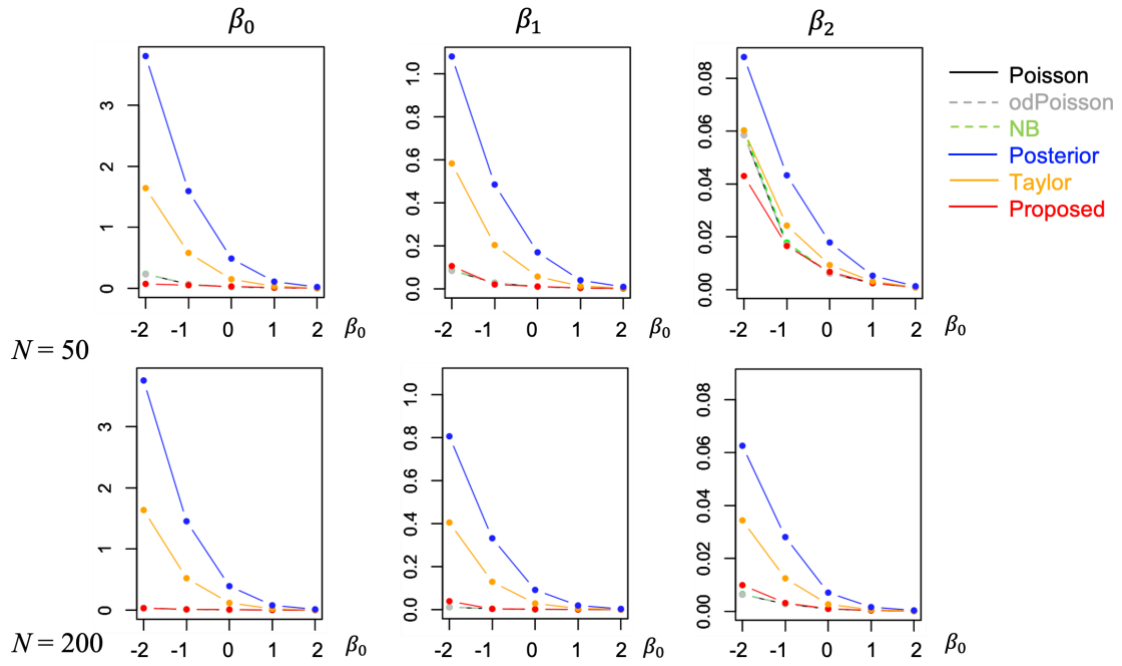


Figure 1: RMSE of the regression coefficients in cases without overdispersion ($\sigma^2 = 1.0$)

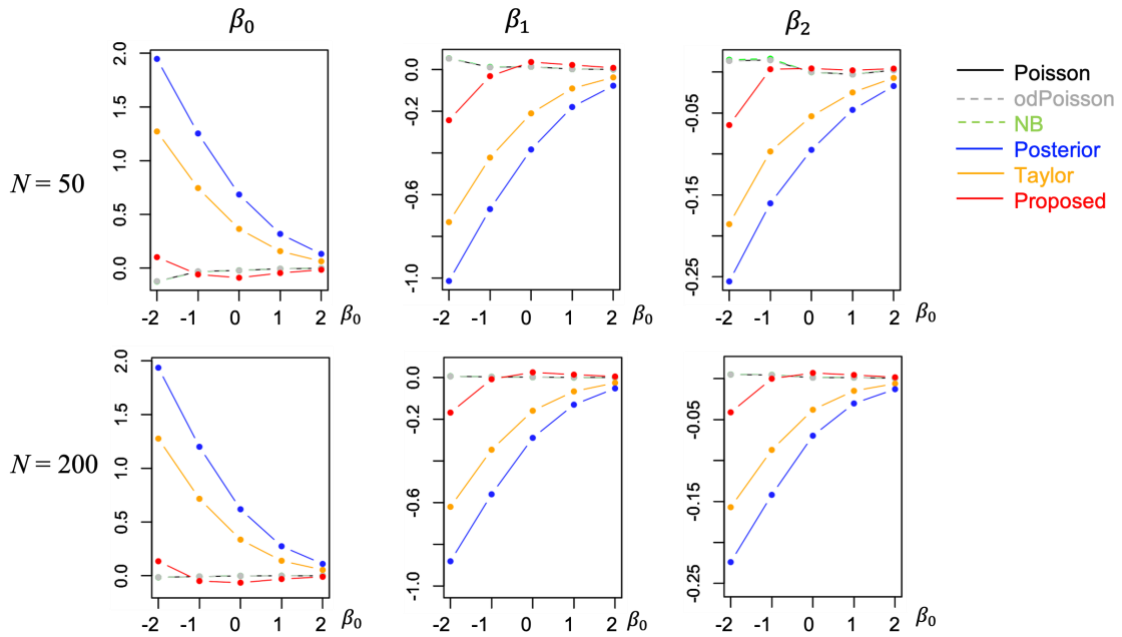


Figure 2: Bias of the regression coefficients in cases without overdispersion ($\sigma^2 = 1.0$)

