MSI Journal of AI and Technology

Assessing the Tolerance of Overdispersion in Geographically Weighted Poisson Regression for Spatial Count Data

Muhammad Nur Aidi[1*], Indahwati[2], Puput Cahya Ambarwati[3]

[1*,2] Lecturers at School of Data Sciences, Mathematics and Informatics, IPB University, Indonesia

[3] Alumny master's degree at School of Data Sciences, Mathematics and Informatics, IPB University, Indonesia

*** Correspondence: Muhammad Nur Aidi**

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 22-April-2025

Accepted: 04-May-2025

Published: 08-May-2025

This article is published by **MSI Publishers** in **MSI Journal of AI and Technology**

ISSN: xxxx-xxxx (Online)

Volume: 1, Issue: 1 (April-Jun) (2025)

ABSTRACT: Count data, particularly in spatial contexts, often exhibits overdispersion and spatial heterogeneity, challenging the assumptions of traditional Poisson regression. Geographically Weighted Poisson Regression (GWPR) extends Poisson regression by accommodating spatial variability in regression parameters, but it assumes equidispersion—an assumption frequently violated in practice. An alternative, the Geographically Weighted Negative Binomial Regression (GWNBR), accounts for overdispersion but is computationally intensive. This study evaluates the robustness of GWPR under varying levels of overdispersion through simulation. Data were generated across 49 spatial locations with two explanatory variables and three levels of overdispersion: negligible, moderate, and severe. Root Mean Square Error (RMSE) was used to assess model performance. Results indicate that GWPR performs reliably when overdispersion is low to moderate, with only a marginal increase in RMSE. However, as overdispersion becomes severe, GWPR's accuracy declines substantially. The findings suggest that GWPR remains appropriate for spatial count data under mild overdispersion

but should be replaced by GWNBR in high-overdispersion contexts.
Keywords: *Count data; Overdispersion; Spatial heterogeneity; Geographically Weighted Poisson Regression (GWPR); Geographically Weighted Negative Binomial Regression (GWNBR); Simulation; Root Mean Square Error (RMSE).*

1. INTRODUCTION

Count data refers to data obtained through enumeration or counting. Examples include the annual number of cases of malnourished children, the number of storm events in a given year, or the number of deaths due to lung cancer in a specific year. According to Rogers (1974), count data generally follows a binomial, Poisson, or negative binomial distribution, depending on the variance-to-mean ratio (VMR). If the VMR is less than 1, the data tends to be systematically dispersed, suggesting a binomial distribution. If the VMR equals 1, the data tends to be randomly dispersed and follows Poisson distribution. If the VMR exceeds 1, the data tends to be clustered and follows a negative binomial distribution.

Agresti (2002) states that one regression model commonly used to describe the relationship between a count response variable and explanatory variables is the Poisson regression model. This model assumes that the variance equals meaning, a condition known as equidispersion. Ignoring this assumption can lead to overdispersion, a condition in which the variance exceeds the mean (McCullagh and Nelder, 1989). Using the Poisson regression model under overdispersion results in underestimated standard errors, which may lead to misleading significance tests and potentially incorrect rejections of the null hypothesis. A classical approach to handling overdispersion in Poisson-based models involves deriving a distribution that combines Poisson and gamma distributions, yielding a form similar to the negative binomial distribution.

Spatial data is observational data that includes not only information about the variables of interest but also the coordinates of the locations where the data were collected. Regression analysis involving spatial data requires special attention due to the potential presence of spatial dependence and spatial heterogeneity. One method that can address spatial heterogeneity is *Geographically Weighted Regression*

(GWR). GWR is a point-based linear regression model that provides local parameter estimates for each location where data is collected (Fotheringham et al., 2002). While GWR is suitable for normally distributed response variables with continuous data, in practice, response variables are often counted.

Nakaya et al. (2005) proposed an extension of GWR for Poisson-distributed count data, known as *Geographically Weighted Poisson Regression* (GWPR). However, GWPR often fails to address the issue of overdispersion. An alternative approach for modeling overdispersed count data with spatial variation is the *Geographically Weighted Negative Binomial Regression* (GWNBR), as introduced by Da Silva and Rodrigues (2013).

While GWNBR can handle both overdispersion and spatial heterogeneity, it is computationally more intensive than GWPR due to the inclusion of additional dispersion parameters for each location. Therefore, it is important to assess the extent to which GWPR remains appropriate for modeling spatial count data under various levels of overdispersion. In this study, a simulation approach is employed to examine different levels of overdispersion—ranging from negligible (approaching zero), mild (approaching one), to severe (significantly greater than one)—in order to determine the threshold at which GWPR can still produce reliable results.

