Geographically Weighted Polynomial Regression: Selection of the Optimal Bandwidth and the Optimal Polynomial Degrees and Its Application to Water Quality Index Modelling

Toha Saifudin[1], Fatmawati[2] and Nur Chamidah[1]
[1]*Department of Statistics, Airlangga University, Surabaya, Indonesia*
[2]*Department of Mathematics, Airlangga University, Surabaya, Indonesia*

Keywords: Spatial Modelling, Varying Coefficients, Nonlinear Relationships, Model Comparison, Water Quality Index.

Abstract: In this paper we introduce geographically weighted polynomial regression (GWPolR) model as a generalization of GWR model. It is an alternative solution to overcome the existence of nonlinear relationships between response variable and one or more explanatory variables involved in spatial modelling. This study aims to provide a procedure for finding the optimal bandwidth and polynomial degrees in the GWPolR technique. This procedure is applied to Water Quality Index (WQI) modelling based on several factors. Because GWR method does not account for nonlinearity relationships of the spatial data type, we hypothesize that a GWPolR model will help us better understand how the factors are related to WQI patterns. Both types of models were applied to examine the relationship between WQI and various explanatory variables in 33 provinces of Indonesia. The goal was to determine which approach yielded a better predictive model. Based on three explanatory variables, i.e. percentage of untreated waste, population density, and number of micro industries, the GWR produced a spatial precision, i.e. R^2, of 35.28%. GWPolR efforts increased the value explained by explanatory variables with better spatial precision (R^2 = 50.12%). The results of GWPolR approach provide more complete understanding of how each explanatory variable is related to WQI, which should allow improved planning of explanatory management strategies.

1 INTRODUCTION

The geographically weighted regression (GWR) model has been one of the useful methods in spatial analysis (Fotheringham et al., 2002). The GWR technique has been studied both in theory and application. In the scope of theory, many authors have studied the GWR technique, for example: Brunsdon et al. (1999), and Fotheringham et al. (1998; 2002). In application, the GWR technique has been also widely applied to different areas, for example: in climatology (Al-Ahmadi& Al-Ahmadi, 2013; Brunsdon et al., 2001; and Wang et al., 2012), in econometric (Mittal et al., 2004; Lu et al., 2014), in social field (Fotheringham et al., 2001; Han & Gorman, 2013). From those studies, some procedures relating to the GWR model have been established.

The GWR coefficients are spatially varying. However, it is important to remember that the GWR model is an expansion of global linear regression (GLR) model, so the response variable in each location is fitted as a linear function of a set of explanatory variables. It may be unrealistic in some real-life situations. There are many possibilities of nonlinearity cases in the relationships between one or more explanatory variables and the response. In a health study, the relationship between age and child weight tend to be nonlinear. In economic study, the relationship between advertising finance and the revenue is commonly nonlinear. In application of spatial analysis, Chamidah et al. (2014) inspected the vulnerability modelling of dengue hemorrhagic fever (DHF) disease in Surabaya based on geographically regression. The results obtained have not been satisfactory. The existence of nonlinear relationships between one or more explanatory variables and the DHF level is suspected to be the cause. In other example, Chiang et al. (2015) showed that the influence of the convenience factor (access to public facilities) is nonlinear over the housing prices in Taipei, Taiwan. If the nonlinear relationships are present in the real situation, then the linear approach may be unrealistic. Therefore, some approach models

Saifudin, T., Fatmawati, . and Chamidah, N.
Geographically Weighted Polynomial Regression:Selection of the Optimal Bandwidth and the Optimal Polynomial Degrees and Its Application to Water Quality Index Modelling.
DOI: 10.5220/0008517700930100
In *Proceedings of the International Conference on Mathematics and Islam (ICMIs 2018)*, pages 93-100
ISBN: 978-989-758-407-7

which accommodate the actual pattern of the real data are required to improve the basic GWR model.

Some expansions of GWR have been proposed in recent years. One of the important expansions is geographically weighted generalized linear models (GWGLM) covering geographically weighted poisson regression (GWPR) and geographically weighted logistic regression (GWLR). Although there was GWGLM, an extension of GWR which accommodates response in continuous variable and has nonlinear relationships with the explanatory variables has not been found. Thus, an extension of the GWR model which can overcome the problems described above is needed.