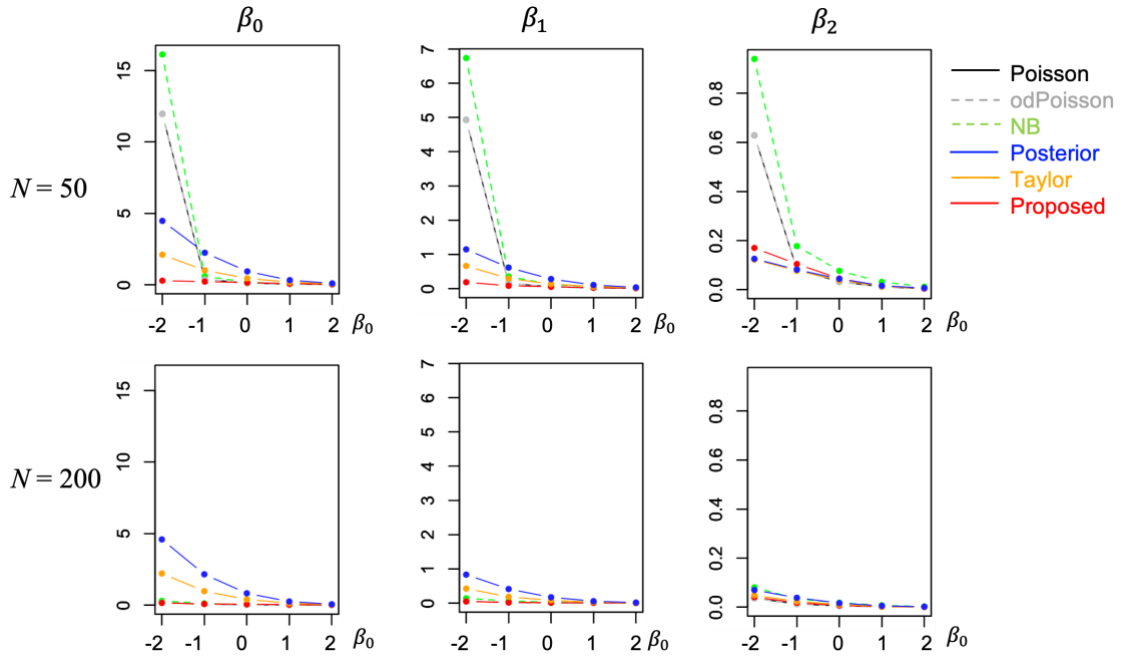


Figure 3: RMSE of the regression coefficients in cases with overdispersion ($\sigma^2 = 5.0$)

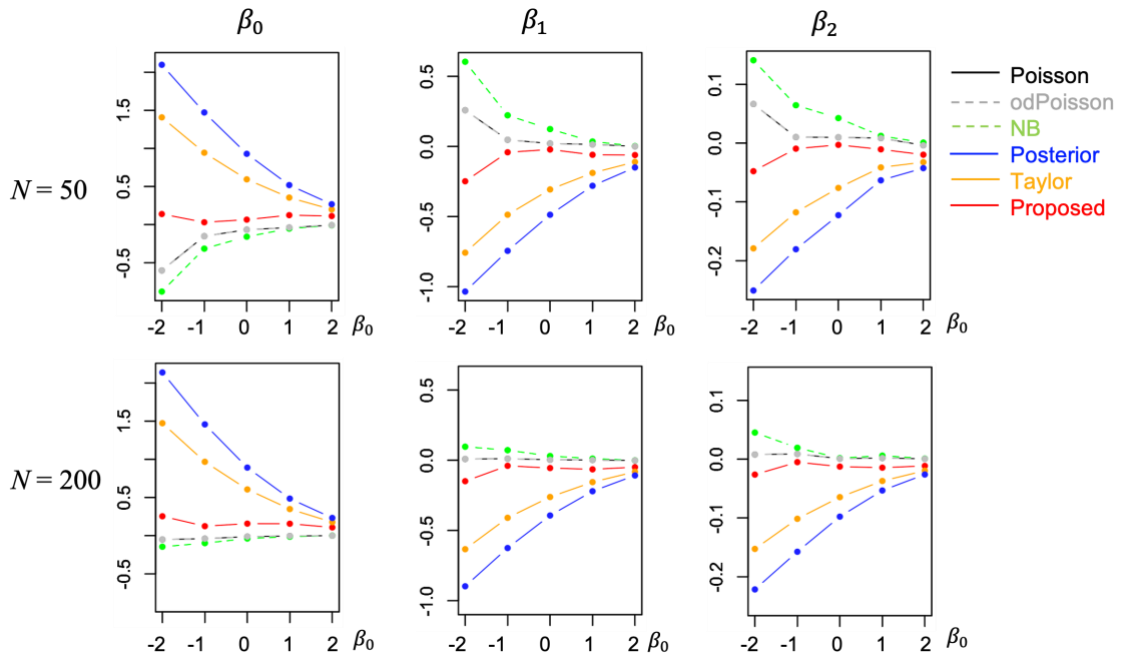


Figure 4: Bias of the regression coefficients in cases with overdispersion ($\sigma^2 = 5.0$)

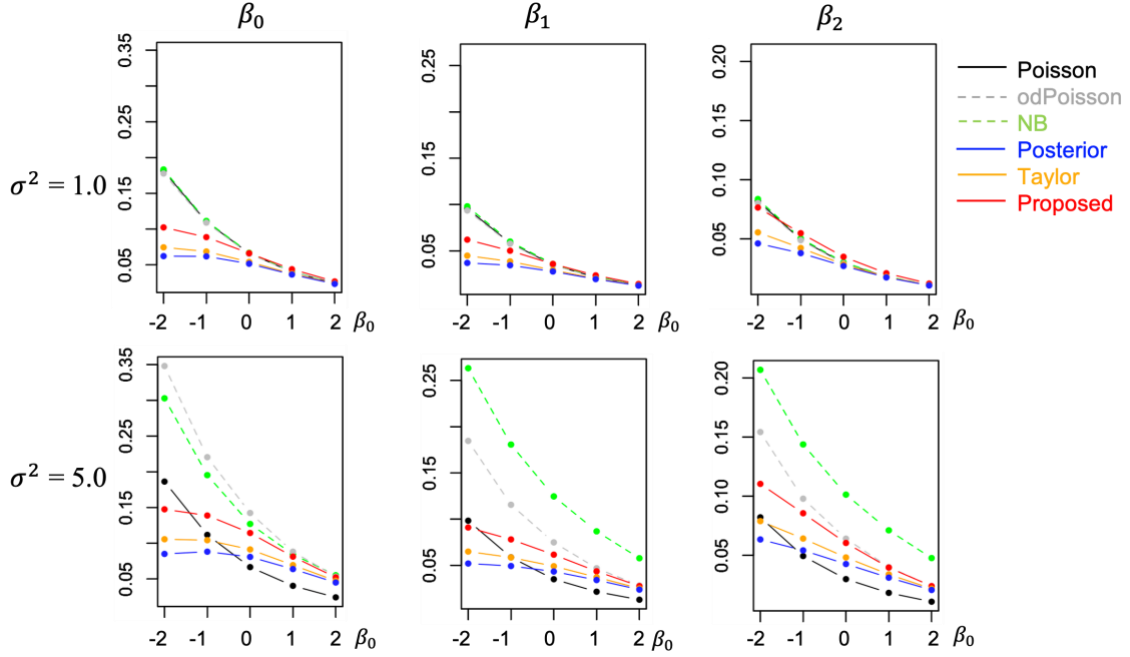


Figure 5: Means of the coefficient standard errors ($N = 200$)