The objective of this study is to evaluate the level of overdispersion that can still be adequately modeled using *Geographically Weighted Poisson Regression* (GWPR).

2. LITERATURE

Geographically Weighted Regression (GWR) is a spatial statistical modeling technique used to explore and analyze the relationship between a response variable and one or more explanatory variables, while accounting for spatial heterogeneity or variation across geographic space (Fotheringham et al., 2002). Unlike traditional (global) regression models, which assume that the relationships between variables are constant across the study area, GWR allows these relationships to vary locally by estimating separate regression parameters at each location.

This local modeling approach is particularly useful in spatial data analysis where the assumption of spatial stationarity does not hold—that is, when the strength and direction of relationships differ from one location to another.

The general form of the GWR model is:

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{1i} + \cdots + \beta_p(u_i, v_i)\, x_{pi} + \varepsilon_i$$

Where:

- y_i is the response variable at location i,
- x_{ki} is the value of the k-th explanatory variable at location i,
- $\beta_k(u_i, v_i)$ represents the location-specific regression coefficient for the k-th variable at coordinates (u_i, v_i)
- ε_i is the random error term at location iii,
- p is the number of explanatory variables.

The coefficients $\beta_k(u_i, v_i)$ are estimated using a weighted least squares method, where observations nearer to the location (u_i, v_i) have greater influence on the local parameter estimates than those farther away. This is achieved using a spatial weighting function, often based on a kernel function such as Gaussian or bisquare, and a bandwidth that determines the spatial extent of the weighting. By providing local parameter estimates, GWR helps identify spatial non-stationarity in relationships, offering richer insights into spatial patterns and potentially improving model accuracy in spatial datasets.

According to Fotheringham et al. (2002), selecting the appropriate spatial weighting in Geographically Weighted Regression (GWR) is essential because it determines how much influence nearby locations have on the regression estimates. The spatial weights are generated using kernel functions, which define how the influence of neighboring observations decreases as distance increases.

There are several types of kernel functions commonly used in GWR. The **Gaussian kernel** assigns weights using an exponential function of the squared distance between locations. The **exponential kernel** also uses an exponential function but is

based on the linear distance. The **bisquare kernel** gives higher weight to locations closer to the target point and assigns zero weight to points beyond a certain distance. Similarly, the **tricube kernel** provides smooth weighting that gradually drops to zero outside the defined bandwidth.

The distance between locations is typically calculated using Euclidean distance based on their spatial coordinates. The **bandwidth** (denoted by h) represents the spatial range or window used to determine which nearby points influence the estimation at a given location.

The choice of bandwidth is very important in GWR because it affects the accuracy of the local parameter estimates. A smaller bandwidth includes fewer neighboring points and captures more local variation, while a larger bandwidth smooths the estimates by including more distant observations.

Nakaya et al. (2005) emphasizes that the optimal bandwidth can be determined through a method called **cross-validation**. This method selects the bandwidth that minimizes the prediction error by excluding each observation in turn and comparing the predicted value with the actual value. The optimal bandwidth is the one that results in the lowest cross-validation score. In summary, spatial weighting and bandwidth selection are key components in building a reliable GWR model, as they directly influence how spatial relationships are captured in the regression analysis.

Geographically Weighted Poisson Regression (GWPR) is an extension of the Poisson regression model and Geographically Weighted Regression (GWR). Therefore, GWPR inherits the same assumption as standard Poisson regression, namely that the mean and variance of the response variable are equal (equidispersion). However, in practice, this assumption is often violated because the variance tends to be greater than the mean, a condition known as overdispersion. Applying Poisson regression to overdispersed data can result in underestimated standard errors, which in turn affects hypothesis testing by increasing the likelihood of incorrectly rejecting the null hypothesis (McCullagh and Nelder, 1989).

GWPR estimates model parameters locally, meaning that each location where data is collected has its own set of parameter estimates (Nakaya et al., 2005). This allows

the model to account for spatial heterogeneity and better capture local variations in the relationship between the explanatory variables and the response variable.

