Forum Statistika dan Komputasi : *Indonesian Journal of Statistics*
ISSN : 0853-8115
Vol. 20 No.2, Oktober 2015, p: 5-8

available online at:
journal.ipb.ac.id/index.php/statistika

GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) INCLUDED THE DATA CONTAINING MULTICOLLINEARITY

Ira Yulita, Anik Djuraidah, Aji Hamin Wigena

Department of Statistics, Bogor Agricultural University, Indonesia

ABSTRACT

One of the reasons of spatial effect of each location is spatial variety. Beside of spatial variety, number of independent variable (X) causes local multicolinearity, that is one or more independent variable, which collaborated with other variable in each location of observation. The methods can be used to solve spatial diversity problem and local multicollinearity in Geographically Weighted Regression (GWR) model that is GWPCA. This research aim to examine GWPCAR feasibility model for PDRB data in 2010 at 113 districts/cities in Java. analysis indicate that GWPCA method can overcome local multicollinearity problem, it can be seen from the characteristic value of VIF which is smaller than 10.

Key words : Local Multicollinearity, Geographically Weighted Principal Components Analysis.

INTRODUCTION

Background

Spatial data is geographically oriented data which has a particular coordinate system as its basic reference. The first law of geography expressed by W Tobler's in Anselin (1988) states that everything depends on everything else, but closer things more so. On the spatial data, the conditions of a location to other location are not the same, which in terms of geography, socio-cultural circumstances, or anything else that may cause the condition of spatial heterogeneity in the location studied. So that if the data still modeled by classical regression methods, it would lead to a model obtained unrepresentative and diversity among regions that cannot be detected. Therefore, modeling to address the spatial heterogeneity has been done using Geographically Weighted Regression (GWR).

In the spatial model, the number of independent variables (X) causes local multicollinearity which mean there are one or more variables are correlated with other variables at each observation location. Detecting cases of multicollinearity by O'Brien 2007 can be done with VIF *(Variance Inflation Factor)*. The method can overcome multicollinearity in GWR among others by Geographically Weighted Principal Components Analysis (GWPCA). GWPCA is the development of a regression analysis that uses the basis of principal component analysis (PCA), the data used contain spatial effects (local). The purpose of this study is to determine GWPCA model using Gaussian-weighted and apply it on the data of GRDP in Java in the year of 2010.

LITERATURE REVIEW

A. *Local Multicollinearity*

Multicollinearity is a condition in which the independent variables (X) are correlated causing the estimation parameters from the regression model generated that has a very large remnant. O'brien (2007) explained that multicollinearity can be detected by looking at the value of VIF *(Variance Inflation Factor)* with the weighted matrix as the collinear detection area at GWR model. Explanatory variables is said to be correlated if its VIF value is larger than 10. At GWR modeling, VIF is possible to be calculated for each explanatory variable. VIF value is expressed as follows:

$$VIF_k(u_i, v_i) = \frac{1}{1 - R_k^2(u_i, v_i)}$$

with $R_k^2(u_i, v_i)$ as coefficient of determination between X_k with other explanatory variables for each location (u_i, v_i). If the value of VIF is more than 10, then this indicates the data having multicollinearity problems (Myers 1990).

B. Geographically Weighted Principal Components Analysis (GWPCA)

According to Fotheringham *et al.* 2002, the working of GWPCA equals PCA in general. The only difference between GWPCA and PCA is GWPCA uses weighted location while PCA is not using weighted location. At GWPCA, the way to tackle multicollinearity problem is by forming KU at the weighted locations, so it can be formulated as follows:

$$\sum(u_i, v_i) = \boldsymbol{X^T W}(u_i, v_i)\boldsymbol{X}$$
$$\left|\sum(u_i, v_i) - \lambda_i \boldsymbol{I}\right| = 0$$
$$\left(\sum(u_i, v_i) - \lambda_i \boldsymbol{I}\right)\boldsymbol{a_i} = \boldsymbol{0}$$

with :

$\sum(\boldsymbol{u_i}, \boldsymbol{v_i})$= variance-covariance matrix from the weighted location (u_i, v_i).