Case 2: Model with spatial effects

To verify the expandability of the proposed model, this section applies the proposed method to estimate a spatial regression model, which has been widely used to analyze spatial phenomena in the environment, economy, and epidemic. We consider the following model:

$$y_i \sim \text{odPoisson}(\lambda_i, \sigma^2), \quad \lambda_i = \exp(\beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + s_i), \quad (12)$$

where $\{\beta_1, \beta_2\} = \{2, 0.5\}$ and $\sigma^2 = 5$. s_i is a process capturing a spatially dependent pattern of the data. Following Murakami and Griffith (2015), it is modeled by assuming random effects whose spatial dependence exponentially decays relative to the Euclidean distance between the geometric centers of the two zones. Equation (11) is an over-dispersed Poisson mixed-effects model (MEM) that considers spatial dependence. The model is estimated by applying the maximum likelihood (ML) estimation for the Poisson MEM (Poisson), an over-dispersed Poisson MEM (odPoisson), the Taylor approximate Poisson MEM (Taylor), and our specification (Proposed). Taylor and Proposed fitted linear MEMs using the transforming explained variables and weight variables (see Table 1). All models were estimated using the R package `mgcv` (<https://cran.r-project.org/web/packages/mgcv/index.html>).

We assumed $\beta_0 \in \{-2, -1, 0, 1, 1\}$ and $N \in \{50, 200\}$. In each case, the models were estimated 500 times, and the estimation accuracies were compared. Figs 6 and 7 display the estimated

RMSEs and biases, respectively. When $N = 50$, odPoisson took extremely large RMSEs due to its singular estimation. Poisson and Taylor also had large RMSEs. In contrast, the proposed method tends to have smaller RMSE values. The proposed method may be a better choice for small samples. Even for $N = 200$, the RMSEs and biases of Proposed were as small as those of Poisson and odPoisson. The estimation accuracy of the proposed method was verified in the case of spatial regression.

Fig 8 compares the coefficient standard errors. For a large β_0 , the SEs obtained from the proposed method are similar to odPoisson, which proposes approximates. While odPoisson has larger SEs for small β_0 , it is attributable to the singular fit. The proposed method stabilizes it and make the SEs small. Finally, Fig 9 compares the estimation accuracy for the spatially dependent process s_i . This figure shows that the proposed method tends to estimate the process more accurately than alternatives.

Note that we performed another Monte Carlo experiment assuming group effects, which model heterogeneity across groups instead of the spatially dependent effects. As summarized in Appendix S1, the RMSEs and biases are as small as Poisson and odPoisson for $N = 200$ and smaller for $N = 50$. The SEs are similar to odPoisson for large β_0 and smaller for small β_0 .

In short, the proposed method provides an accurate and stable approximation for an over-dispersed Poisson MEM.

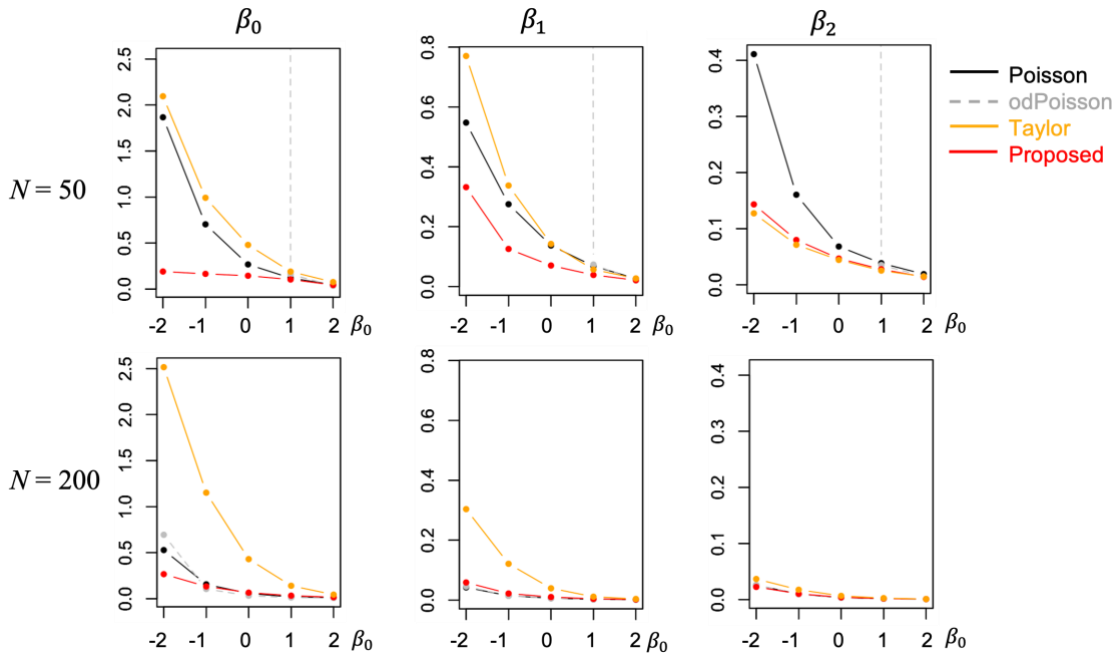


Figure 6: RMSE of the regression coefficients (model with spatial effects)

