Journal of Environmental Informatics Letters 9(1) 1-6 (2023)

Journal of Environmental Informatics Letters

www.iseis.org/jeil

Regional $PM_{2.5}$ Estimation for Southern Ontario through Geographically Weighted Regression

K. Huang[1 *]

[1] *Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada*

Received 20 January 2023; revised 01 March 2023; accepted 18 March 2023; published online 28 March 2023

ABSTRACT. In this study, a geographically weighted regression (GWR) approach was adopted to forecast regional concentration of particulate matter 2.5 ($PM_{2.5}$) for the southern Ontario based on both in situ meteorological measurement and Satellite retrievals of aerosol optical depth (AOD). The correlation between monitored concentration of $PM_{2.5}$ and Satellite-retrieved AOD would be quantified. The ground-level $PM_{2.5}$ for South Ontario area was then predicted using GWR with AOD and meteorological variables considered as inputs. The results indicated that performance of GWR was slightly better than the ordinary least squares (OLS) model, indicating spatial variations between independent and dependent variables. Consequently, the GWR model can help us to predict the $PM_{2.5}$ concentration in terms of time or region with satellite data, and also help improve satellite data inversion.

Keywords: geographically weighted regression, air pollution, particulate matter, Southern Ontario

1. Introduction

At present, the fine particulate matter 2.5 ($PM_{2.5}$) has become a hot spot and frontier in the field of global air pollution research. The atmospheric particle size ranges from 0.01 to 100 μm, collectively referred to as total suspended particles (TSP), and $PM_{2.5}$ is the atmospheric particulate matter with an aerodynamic diameter of 2.5 μm or less. $PM_{2.5}$ belongs to the category of fine particles. The composition of the particles is complex, mainly depends on the source. The main source is from the surface of the dust, containing oxide minerals and other ingredients. Part of the particles is produced by natural processes, derived from volcanic eruptions, dust storms, forest fires and so on. $PM_{2.5}$ can also be converted from sulfur and nitrogen oxides. Fine particles $PM_{2.5}$ can cause significant impacts on the human health. Many studies have reported that $PM_{2.5}$ increased the carcinogenicity and morbidity of the human body (Boldo et al., 2006; Fann et al., 2012; Hu et al., 2012), while mortality was also increased (Boldo et al., 2006; Mar et al., 2006; Fann et al., 2012). The impact on children's health is also worthy of our attention (Hu et al., 2012). However, characterization of air quality and the resulting public health concerns requires accurate prediction of long-term $PM_{2.5}$ concentration prediction.

At present, many research works have been reported for concentration prediction of $PM_{2.5}$ in different areas and different countries. Hu et al. (2013) applied geographically weighted regression to estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. Centered at the Atlanta Metro area, the whole study area is 750 × 750 km^2. In recent years, China's air quality is getting worse, we pay more attention to $PM_{2.5}$. Song et al. (2014) proposed a satellite-based geographically weighted regression model for estimating regional $PM_{2.5}$ estimation over the Pearl River Delta region in China. Su et al. (2014) used the proxy-based approach and geographically weighted regression (GWR) to describe the change of ecosystem service values (ESV) and their relationship with urbanization in Shanghai suburb, China. Zou et al. (2016) using geographically weighted regression to produce high-resolution satellite map of fine particulates over Jing-jin-Ji region in China. Luo et al. (2017) used geographically weighted regression to characterize spatiotemporal patterns of $PM_{2.5}$ concentrations in Mainland China and then reveal its influencing factors. There are two objectives in this study. The first primary objective is to analyze the spatiotemporal pattern of $PM_{2.5}$ concentrations in China using more than a decade of data, and the other one is to explore the contributions of the influencing factors, including natural geographical and socioeconomic factors on $PM_{2.5}$ using geographically weighted egression model. Ma et al. (2014) and You et al. (2016) applied geographically weighted regression to estimate national-scale ground-level $PM_{2.5}$ concentration based on Moderate Resolution Imaging Spectroradiometer (MODIS) and Multi-angle Imaging Spectro Radiometer (MISR) aerosol optical depths (AOD) to estimate $PM_{2.5}$ concentration in China.

