Modelling the number of family planning participants in Southeast Sulawesi using geographically weighted regression model

Mukhsar¹, Asnidar¹, Bahriddin Abapihi¹*, Agusrawati¹, Lilis Laome² and Sitti Wirdhana Ahmad³

¹Department of Statistics, FMIPA, Halu Oleo University, Kendari-Indonesia
²Department of Mathematics, FMIPA, Halu Oleo University, Kendari-Indonesia
³Department of Biology, FMIPA, Halu Oleo University, Kendari-Indonesia
bahriddinabapihi@yahoo.com (corresponding author)

Abstract. The number of family planning participants in Southeast Sulawesi are influenced by spatial factor, for example geographically, socio culturally, and others. Geographically Weighted Regression (GWR) method is needed. Parameter estimates are only for each location and are different from other locations. The number of fertile age couples (X₁) is significant in the regions of Buton, Muna, Konawe, Konawe Selatan, Wakatobi, Buton Utara, Konawe Utara, Konawe Kepuluan, Kolaka Utara, Muna Barat, Buton Tengah, Buton Selatan, Kendari and Bau-Bau. The number of childbearing age couples (X₂) and number of counselling officer (X₃) is significant in the regions of Kolaka, Bombana and Kolaka Timur.

1. Introduction
The population density control program through the number of family planning participants in Southeast Sulawesi are a priority because it impacts on various factors, such as economic, social, and others. The family planning program is also influenced by environmental or spatial factors [1, 5, 9, 12, 13]. The handling of the number family planning participants is different between regions (in urban and rural areas). It is necessary to pay attention to spatial factors or spatial heterogeneity because the data between locations differ geographically, socio culturally, and others [10, 11]. Regression parameters for heterogeneity data are spatially vary and non-stationary problem, a Geographically Weighted Regression (GWR) method is needed [2, 3, 4]. Parameter estimates are only for each location and are different from other locations [6, 7, 8]. The GWR is a spatial analysis technique and models the local relationships between these predictors and an outcome of interest.

Southeast Sulawesi Province has 17 regencies and cities, namely Kabupaten Buton, Muna, Konawe, Kolaka, Konawe Selatan, Bombana, Wakatobi, Kolaka Utara, Buton Utara, Konawe Utara, Kolaka Timur, Konawe Kepuluan, Muna Barat, Buton Tengah, Buton Selatan, Kendari and Bau. Each district has a family planning program according to regional conditions, such as traditional traditions, religion, economy, side effects of family planning, and other reasons.
2. Methodology

2.1. Spatial Data and Heterogeneity Test

The regression method describes the relationship between the response variable \( y \) with several predictor variables \( x \). Regression models use data with a normal distribution. Kolmogorov-Smirnov test, \( P \leq 0.05 \), is used to check the distribution of data. Furthermore, heteroscedasticity test is needed to find out whether there is an unequal variance of residuals between observations. In this case, the Glejser test is used to find out heteroskedasticity [3].

Multicollinearity test is to determine the existence of a linear relationship between all variables. Some methods can be used to detect multicollinearity, such as Variance Inflation Factors (VIF):

\[
(VIF) = \frac{1}{1 - R_j^2}
\]

with \( R_j^2 \) is determinant coefficient. If \( (VIF)_j > 10 \) then there is multicollinearity.

The autocorrelation test is used to determine the correlation between confounding errors in the period \( t \) with \( t-1 \). The Durbin Watson rule requires intercepts in the regression model and there are no variables among the independent variables. The Durbin Watson (d) Statistical Equation is as follows:

\[
d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}
\]

with

- if \( d < d_{u} \) then there is autocorrelation
- if \( d > d_{u} \) then there is not autocorrelation
- if \( d < d_{l} \) then there is not autocorrelation
- if \( d > d_{l} \) then there is not summary.

