Modelling the number of HIV/AIDS in Central Sulawesi

I U Utami¹, I Setiawan¹ and D Daniaty²,
¹ Statistics Department, Tadulako University, Palu, Indonesia
² Statistics Indonesia, Jakarta, Indonesia
*Email: triutami.iut@gmail.com

Abstract. The spread of HIV/AIDS in Central Sulawesi is centralized and spreads in certain districts or cities so that there are indications of a spatial effect in the spread of HIV/AIDS. Geographically Weighted Negative Binomial Regression (GWNBR) is one of the right solutions for modelling the relationship between response variables and explanatory variables on counted data that is local for each observation location with overdispersion and the influence of location or spatial aspects on the data. Spatial aspects can be caused by geographic, socio-cultural, economic conditions, as well as different people's knowledge between regions. Overdispersion is a condition where the variance of the data is greater than the average data. This study aims to determine the GWNBR model and the factors that influence the number of HIV/AIDS cases in the province of Central Sulawesi. The obtained result based on GWNBR model shown that the districts with the most number of districts that have the most significant similarities in variables are divided into nine subdistrict groups. Factor affecting the number of HIV/AIDS cases in Central Sulawesi Province in 2018 was population density.

1. Introduction
The number of people living with HIV/AIDS in Indonesia in 2018 has increased by 30% from the previous year, namely 15%. Currently, the spread of HIV has reached all districts in Central Sulawesi, from 13 districts or cities in Central Sulawesi, all HIV cases have now been found. The results of the 2016 sero survey in 13 districts or cities in Central Sulawesi, the prevalence of HIV among female sex workers was 0.3%. Although the HIV prevalence in Central Sulawesi is still below 0.5% according to program indicators, this case continues to increase from year to year. The increasing number of AIDS cases in Central Sulawesi needs serious attention going forward [6].

People with HIV/AIDS tend to occur and are mostly infected in the productive age group with an age range of 25-49 years. The spread of HIV is influenced by certain factors such as socioeconomic factors/poverty, gender, behaviour and lifestyle, socio-culture, biology and health services [9]. The spread of HIV infection which tends to increase shows that the implementation of HIV/AIDS prevention and control programs in Central Sulawesi is not optimal yet, so a study is needed to determine and analyze the factors that influence the number of HIV/AIDS cases in Central Sulawesi.

A number of HIV/AIDS cases is a count data which is suspected of having a spatial aspect with overdispersions. One of the methods for the modelling of counting spatial data with overdispersions is Geographically Weighted Negative Binomial Regression (GWNBR) [10]. GWNBR is the appropriate solution to form a regression analysis that is specific to each observation’s location. The method allows the model parameters to vary spatially and produces non-parametric surfaces of their estimates.
2. Material and method

2.1. Data analysis

This research is a case study. This research aims to determine the GWNBR model and the factors that influence the number of HIV/AIDS cases in the province of Central Sulawesi. The results of the analysis will produce a Poisson regression model, negative binomial regression model, GWNBR model and factors affecting the number of HIV/AIDS cases in Central Sulawesi province. To determine the geographical weights using adaptive bisquare kernel functions.

The data used in this study are secondary data obtained from the publication of Statistics Indonesia (BPS). This study used the GWNBR method with a research unit of 140 districts in Central Sulawesi province. The variables used in this study are given in Table 1.

| Variables                                      | Unit     |
|------------------------------------------------|----------|
| The number of HIV/AIDS (Y)                     | cases    |
| The number of condom users (X₁)                | Person   |
| The number of health workers (X₂)              | Person   |
| The number of health facilities (X₃)           | Unit     |
| Population density (X₄)                        | km²      |
| The number of areas with village status (X₅)   | Village  |
| The number of active family planning participants (X₆) | Participant |

The analysis steps used for answer the objectives of this study are given as follows:
1. Describing the characteristics of the number of HIV/AIDS in Central Sulawesi Province uses area mapping with thematic maps.
2. Identifying the correlation between the predictor variables to detect any cases of multicollinearity by means of the Variance Inflation Factor (VIF).
3. Modelling of the Poisson regression.
4. Detecting the presence of over dispersion on the data
5. Modelling of the Negative Binomial regression.
6. Testing of spatial aspects. Breusch Pagan test is to determine the spatial heterogeneity of data and Moran Global Index (I) to determine the spatial dependency of the data.
7. Modelling data using the GWNBR model
8. Draw Conclusions and Suggestions.