In this study, the objective is to forecast regional concentration of $PM_{2.5}$ for the southern Ontario through geographically weighted regression approach based on both in-situ meteorological measurement and Satellite retrievals of AOD. In detail, correlation between monitored concentration of $PM_{2.5}$ and Satellite-retrieved AOD would be quantified. Then the ground-level $PM_{2.5}$ for South Ontario area will be predicted using geographi-

[*] Corresponding author. Tel.: +1 306-501-8706.
E-mail address: huang22k@uregina.ca (K. Huang).

ISSN: 2663-6859 print/2663-6867 online
 doi:10.3808/jeil.202300100

cally weighted regression, in which AOD and meteorological variables are considered as inputs.

2. Methodology

Introduced by Brunsdon et al. (1996), geographically weighted regression was developed to allow relationships in a regression model to vary over space. In comparison to traditional regression approaches with constant coefficients over space, GWR can estimate model coefficient locally according to spatially distributed data points. GWR has been applied to a number of research topics related to ecology, environment, geography and regional sciences. Wang et al. (2016) used the model of GWR to analyze the explanatory factors of urban floods, and the spatially explicit relationships between explanatory factors and a dependent variable. At the same time there are researchers on the development and application of GWR. Zhang et al. (2016) developed a semi-physical geographically weighted regression model for real-time estimating of satellite-derived $PM_{2.5}$. This study firstly fused Aqua AOD products based on two widely used algorithms [i.e., Dark Target (DT) and Deep Blue (DB)] and aligned them to real-time $PM_{2.5}$ based on five time zones. Then the semi-physical GWR model combined with reanalysis of meteorological variables was developed to quantify the spatial distribution of $PM_{2.5}$. Geographically weighted regression is a new spatial analysis method proposed in recent years, which can detect the nonstationary of spatial relations by embedding the spatial structure in the linear regression model. Geographically weighted regression is a method of studying the quantitative relationship between two or more variables with spatial (or regional) distribution characteristics by regression theory, taking local features as weights when dealing with data. Geo-weighted regression is characterized by the fact that the regression function is assumed to be the position function of the geometric position of the observation point, and the spatial characteristics of the data are included in the model, which creates the conditions for the analysis of the spatial characteristics of the regression relation. van Donkelaar et al. (2015) applied geographically weighted regression to produce high-resolution satellite-derived $PM_{2.5}$.

As a local regression technique, GWR is an extension of global regression technology (Li et al., 2010). Fotheringham et al. (2002) give a detailed description of the algorithm and principle. Here, we only briefly introduce GWR. Consider the global regression model given as follows:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \tag{1}$$

where y_i is the dependent variable, x_{ik} represent the independent variables, ε_i is random error term at different spatial points (the subscripts i and k stand for the spatial locations and the independent variable number, respectively), β_0 is the model intercept, and β_k is the slope coefficient for kth independent variable.

This type of model is aspatial, i.e., no geographical location information is considered in the estimation of the model parameters, and all parameters are averages across the whole data set.

The GWR technique extends the conventional global regression of Equation 1 by adding a geographical location parameter, with the model rewritten as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2}$$

where (u_i, v_i) denotes the coordinates of the ith point in space, β_0 and β_k are parameters to be estimated, and again ε_i is the random error term at point i.

To estimate the parameters in Equation 2, an observation is weighted according to its proximity to a specific point i, i.e., the distance between observation and point i determines the weight given to the observation; larger weights are assigned to observations closer to point i. Therefore, the weighting of an observation in the analysis is not constant, but a function of geographical location. The parameters in Equation (2) may be estimated by solving the following matrix equation:

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

where $\hat{\beta}(u_i, v_i)$ represents the unbiased estimate of β, and $W(u_i, v_i)$ is the weighting matrix, whose role is to ensure that observations near to the specific location have larger weight. In this study, we used the following Gaussian weighting kernel function form:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{b^2}\right) \tag{4}$$

where d_{ij} is the Euclidean distance between regression point i and neighboring observation j, and b represents a basal width of the kernel function, called bandwidth.