Spatial data is data collected from different spatial locations and has a location dependency. Spatial data are assumed to be normally distributed and have a spatial relationship. An important aspect of defining spatial autocorrelation is the determination of the relationship of the closest region [2, 6, 7]. The range of values \( 0 < I < 1 \) indicates a positive autocorrelation which means that there are similarities of data, while \( 1 < I < 0 \) is a negative autocorrelation indicating that adjacent locations have different values and tend to spread. Moran’s I test is used for spatial autocorrelation test.

\[
R_0: I = 0 \text{ (it is not spatial autocorrelation)}
\]

\[
R_1: I \neq 0 \text{ (it is spatial autocorrelation)}
\]

Statistic Test:

\[
Z_{hit} = \frac{I - E(I)}{\sqrt{Var(I)}}
\]

Spatial heterogeneity is also reflected in errors. In other words the resulting error variance is not constant. The Breusch-Pagan test is a statistical test that can detect spatial heterogeneity. The hypothesis used in the Breusch-Pagan test is:

\[
H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2 \text{ (there is not heterogeneity between locations)}
\]

\[
H_1: \sigma_1^2 \neq \sigma^2 \text{ (there is heterogeneity between locations)}
\]

Statistic test:

\[
BP = \frac{1}{2}f^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T f
\]

with the vector \( f \) is written as \( f = \left( \frac{e_i^2}{\sigma^2} - 1 \right) \), \( BP \) is Breusch-Pagan test, \( e~ IIDN(0, \sigma^2) \), \( \bar{y} \) is response variable, \( \mathbf{Z} \) is standardized matrix \( n \times (p + 1) \).
2.2. Bandwidth

Bandwidth is having a radius $h$ from the centre of the location to determine the weight of each observation of the regression model for each location. Observations within radius $h$ are still considered to affect the model. The smaller bandwidth can produce large variances. The greater the bandwidth is the greater the estimated bias. The kernel weighting function contains bandwidth parameters whose values are unknown. Therefore, it is necessary to estimate the bandwidth parameters [8]. The selection of optimum bandwidth in the GWR is important because it affects the accuracy of the model to the data.

The choice of bandwidth is very important for the proper estimation of the Kernel function. Bandwidth with a very small value is the result of greater variance. Bandwidth value is very small resulting in slightly different observations on the radius $h$. The bandwidth value is very large resulting in smaller variances. To avoid inhomogeneous variance due to the increased value of the parameter coefficient, we need a way to choose the right bandwidth. The preferred method for selecting optimum bandwidth is Cross Validation (CV).

$$CV = n \sum_{i=1}^{n} (y_i - \hat{y}_i(b))^2$$  \hspace{1cm} (5)

2.3. Geographically Weighted Regression

Geographically Weighted Regression is the development of ordinary least square into a weighted regression model with regard to spatial effects. As a result it produces parameter estimators that can only be used to predict each point or location where the data is observed and inferred. The assumptions in the GWR model are normally distributed residuals with zero mean and variance $\sigma^2$[7, 8]. The relationship between the dependent variable $y$ and the independent variables $x_1, x_2, x_3, \ldots, x_k$ in the GWR model for the $i$th location is:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \epsilon_i,$$  \hspace{1cm} (6)

With $y_i$ is dependent variable at location-$i$, $(u_i, v_i)$ is longitude-latitude coordinate at location-$i$, $x_{ik}$ is independent variable $k$ in observation-$i$, the $\beta_k(u_i, v_i)$ is parameter at location-$i$, $\epsilon_i$ is residual-$i$ assumed iid with zero mean.

The weighting function in the GWR model is very important because the weighting values represent the location of the observation data with each other. The weighting scheme in the GWR can use several different methods, for example using the Kernel function [10]. According to [8], several types of weighting functions are used, for example the Fixed Gaussian Kernel Function

$$w_j(u_i, v_i) = \exp \left[ -1 \left( \frac{d_{ij}}{h^2} \right)^2 \right]$$  \hspace{1cm} (7)

with $h$ is bandwidth.

Parameter estimation in the GWR model uses the Weighted Least Square (WLS) method, which is by giving a different weighting to each observation location. The weighting model of GWR has a very important role because the weighting values represent the location of the observation data with each other. If the location weighting $(u_i, v_i)$ is $w_k(u_i, v_i)$ with $k = 1, 2, \ldots, p$ then the observation location parameters $(u_i, v_i)$ are estimated by adding the weighting element $w_j(u_i, v_i)$. Further is minimizing the Sum Square Residual (SSR). GWR model parameter estimator:

$$\hat{\beta}(u_i, v_i) = (X^TW(u_i, v_i)X)^{-1}X^TW(u_i, v_i)y$$  \hspace{1cm} (8)

Spatial autocorrelation test is to identify interdependency between districts. The range of value $0 < I \leq 1$ indicates positive autocorrelation which means there is a similarity of values from adjacent locations, while the value of $-1 \leq I < 0$ negative autocorrelation indicates that adjacent locations have different values and tend to spread. The result of Moran’s I is 0.24 which means there is a positive spatial autocorrelation. This means that between regencies has similar values. Spatial heterogeneity is
to determine whether there is diversity between districts or not. If the Breusch-Pagan probability value is less than 0.05 then there is spatial heterogeneity between regions. We obtained 0.043 <0.05, so it can be defined that there is heterogeneity between the regencies in Southeast Sulawesi Province.