Geographically Weighted Negative Binomial Regression (GWNBR) is a statistical method suitable for modeling count data that exhibits overdispersion along with spatial dependence or variability. The GWNBR model is derived from a combination of the Poisson and Gamma distributions, allowing it to account for greater variability in the data. According to Da Silva and Rodrigues (2013), this model estimates regression coefficients and dispersion parameters that vary locally for each geographic location, making it effective in capturing spatial heterogeneity in both the mean and dispersion structures of the data.

3. SIMULATION DATA

The simulation data generation in this study is an extension of the simulation conducted by Liu et al. (2017). The observation locations consist of m×m points, with a distance $l = (\frac{6}{m-1})$ In this case, m=7 and the sample size n=49 observations.

The steps for generating the data are as follows:

1. Determine the coordinates of the locations (u_i, v_i)), where the observation can be expressed as:

   $$(u_i, v_i) = \left(mod\,(i-1, m)\,, mod\,\left(int\left(\frac{i-1}{m}\right), m\right)\right)$$

   for i=1,2,…,m^2. where mod(i−1,m) is the remainder of i−1 divided by m, and $int\left(\frac{i-1}{m}\right)$) is the integer value of $\left(\frac{i-1}{m}\right)$

2. Create a dataset with spatial variability, initialized as follows:

   $$\beta_{0i} = 3 + (u_i, v_i), \beta_{1i} = \frac{1}{2}[3(2-u_i)], \beta_{2i} = 2[9-(6-v_i)]$$

3. Generate the explanatory variables X1 and X2, which follow a uniform distribution between (0, 100).
4. Generate the response variable Y according to equation

   $$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{1i} + \beta_2(u_i, v_i)\, x_{2i} + \varepsilon_i$$

5. Repeat step 4 for 100 iterations.

4. RESULT

In this simulation, the generated data was designed to resemble the structure of GWR data, which is characterized by overdispersion and spatial variability. The simulation involved 49 locations, each with longitudinal and latitudinal coordinates. Two explanatory variables were used, both generated from a uniform distribution with a lower bound of 0 and an upper bound of 1. Although the appropriate model for this case is GWNBR, this study utilizes the GWPR model to evaluate its tolerance for overdispersion.

Three overdispersion conditions were considered: approaching zero (ranging from $1*10^{-06}$ to $1*10^{-42}$), approaching one (ranging from 1.110 to 2.076), and far from one (ranging from 4.610 to 7.450). The computational time for the GWPR model was approximately 5 minutes, whereas the GWNBR model required around 1 hour for each overdispersion condition.

RMSE (Root Mean Square Error) was used as a criterion to assess model performance. A lower RMSE indicates a better-performing model. Figure 1 shows a comparison of the average RMSE for the intercept (beta 0) across the three overdispersion conditions for both the GWPR and GWNBR models.For the GWPR model, the average RMSE under the condition where overdispersion approaches zero was the lowest among all scenarios. The average RMSE values for the conditions where overdispersion approached zero and approached one were relatively close, with values of 1.348 and 1.398 respectively—a difference of 0.050. In contrast, the condition where overdispersion was far from one yielded the highest RMSE of 1.254.

In the GWNBR model, the average RMSE across the three overdispersion conditions was quite consistent, with values of 3.749, 3.718, and 3.735. Overall, the GWPR model demonstrated lower average RMSE values for beta 0 across all three conditions when compared to the GWNBR model.