In Equation (4), if j coincides with i, the weighting value of the data at that point is set to 1, while w_{ij} decreases according to a Gaussian curve as the distance d_{ij} increases (Fotheringham et al., 2002; Shi et al., 2006). In this research, GWR analysis was carried out using ArcGIS 10.2 software and all maps were produced using the same software.

3. Overview of Study Area

The study area for this analysis is the Southern Ontario. There are thirteen cities located in this area, including Toronto, Ottawa, Hamilton, Kitchener, London, St. Catharines, Oshawa, Windsor, Barrie, Kingston, Guelph, Brantford and Peterborough. Southern Ontario has a humid continental climate with four distinct seasons. In July, the average high temperature is up to 25 ~ 28 °C (81 ~ 86.4 °F). In December, high temperatures range from 0 to 2 °C (32 ~ 35.6 °F). The maximum temperature in the history of Southern Ontario is 45 °C (113 °F) and the humidity index is about 52 °C (129.6 °F). When the cold spell struck, the winter temperature in the central and eastern parts of Southern Ontario could drop to −30 °C (−22 °F), while temperatures in the southwest and the Niagara region were below −20 °C (−4 °F).

3.1. In Situ $PM_{2.5}$ Measurements

The monthly average $PM_{2.5}$ concentrations in South Ontario are downloaded from the Air Quality Ontario website. There are 19 $PM_{2.5}$ observation sites in those thirteen cities in southern Ontario (http://airqualityontario.com/history/index.php). The $PM_{2.5}$ concentrations in December 2012 are applied for this analysis. These in situ measurements are considered as "True" values despite the possibility of measurement errors or representativeness issues arising between these point measurements and area averaged concentrations, such as observed from satellite.

3.2. Aerosol Optical Depth Data

Satellite retrievals of AOD provide insight into the total atmospheric aerosol column, as represented by the extinction of light (van Donkelaar et al., 2015). These AOD retrievals are being applied to increase the spatial extent of observational constraints on $PM_{2.5}$ for application to exposure assessment and epidemiologic studies (van Donkelaar et al., 2015; You et al., 2016; Luo et al., 2017). In this study, MODIS aerosol data are applied for regional $PM_{2.5}$ prediction. MODIS is an instrument aboard the Terra and Aqua satellites operated by National Aeronautics and Space Administration (NASA) (Remer et al., 2005).

MODIS measures the abundance of atmospheric particles on the global scale at a moderate spatial resolution (10 km) (Hu et al., 2013). In this study, the MODIS aerosol data in December 2012 are download from NASA Earth Observations website (https://neo.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_RA). The quality flag of AOD is 0.5. The AOD data (unitless) for each site is selected by latitude and longitude. The latitude range is from 42°29.3′ to 45°43.4′, longitude range is from −83°07.3′ to −75°67.6′.

3.3. Meteorological Data

A number of research works have revealed that the concentration of $PM_{2.5}$ is also influenced by meteorological conditions, such as wind speed, temperature, and so on. In this study, the monthly average data of wind speed, temperature will be used for prediction of $PM_{2.5}$ concentration over South Ontario. These data are downloaded from Environmental Canada (http://climate.weather.gc.ca/historical_data/search_historic_data_e.html).

4. Results Analysis

Regression analysis is probably the most commonly used statistic method in the natural and social sciences. Regression is used to evaluate relationships between two or more feature attributes. Identifying and measuring relationships help better understand what is going on in a place, predict where something is likely to occur, or examine causes of why things occur. Ordinary least squares (OLS) model is one of the most used techniques for parameter estimation in regression models. It is also the proper starting point for all spatial regression analyses. It can provide a global model of the variable or process to be understood and create a single regression equation to represent that process. In this study, the independent variables include AOD, temperature, and wind speed, while the dependent variable is the concentration of $PM_{2.5}$. Therefore, the initial linear regression model can be formulated as:

$$PM_{2.5} \sim AOD + TEMP + WS \tag{5}$$

where $PM_{2.5}$ refers to the monthly ground-level $PM_{2.5}$ concentrations (μg/m^3), AOD is the MODIS aerosol optical depth value (unitless), TEMP is the air temperature (°C), and WS refers to the surface wind speed (km/h). The predictor variables explained in Table 1.