As a solution to the above problem, we introduce a generalization of the basic GWR model by using polynomial function approach. Then, it is called the geographically weighted polynomial regression (GWPolR) model. Here, we provide an analytical formula for the coefficient estimator which still depends on a bandwidth and several polynomial degrees. Next, the purpose of this paper is to provide a procedure for selecting the optimal bandwidth and the optimal polynomial degree of each explanatory variable involved in the model. Then, as an example of application we provide a modelling of water quality index in Indonesia.

2 GEOGRAPHICALLY WEIGHTED POLYNOMIAL REGRESSION

In this section, we will introduce the GWPolR model and a procedure to get its optimal estimator.

2.1 Model and Weighted Least Squares Estimation

We briefly review the basic GWR model from Brunsdon et al. (1996; 1999) and Fotheringham et al. (1998; 2002). The GWR model is in the form of

$$y_i = \sum_{j=1}^{p} \beta_j(u_i, v_i)x_{ij} + \varepsilon_i, i = 1, 2, \dots, n, \qquad (1)$$

where $\varepsilon_i$is distributed $N(0, \sigma^2)$.

We expand the linear relationship in equation (1) by using polynomial function approach as follows

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p}\sum_{j=1}^{d_k} \beta_{k,j}(u_i, v_i){x_{ik}}^{j} + \varepsilon_i, \qquad (2)$$

where $\varepsilon_i \sim N(0, \sigma^2)$. In a matrix form, it can be written as

$$y_i = \boldsymbol{x}_{\boldsymbol{i}}^{*\mathrm{T}}\boldsymbol{\beta}_{pol}(u_i, v_i) + \varepsilon_i\,, \quad i = 1,2, \dots, n, \qquad (3)$$

where

$$\boldsymbol{x}_{\boldsymbol{i}}^{*\mathrm{T}} = \left(1x_{i1}x_{i1}^2 \;\cdots\; x_{i1}^{d_1} \;\cdots\; x_{ip}x_{ip}^2 \;\cdots\; x_{ip}^{d_p}\right), \qquad (4)$$

and

$$\boldsymbol{\beta}_{Pol}^{\mathrm{T}}(u_i, v_i) = (\,\beta_0(u_i, v_i)\beta_{1,1}(u_i, v_i)\beta_{1,2}(u_i, v_i)\cdots\beta_{1,d_1}(u_i, v_i)\cdots \beta_{p,1}(u_i, v_i)\beta_{p,2}(u_i, v_i)\cdots\beta_{p,d_p}(u_i, v_i)\,). \qquad (5)$$

For a given location (u_0, v_0), we can estimate $\boldsymbol{\beta}_{Pol}(u_0, v_0)$by minimizing the weighted least square function as follows

$$\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_{\boldsymbol{i}}^{*\mathrm{T}}\boldsymbol{\beta}_{pol}(u_0, v_0)\right)^2 w_i(0), \qquad (6)$$

with respect to each element of $\boldsymbol{\beta}_{pol}(u_0, v_0)$. An explicit expression of the solution is

$$\hat{\boldsymbol{\beta}}_{pol}(u_0, v_0) = \left[\boldsymbol{X}_{pol}^{\mathrm{T}}\boldsymbol{W}(u_0, v_0)\boldsymbol{X}_{pol}\right]^{-1}\boldsymbol{X}_{pol}^{\mathrm{T}}\boldsymbol{W}(u_0, v_0)\boldsymbol{y}, \qquad (7)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^{\mathrm{T}}$,

$$\boldsymbol{X}_{pol} = \begin{bmatrix} 1x_{11}x_{11}^2 & \cdots & x_{11}^{d_1} & \cdots & x_{1p}x_{1p}^2 & \cdots & x_{1p}^{d_p} \\ 1x_{21}x_{21}^2 & \cdots & x_{21}^{d_1} & \cdots & x_{2p}x_{2p}^2 & \cdots & x_{2p}^{d_p} \\ & & & \vdots & & & \\ 1x_{n1}x_{n1}^2 & \cdots & x_{n1}^{d_1} & \cdots & x_{np}x_{np}^2 & \cdots & x_{np}^{d_p} \end{bmatrix}, \qquad (8)$$

and

$$\mathbf{W}(u_0, v_0) = \mathrm{diag}[w_1(0), w_2(0), \dots, w_n(0)], \qquad (9)$$

is a diagonal weighting matrix with the elements $w_i(0)$for $i = 1, 2, \dots, n$.