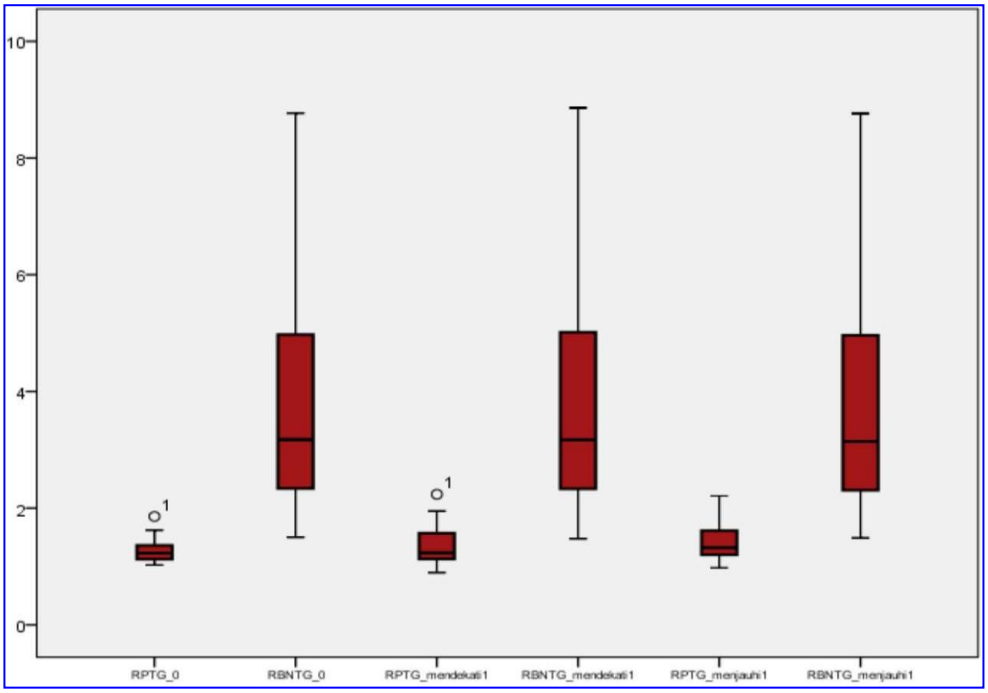

Figure 1. Boxplot Comparison of the Average RMSE for Beta 0

Figure 2 shows a comparison of the average RMSE values for beta 1, which appear to be nearly the same under two conditions: when overdispersion approaches zero and when it approaches one. Based on the calculations, the average RMSE under the condition where overdispersion approaches zero was slightly lower at **1.278**, compared to **1.285** when overdispersion approached one. Meanwhile, the condition where overdispersion was far from one yielded the highest average RMSE value of **6.341**, significantly larger than the other two conditions.

For the **GWNBR** model, the average RMSE values across the three overdispersion conditions were quite similar, namely **2.144**, **2.151**, and **2.153**. Among these, the overdispersion condition approaching zero had the lowest average RMSE, slightly lower than that for overdispersion approaching one (**2.151**) and far from one (**2.153**).

Overall, across all conditions, the GWPR model under the overdispersion condition far from one resulted in the **highest RMSE value** for beta 1. When comparing both models, the average RMSE for **beta 1** in the GWPR model is lower than in the GWNBR model under most conditions, except when overdispersion is far from one, where GWPR performs worse.

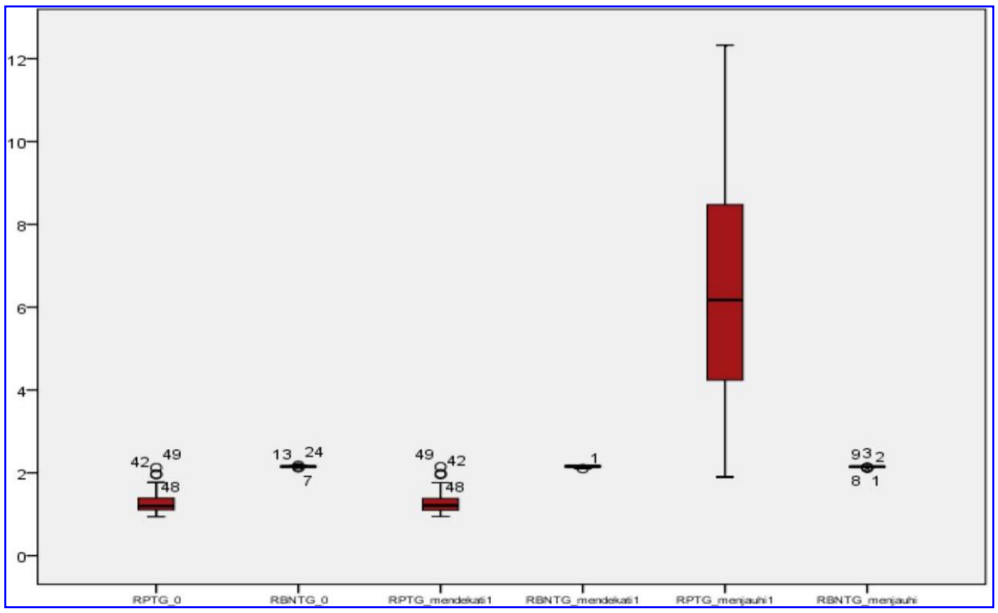

Figure 2. Boxplot Comparison of the Average RMSE for Beta 1

Figure 3 presents the comparison of the average RMSE values for beta 2, which are nearly identical under two conditions: when overdispersion approaches zero and when it approaches one. Based on the calculations, the average RMSE for the overdispersion condition approaching zero is slightly lower at **1.185**, compared to **1.189** for the condition approaching one. The difference between these two conditions is minimal, at only **0.004**. However, under the conditions where overdispersion is far from one, the average RMSE increases significantly to **6.130**, making it the highest among the three conditions.