Table 1. Definitions of Predictor Variables in Equation (5)

Name	Unit	Description
$PM_{2.5}$	μg/m^3	$PM_{2.5}$ concentrations
AOD	Unitless	Terra/Aqua MODIS AOD
TEMP	°C	Temperature
WS	km/h	Wind speed

Table 2. Parameterization of OLS Model for the Estimation of $PM_{2.5}$ Concentrations in December 2012

Parameter	Value
Number of observations	19
Akaike's Information Criterion (AIC)	33.55
Multiple R^2	0.82
R^2 adjusted	0.78

Table 3. Parameterization of GWR Model for the Estimation of $PM_{2.5}$ Concentrations in December 2012

Parameter	Value
Neighbors	19
Residual squares	2.55
Effective number	5.27
Sigma	0.43
AIC	32.60
R^2	0.83

Table 2 shows the parameterization of OLS model for the estimation of December 2012 $PM_{2.5}$ concentrations. The results shows that the linear regression model can well capture the relationship between the selected independent variables and the $PM_{2.5}$ concentration in South Ontario, with the value of R^2 reaching 0.82. Table 3 shows the performance of GWR model. The results suggestion the prediction from GWR is slight better than OLS model, indicating that model parameters are spatially varied and the GWR model is generally more accurate and appropriate for data analysis in this study.

The spatial distribution of December 2012 $PM_{2.5}$ concentrations in Southern Ontario for the GWR model shows in Figure 1. The overall trend in the $PM_{2.5}$ concentration of Southern Ontario is getting higher from east to west. In detail, Peterborough is located in the central Ontario in the Kawartha Lakes area in Ontario and is situated in the St. Lawrence lowland ecological zone, which is the south of the Canadian shield, and about 35 kilometers north of Lake Ontario. In Peterborough, the $PM_{2.5}$ concentration is the lowest and the air quality is the best. This may because the city endless surrounds by the Little Lake at