By taking (u_0, v_0) to be each of the designed locations(u_i, v_i), $i = 1, 2, \dots, n$, we can obtain the vector of the fitted values for the response y at n designed locations as

$$\hat{\boldsymbol{y}}_{Pol} = (\hat{y}_1^*, \hat{y}_2^*, \dots, \hat{y}_n^*)^{\mathrm{T}} = \mathbf{G}\,\mathbf{y}, \qquad (10)$$

where

$$\mathbf{G}=\begin{bmatrix} \boldsymbol{x}_1^{*T}\left[\mathbf{X}_{pol}^T\mathbf{W}(u_1,v_1)\mathbf{X}_{pol}\right]^{-1}\mathbf{X}_{pol}^T\mathbf{W}(u_1,v_1) \\ \boldsymbol{x}_2^{*T}\left[\mathbf{X}_{pol}^T\mathbf{W}(u_2,v_2)\mathbf{X}_{pol}\right]^{-1}\mathbf{X}_{pol}^T\mathbf{W}(u_2,v_2) \\ \vdots \\ \boldsymbol{x}_n^{*T}\left[\mathbf{X}_{pol}^T\mathbf{W}(u_n,v_n)\mathbf{X}_{pol}\right]^{-1}\mathbf{X}_{pol}^T\mathbf{W}(u_n,v_n) \end{bmatrix}, \quad (11)$$

is called a hat matrix of GWPolR model. Based on the hat matrix **G**, the residual vector is

$$\hat{\boldsymbol{\varepsilon}}_{Pol}=\boldsymbol{y}-\hat{\boldsymbol{y}}_{Pol}=(\mathbf{I}-\mathbf{G})\boldsymbol{y}, \qquad (12)$$

and the residual sum of squares(RSS$_{\text{pol}}$) is

$$\text{RSS}_{\text{pol}}=\hat{\boldsymbol{\varepsilon}}_{pol}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_{Pol}=\ \boldsymbol{y}^{\mathrm{T}}(\mathbf{I}-\mathbf{G})^{\mathrm{T}}(\mathbf{I}-\mathbf{G})\boldsymbol{y}\ , \quad (13)$$

where **I** is identity matrix of order *n*.

2.2 Spatial Weighting Functions

In spatial analysis, the weighting matrix $\mathbf{W}(u_i,v_i)$ contains of different emphases on different observations in producing the estimated parameters. The first approach of the weighting matrix at location *i* can be expressed as

$$w_j(i)=\begin{cases}1,\ if d_{ij}\le h, & j=1,2,\dots,n.\\ 0,\ if d_{ij}>h.\end{cases} \qquad (14)$$

The weighting function written above has a discontinuity problem. One method to solve the problem is to specify $w_j\,(i)$ as a continuous function of d_{ij}. One preference might be

$$w_j\,(i)=exp\left(-\tfrac{1}{2}\left(\tfrac{d_{ij}}{h}\right)^2\right),\ \ j=1,2,\dots,n, \qquad (15)$$

where *h* denotes the bandwidth. It is the most common choice in practice (Fotheringham et al., 2002).

An alternative weighting function can be created by the bisquare kernel weighting function as follows

$$w_j\,(i)=\begin{cases}\left(1-\left(\frac{d_{ij}}{h}\right)^2\right)^2,\ if\ d_{ij}\le h, & j=1,2,\dots,n.\\ 0, \qquad if\ d_{ij}>h.\end{cases} \quad (16)$$

2.3 Cross-Validation Criterion for Choosing the Optimal Model

The estimator of GWPolR depends on the weighting function selected and on the polynomial degree of each explanatory variable. So, the bandwidth *h* and the number of the polynomial degrees should be determined.