For the **GWNBR** model, the average RMSE values across the three overdispersion conditions are quite similar, with values of **1.859**, **1.831**, and **1.844**. Among the GWPR model results, the overdispersion condition, far from once again produces the highest RMSE for beta 2 compared to the other conditions.

Overall, the **average RMSE values for beta 2** indicate that the GWPR model performs well under conditions where overdispersion is close to zero or one, but its performance declines significantly when overdispersion is far from one, unlike the GWNBR model which remains more stable across all conditions.

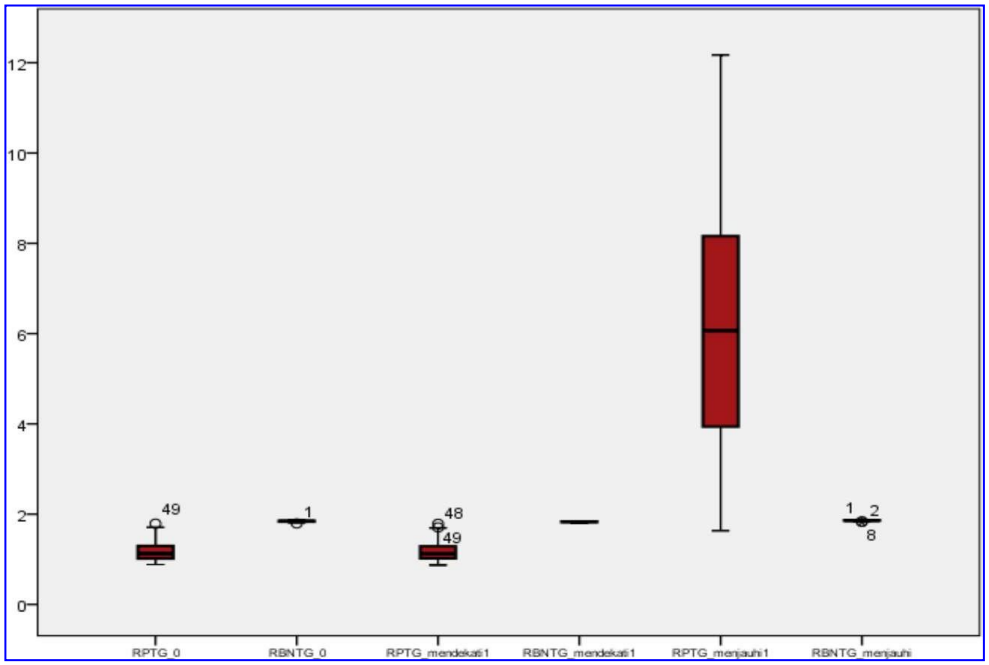

Figure 3. Boxplot Comparison of the Average RMSE for Beta 2

Bias in statistics is also one of the key criteria used to assess model performance. A smaller bias value indicates a better-performing model. **Figure 4** illustrates a comparison of the average bias for the beta 0 parameter under three different overdispersion conditions for both the GWPR and GWNBR models.

For the **GWPR** model, the average bias values under the conditions where overdispersion approaches 1 and where it moves far from 1 are relatively similar. According to the calculations, the lowest average bias was found when overdispersion approached zero, with a value of **0.698**, compared to **1.037** when overdispersion approached one. However, when overdispersion moved far from one, the bias increased significantly to **4.549**, the highest among all conditions.

In the **GWNBR** model, the average bias across all three overdispersion conditions was relatively stable, with values of **2.525**, **2.467**, and **2.441**. Notably, under the condition where overdispersion approached one, the GWPR model exhibited a higher average bias compared to its other conditions.

Overall, the comparison of average bias values for **beta 0** shows that the GWPR model performs best under conditions of low overdispersion, while the GWNBR model demonstrates more consistent performance across varying levels of overdispersion.

