em.Int-05-2018-Warsito_2018_J._Phys.-_Conf._Ser._1025_012096.pdf

by

PAPER • OPEN ACCESS

Robust geographically weighted regression of modeling the Air Polluter Standard Index (APSI)

IOP Conf. Series: Journal of Physics: Conf. Series **1025** (2018) 012096 doi:10.1088/1742-6596/1025/1/012096

Robust geographically weighted regression of modeling the Air Polluter Standard Index (APSI)

Budi Warsito [1], Hasbi Yasin[1], Dwi Ispriyanti[1], Abdul Hoyyi[1]

[1]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia

E-mail: budiwrst2@gmail.com

Abstract. The Geographically Weighted Regression (GWR) model has been widely applied to many practical fields for exploring spatial heterogenity of a regression model. However, this method is inherently not robust to outliers. Outliers commonly exist in data sets and may lead to a distorted estimate of the underlying regression model. One of solution to handle the outliers in the regression model is to use the robust models. So this model was called Robust Geographically Weighted Regression (RGWR). This research aims to aid the government in the policy making process related to air pollution mitigation by developing a standard index model for air polluter (Air Polluter Standard Index - APSI) based on the RGWR approach. In this research, we also consider seven variables that are directly related to the air pollution level, which are the traffic velocity, the population density, the business center aspect, the air humidity, the wind velocity, the air temperature, and the area size of the urban forest. The best model is determined by the smallest AIC value. There are significance differences between Regression and RGWR in this case, but Basic GWR using the Gaussian kernel is the best model to modeling APSI because it has smallest AIC.

1. Introduction

Brunsdon et al. [1] proposes Geographically Weighted Regression (GWR). This method has been applied to several fields such as geography and environmental science, to explore the relationship of spatial regression. Since the GWR technique was introduced, it has been extensively studied in its methodology except for many applications. The earlier research includes, for example, Brunsdon et al. [1,2,3,4], Fotheringham et al. [5,6,7], Leung et al. [8,9], Pa´ez et al. [10,11], Mei et al. [12,13], Huang et al. [14] and Harini et al. [15]. Recently, Winarso et al. [16] discussed the Mixed Geographically and Temporally Weighted Regression (MGTWR). Harris et al. [17] proposed an improved GWR, method, called Robust GWR (RGWR), to handle the outliers in the GWR model. Zhang and Mei [18] also proposed least absolute deviation (LAD) to estimate the parameter of Robust GWR.

One of the real problems that threaten the environment and even threaten human life is air pollution. It is characterized by a decrease in air quality, especially in the big cities in recent years. Factors to be a major source of air pollution in large cities is transportation-engined vehicles, the exhaust gas industries, population density, shopping centers, air humidity, air temperature and wind speed, and so on. The factors which may prevent or inhibit the emergence of air pollution is the presence of many green areas and trees in city parks [19,20]. Elements of pollutant major accordance with Air Polluter Standard Index (APSI) are Carbon Monoxide (CO), Nitrogen Dioxide (NO2), sulfur dioxide (SO2), particulate matter (PM), and ozone (O3) [20]. The cause and impact of air pollution will be related to the location of the observation. Each location will have different impacts according to the character of each location. Demographically, the potential impacts and causes of air pollution will differ between regions, in addition to the impact and causes of air pollution cannot use a global approach, because by using a global approach there are local variations invisible influence [21,22]. Spatial regression method frequently used method is to use Geographically Weighted Regression (GWR), which is a regression method involving the effect of the location into the predictor [23]. The parameters of the linear regression model apply globally, while the GWR model parameters are local to each location of observation [24]. In this study we modelling the RGWR model to investigate the relationship of The traffic velocity, The population density, The air humidity, The business center aspect, The air temperature, The wind velocity, and The size of the urban forest to The APSI locally.