The problem is how to select the optimal bandwidth and the optimal number of polynomial degree of each explanatory variable. C*ross-validation* (CV) approach is a solution to this problem (Fotheringham et al, 2002). Here, we adopt such procedure by adding polynomial degrees for explanatory variables which should be selected. So, we have

$$\text{CV}(h,d_1,d_2,\dots,d_p)=\textstyle\sum_{i=1}^{n}\left(y_i-\hat{y}_{(i)}(h,d_1,d_2,\dots,d_p\,)\right)^2 \quad (17)$$

as an objective function, where $\hat{y}_{(i)}(h,d_1,d_2,\dots,d_p)$ is the fitted value of $y_i$under bandwidth *h* and degrees of polynomial $d_1,d_2,\dots,d_p$ with the observation at location (u_i,v_i) omitted from the process of estimation. Select $h_0,\ d_{1\text{o}},d_{2\text{o}},\dots,d_{p\text{o}}$ as the optimal values, such that the equation (17) is minimum.

2.4 An Algorithm for Finding Bandwidth and Polynomial Degrees in Optimal Condition

Here, we provide an algorithm to select the optimal bandwidth and the optimal number of polynomial degrees based on the CV criterion as follows:

1) Determine the number of explanatory variables involved in the model, denoted by p.
2) Specify the maximum polynomial degree for each explanatory variable, denoted by d_j for $j=1,2,\dots,p$.
3) Find all arrays of numbers obtained from the existing polynomial degrees for all explanatory variables. Let s be the number of arrays, then $s=\prod_{j=1}^{p}d_j$.
4) Find the minimum CV value of GWPolR modelling in each array.
5) Find the smallest CV value among the minimum CV values generated from the entire arrays.
6) Select the bandwidth and polynomial degrees that yield the smallest CV value obtained in step 5 as the optimal solution.

To explain the above algorithm, we will give the following illustration. For example, based on a given spatial dataset we use three explanatory variables in the model, so we have $p=3$. Then, we specify the maximum polynomial degree for each explanatory variable, for instance we set $d_1=2$, $d_2=3$ and $d_3=2$ for the 1st, 2nd and 3rd explanatory variables, respectively. Based on the setting, we have several $s=2\times3\times2=12$ arrays of polynomial degrees. These arrays are listed in Table 1.

Table 1: All possible arrays of polynomial degrees for the setting of $\square_1 = 2$, $\square_2 = 3$ and $\square_3 = 2$.

Number	Array
1	(1, 1, 1)
2	(1, 1, 2)
3	(1, 2, 1)
4	(1, 2, 2)
5	(1, 3, 1)
6	(1, 3, 2)
7	(2, 1, 1)
8	(2, 1, 2)
9	(2, 2, 1)
10	(2, 2, 2)
11	(2, 3, 1)
12	(2, 3, 2)

An array represents the polynomial degree of the 1st, 2nd and 3rd explanatory variables, respectively. For example, in number 3 we have an array of (1, 2, 1) which means that the polynomial degree of the 1st, 2nd and 3rd explanatory variables are 1, 2, and 1, respectively. This leads to a model as follows

$$y_i = \beta_0(u_i, v_i) + \beta_{1,1}(u_i, v_i)x_{i1} + \beta_{2,1}(u_i, v_i)x_{i2} + \beta_{2,2}(u_i, v_i)x_{i2}^2 + \beta_{3,1}(u_i, v_i)x_{i3} + \varepsilon_i. \quad (18)$$

Using the model in (18), we select the minimum CV value based on GWPolR estimation procedure. The same selection is conducted on other arrays. So, we have 12 minimum CV values. Then, we select the smallest CV value among the existing 12 minimum CV values. In the last step, we take the bandwidth and polynomial degrees corresponding to the smallest CV value obtained.

3 AN EMPIRICAL EXAMPLE: WATER QUALITY INDEX MODELLING

In this section, we will apply the algorithm explained above to water quality index modelling. In addition, we also provided modelling results to such data using global linear regression (GLR) and GWR for comparison.