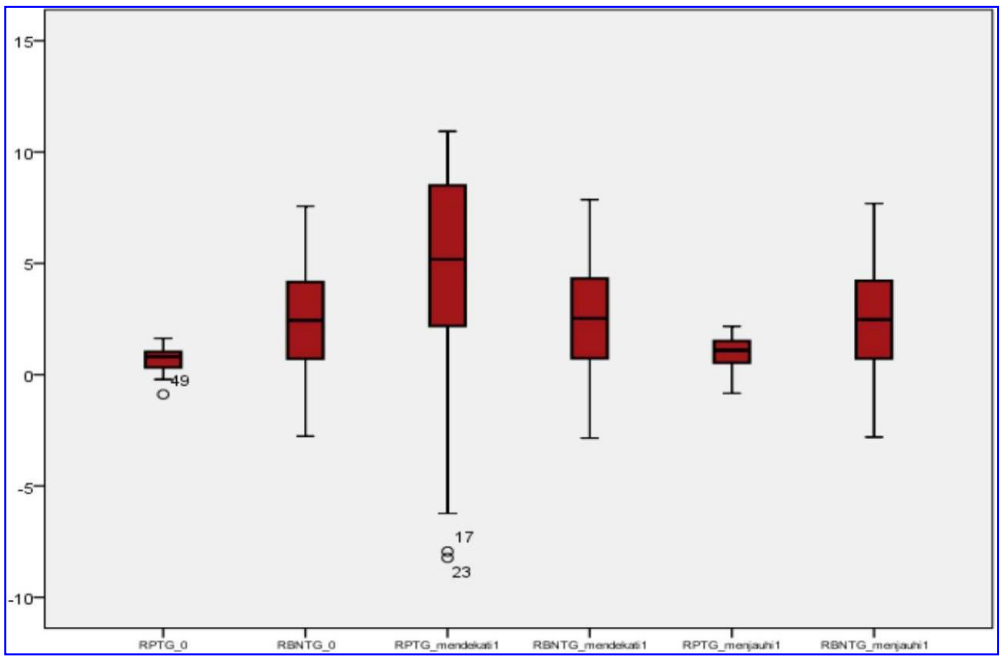

Figure 4. Boxplot Comparison of the Average Bias for Beta 0

Figure 5 shows the average bias of the **beta 1** parameter, which is nearly the same under two conditions: when overdispersion approaches zero and when it approaches one. Based on the calculations, the average bias under the condition of overdispersion approaching zero is slightly smaller, at **–0.187**, compared to **–0.188** when overdispersion approaches one. Meanwhile, the average bias under the condition where overdispersion moves far from one is **–6.259**, which is the largest (in absolute value) among the three conditions.

The negative values indicate **downward bias** in the estimation of beta 1. In the **GWNBR** model, the average bias across the three overdispersion conditions is also negative, with values of **–10.168**, **–8.626**, and **–15.812**, indicating a more severe downward bias.

Among the conditions for the **GWPR** model, the largest average biases (in absolute terms) appear when overdispersion approaches or moves away from one. Overall, the average bias of **beta 1** in both the GWPR and GWNBR models reveals that both tend to underestimate beta 1, particularly under the condition where overdispersion is far from one.