2. Methodology

2.1. Linear Regression

Linear regression is a method that models the relationship between response variables and predictor variables. Linear regression model for p predictor variables are generally written as follows:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \tag{1}$$

where i = 1, 2, ..., n; $\beta_0, \beta_1, ..., \beta_p$ are the model parameters and $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are the error term with zero mean and homogeneous variance σ^2. Estimation of regression parameters are done by Ordinary Least Squares (OLS) method. Testing of parameter regression model using the F distribution approach and the partial use of the t distribution approach [25].

2.2. Geographically Weighted Regression

The GWR model is the development of a linear regression model based on nonparametric regression ideas [13]. This model is a locally linear regression that produces the model parameters that are local to each location where the data was collected. GWR model can be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2}$$

In this case, y_i : observation of response ; (u_i, v_i): geographical point (*longitude*, *latitude*); $\beta_k(u_i, v_i)$: p unknown functions of geographical locations (u_i, v_i), k = 0,1,...,p; x_{ik}: explanatory variable at location (u_i, v_i) and ε_i : error term with zero mean and homogeneous variance σ^2.

Estimation of GWR model parameter are using the Weighted Least Squares (WLS) method that give a different weighting for each observation. The estimator of model parameters (3) for each location are:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{y} \tag{3}$$

Weighting function that used to estimate the parameters in the GWR model are the gaussian kernel functions [23], which can be written as follows:

$$w_j(u_i, v_i) = \exp\left[\left(-\tfrac{1}{2}\right)\left(d_{ij}/h_i\right)^2\right]$$

where d_{ij} denotes the distance between the location (u_i, v_i) to location (u_j, v_j) and h_i are nonnegative parameters are known and are usually called smoothing parameter (*bandwidth*) for location (u_i, v_i). So $\mathbf{W}(u_i, v_i) = \mathrm{diag}\left(w_1(u_i, v_i), w_2(u_i, v_i), \ldots, w_n(u_i, v_i)\right)$. One method that is used to select the optimum bandwidth is the of Cross Validation (CV) method which defined by:

$$CV(h_i) = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(h_i)\right)^2 \tag{4}$$

IOP Conf. Series: Journal of Physics: Conf. Series **1025** (2018) 012096 doi:10.1088/1742-6596/1025/1/012096

where $\hat{y}_{\neq i}(h)$ is the fitted value of y_i with the observation at location (u_i, v_i) omitted from fitting process.

2.3. The algorithm for selecting GWR Models

The GWR algorithms to choose the model consists of four steps:

Step 1. Compare all the possible GWR models by including one predictor variable at a time;

Step 2. Select the best model that generates a minimum AICc value, and include the corresponding predictor variable in the next model permanently;

Step 3. Enter in sequence a remaining predictor variable to build a new model with permanently installed predictor variables, and specify the next permanent variable of the best model that has a minimum AICc value;

Step 4. Repeat step 3 up to all predictor variables are entered permanently into the model.

In this algorithm, the predictor variables are iteratively incorporated into the model in a "forward" direction.

2.4. Robust GWR Models

There are two approaches to the powerful GWR that is directly borrowed from the powerful Maximum Likelihood Ratio paradigm [23]. The first approach is to improve the GWR model by removing observations with large residual values. Absolute errors that are valued above three are considered outliers. Another approach, an automated approach where observations with large raw errors may be reversed. The automated approach has its drawbacks, because it uses raw residues and does not easily allow unusual observation checks. Preferably, this approach is not as effective as computation as a filtered data approach. Both approaches have an element of subjectivity, in which the filtered data approach depends on the selected residual cutoff and the automated approach depends on the selected down weighting function. With this observation, only the first approach was taken in this study and applied in a global and local context as in Haris et al. [17]. The rationale for the localized version is to allow the identification of outliers on the same spatial scale as the selected GWR model [17].

2.5. Selection of the Best Model

The method that is used to select the best model is Akaike Information Criterion (AIC) which is defined as follows:

$$AIC_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right\}$$

where $\hat{\sigma}$ is the estimator of standard deviation of the error and $\mathbf{S}$ is the hat matrix, where $\hat{\mathbf{y}} = \mathbf{Sy}$. The best model selection is done by determining the model with the smallest AIC value [26].