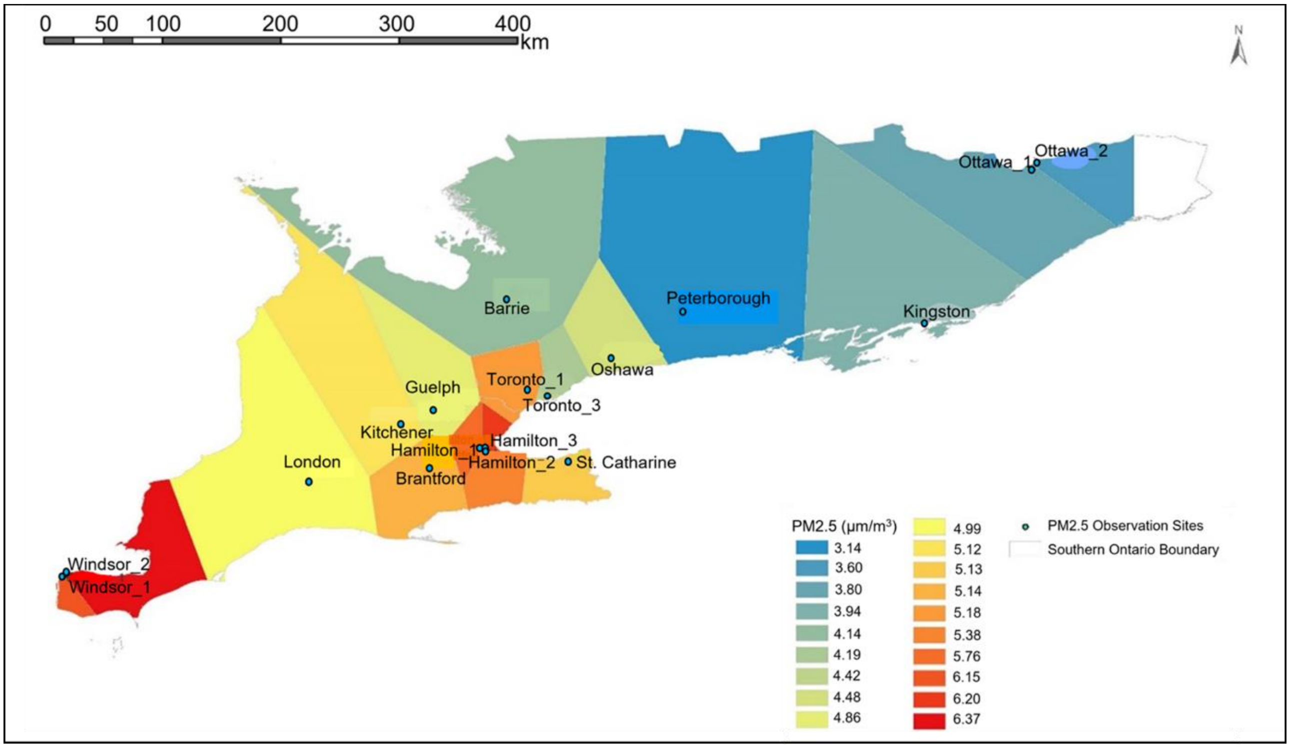

Figure 1. The spatial distribution of $PM_{2.5}$ concentrations in Southern Ontario for the GWR model in December 2012.