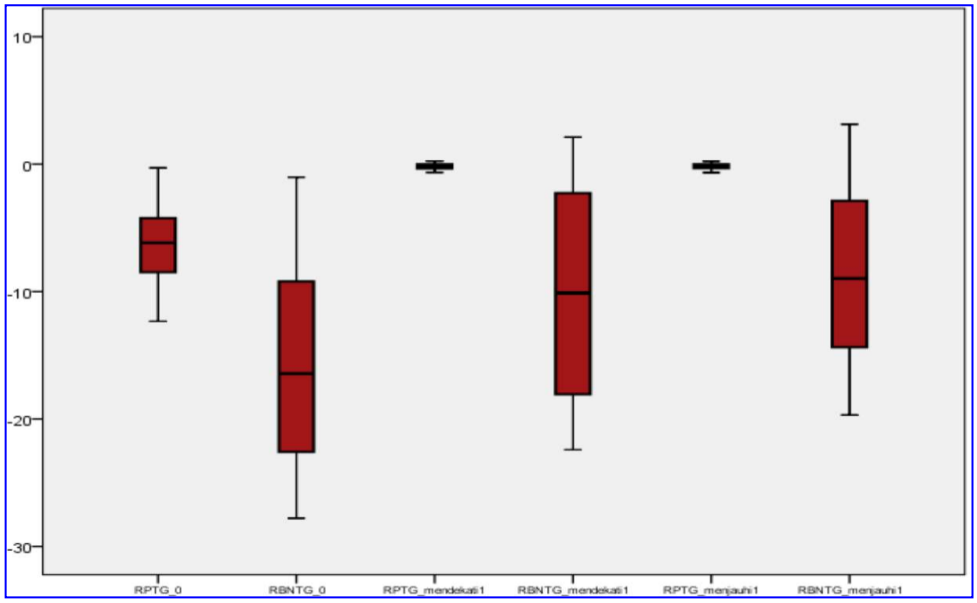

Figure 5. Boxplot Comparison of the Average Bias for Beta 1

Figure 6 presents the average bias of the **beta 2** parameter, which is nearly identical under two conditions: when overdispersion approaches zero and when it approaches one. Under the condition where overdispersion moves far from one, the bias becomes negative, with a value of **–6.038**, which is the largest (in absolute value) among the three conditions.

For the **GWNBR** model, the average bias under two overdispersion conditions is negative, while under the condition where overdispersion moves far from one, the bias is **positive**. The highest average bias in the GWNBR model is observed when overdispersion approaches zero.

In summary, the average bias for **beta 2** in both the **GWPR** and **GWNBR** models varies depending on the overdispersion condition, with the **GWPR model** showing the largest downward bias when overdispersion is far from one, and the **GWNBR model** showing its highest bias when overdispersion is near zero.

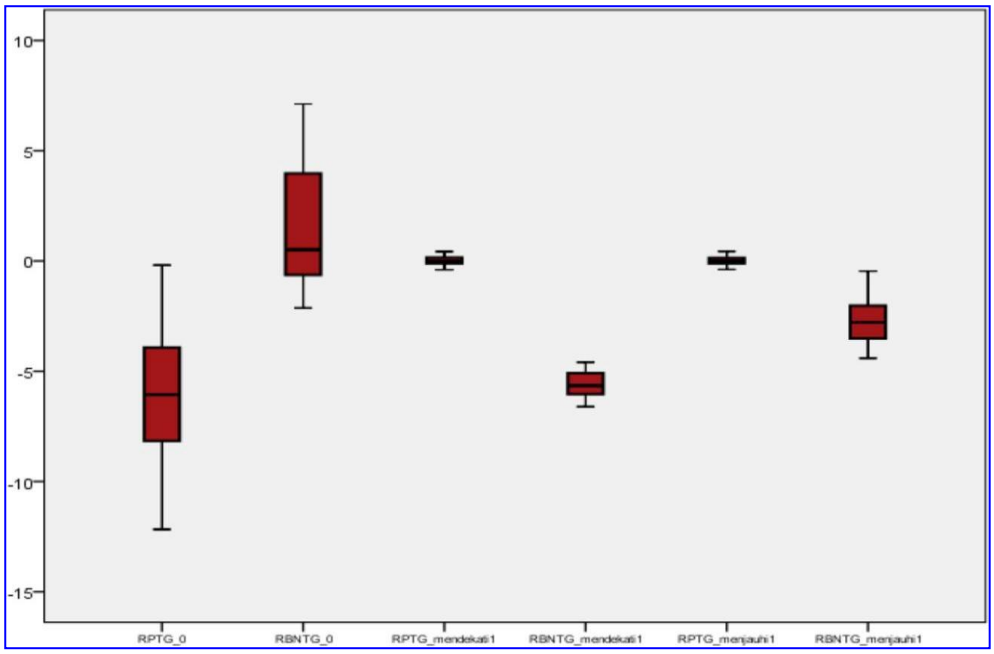

Figure 6. Boxplot Comparison of the Average Bias for Beta 2

Overdispersion occurs when the variance exceeds the meaning. One of its effects is increasing the likelihood of rejecting the null hypothesis (H_0), or accepting the alternative hypothesis (H_1), especially if the data is still modeled using the **GWPR** model. This can lead to an inflated number of significant p-values, which is a sign of model inadequacy.

Table 1. Comparison of the Number and Average of Significant p-values in GWPR and GWNBR Models

Model	Overdispersion Approaching 0			Overdispersion Approaching 1			Overdispersion Moving Away		
	Betha=0	Betha=1	Betha=2	Betha=0	Betha=1	Betha=2	Betha=0	Betha=1	Betha=2
Number RPTG	2624	1776	1214	4888	4323	4251	4887	4327	4255
Number RBNTG	2714	2233	1652	2764	2272	1776	2860	2363	1855
Average RPTG (%)	53.55	36.24	24.77	99.75	88.22	86.75	99.73	88.31	86.84
Average RBNTG (%)	55.39	45.57	33.71	56.41	46.37	36.25	58.37	48.22	37.86

Based on Table 1, the **GWPR model** shows the smallest number and average of significant p-values when overdispersion is **near zero**, indicating that GWPR is appropriate for data with **no or minimal overdispersion**.