3. Research Variables

This study aims to build APSI relationship model using RGWR model in five locations: Centre Surabaya (Jalan Ketabang kali//SUF1), South Surabaya (Jalan Masjid Al Akbar Gayungan/SUF4), West Surabaya (Jalan Simomulyo/SUF3), East Surabaya (Jalan Arief Rahman Hakim/SUF5), and Wonorejo Surabaya (Jalan Kendal Sari 117/SUF6). The respon variable is the APSI data, and the predictor variables are: The traffic density (X_1), The population density (X_2), The business center aspect (X_3), The humidity of air (X_4), The wind velocity (X_5), The air temperature (X_6), and The size of the urban forest (X_7). Research and data collection was conducted for 1 month, consisting of 2 weeks in dry season and 2 weeks in rainy season. Seven days of observation in a week, a day there are 3 taking time that is morning at 08.00-09.00 a.m, afternoon at 04.00-05.00 p.m and night at 10.00-11.00 p.m. The taking of secondary data of elemental content of PM, CO, SO_2, NO_2 and O_3 is done at five location point of APSI.

IOP Conf. Series: Journal of Physics: Conf. Series **1025** (2018) 012096 doi:10.1088/1742-6596/1025/1/012096

4. Results and Discussion

4.1. Selection GWR Models

The first step is the selection of variables in GWR modeling. Figure 1 shows that the most influential variable based od AIC value is X_4, then X_7, X_6, X_5, X_1, X_3, and X_2.

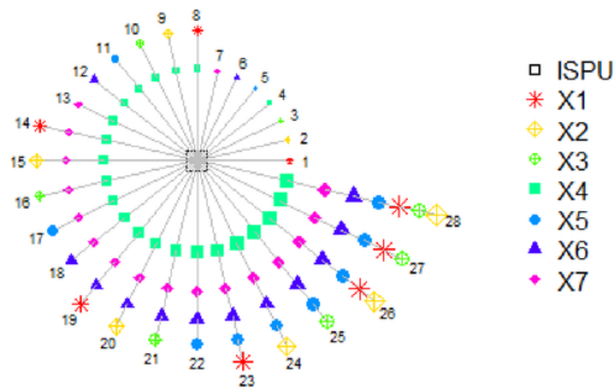

Figure 1. GWR Model Selection with Different Variables

4.2. Modeling APSI Using RGWR

Using the GWmodel R Package [27] based on AIC value, the optimum bandwidth for each location using Gaussian kernel is 0.01575. Then, using this bandwidth we estimate the RGWR model. The goodness of fits for RGWR model can be stated by the following hypothesis:

H_0 : $\beta_k(u_i, v_i) = \beta_k \; k = 0, 1, 2, \ldots, q$, and $i = 1, 2, \ldots, n$

(RGWR model is not significantly different from the Regression model)

H_1 : at least one $\beta_k(u_i, v_i) \neq \beta_k$

(RGWR model is significantly different from the regression model)

Based on Fotheringham et al. (GWR book p92), the F4 test statistical value is 4.598 (p-value = 0.046) so we can reject H_0 at level 5% and conclude that the RGWR model with Gaussian kernel bandwidth is significantly different from the regression model. Therefore, it can be said that the RGWR model is more suitable for APSI relationship model locally.

The next step is to perform the monte-carlo simulation to test the regression coefficient non-stationarity. This test conducted to select the global regression part and the GW regression part. Table 1 shows that The population density (X_2), The wind velocity (X_5), The air temperature (X_6), and The size of the urban forest (X_7) are the global regression part. Meanwhile, three other predictor variables are the GWR part because these variables have the p-value less than 0.05.