3.1 Water Quality Index Dataset

Some researchers have stated that WQI is influenced by many factors, including untreated waste (Dhawde et al., 2018), population density (Opaminola and Jessie, 2015; Liyanage and Yamada, 2017), and the number of micro industries (Bhutiani et al., 2018). The explanation from these researchers led the authors to include the variables in the example here.

The involved variables in this study are water quality index (WQI) as the response variable, and three explanatory variables, i.e. percentage of untreated waste (PUW), population density (PD), and number of micro industries (NMI). Data was provided by Ministry of Environment and Forestry Republic of Indonesia (Kementerian Lingkungan Hidup, 2015) and Statistics of Indonesia (BPS, 2016; 2017a; 2017b). Observation units are 33 provinces of Indonesia in 2014.

Logically in public opinion, WQI variable has contrary relationship with each explanatory variable here. If the value of each explanatory variable increases, it will cause in decreasing of the WQI value. It means that the parameter of each explanatory variable should have negative sign. The conformity of the estimator signs will also be considered in the comparison of models.

3.2 Global Linear Regression Results

Firstly, WQI is fitted by using GLR model. Based on three explanatory variables declared above, the simultaneous test for parameters yields a *p*-value of 0.007. It means that the parameters affect to the response variable simultaneously. Here, we have four global parameters including the intercept. Furthermore, the results of GLR estimation on the WQI dataset is listed in Table 2.

Table 2: Summary of global linear regression results on the WQI dataset.

Predictor	Coef	SE Coef	T	*p*-value
Intercept	36.92	16.67	2.21	0.035
PUW	0.2228	0.2020	1.10	0.279
PD	-0.0015	0.0004	-3.45	0.002
NMI	-0.0079	0.0064	-1.24	0.226

From table 2, based on partial test with significance level of 0.05 we know that only PD variable affect significantly to the WQI (*p*-value < 0.05), whereas PUW and NMI variables are not significant (*p*-value > 0.05). We suspect that PUW and NMI variables may not have linear relationships with the response, but they may have nonlinear relationships. Here we don’t continue with remodeling that uses significant explanatory variables, but we let the results. We would like to see the comparison of results with other modelling (especially using polynomial approach) involving the same explanatory variables. In other words, if a model yielded better results while we didn't apply the same variables to all models in a comparison, we would not know if the improvement was due to the modelling approach or the explanatory variables that was used to build each model.

To clarify our allegation, we examine the misspecification function using Ramsey RESET. For this test, we hypothesize

H0: there is not misspecification function in global linear regression
H1: there is misspecification function in global linear regression

By using syntax of `resettest()` in R statistical software with optional input `power=2` (for testing the existence of polynomial degree of 2), we find that the *p*-value of Ramsey RESET test is 0.0415. If we take a significance level of 0.05, we decide to reject the null hypothesis and conclude that there is misspecification function in GLR modelling. Then, this conclusion becomes an early reason for the need to extend the linear modelling with polynomial modelling on WQI dataset.

3.3 Geographically Weighted Regression Results

This WQI dataset was drawn based on spatial area, therefore proceeding with GWR model was warranted. The GWR model in this study was implemented using the following model:

$$WQI(u_i, v_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)PUW_i \\ +\beta_2(u_i, v_i)PD_i + \beta_3(u_i, v_i)NMI_i + \varepsilon_i. \quad (19)$$

Based on this model and using weighting function in equation (15), we found that the R^2 of GWR was 0.3528 with a bandwidth of 31.3° (or equals to 3,484.378 km) when the minimum CV was 10232.400. Our preferred measure of model fit, RSS, gave a value of 1,253.488. Here, we have several four parameters that vary in 33 locations. The summary of the GWR estimators is listed in Table 3.

We can see in Table 3 that the estimated parameter of PD and NMI variables have negative sign in every quartile, which is suitable with the public opinion mentioned above. On the other hand, the estimated parameter of PUW does not have negative value at all. It is a contradiction to the public opinion. We will try to correct this case by using GWPolR modelling in the next subsection.

Table 3: Summary of geographically weighted regression results on the WQI dataset.