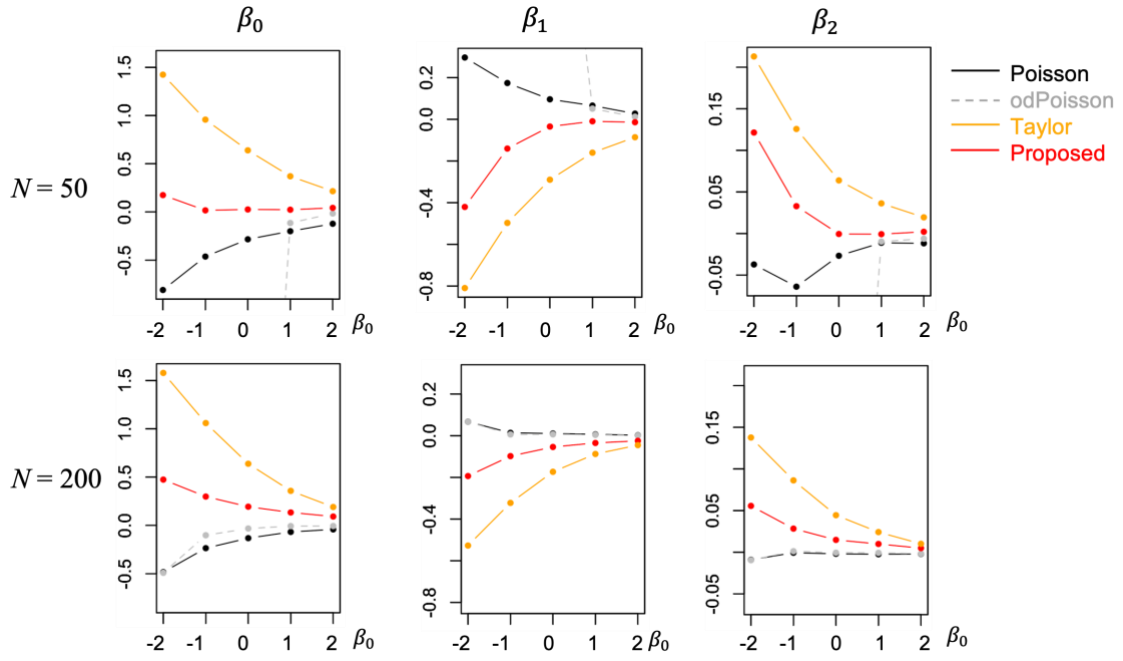


Figure 7: Bias of the regression coefficients (model with spatial effects)

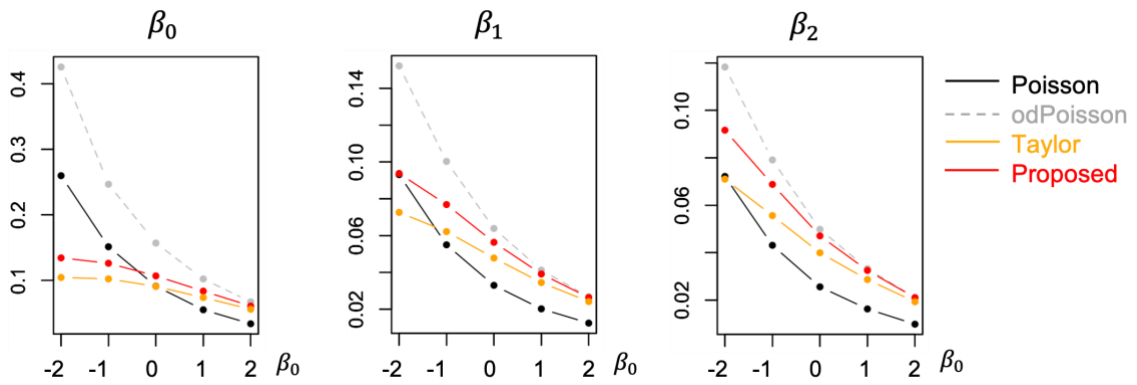


Figure 8: Means of the coefficient standard errors ($N = 200$)

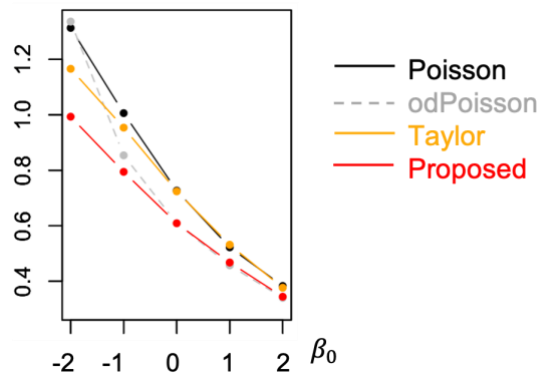


Figure 9: RMSE of the estimated spatial effects

Results: COVID-19 analysis

Outline

This section employs the developed approximation to an analysis of the COVID-19 (coronavirus disease 2019) pandemic. Since the first case was detected in Wuhan, China, in December 2019, the coronavirus spread. As of February 1, 2021, the cumulative number of confirmed cases is 103.41 million, while the confirmed death toll is 2.25 million. To achieve effective infection control for not only COVID-19 but also pandemics/endemics in the future, it is important to investigate the determinants behind the disaster.

Fig 10 plots the number of daily cases in Japan between February 1, 2020, and January 29, 2021. The number peaked around April 2020, August 2020, and January 2021, respectively. Based on the time trend, we refer to February 1 – May 31 as the first wave, June 1 – September 30 as the second wave, and October 1 – January 29, 2021, as the third wave. Fig 11 displays the spatial plots of the daily new cases by prefecture. This figure shows the tendency of the number of infected people to become large near Tokyo and Osaka, which are major urban areas.