Table 1. Monte Carlo test of regression coefficient non-stationerity

Coefficient	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
p-value	0.00*	0.00*	0.22	0,00*	0.00*	1.00	1.00	1.00

Note: * significant at α=5%

IOP Conf. Series: Journal of Physics: Conf. Series **1025** (2018) 012096 doi:10.1088/1742-6596/1025/1/012096

Table 2 shows the comparison of the GWR basic model and Robust GWR using the fixed Gaussian kernel and fixed exponential kernel weighting function. The result shows that the GWR using fixed Gaussian kernel is the best model for modeling APSI in Surabaya City because it has the smallest AIC and the greatest R-Square. Thus the GWR model in the case of APSI is not affected by the outlier as in as in the results of research by Ispriyanti et al. [28].

Tabel 2. Comparison of Models

Model	Weighting Function	Bandwidth	R-Square	AIC
GWR	Gaussian*)	0.01575683	0.5177332	3562.747
	Exponential	0.01485379	0.5165772	3563.243
RGWR	Gaussian	0.01575683	0.5045795	3574.053
	Exponential	0.01485379	0.5023924	3575.389

5. Conclusion

APSI modeling is more suitable to use the RGWR model because it is influenced significantly by geographical factors and the possibility of outliers. But, Basic GWR using Gaussian kernel is the best model for APSI modeling because it has the smallest AIC.

Acknowledgment

We would like to thank to Directorate of Research and Public Services, The Ministry of Research, Technology and Higher Education, Republic of Indonesia for their support. This research is funded by "PTUPT" Research Grant 2017 based on assignment letter of research implementation number: 344-57/UN7.5.1/PP/2017.

References

[1] Brunsdon C, Fotheringham AS and Charlton M 1996 Geographically weighted regression: a method for exploring spatial non-stationarity *Geographical Analysis* **28** p.281–298
[2] Brunsdon C, Fotheringham AS and Charlton M 1998 Geographically weighted regression: modelling spatial nonstationarity *The Statistician* **47** 431–443
[3] Brunsdon C, Fotheringham AS and Charlton M 1999a Some notes on parametric significance tests for geographically weighted regression *Journal of Regional Science* **39** p.497–524
[4] Brunsdon C, Fotheringham AS and Charlton M 1999b A comparison of random coefficient modelling and geographically weighted regression for spatially non-stationary regression problems *Geographical and Environmental Modelling* **3** p.47–62
[5] Fotheringham AS, Charlton M and Brunsdon C 1997a Measuring spatial variations in relation-ships with geographically weighted regression In: M.M. Fischer and A. Getis, eds. *Recent developments in spatial analysis*. London: Springer, 60–82
[6] Fotheringham AS, Charlton M and Brunsdon C 1997b Two techniques for exploring non-stationarity in geographical data *Geographical System* **4** p.59–82
[7] Fotheringham AS, Charlton M and Brunsdon C 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis *Environment and Planning A* **30** p.1905–1927
[8] Leung Y, Mei CL and Zhang WX 2000a Statistical tests for spatial nonstationarity based on the geographically weighted regression model *Environment and Planning A* **32** p. 9–32
[9] Leung Y, Mei CL and Zhang WX 2000b Testing for spatial autocorrelation among the residuals of the geographically weighted regression *Environment and Planning A* **32** p. 871–890
[10] Pa´ez A, Uchida T and Miyamoto K 2002a A General Framework for Estimation and Inference of Geographically Weighted Regression Models: 1. Location-Specific Kernel Bandwidths and a Test for Locational Heterogeneity *Environment and Planning A* **34** p. 733–754
[11] Pa´ez A, Uchida T and Miyamoto K 2002b A General Framework for Estimation and Inference