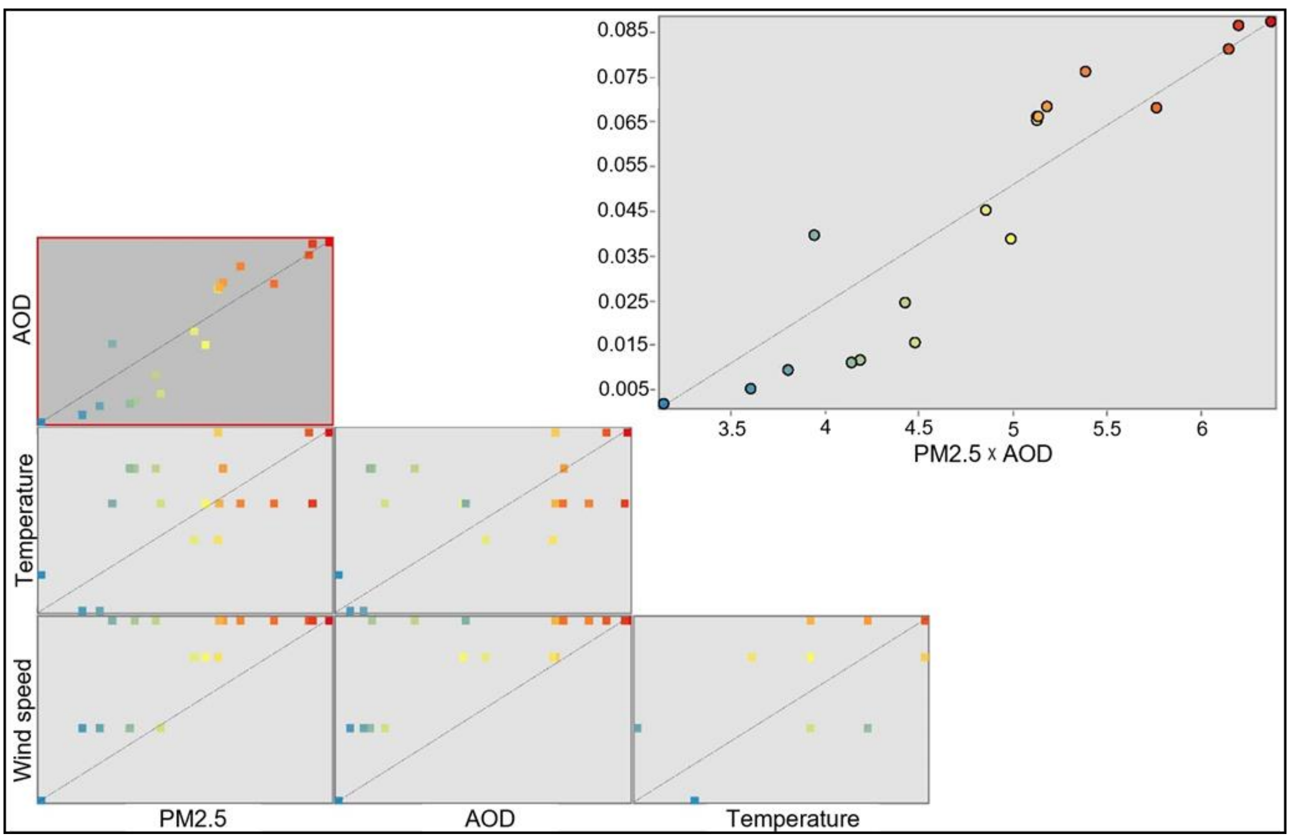

Figure 2. Scatter plot of AOD vs. $PM_{2.5}$ (top-right corner of Figure 2) and relationship with temperature and wind speed (bottom-left corner of Figure 2).

Otonabee, and the Trent Canal runs along the eastern edge of the city, connecting the Little Lake to the Otonabee area. In comparison, The City of Windsor experienced the highest $PM_{2.5}$ concentration in the area of South Ontario, and also has the worst air quality. This mainly because that Windsor is located downwind from several strong polluters, notably coal-burning power plants in the United States. The City of Toronto is also in an area with relatively high $PM_{2.5}$ concentration, since the City of Toronto has biggest population as well as a great volume of vehicles and manufacturing factories, which are main emitters of $PM_{2.5}$.

In terms of the correlation among independent variables and the dependent variable, AOD and $PM_{2.5}$ have a great relevance, as shown in Figure 2. Also, it can be observed that temperature and wind speed also have a certain impact on $PM_{2.5}$. But the impacts are not obvious. Overall, AOD, temperature, wind speed and $PM_{2.5}$ are positively correlated.

5. Conclusions

In this study, the geographically weighted regression (GWR) was applied for analysis and prediction of $PM_{2.5}$ concentration in consideration of both satellite retrievals of AOD and in situ meteorological variables. GWR is one of most widely used spatial regression techniques, especially in geography and environment disciplines. GWR provides a local model of the variable or process by fitting a regression equation to every feature in the dataset. In this study, both ordinary least square (OLS) model and GWR model have been applied for concentration prediction of $PM_{2.5}$. The results indicated that performance of GWR was slightly better than the OLS model, indicating spatial variations between independent and dependent variables. The GWR model can help us to predict the $PM_{2.5}$ concentration in terms of time or region with satellite data, and also help improve satellite data inversion.

References

Boldo, E., Medina, S., Le Tertre, A., Hurley, F., Mücke, H.G., Ballester, F. and Aguilera, I. (2006). Apheis: Health impact assessment of long-term exposure to $PM_{2.5}$ in 23 European cities. *European journal of epidemiology*. 21(6), 449-458. https://doi.org/10.1007/s10654-006-9014-0

Brunsdon C., Fotheringham A.S.and Charlton M. (1996). Geographically weighted regression: A method for exploring spatial non-stationarity. *Geographical Analysis*. 28(4): 281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Fann, N., Lamson, A.D., Anenberg, S.C., Wesson, K., Risley, D. and Hubbell, B.J. (2012). Estimating the national public health burden associated with exposure to ambient $PM_{2.5}$ and ozone. *Risk analysis*. 32(1), 81-95. https://doi.org/10.1111/j.1539-692 4.2011.01630.x

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2003). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley and Sons. ISBN: 0471496162

Hu, X., Zhang, Y., Ding, Z.H., Wang, T.J., Lian, H.Z., Sun, Y.Y. and Wu, J.C. (2012). Bioaccessibility and health risk of arsenic and heavy metals (Cd, Co, Cr, Cu, Ni, Pb, Zn and Mn) in TSP and $PM_{2.5}$ in Nanjing, China. *Atmospheric Environment*. 57, 146-152. https://doi.org/10.1016/j.atmosenv.2012.04.056

Hu, X.F., Waller, L.A., Al-Hamdan, M.Z., Crosson, W.L., Estes Jr, M.G., Estes, S.M., Quattrochi, D.A., Sarnat, J.A. and Li, Y. (2013). Estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. using geographically weighted regression. *Environmental Research*. 121, 1-10. https://doi.org/10.1016/j.en vres.2012.11.003