2.2. Poisson regression

Poisson regression is a nonlinear regression model that is used to discrete (count) data types where the response variables follow a Poisson distribution. Predictor variables are linked to the outcome via a natural log transformation [2]. For a simple Poisson regression, the model is

\[ \ln(\lambda_i) = \mu_i = a + bX_i \]  \( (1) \)

The distribution of \( y_i \) or the expected value \( y_i \) by the set of predictor variables \( x_i \). Let assume that the expected value of \( y_i \) is given by

\[ E(y_i | x_i) = \exp(\kappa_i^T \beta) \]  \( (2) \)

Multicollinearity is a statistical phenomenon in which two or more predictor variables are highly correlated. The VIF is used as an indicator of multicollinearity. Computationally, it is defined as the reciprocal of tolerance.
\[ VIF = \frac{1}{1 - R^2} \]  

if the value of VIF > 10 indicates multicollinearity [3].

For detecting the overdispersion, can be seen from the value of deviance divided by degree of freedom (df) or Pearson chi-square/ df. If the quotient is greater than 1, it can be said to be a case of overdispersion, whereas if it is less than 1 then there underdispersion [8].

2.3. Negative Binomial regression (NB Regression)
The Poisson distribution assumes that the mean and variance of the variable are equal. Sometimes count variables do not meet this assumption, especially when there are more zeros or more high values than expected. This is called overdispersion and results in a variable’s variance (v) being much larger than its mean (\( \lambda \)). One way of doing the overdispersion data by using a negative binomial (NB) distribution for the residuals. The form of a negative binomial probability distribution is given as:

\[
P(y_i) = \frac{\Gamma\left(\frac{1}{\alpha} + y_i!\right)}{\left(\frac{1}{\alpha}\right)^{y_i!} \left(\frac{1}{\alpha} + \lambda_i\right)^{1/\alpha}} \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)
\]

where \( \Gamma(.) \) is a gamma function [5].

2.4. Geographically Weighted Negative Binomial Regression (GWNBR)
Silva and Rodrigues (2014) developed a procedure to estimate GWNBR to allow discrete data to be modeled in a non-stationary way and include the overdispersion of the data. The general form of the GWNBR model:

\[ y_j \sim NB \left[ t_j \exp \left( \sum_k \beta_k(u_i, v_j)x_{jk} \right), \alpha \right] \]

where \((u_i, v_j)\) are the location coordinates of data points \( j \), for \( j = 1, ..., n \), \( t_j \) is an offset variable, \( \alpha \) is the parameter of overdispersion, \( \beta_k \) is the parameter related to the predictor variable \( x_k \), for \( k = 1, ..., K \) and \( v_j \) is the \( j \)-th dependent variable.

The parameters \( \beta_k \) and \( \alpha \) can be estimated using a modified Iteratively Reweighted Least Squares (IRLS) procedure and carrying out a subroutine with the maximum likelihood (ML) method by means of the Newton-Raphson (NR) algorithm [10].