Coef. of Predictor	Min	1st Quartil	Med	3rd Quartil	Max
Intercept	30.900	33.490	35.4000	38.160	43.6300
PUW	0.13850	0.20720	0.24160	0.26540	0.29780
PD	-0.00161	-0.00159	-0.00158	-0.00156	-0.00153
NMI	-0.00797	-0.00794	-0.00791	-0.00789	-0.00775

3.4 Geographically Weighted Polynomial Regression Results

Because of the presence of nonlinear relationships in WQI dataset, we suspected that a GWPolR model would help us better understand how explanatory variables were related to WQI patterns. In this subsection, we use an algorithm for selecting the optimal bandwidth and polynomial degrees explained above.

Based on WQI dataset we involved three explanatory variables in the model, so we have $p = 3$. Then, we specified the maximum polynomial degree for each explanatory variable. Here, we set the same value, i.e. $d_1 = d_2 = d_3 = 2$ for the maximum polynomial degree of PUW, PD, and NMI variables. Based on the setting, we had number of $s = 2 \times 2 \times 2 = 8$ arrays of polynomial degrees. Further, we selected the minimum CV value based on GWPolR estimation procedure in each array. The minimum CV value and the accordingly optimal bandwidth are listed in Table 4.

The smallest CV value among eight minimum CV values was 4878.902. It is found in the row of number six, according to the optimal bandwidth of 9 (or equals to 1,001.898 km) and array of (2, 1, 2). This array means that the optimal polynomial degree for PUW, PD, and NMI variables are 2, 1, and 2, respectively. So, the GWPolR model under optimal condition in this study was

$$WQI(u_i, v_i) = \beta_0(u_i, v_i) + \beta_{1,1}(u_i, v_i)PUW_i \\ +\beta_{1,2}(u_i, v_i){PUW_i}^2 + \beta_{2,1}(u_i, v_i)PD_i \\ +\beta_{3,1}(u_i, v_i)NMI_i + \beta_{3,2}(u_i, v_i){NMI_i}^2 + \varepsilon_i. \quad (20)$$

Based on model in equation (20) and weighting function in equation (15) we found that the R^2 of GWPolR was 0.5012.A goodness indicator of model fit, RSS, gave a value of 966.1458. Here, we have several six parameters that vary in 33 locations. Furthermore, the summary of the GWPolR estimators is listed in Table 5.

Table 4: Optimal bandwidth and minimum CV for each array of polynomial degrees on the WQI dataset.

Number	Array	Opt *h*	Minimum CV
1	(1, 1, 1)	31	10253.855
2	(1, 1, 2)	42	11032.556
3	(1, 2, 1)	1	95336.878
4	(1, 2, 2)	1	95336.878
5	(2, 1, 1)	42	20854.561
6	(2, 1, 2)	9	4878.902
7	(2, 2, 1)	1	95336.878
8	(2, 2, 2)	1	95336.878

Table 5: Summary of geographically weighted polynomial regression results on the WQI dataset.

Coef. Of Predictor	Min.	1st Quartile	Median	3rd Quartile	Max.
Intercept	-35.0200	64.4100	203.2000	273.0000	322.5000
PUW	-7.1840	-5.9930	-4.1800	-0.4828	2.4000
PUW^2	-0.01593	0.00331	0.02799	0.03972	0.04789
PD	-0.00182	-0.00179	-0.00174	-0.00147	-0.00122
NMI	-0.04314	-0.04153	-0.00371	-0.02613	-0.01414
NMI^2	0.00001	0.00002	0.00004	0.000046	0.000049

From Table 5, the estimated parameter for the first degree of explanatory variables involved in the model have negative sign in almost of all positions. It means that the estimated parameters of GWPolR are according to the public opinion explained above. In addition, another information can be obtained from the GWPolR modelling. The estimated parameters of the second-degree polynomial variables have positive sign in almost of all positions. It interprets that the PUW and NMI variables have an accelerating effect to the WQI decrease.

3.5 A Comparison on Water Quality Index Modelling

In this subsection, we make comparison about the results obtained under the optimal condition of the GWPolR, GWR, and GLR modelling on the WQI dataset. The comparison is based on some goodness of fit indicators (including CV, RSS, and R^2) and several other criteria. The results are presented in Table 6.