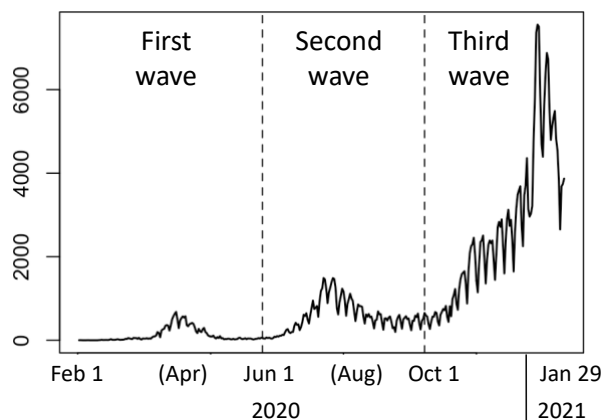


Figure 10: Daily number of cases across Japan

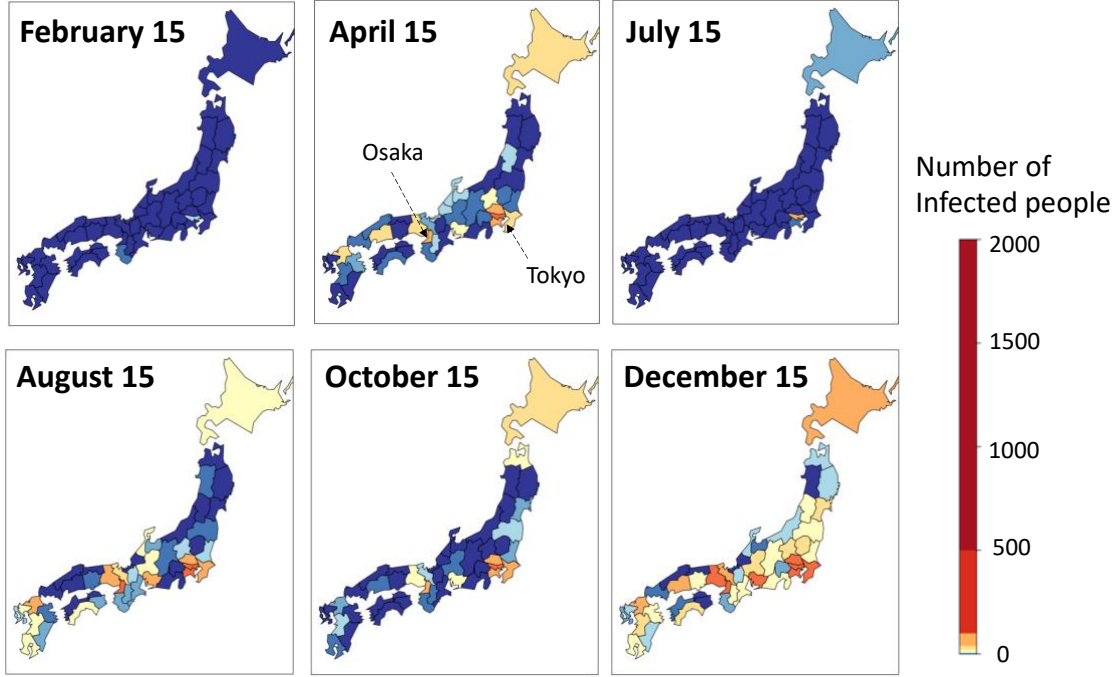


Figure 11: Number of cases by prefecture

We performed a regression analysis exploring the determinants of the increase/decrease in each wave. The explained variables are the number of daily cases by prefecture by generation every decade. The sample sizes were 51,336, 50,508, and 50,094 for the three waves respectively. Unfortunately, the counts were zero-inflated; 89.0 % (45,696 samples), 77.1 % (38,954 samples), and 49.7 % (24,873 samples) of samples are zeros.

For the zero-inflated COVID-19 data, we fit Eq. (13) approximating an over-dispersed Poisson additive mixed model:

$$\log(y_i^*) = \beta_0 + x_i \beta_1 + \sum_{l=1}^4 g_{i,l} + s_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \frac{\sigma^2}{y_i + 0.5}\right), \quad (13)$$

where $y_i^* = \frac{y_i + 0.5}{z_i} \exp\left(-\frac{1.5}{y_i + 0.5}\right)$, where y_i is the number of daily new cases. β_0 and β_1 are regression coefficients. To scale the mean function according to the population, the offset variable z_i is given by the prefectural population. The explanatory variable x_i is the prefectural pedestrian density by day, which is relative to January 13, 2020 (source: Apple Mobility Trends: <https://covid19.apple.com/mobility>). The density is estimated based on the number of route searches by Apple map users. For further detail, see the source page. $g_{i,l}$ represents the l -th group-wise random effect. We consider the effects by week ($g_{i,1}$), days of the week ($g_{i,2}$), generation ($g_{i,3}$), and prefecture

($g_{i,4}$). In addition, a Moran coefficient-based spatial random effect s_i is included to eliminate residual spatial dependence (Murakami and Griffith, 2015).

The model was estimated using the R package `spmoran` (Version 0.2.1; <https://cran.r-project.org/web/packages/spmoran/index.html>).

Results

Table 1 summarizes the estimated parameters. The estimated coefficients of pedestrian density become positively significant in the second and third waves. Self-restraint was estimated to reduce the number of cases after June. Based on the estimated residual standard error (σ^2), the residual variance was over-dispersed, and the tendency became stronger over time.

Table 1: Parameter estimates. See Figure 12 for the fitting on the number of cases.

	First wave		Second wave		Third wave	
	Est.	t value	Est.	t value	Est.	t value
Const (β_0)	-16.60	-220.24 ***	-16.68	-114.03 ***	-14.88	-115.94 ***
Pedestrian density (β_1)	-0.38	-7.68 ***	0.23	3.48 ***	0.15	3.34 ***
Residual S.E. (σ^2)	1.40		1.67		1.91	
Log-Likelihood	-71375		-100418		-93152	