The initial step in the GWR analysis is determining the coordinates of the latitude and longitude in each Regency of Southeast Sulawesi Province. Then calculate the Euclidean distance between locations \((u_i, v_i)\) to \((u_j, v_j)\) with the formula 
\[d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}\]. Based on the formula, the distance between the locations of Southeast Sulawesi Province is obtained, for example in Table 1. Determination of bandwidth \((h)\) used in the GWR model uses the minimum Cross Validation (CV) between the weighting functions. Determination of optimum bandwidth using the Cross Validation method is done using GWR4 software. Optimum bandwidth uses the Fixed Gaussian weighting function, meaning that the optimum bandwidth value is the same in every location, so the bandwidth value used to determine the weighting value is 20 km divided by 1 degree of earth. The optimum bandwidth obtained is 0.17966385, then substituted into the Gaussian Kernel weighting function formula. The weighting function can be written as:

\[w_j(u_i, v_i) = \exp\left(-\frac{d_{ij}^2}{(0.17966385)^2}\right)\]

| Regency          | Euclidean distance |
|------------------|--------------------|
| Buton            | 0                  |
| Muna             | 0.560946           |
| Konawe           | 1.698204           |
| Kolaka           | 1.945848           |
| Konawe Selatan   | 1.254076           |
| Bombana          | 1.190293           |
| Wakatobi         | 0.539044           |
| Kolaka Utara     | 2.855908           |
| Buton Utara      | 0.574564           |
| Konawe Utara     | 2.185293           |
| Kolaka Timur     | 1.564871           |
| Konawe Kepulauan| 1.121425           |
| Muna Barat       | 0.523394           |
| Buton Tengah     | 0.530107           |
| Buton Selatan    | 0.056425           |
| Kendari          | 1.329312           |
| Baubau           | 0.508755           |

### 3. Parameter Estimation GWR Model

The data used in this study are secondary data collected from Health Office of Southeast Sulawesi Province, period 2017-2019. Data collected are fertile age couples and extension workers number for 17 regencies and cities, namely Kabupaten Buton, Muna, Konawe, Kolaka, Konawe Selatan, Bombana, Wakatobi, Kolaka Utara, Buton Utara, Konawe Utara, Kolaka Timur, Konawe Kepulauan, Muna Barat, Buton Tengah, Buton Selatan, Kendari, and Baubau.

The estimated parameter in the GWR model is performed using the Weighted Least Square (WLS) method, with the help of Minitab 17 software (Table 2).
| Function | Max  | Min  | T-hit | Explanation   |
|----------|------|------|-------|---------------|
| $\beta_0$ | 38913 | -2680 |       |               |
| $\beta_1$ | 0.99315 | 0.073 | 2.22  | Significant   |
| $\beta_2$ | 252.5  | -1899 | 1.89  | Not significant |

Table 3. Estimation GWR Model

| Regency          | GWR Model                              |
|------------------|----------------------------------------|
| Buton            | $\hat{y} = 0.419X_1$                   |
| Muna             | $\hat{y} = 0.22281X_1$                 |
| Konawe           | $\hat{y} = 0.7011X_1$                  |
| Kolaka           | $\hat{y} = 0.755331X_1 + 19.7384X_2$  |
| Konawe Selatan   | $\hat{y} = 0.0820X_1$                  |
| Bombana          | $\hat{y} = 0.330X_1 + 252.5X_2$       |
| Wakatobi         | $\hat{y} = 0.99315X_1$                 |
| Kolaka Utara     | $\hat{y} = 0.5541X_1$                 |
| Buton Utara      | $\hat{y} = 0.983X_1$                   |
| Konawe Utara     | $\hat{y} = 0.8009X_1$                  |
| Kolaka Timur     | $\hat{y} = 0.6721X_1 + 47.8X_2$       |
| Konawe Kepulauan | $\hat{y} = 0.8774X_1$                  |
| Muna Barat       | $\hat{y} = 0.2197X_1$                  |
| Buton Tengah     | $\hat{y} = 0.2348X_1$                  |
| Buton Selatan    | $\hat{y} = 0.553X_1$                   |
| Kendari          | $\hat{y} = 0.0730X_1$                  |
| Baubau           | $\hat{y} = 0.749X_1$                   |