3. Results and discussion
3.1. Characteristics of the number of HIV/AIDS in Central Sulawesi Province in 2018
The initial description of the number of HIV/AIDS cases in Central Sulawesi will be explained using thematic maps. The method used in making thematic maps is the Natural Breaks method which is divided into 3 groups, namely low, medium, and high. The Natural Breaks method was used because the data on the number of HIV/AIDS cases in each district had an uneven value. The distribution of the number of HIV/AIDS cases in Central Sulawesi in 2018 is presented in Figure 1.
Based on Figure 1, it can be seen that the regions that have a low number of HIV/AIDS cases are 233 districts, depicted in green. In the medium group, there are 19 districts depicted in yellow. While the high group consists of two districts, namely South Palu and Mantikulore which are depicted in red. The trend of reported HIV/AIDS cases from year to year tends to increase and in 2018 the number of HIV/AIDS cases was reported as many as 254 cases. Mantikulore district had the highest number of HIV/AIDS cases with a total of 36 cases. The number of HIV/AIDS cases had a mean and standard deviation that is 5 cases.

3.2. Diagnostic of multicollinearity
Detecting multicollinearity were carried out prior to build the model. Multicollinearity is usually regarded as a problem arising out of the violation of the assumption that predictor variables are linearly independent in modelling with regression analysis. The VIF estimates how much the variance of a
regression coefficient is inflated due to multicollinearity in the model. One way to assess multicollinearity is by means of the Variance Inflation Factor (VIF). In general, values of VIF that exceed 10 indicates multicollinearity, so that the parameter estimates obtained will be biased. The VIF value for each predictor variables were shown in Table 2.

Table 2 VIF for Six Predictor Variables

| Variables | VIF   |
|-----------|-------|
| x1        | 1.040562 |
| x2        | 1.140185 |
| x3        | 1.398633 |
| x4        | 1.655415 |
| x5        | 1.749358 |
| x6        | 1.554625 |

The VIF for the six predictors are in the range of 1.04–1.75. These results indicate that no multicollinearity among predictor variables. Thus all the variables can be included in the subsequent analysis modelling with Poisson regression, negative binomial regression and GWNBR model.

3.3. Modelling for the Poisson regression

Data on the number of HIV/AIDS cases are assumed to have a Poisson distribution because they are discrete data (count). After detecting multicollinearity among predictor variables and obtained that the six predictor variables will perform modelling of Poisson regression. Maximum Likelihood Estimation (MLE) method is used to obtain the estimation of the Poisson regression model parameters as shown in Table 3.

Table 3 Parameter estimates of the Poisson regression

| Parameter | Estimate | Standard Error | Z Value | p-value |
|-----------|----------|----------------|---------|---------|
| β0        | -1.590e+00 | 2.165e-01    | -7.344  | 2.08e-13 |
| β1        | -3.219e-04 | 1.039e-03    | -0.310  | 0.7567  |
| β2        | -3.298e-04 | 7.293e-04    | -0.452  | 0.6511  |
| β3        | 3.201e-02  | 6.192e-03    | 5.170   | 2.35e-07 |
| β4        | 2.162e-05  | 2.972e-05    | 0.728   | 0.4668  |
| β5        | 2.936e-02  | 2.219e-03    | 13.233  | < 2e-16 |
| β6        | 1.218e-04  | 7.039e-05    | 1.731   | 0.0835  |

Residual Deviance 337.59 Df = 133
AIC 526.5

Table 3 gives the results of the goodness of fit of a Poisson Regression that has been formed. It can be seen that the residual deviance is 337.59. This value is greater than the value of $\chi^2_{(0.05,6)}$ (12.59). This indicates that the model is good fit. The testing is continued partially.

Partial parameter test is used to know which parameters give significant influence with the response variable on each location. Based on Table 3, it is known that the number of health facilities ($X_1$) dan population density ($X_4$) have a p-value less than α by 5%, it is decided to reject the hypothesis. This means that $X_3$ dan $X_4$ have an influence on model, so that the variables have a significant effect against the number of HIV/AIDS cases. The estimated equations of Poisson Regression is

$$\hat{\mu} = \exp(-1.590e+00 - 3.219e-04X_1 - 3.298e-04X_2 + 3.201e-02X_3 + 2.162e05X_4 + 2.936e-02X_5 + 1.218e-04X_6)$$
3.4. Detecting overdispersion

Overdispersion describes the observation that variation is higher than would be expected. Overdispersion can be detected by its residual deviance divided by the degrees of freedom. When this quotient is larger than 1 indicating severe overdispersion [8]. Based on Table 3, the quotient was 2.54 (337.59/133) for the Poisson Regression. It is indicate overdispersion on data, the negative binomial regression should be considered.