However, under conditions where overdispersion **approaches or moves far from 1**, the number and average of significant p-values in the GWPR model are **much higher** than those in the GWNBR model. This indicates that **GWPR is not suitable for modeling data with overdispersion**.

On the other hand, the **GWNBR model** yields a **smaller number and average of significant p-values** than GWPR under both moderate and severe overdispersion. The difference in counts between GWPR and GWNBR models under these two conditions ranges between **1960 and 2479 cases**, or approximately **40% to 50.59%**, highlighting a notable impact of overdispersion in inflating false positives under GWPR.

Interestingly, under the **near-zero overdispersion** condition, GWNBR actually has a **larger number and average of significant p-values** than GWPR, confirming that GWNBR is specifically tailored for data **with overdispersion**.

Finally, within the GWNBR model itself, the number and average of significant p-values under **high overdispersion (>>1)** are **greater** than those under **moderate overdispersion (≈1)**. This suggests that the GWNBR model is most reliable **when overdispersion is moderate**, i.e., **approaching 1**.

5. CONCLUSION

Based on the simulation data, it can be concluded that the tolerance limit still suitable for modeling using **GWPR** is when the overdispersion is close to 1, within the overdispersion range of **1.110 to 2.067**. In this condition, the GWPR model provides more stable and appropriate results.

However, as the overdispersion moves further away from 1, within the overdispersion range of **4.610 to 7.450**, there is a significant increase in the number of significant p-values or rejections of H_0. This indicates that the GWPR model becomes less suitable for data with higher levels of overdispersion, as the higher the overdispersion, the more significant p-values appear, leading to an increased risk of Type I errors (false positives).

6. REFERENCES

1. Anselin, L. (1988). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.
2. Badan Pusat Statistik (BPS). (2017). East Java Province in Figures 2017. East Java, Indonesia: BPS.

3. Da Silva, A. R., & Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression—incorporating overdispersion. *Statistics and Computing*, *24*, 769-783.
4. Dinas Kesehatan Provinsi Jawa Timur. (2017). Health Profile of East Java Province 2017. East Java, Indonesia: Dinas Kesehatan.
5. Famoye, F., Wulu, J. T., & Singh, K. P. (2004). On the generalized Poisson regression model with an application to accident data. *Journal of Data Science*, *2*(3), 287-295.
6. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2009). Geographically weighted regression. *The Sage handbook of spatial analysis*, *1*, 243-254.
7. Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, *99*(3), 585-590.
8. Gujarati, D. (2006). Basic Econometrics (Indonesian ed.). Jakarta, Indonesia: Erlangga.
9. Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. Stata press.
10. Kementerian Kesehatan Republik Indonesia. (2018). Pocketbook for Monitoring Nutritional Status 2017. Jakarta, Indonesia: Ministry of Health.
11. Liu, J., Zhao, Y., Yang, Y., Xu, S., Zhang, F., Zhang, X., ... & Qiu, A. (2017). A mixed geographically and temporally weighted regression: Exploring spatial-temporal variations from global and local perspectives. *Entropy*, *19*(2), 53.
12. McCullagh, P. (2019). *Generalized linear models*. Routledge.
13. Miranti, Z., & Purhadi. (2016). Mapping the Number of Severely Malnourished Under Five Children in Surabaya Using GWNBR and the Flexibly Shaped Spatial Scan Statistic. Journal of Science and Arts ITS, Surabaya.
14. 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.
15. Rogers, A. (1974). Statistical analysis of spatial dispersion: the quadrat method. *(No Title)*.

16. Sofia, A. (2018). Poisson Regression and Spatial Autoregressive Poisson in Estimating Factors of Severe Child Malnutrition Cases on Java Island [Master's thesis, Bogor Agricultural Institute]. Bogor, Indonesia: Institut Pertanian Bogor.
17. Soekirman. (2012). A New Paradigm to Address Macronutrient Malnutrition in Indonesia. Bogor, Indonesia: Bogor Agricultural Institute. Retrieved March 16, 2019, from gizi.depkes.go.id/wp-content/uploads/2012/05/prof-soekirman.pdf.