of Geographically Weighted Regression Models: 2. Spatial Association and Model Specification Tests *Environment and Planning A* **34** p. 883–904
[12] Mei CL, He SY and Fang KT 2004 A note on the mixed geographically weighted regression model *Journal of Regional Science* **44** p. 143–157
[13] Mei CL, Wang N and Zhang WX 2006 Testing the importance of the explanatory variables in a mixed geographically weighted regression model *Environment and Planning A* **38** p. 587–598.
[14] Huang B, Wu B and Barry M 2010 Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House Prices *International Journal of Geographical Information Science* Vol. **24**, No. **3**, p. 383-401
[15] Harini S, Purhadi, Mashuri M and Sunaryo S 2012 Statistical Test for Multivariate Geographically Weighted Regression Model Using the Method of Maximum Likelihood Ratio Test *International Journal of Applied Mathematics and Statistics*, ISSN: 0973-7545, Vol. **29**, No. **5**, p. 110-115
[16] Winarso K, Notobroto HB and Fatmawati 2014 Development of Air Pollutant Standard Index Model Based On Mixed Geographically Temporal Weighted Regression Approach *Applied Mathematical Science* **8 (118)** p. 5863-5873. DOI 10.12988/ams.
[17] Harris P, Fotheringham AS and Juggins S 2010 Robust Geographically Weighted Regression: A Technique for Quantifying Spatial Relationships Between Freshwater Acidification Critical Loads and Catchment Attributes *Annals of the Association of American Geographers* **100:2** p. 286-306 http://dx.doi.org/10.1080/00045600903550378
[18] Zhang H and Mei C 2011 Local Least Absolute Deviation Estimation of Spatially Varying Coefficient Models: Robust Geographically Weighted Regression Approaches *International Journal of Geographical Information Science* Vol. **25**, No. **9** p. 1467–1489
[19] Atash F 2007 The Deterioration of Urban Environments in Developing Countries: Mitigating the Air Pollution Crisis in Tehran, Iran. *Cities* **24 (6)**:399-409.
[20] Fahimi M, Dharma B, Fetararayani D and Baskoro 2012 Asosiasi antara polusi udara dengan IgE total serum dan tes faal paru pada polisi lalul intas *Jurnal Penyakit Dalam* p. 1-9
[21] Gilbert A and Chakraborty J 2011 Using Geographically Weighted Regression for Environmental Justice Analysis: Cumulative Cancer Risks from Air Toxics in Florida *Social Science Research* Vol. **40** p. 273-286
[22] Robinson D and Lloyd JM 2013 Increasing the Accuracy of Nitrogen Dioxide (NO2) Pollution Mapping Using Geographically Weighted Regression & Geostatistic *International Journal of Applied Earth Observation and Geoinformation* Vol. **21** p. 374-383.
[23] Fotheringham AS, Brunsdon C and Charlton M 2002 Geographically weighted regression Chichester, UK: John Wiley and Sons
[24] Purhadi and Yasin H 2012 Mixed geographically weighted regression model (Case study: The percentage of poor households in Mojokerto 2008) *European Journal of Scientific Research* **69(2)**: 188-196 ISSN: 1450-202X
[25] Rencher AC 2000 Linear Models in Statistics John Wiley & Sons New York.
[26] Nakaya T, Fotheringham AS, Brunsdon C and Charlton M 2005 Geographically Weighted Poisson Regression for Disease Association Mapping *Statistics in Medicine* Vol. **24** Issue **17** p. 2695-2717
[27] Lu B, Harris P, Charlton M, Brunsdon C, Nakaya T and Gollini I 2016 Package 'GWmodel'.
[28] Ispriyanti D, Yasin H, Warsito B and Winarso K 2017 Mixed Geographically Weighted Regression using Adaptive Bandwidth to Modeling of Air Polluter Standard Index *ARPN Journal of Engineering and Applied Sciences* **12(15)** p. 4477-4482

em.Int-05-2018-Warsito_2018_J._Phys.-_Conf._Ser._1025_012096.pdf

ORIGINALITY REPORT

10%	8%	13%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to iGroup Student Paper	6%
2	china.iopscience.iop.org Internet Source	4%

Exclude quotes	On	Exclude matches	< 3%
Exclude bibliography	On		