3.5. Modelling for the Negative Binomial (NB) regression

Negative binomial regression is a common solution for addressing overdispersion data. The scalar heterogeneity parameter in the NB model can often appropriately adjust for the extra correlation in the data. Thus it is necessary to assess the heterogeneity parameter to determine whether it is different from zero. By using R software, the parameter estimates of the NB regression is obtained in Table 4.

| Parameter | Estimate  | Standard Error | Z value | p-value |
|-----------|-----------|----------------|---------|---------|
| $\beta_0$ | -1.0446090 | 0.3604558 | -2.898 | 0.00376 |
| $\beta_1$ | -0.0017998 | 0.0025222 | -0.714 | 0.47548 |
| $\beta_2$ | 0.0007986 | 0.0015319 | 0.521 | 0.60215 |
| $\beta_3$ | 0.0181425 | 0.0120349 | 1.507 | 0.13168 |
| $\beta_4$ | 0.0001037 | 0.0001298 | 0.800 | 0.42397 |
| $\beta_5$ | 0.0249038 | 0.0063206 | 3.940 | 8.15e-05 |
| $\beta_6$ | 0.0001030 | 0.0001299 | 0.793 | 0.42775 |

Table 4 gives the results of the goodness of fit of a NB Regression. It can be seen that the residual deviance is $129.57$. This value is greater than the value of $\chi^2_{(0.05;6)} (12.59)$. This indicates that the model is good fit. The testing is continued partially.

Partial parameter testing is used to know which parameters give significant influence on the response variable on each location. Based on Table 4, it is known that The number of areas with village status ($X_5$) have a p-value less than $\alpha$ by 5%, it is decided to reject the hypothesis. This means that $X_5$ have an influence on model, so that the variables have a significant effect against the number of HIV/AIDS cases. The estimated equations of NB Regression are summarized as follows

$$\hat{\mu} = \exp (-1.0446090 - 0.0017998X_1 + 0.0007986X_2 + 0.0181425X_3 + 0.0001037X_4 + 0.0249038X_5 + 0.0001030X_6)$$

According to Table 3 dan 4, the NB Regression model is better than the Poisson Regression model. This can be seen that NB Regression has the Akaike Information Criterion (AIC) values smaller than the Poisson regression. This suggests that the Negative Binomial regression model is more appropriate in the case overdispersion Poisson regression.

3.6. Spatial aspect testing

Spatial dependence and spatial heterogeneity are two characteristics of spatial data that need to be taken into account when modelling of Geographically Weighted Regression (GWR). The spatial heterogeneity in data can be detected by Breusch-Pagan (BP) test. Based on the test results with R software obtained the BP test value is 42.273 and the p-value is 1.624e-07. If the value of BP is compared with the value of $\chi^2_{(0.05;6)} = 12.59$, it is known that the value of BP $> \chi^2_{(0.05;6)}$. It can be concluded that there are differences in characteristics between one point of observation and another point of observation.
Dependency testing can be analyzed by Moran’s *I* statistics. Based on test results with R software obtained p-value is 2.220446e-16. When compared with a significance level (\( \alpha \)) of 0.05, the p-value > \( \alpha \), it can be concluded that there are no dependencies spatially between regions. The results of the spatial dependency test which states that the observation of a location does not depend on observations in other locations that are located nearby.