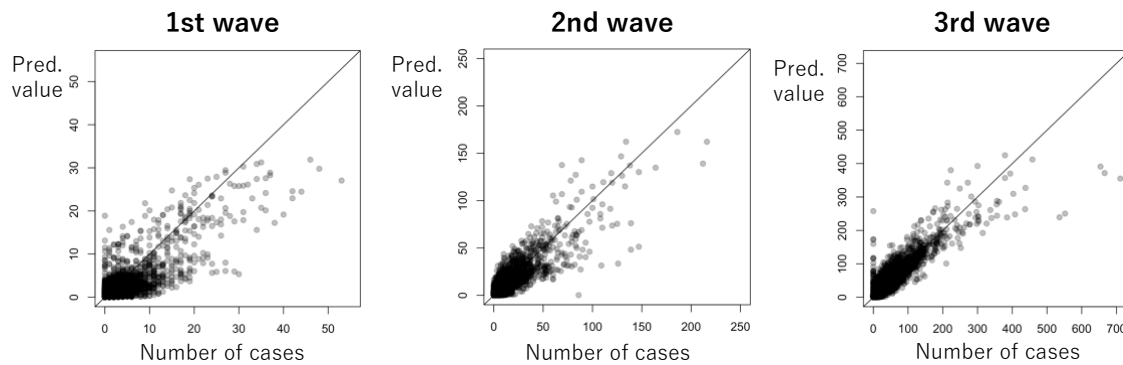


Figure 12: Comparison of the observed and predicted number of cases

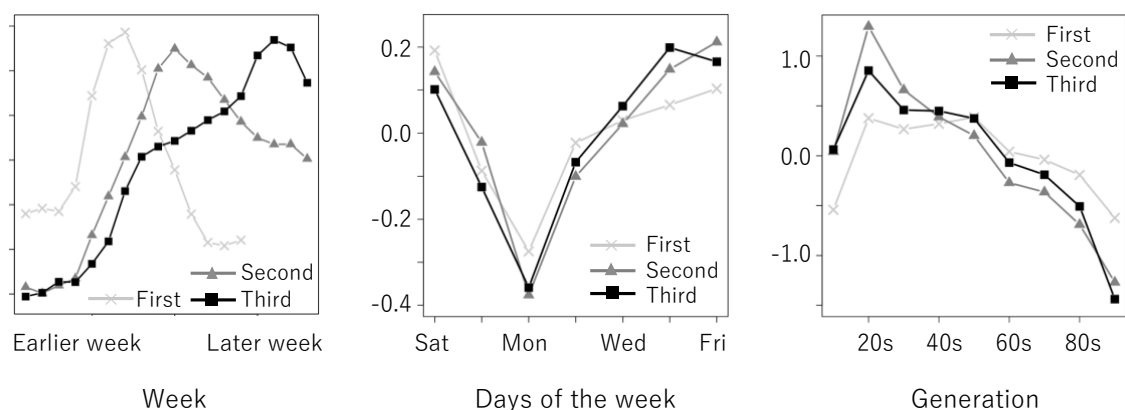


Figure 13: Estimated group effects (week, days of the week, generation)

Fig 13 plots the estimated group effects by week, days of the week, and generation. The estimated week-wise effects show that the increase in cases lasts longer in later waves. Control of the infection spread might be getting more difficult over waves. Regarding the days of the week, Monday has the lowest while Thursday, Friday, and Saturday have higher values. The difference is attributable to some business reasons such as the closing of hospitals and PCR test sites. The estimated generation effects have considerable differences across waves. In the first wave, people who are in the working generation (the 20s - 50s) tend to be infected. Commuting and/or meeting in the office might trigger the infection. In the second wave, the 20's group has a strong tendency of being infected as compared to the elders, therefore, more self-restriction is needed. In the third wave, not only the 20s but also the 30s - 50s have high chances of being infected. Infection might spread again across the working generation.

Fig 14 plots the estimated prefecture-wise independent effects and spatially dependent effects. The former estimates local hotspots while the latter, global hotspots. The estimated prefecture-wise effects suggest that prefectures including major cities (Tokyo, Osaka, Fukuoka) and Hokkaido are local hotspots. More countermeasures might be required in these prefectures. On the other hand, based on the estimated spatially dependent effects, there is a global hotspot around Tokyo, and the influences grow over waves. Control of the infection spread from Tokyo might have been important to mitigate the third wave.

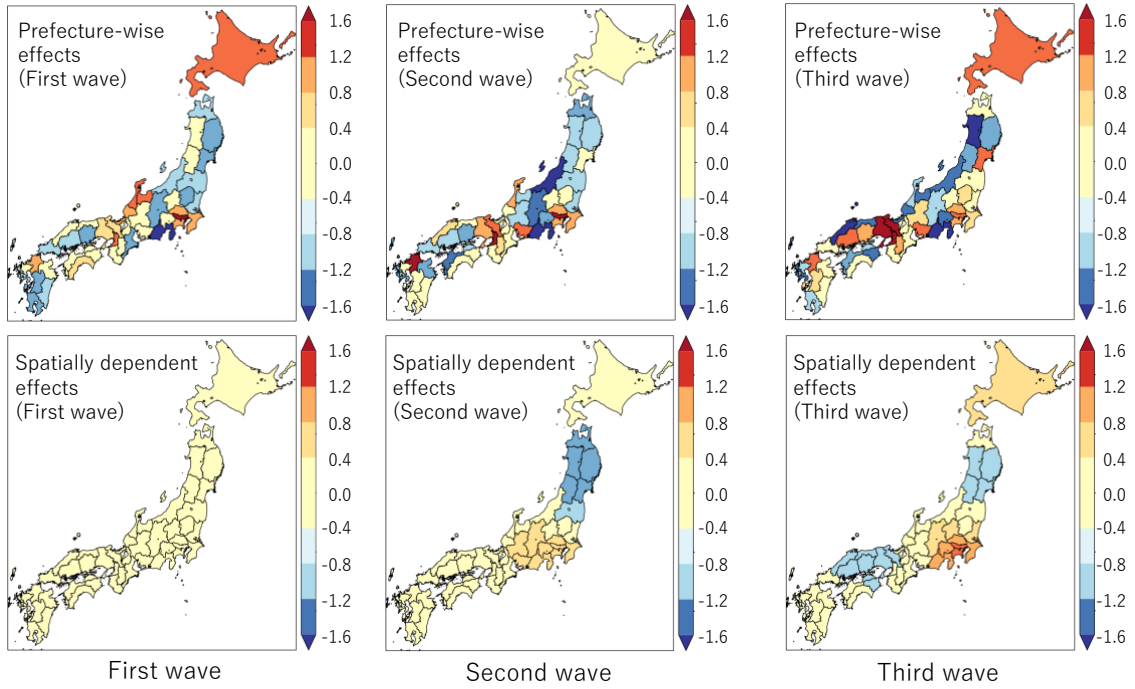


Figure 14: Estimated group effects by prefecture (top) and spatially dependent effects (bottom).

Discussion

This study develops a practical log-Gaussian approximation for Poisson regression models. Considering its simplicity, computational efficiency, and applicability to zero-inflated data, it will be useful for researchers as well as practitioners.