We know that the minimum CV is a criterion for finding the best fitted model. From Table 6, the CV value of the GWPolR model is much lower than that of the GWR model.

Table 6: The comparison of the GWPolR, GWR, and GLR modelling on the WQI dataset.

Indicator	Model		
	GLR	GWR	GWPolR
Optimal *h*	-	31.3	9.0
Minimum CV	-	10232.400	4878.902
RSS	1289.380	1253.488	966.145
R^2	33.40%	35.28%	50.12%
Conformity to public opinion	No	No	Yes
Number of parameters	Simple (Parsimony)	Complex	More complex

Table 6 also presents that the RSS of GWPolR is the lowest among the RSS of involved models here. In view of GWPolR model, there are RSS decrease of 323.235 and 287.343 from GLR and GWR, respectively. The coefficient of determination (R^2) of the GWPolR model can capture the largest amount (50.12%) of variance of water quality index based on the explanatory variables. There are R^2 increase of 16.72 and 14.84 from GLR and GWR, respectively. These are relatively strong evidence of an improvement in the model fit to the data. It means that GWPolR model is the best model among the studied models here based on goodness of fit indicators.

Furthermore, the estimated parameters of GWPolR confirm to the public opinion about the relation of WQI and the explanatory variables. The other two models give inappropriate results. In addition, here GWPolR has a parameter for the 2nd polynomial degree which interprets that the speed of response change caused by an explanatory variable is not constant. In other word, there is an acceleration here. So, the GWPolR modelling has more complete interpretation. On the other hand, The GWPolR model has several more complex parameters. In modelling, people often prefer to use a model with a small number of parameters to be easily interpreted. Therefore, a statistical test especially inter spatial modelling is needed to examine whether a GWPolR is significantly better than a GWR or not. If GWPolR is proven to be significantly better than GWR even though the number of parameters is more complex, then GWPolR should be selected.

4 CONCLUSIONS

The algorithm written in this paper yields a bandwidth and some polynomial degrees for the GWPolR model in optimal condition. It means that the result is the best condition based on the used criterion. However, computational programming based on this algorithm still takes a long time. The time will be longer when the number of variables involved in model increases or the maximum degree of polynomial is set greater. An efficient algorithm in term of execution time is needed even though the results may be only sub-optimal, for example, genetic algorithm or neural network. Based on the goodness of fit criteria(inter alia: CV, RSS, and R^2) and consideration of conformity with public opinion, we can empirically conclude that the GWPolR model is the best model among some models used here on WQI dataset. Nevertheless, statistical tests between spatial modelling need to be developed to determine whether a GWPolR is significantly better than a GWR or not.

ACKNOWLEDGEMENTS

The authors would like to thank Directorate General of Science and Technology Resources and Higher Education of The Ministry of Research, Technology and Higher Education of Indonesia for scholarship that be received by the first author throughout undergoes doctoral program. The authors also thank to Ministry of Environment and Forestry of Indonesia and Statistics of Indonesia for kindly providing the data.

REFERENCES

Al-Ahmadi, K., Al-Ahmadi, S., 2013. Rainfall-Altidude Relationship in Saudi Arabia. *Advances in Meteorology*, [Internet] [Citation: 25 January 2017]. Obtained from: http://dx.doi.org/10.1155/2013/363029.

BPS, 2016. Jumlah Perusahaan Industri Mikro dan Kecil Menurut Provinsi, 2013–2015. [Internet] [Citation: 20 June 2018]. Obtained from: https://bps.go.id.

BPS, 2017a. Kepadatan Penduduk menurut Provinsi, 2000–2015. [Internet] [Citation: 20 June 2018]. Obtained from: https://bps.go.id.

BPS, 2017b. Persentase Rumah Tangga Menurut Provinsi dan Perlakuan Memilah Sampah Mudah Membusuk dan Tidak Mudah Membusuk, 2013–2014. [Internet] [Citation: 20 June 2018]. Obtained from: https://webapi.bps.go.id.