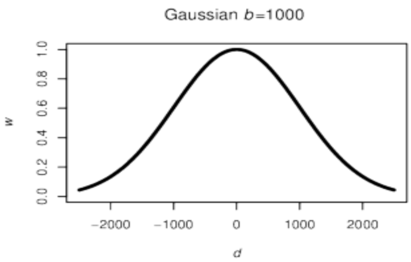
Weighted location is used in this study was Kernel Normal (Gaussian), which has fuction as follows:

$$w_j(i) = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]$$

with :

$d_{ij} = \sqrt{\left(u_i - u_j\right)^2 + (v_i - v_j)^2}$

is the euclidean distance between between location of the-ith to the location of the-jth and b is bandwidth which is the function of smoothing parameter value whose value is always positive.



METHODOLOGY

GWPCA method will be applied to the secondary data derived from Central Bureau of Statistics (CBS), namely Village Potential (PODES) Dataset, Gross Regional Domestic Product (GRDP) of district / city, and the population of the district / city. Observational data in this research is the data of 113 districts / cities in Java Island in 2010. Bound variables in this study are the GDP districts / cities in Java.

RESULT AND DISCUSSION

GWPCA method is used on the grounds that data contains spatial effects and local multicollinearity. Local multicollinearity is used to see the relationship between the explanatory variables in each location. Local collinearity detection was conducted using *Variance Inflation Factor* (VIF) method.

Examples of VIF value for several districts / cities in Java can be seen in the Table 1.

Table 1. VIF value of several independent variables at every location

Districts/Cities	X4	X5	X7
South Jakarta	37.3813	11.3282	13.8838
East Jakarta	37.2131	11.2789	13.8244
Central Jakarta	37.358	11.3181	13.8722
West Jakarta	37.5034	11.3595	13.9221
North Jakarta	37.3786	11.3226	13.8779
Bogor	36.9397	11.2093	13.7387
Sukabumi	37.0811	11.2683	13.8069
Cianjur	36.5015	11.0999	13.603
Bandung	35.9347	10.9345	13.4019
Garut	35.4873	10.8169	13.2558
….			
Madiun City	31.5264	9.71995	11.8519
Surabaya City	30.7817	9.52102	11.5739
Batu	30.7359	9.5247	11.5744
Pandeglang	38.9229	11.7844	14.4299
Lebak	38.1077	11.5515	14.1503
Tangerang	37.9161	11.48	14.067
Serang	38.5134	11.6504	14.2716
Tangerang City	37.6445	11.401	13.9721
Cilegon	38.7696	11.7208	14.3564
Serang City	38.4596	11.6351	14.2533
South Tangerang	37.4812	11.3575	13.919

VIF values X tend to be high in some districts / cities as described as in table 4.1. This indicates that multikolinieritas cases occur in the model for each observation location.

One of the ways to solve the local multicollinearity problem is using GWPCA method. This method will will generate a new variable or called by the main component which is a linear combination of the original variables. The main component is formed on the whole are as many as 11 components.

Examples of the main components that form to the area of South Jakarta:

$$KU1 = 0.196134X_1 + 0.034273X_2 - 0.37769X_3 + \cdots + 0.054623X_{11}$$

Besides the main component which is formed for each location, the criteria of eigen valuecan be used to determine the number of major components which represent the whole original variables.

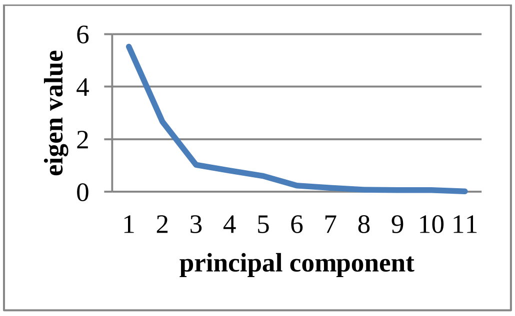


Figure 1. Eigen value of South Jakarta location

Figure 1 described eigen value which has value more than 1 are the first three eigen traits, this case shows that there are three main component pieces to be used to represent the entire original variables. Or have *r* pieces major components as the largest contributor to the diversity of data that provides a total diversity of more than 0.75 or $\sum_{i=1}^{r} \lambda_i > 0.75$ presented in Figure 2 below:

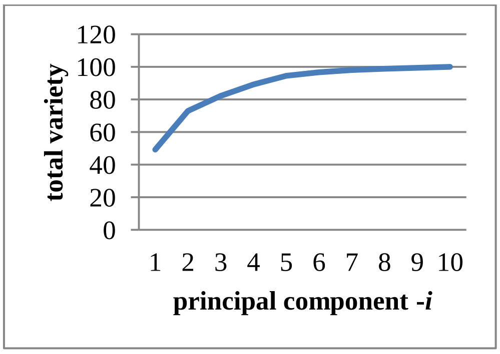


Figure 2. Total variety of South Jakarta location

From the three major components obtained, then it will be determined the value of newly formed or variables. The main components value that was formed which will be used as new variables for the benefit of building the analysis model of GWPCA.

Model derived from the analysis is 113 models which are as many districts / cities in Java. Examples of models that form for the location of South Jakarta are:

$$y = 0.131775 - 0.41591\, W_1 - 0.54559\, W_2 + 0.20678\, W_3$$

Having analyzed using GWPCA, the VIF values obtained are as follows:

Table 2. VIF Value of analyzed using GWPCA

Location	W1	W2	W3
South Jakarta	1.404587	1.214101	1.593401
East Jakarta	1.368899	1.189852	1.536564
Central Jakarta	1.403975	1.209267	1.589122
West Jakarta	1.436703	1.229679	1.640061
North Jakarta	1.410276	1.211599	1.597714
…			
Lebak	1.526046	1.317114	1.807274
Tangerang	1.521398	1.287118	1.777056
Serang	1.634496	1.358343	1.961414
Kota Tangerang	1.46604	1.249885	1.687433
Cilegon	1.682185	1.383456	2.037309
Serang City	1.625076	1.3526	1.945874
South Tangerang	1.425599	1.228461	1.627078

CONCLUSION

Results of the analysis indicate that GWPCA method can overcome local multicollinearity problem, it can be seen from the characteristic value of VIF which is smaller than 10.

REFERENCES

Anselin L. 1988. *Spatial Econometrics. Method and Model*. Kluwer Academic Publisher. Netherland.

Arbia G. 2006. *Spatial Econometrics: Statistical Foundation and Application to Regional Convergence*. Berlin: Springer.

[BPS]. Badan Pusat Statistik. 2014. Produk Domestik Regional Bruto (PDRB) Kabupaten/Kota di Indonesia 2009-2010. BPS. Jakarta

Efron B, Hastie T, Johnstone I, Tibshirani R. (2004). *Least angle regression*. Annals of Statistics 32 (2) : 407 – 451

Fatulloh 2013. Penerapkan metode Regresi Terboboti Geografis (RTG) untuk data Produk Domestik Regional Bruto (PDRB) di Pulau Jawa tahun 2010. [skripsi]. Institut Pertanian Bogor. Bogor

Fotheringham S, Brunsdon C, and Charlton M. 2002. *Geographically Weighted Regression*: The Analysis of Spatially Varying Relationships. John Willey and Sons. New York.

Gollini I, Lu B, Charlton M, Brunsdon C, and Harris P. 2013. GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. (http://arxiv.org/pdf/1306.0413.pdf.)

Harris P, Brunsdon C, Charlton M. 2011. Geographically weighted principal components analysis. *International Journal of Geographical Information Science*. 25(10): 1717-1736.

Johnson RA, and Wichern DW. 2007. *Applied Multivariate Statistical Analysis*. Sixth edition, Prentice Hall. New Jersey

Leung Y, Mei CL, dan Zhang WX. 2000. *Statistical tests for spatial nonstationarity based on the Regresi Terboboti Geografismodel, Journal of Environ Plan A*, 32,: 9-32.

Myers RH. 1990. Classical and Modern Regression With Application. Second Edition. PWS-Kent Publishing Company, Boston

Tibshirani, R. 1996. Regression Shrinkage and Selection Via The Lasso. *Journal of the Royal Statistical Society*. 58(1) :267-288.

Wheeler D, Tiefelsdorf M. 2005. *Multicollinearity and correlation among local regression coefficients in geographically weighted regression*. Journal of Geographical Systems 7 :161 – 187