Exploring the expandability of our approach is an important future task. For example, our approach might be useful for spatial and spatiotemporal interpolation of count data by combining it with Gaussian process models without additional computation and implementation costs. Our approach might also be useful for fast count data assimilation by combining it with a state-space model. Improving approximation accuracy would also be an interesting research endeavor. If a closed-form approximation comparable with non-closed-form alternatives regarding not only RMSE but also bias is developed, it will be extremely useful in practice.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18H01556 and 18H03628, and JST-Mirai Program Grant Number JP1124793, Japan.

Reference

1. Breslow NE. Extra-Poisson variation in log-linear models. *J Roy Stat Soc C* 1984; 33(1): 38-44.
2. Bellego C, Pape LD. Dealing with logs and zeros in regression models. SSRN: 3444996 [Preprint]. 2019 Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3444996
3. Chan AB, Dong D. Generalized Gaussian process models. Proceedings of the IEEE conference on computer vision and pattern Recognition 2011; 5995688: 2681-8.
4. Chan AB, Vasconcelos N. Counting people with low-level features and Bayesian regression. *IEEE Trans Image Process* 2011; 21(4): 2160-77.
5. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *J Roy Stat Soc C* 1998; 47(3): 299-350.
6. El-Sayyad GM. Bayesian and classical analysis of Poisson regression. *J Roy Stat Soc B* 1973; 35(3): 445-51.
7. Lee D, Neocleous T. Bayesian quantile regression for count data with application to environmental epidemiology. *J Roy Stat Soc C* 2010; 59(5): 905-20.
8. Lindén A, Mäntyniemi S. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* 2011; 92(7): 1414-21.
9. Murakami D, Griffith DA. Random effects specifications in eigenvector spatial filtering: a simulation study. *J Geogr Syst* 2015; 17(4): 311-331.
10. Osgood DW. Poisson-based regression analysis of aggregate crime rates. *J Quant Criminol* 2000; 16(1): 21-43.
11. Oztig LI, Askin OE. Human mobility and coronavirus disease 2019 (COVID-19): a negative binomial regression analysis. *Public health* 2020; 185: 364-367.
12. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. New York: Springer; 2000.
13. Rodríguez-Álvarez MX, Lee DJ, Kneib T, Durbán M, Eilers P. Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Stat Comput* 2015(5); 25: 941-957.
14. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Roy Stat Soc B* 2009; 71(2): 319-92.
15. Silva JS, Tenreiro S. On the existence of the maximum likelihood estimates in Poisson regression. *Econ Lett* 107(2): 310-312
16. Soomro K, Bhutta MN M, Khan Z, Tahir MA. Smart city big data analytics: An advanced review. *Wiley Interdiscip. Rev Data Min Knowl Discov* 2019(5); 9: e1319.

17. Ver Hoef JM, Boveng PL. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology* 2007; 88(11): 2766-2772.
18. Viner RM, Russell S, Croker H, Packer J, Ward J, Stansfield C, et al. School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *Lancet Child Adolesc Health* 2020; 4(5): 397-404.
19. Vokó Z, Pitter JG. The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis. *GeroScience* 2020; 42(4): 1075-1082.
20. Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics* 2007; 8: 158-183.
21. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J Roy Stat Soc B* 2011; 73(1): 3-36.
22. Wood SN. *Generalized additive models: an introduction with R*. Boca Raton: CRC press; 2017.

Supporting information

Appendix S1: Monte Carlo experiments assuming group-wise random effects

This section examines the estimation accuracy of the proposed model, assuming a Poisson MEM with group-wise random effects. The synthetic data is generated by an over-dispersed Poisson MEM defined as

$$y_i \sim \text{odPoisson}(\lambda_i, \sigma^2), \quad \lambda_i = \exp(\beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + g_i), \quad (\text{A-1})$$

where $\{\beta_1, \beta_2\} = \{2, 0.5\}$ and $\sigma^2 = 5$. g_i represents the group-wise random intercept. Samples were randomly assigned to three groups for $N = 50$ (16.7 samples per group) and 10 groups for $N = 200$ (20 samples per group). The value of g_i was sampled from a standard normal distribution.

Following the Monte Carlo experiments section, the Poisson MEM (Poisson), an over-dispersed Poisson MEM (odPoisson), the Taylor approximate Poisson MEM (Taylor), and our specification (Proposed) are estimated by the likelihood maximum, and their estimation accuracies are compared while varying $\beta_0 \in \{-2, -1, 0, 1, 1\}$ and $N \in \{50, 200\}$. In each case, the models were estimated 500 times.

Figs A-1 and A-2 summarize the estimated RMSEs and biases. Proposed tends to have smaller RMSEs than Poisson, odPoisson, and Taylor in cases with small samples ($N = 50$) while as small as Poisson and odPoisson in cases with $N = 200$. Although Bias tends to be larger than Poisson

and odPoisson, they are much smaller than Taylor. Fig A-3 compares SEs. The result is consistent with another simulation assuming spatial dependence (see the Monte Carlo experiment section); the SEs estimated from Proposed are similar for odPoisson for large β_0 and smaller for small β_0 . Regarding the group effects estimates, the accuracy of Proposed is as same as odPoisson and better than Taylor and Poisson.

Overall, the results suggest that the proposed method accurately approximates the over-dispersed Poisson MEM with group effects.