Brunsdon, C., Fotheringham, A. S., Charlton, M., 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4): 281–298.

Brunsdon, C., Fotheringham, A. S., Charlton, M., 1999. Some Notes on Parametric Significance Tests for Geographically Weighted Regression. *Journal of Regional Science*, 38(3): 497–524.

Brunsdon, C., McClatchey, J., Unwin, D. J., 2001. Spatial Variations in the Average Rainfall–Altitude Relationship in Great Britain: An Approach using Geographically Weighted Regression. *International Journal of Climatology*, 21: 455–466.

Bhutiani, Faheem, A., Varun, T., Khushi, R., 2018. Evaluation of water quality of River Malin using water quality index (WQI) at Najibabad, Bijnor (UP) India. *Environment Conservation Journal*, 19 (1&2): 191–201. https://www.researchgate.net/publication/325986999

Chamidah, N., Saifuddin, T., Rifada, M., 2014. The Vulnerability Modelling of Dengue Hemorrhagic Fever Disease in Surabaya Based on Spatial Logistic Regression Approach. *Applied Mathematical Sciences*, 8(28): 1369 – 1379.

Chiang, Y-H., Peng, T-C., Chang, C-O., 2015. The nonlinear effect of convenience stores on residential property prices: A case study of Taipei, Taiwan. *Habitat International*, 46: 82–90.

Dhawde, R., Surve, N., Macaden, R., Wennberg, A.C., 2018. Physicochemical and Bacteriological Analysis of Water Quality in Drought Prone Areas of Pune and Satara Districts of Maharashtra, India. *Environment*, 5, 61, doi: 10.3390/environments5050061. [Internet] [Citation: 20 January 2018]. Obtained from: http://dx.doi.org/10.3390/ environments5050061.

Fotheringham, A. S., Charlton, M. E., Brunsdon, C., 1998. Geographically Weighted Regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30: 1905–1927.

Fotheringham, A. S., Charlton, M. E., Brunsdon, C., 2001. Spatial Variation in School Performance: A Local Analysis using Geographically Weighted Regression. *Geographical & Explanatory Modelling*, 5(1): 43–66.

Fotheringham, A. S., Brunsdon, C., Chartlon, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley and Sons, USA.

Han, D., Gorman, D. M., 2013. Exploring Spatial Associations between On-Sale Alcohol Availability, Neighborhood Population Characteristic, and Violent Crime in a Geographically Isolated City. *Journal of Addiction*, [Internet] [Citation: 20 May 2017]. Obtained from: http://dx.doi.org/10.1155/2013/ 356152.

Kementerian Lingkungan Hidup, 2015. Indeks Kualitas Lingkungan Hidup Indonesia 2014. Kementerian Lingkungan Hidup. [Internet] [Citation: 20 June 2018]. Obtained from: www.menlhk.go.id/downlot.php?file= iklh2014.pdf.

Liyanage, C. P., Yamada, K., 2017. Impact of Population Growth on the Water Quality of Natural Water Bodies. *Sustainability*, 9, 1405. [Internet] [Citation: 20 January

2018]. Obtained from: http://dx.doi.org/10.3390/su9081405.

Lu, B., Charlton, M., Harris, P., Fotheringham, A. S., 2014. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International journal of Geographical Information Science*, [Internet] [Citation: 30 May 2016]. Obtained from: http://dx.doi.org/10.1080/13658816.2013.865739.

Mittal, V., Kamakura, W. A., Govind, R., 2004. Geographic Pattern in Customer Service and Satisfaction: An Empirical Investigation. *Journal of Marketing*, 68: 48–62.

Opaminola, D., Jessie, E., 2015. Effects of Population Density on Water Quality in Calabar Municipality Cross River State, Nigeria. *Journal of Environment and Earth Science*, 5(6): 20–31.

Wang, C., Zhang, J., Yan, X., 2012. The Use of Geographically Weighted Regression for the Relationship among Extreme Climate Indices in China. *Mathematical Problems in Engineering*, Vol. 2012, Article ID 369539: 1–15. [Internet] [Citation: 27 January 2017]. Obtained from: http://dx.doi.org/10.1155/2012/369539.