3.7. Modelling of the Geographically Weighted Negative Binomial Regression (GWNBR)

The existence of spatial heterogeneity causes the need for a spatial weighting matrix. The weighting function used in this study is the adaptive bi-square kernel. The first step to form a weighting matrix is to calculate Euclid's distance. Next determine bandwidth, optimal bandwidth selection using the Golden Section Search technique is done iteratively using the Cross Validation (CV) criteria. The CV value minimum indicates the optimal bandwidth value.

The GWNBR model will generate local parameter estimates with each location will have different parameters. The parameter estimator is obtained from Iteratively Newton Raphson method. Testing the significance of the GWNBR model simultaneously aims to find out whether the simultaneously the predictor variables have an effect against the model. Based on the results of testing using R software obtained the deviance value of the GWNBR model amounted to 420.2802. With a significance level of 5%, it is obtained \( \chi^2_{(0.05;6)} = 12.59 \), which means that at least one parameter estimate of the GWNBR model has a significant effect. It is necessary to continue with partial testing.

Partial parameter testing is used to know which parameters give a significant influence on the response variable on each location. Based on the test results, it was found that the Z value was different for each location. Based on the test results, it was found that the Z value was different for each location. The results of grouping according to the GWNBR model are divided into 9 district groups based on the significant similarities in variables. District grouping mapping based on significant similarities in variables, as shown in Figure 2.

![Figure 2. Mapping of district grouping based on significant variables](image-url)
The first group consisted of 14 districts with two variables that had a significant effect on the number of HIV/AIDS cases, namely the number of health workers ($X_2$) and population density ($X_4$). The second group consists of 14 districts with one variable that has a significant effect on the number of HIV/AIDS cases, namely population density ($X_4$). The third group consisted of 13 districts with three variables that were significant to the number of HIV/AIDS cases, namely the number of health workers ($X_2$), the number of health facilities ($X_3$), and the population density ($X_4$). The fourth group consists of 11 districts with four significant variables, namely the number of condom users ($X_1$), the number of health workers ($X_2$), population density ($X_4$) and the number of areas with village status ($X_5$). The fifth group consists of 8 districts with all significant variables, namely the number of condom use ($X_1$), the number of health workers ($X_2$), the number of health facilities ($X_3$), population density ($X_4$), the number of areas with village status ($X_5$) and the number of family planning active ($X_6$).

4. **Conclusion**

The results of modelling with the GWNBR method are formed in nine district groups according to influencing variables significant to the number of HIV/AIDS. The factors that significantly influence the number of HIV/AIDS cases are different in each district in Central Sulawesi Province, of the 140 locations there are only 9 districts have variables that have a significant effect, the rest have no variables which has a significant effect. Factor affecting the number of HIV/AIDS cases in Central Sulawesi Province in 2018 was population density.

**Acknowledgement**

We thank the Faculty of Mathematical and Natural Science who has funded this research, so that this research can be done smoothly. Thanks also to the Statistics Indonesia who assisted in providing data.

**References**

[1] Badan Pusat Statistik Provinsi Sulawesi Tengah 2019 *Kabupaten dalam Angka* (Palu: BPS)
[2] Cameron A C and Trivedi P K 1998 *Regression Analysis of Count Data* (Cambridge: Cambridge University Press)
[3] Fotheringham A S and Oshan T M 2016 *Journal of Geographical Systems* 18 303–329
[4] Gschlößl S and Czado C 2005 *Sonderforschungsbereich* 386 412
[5] Hilbe J M 2011 *Negative Binomial Regression* (Cambridge: Cambridge University Press)
[6] Kementerian Kesehatan Republik Indonesia 2019 *Profil Kesehatan Provinsi Sulawesi Tengah* 2019 (Jakarta: Kemkes)
[7] Mansfield E R and Helms B P 1982 *The American Statistician* 36 158-160
[8] McCullagh P, Nelder J A 1989 *Generalized linear models*. (New York: Chapman and Hall)
[9] Priscilla V 2015 *Majalah Kedokteran Andalas* 32 2
[10] Silva A R and Rodrigues T C V 2014. *Statistics and Computing* 24 769–783