Li, S.C., Zhao, Z.Q., Miaomiao, X. and Wang, Y.L. (2010). Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression. *Environmental Modelling and Software*. 25(12), 1789-1800. https://doi.org/10.1016/j.envsoft.2010.06.011

Luo, J.Q., Du, P.J., Samat, A., Xia, J.S., Che, M.Q. and Xue, Z.H. (2017). Spatiotemporal Pattern of $PM_{2.5}$ Concentrations in Mainland China and Analysis of Its Influencing Factors using Geographically Weighted Regression. *Scientific Reports*. 7, 40607. https://doi.org/10.1038/srep40607

Ma, Z.W., Hu, X.F., Huang, L., Bi, J. and Liu, Y. (2014). Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing. *Environmental Science and Technology*. 48, 7436-7444. https://doi.org/10.1021/es5009399

Mar, T.F., Ito, K., Koenig, J.Q., Larson, T.V., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Neas, L., Stölzel, M., Paatero, P., Hopke, P.K. and Thurston, G.D. (2006). PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ. *Journal of Exposure Science and Environmental Epidemiology*. 16(4), 311-320. https://doi.org/10.1038/sj.jea.7500465

Remer, L.A., Kaufman, Y.J., Tanré, D., Mattoo, S., Chu, D.A., Martins, J.V., Li, R.R., Ichoku, C., Levy, R.C., Kleidman, R.G. and Eck, T.F. (2005). The modis aerosol algorithm, products, and validation. *Journal of the Atmospheric Sciences*. 62(4), 947-973. https://doi.org/10.1175/JAS3385.1

Shi, H.J., Zhang, L. and Liu, J.G. (2006). A new spatial-attribute weighting function for geographically weighted regression. *Canadian Journal of Forest Research*. 36(4), 996-1005. https://doi.org/10.1139/X-05-295

Song, W.Z., Jia, H.F., Huang, J.F. and Zhang, Y.Y. (2014). A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the Pearl River Delta region in China. *Remote Sensing of Environment*. 154, 1-7. https://doi.org/10.1016/j.rse.2014.08.008

Su, S.L., Li, D.L., Hu, Y.N., Xiao, R. and Zhang, Y. (2014). Spatially non-stationary response of ecosystem service value changes to urbanization in Shanghai, China. *Ecological Indicators*. 45, 332-339. https://doi.org/10.1016/j.ecolind.2014.04.031

van Donkelaar, A., Martin, R.V., Spurr, R.J.D. and Burnett, R.T. (2015). High-Resolution Satellite-Derived $PM_{2.5}$ from Optimal Estimation and Geographically Weighted Regression over North America. *Environmental Science and Technology*. 49, 10482-10491. https://doi.org/10.1021/acs.est.5b02076

Wang, C.X., Du, S.Q., Wen, J.H., Zhang, M., Gu, H.H., Shi, Y. and Xu, H. (2016). Analyzing explanatory factors of urban pluvial floods in Shanghai using geographically weighted regression. *Stochastic Environmental Research and Risk Assessment*. 31, 1777-1790. https://doi.org/10.1007/s00477-016-1242-6

You, W., Zang, Z.L., Zhang, L.F., Li, Y. and Wang, W.Q. (2016). Estimating national-scale ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on MODIS and MISR AOD. *Environmental Science and Pollution Research*. 23, 8327-8338. https://doi.org/10.1007/s11356-015-60 27-9

Zhang, T.H., Liu, G., Zhu, Z.M., Gong, W., Ji., Y.X. and Huang, Y. (2016). Real-Time Estimation of Satellite-Derived $PM_{2.5}$ Based on a Semi-Physical Geographically Weighted Regression Model. *International Journal of Environmental Researchand Public Health*. 13, 974. https://doi.org/10.3390/ijerph13100974

Zou, B., Pu, Q., Bilal, M., Weng, Q., Zhai, L. and Nichol, J.E. (2016).

High-resolution satellite mapping of fine particulates based on geographically weighted regression. *IEEE Geoscience and Remote Sensing Letters*. 13(4), 495-499. https://doi.org/10.1109/LGRS.2016.2520480