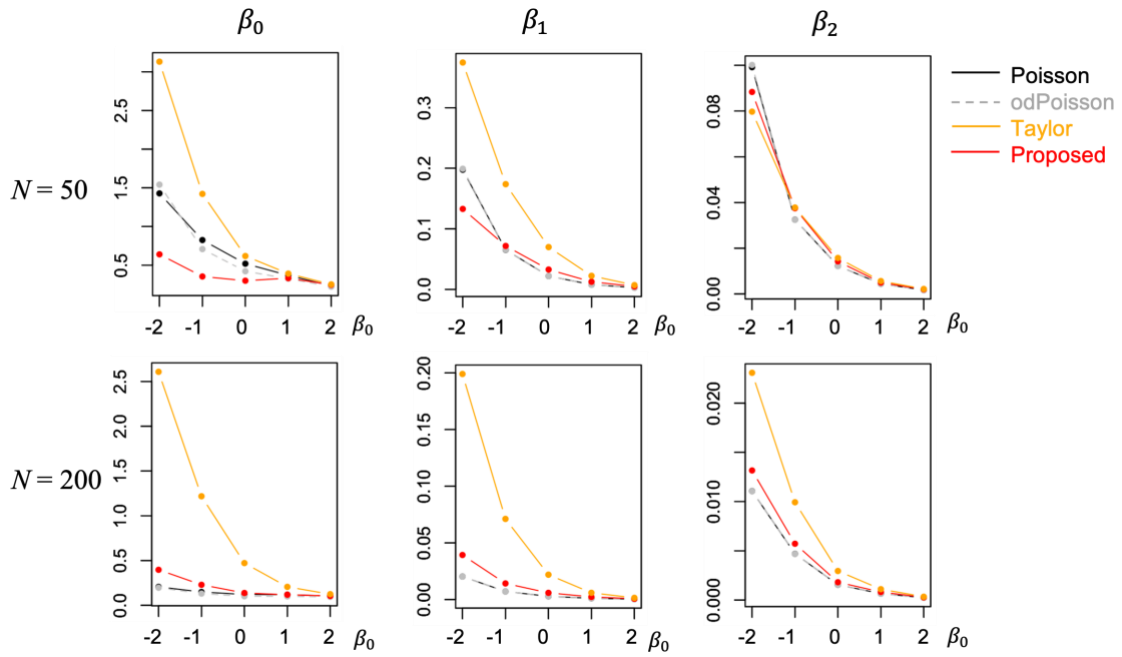


Figure A-1: RMSE of the regression coefficients (model with group effects)

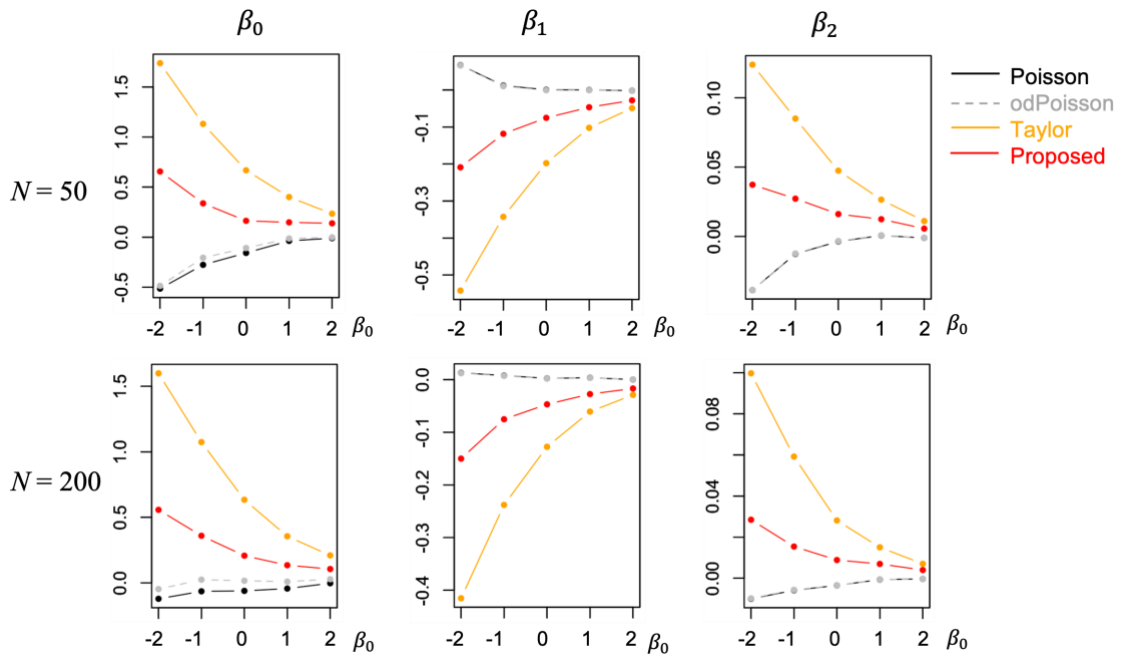


Figure A-2: Bias of the regression coefficients (model with group effects)

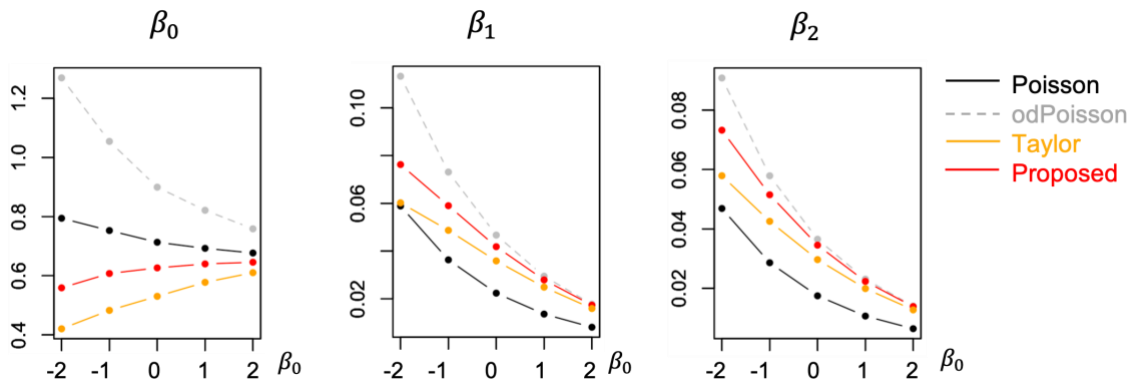


Figure A-3: Means of the coefficient standard errors ($N = 200$)

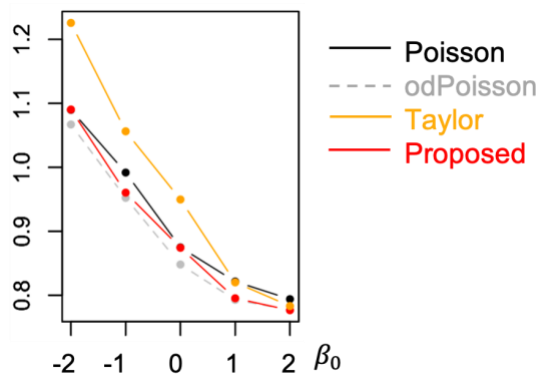


Figure A-4: RMSE of the estimated group effects