The estimated parameter $\beta_1$ corresponding to the number of fertile age couples ($X_1$) ranges from 0.073 to 0.99315, while the estimated parameter $\beta_2$ that corresponds to the number of extension workers ($X_2$) ranges from -1899 to 252.5. Viewed from the maximum and minimum values, the estimated parameters in the GWR model have different values for each location, meaning that geographical factors affect the GWR model. The GWR model for the use of tools or ways of planning family matter in Kolaka District is a better model. The AIC value of the GWR model is 118,501 smaller than the multiple linear regression model of 119,218. The R2 value of the GWR model is 69.04% greater than the multiple linear regression model that is 62.69%, see Table 3 for details. This shows that the GWR model has better diversity in the use of the number family planning participant methods in Southeast Sulawesi Province. Significant variables from the GWR model differ for each region.

4. Conclusion
Couples of childbearing age influence the use of the number family planning participants in Southeast Sulawesi. Based on the GWR model, the number of fertile age couples ($X_1$) is significant in the regions of Buton, Muna, Konawe, Konawe Selatan, Wakatobi, Buton Utara, Konawe Utara, Konawe Kepulauan, Kolaka Utara, Muna Barat, Buton Tengah, Buton Selatan, Kendari, and Baubau. The number of couples of childbearing age ($X_1$) and number of counselling ($X_2$) is significant in the regions of Kolaka, Bombana, and Kolaka Timur.
Acknowledgement
The authors would like to thank the Directorate of Research and Community Service (or DPRM) Ristekdikti Jakarta and Department of Mathematics, Faculty of Mathematics and Sciences, Halo Oleo University. We also would like to thank the International Conference on Science and Mathematics (ICSM) 2019 and anonymous reviewers of ICSM2019 for their valuable suggestions.

References
[1] Anselin, L. 1988 *Spasial Econometrics: Methods and Models*. The Netherlands: Kluwer Academic Publishers.
[2] Burnham, K., A and Anderson, D. R. 2002 *Model Selection and Multimodel Inference: a practical information-theoretic approach*, 2nd edition, New York: Springer
[3] Belsley, DA, Kuh, E and Welsch, R. 1980 *Regression Diagnostics: identifying influential data and sources of collinearity*, Hoboken, NJ: Wiley
[4] Cressie, N.A.C. 1993 *Statistics For Spatial Data Revised ed*. New York: John wiley and Sons.
[5] Comber, A.J.; Brunsdon, C.; Radburn, R. A. 2011 spatial analysis of variations in health access: Linking geography, socio-economic status and access perceptions. *Int. J. Health Geogr*, 10, 44–55. [CrossRef] [PubMed]
[6] Fischer, M., M and Wang, J. 2011 *Spasial Data Analisis: Models, Methods, and Techniques*. New York: Springer.
[7] Fotheringham, A. S. 2006 *Quantification, Evidence and Positivism*, Chapter 32 in Approaches to Human Geography eds. G. Valentine and S Aitken. Sage: London.
[8] Fotheringham, A.S., Brundson, C., and Charlton, M. 2002 *Geographically weighted regression*. Chichester: John Wiley and Sons.
[9] Jivraj, S., Brown, M.,and Finney, N. 2013 Modelling Spatial Variation in the Determinants of Neighbourhood Family Migration in England with Geographically Weighted Regression. *Appl. Spat. Anal. Policy*, 6, 285–304.
[10] Metha, B. 2008 *Introduction to Metadata Second Edition*. Los Angeles: Getty Publications.
[11] Thissen, D, Steinberg, L, and Kuang, D. 2002 Quick and easy implementation of the Benjamini Hochberg procedure for controlling the false positive rate in multiple comparisons, *Journal of Educational and Behavioural Statistics*, 27(1), 77-83
[12] Young, L.J., and Gotway, C.A. 2010 Using geostatistical methods in the analysis of public health data: The final frontier, *Quant. Geol. Geostat.*, 16, 89–98. [CrossRef]
[13] Yu, D.-L., Wei, Y., and Wu, C. 2007 Modelling spatial dimensions of housing prices in Milwaukee, WI. *Environ. Plan. B Plan*.Des., 34, 1085–1102. [CrossRef]