Maejo Int. J. Sci. Technol. **2016**, *10*(03), 256-271

Maejo International Journal of Science and Technology

ISSN 1905-7873
Available online at www.mijst.mju.ac.th

Full Paper

Spatial and temporal air quality analysis of Chinese cities using geographically and temporally weighted regression

Haiyan Xuan [1], *, Qi Li [2], Mahammad Amin [3], * and Anqi Zhang [2]

[1] School of Economics and Management, Lanzhou University of Technology, Lanzhou, China
[2] School of Science, Lanzhou University of Technology, Lanzhou, China
[3] Nuclear Institute for Food and Agriculture (NIFA), Peshawar, Pakistan
* Corresponding authors, e-mail: haiyanxuan@gmail.com, aminkanju@gmail.com

Received: 16 February 2015 / Accepted: 4 August 2016 / Published: 23 September 2016

Abstract: Local modelling techniques are useful for analysis of spatio-temporal data because of their ability to extract underlying patterns. We use geographically and temporally weighted regression (GTWR) to analyse the spatial and temporal characteristics of air quality in 67 Chinese cities. The method employs a two-step estimator to examine the relationship between the indices and the given climatic conditions in the cities. The simulation performance is satisfactory. The mean monthly air quality index (AQI) varies with the spatio-temporal position when the level of the monthly precipitation and the mean monthly temperature are controlled at one station. The AQI is highest in northern China, moderate on the south-eastern coast and low in southern China. Cities in northern China, eastern coastal cities, north-eastern cities and the city of Urumqi have a maximum rate of change when the monthly precipitation increases by one unit. The monthly mean AQI of most cities in Anhui, Zhejiang and Jiangsu provinces exhibits the greatest decrease with an increase in the monthly average temperature when the monthly precipitation remains unchanged.

Keywords: GTWR model, two-step estimator, air quality, spatial and temporal air quality analysis, Chinese cities

INTRODUCTION

Air pollution is a serious threat to human health. The association between pollutant concentration and mortality and morbidity has been established in many studies [1-3]. Industrialisation and modernisation have compromised air quality and population health in many areas [4, 5]. The environment in many areas of China has been badly contaminated. Cities often have serious air pollution that exceeds reasonable health standards. Environmental protection has become more critical as the air pollution issue exacerbates. Air quality issues need to be better understood, especially in the context of climate change. Air pollution patterns are usually heterogeneous in space as well as in time and can change rapidly based on geographical location, temporal scale and environmental factors [6, 7]. Air quality measurements allow the characterisation of local pollution levels, and the data can be used by authorities to justify the implementation of measures to protect human health.

Several air quality indices are used worldwide. The air quality index (AQI), introduced by the US Environmental Protection Agency, is commonly used. The AQI is calculated using five major air-regulated pollutants: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide and nitrogen dioxide [8]. Several European countries including Belgium, UK, France and Germany have introduced national air quality indices. The common air quality indices (CAQI) was developed as part of the CITEAIR project and has been operating via the Internet since 2006 [9]. Some studies have evaluated the factors that affect air quality and air circulation in Chinese cities using air pollution indices [10]. These studies have mainly focused on large cities such as Beijing and Lanzhou or small cities that were used in some analyses. Almost all existing studies using AQI refer to Chinese small cities [11].

Most air pollution models are regression models or their extensions. In regression models data are assumed to be independent and uniformly distributed. However, air pollution features and their related causative factors are strongly spatially and temporally dependent and unstable. They have significant spatial heterogeneity. Local summary statistics and modelling methods have been used to explore the effect of spatial heterogeneity. Geographically weighted regression (GWR) is a common local smoothing method applied to explore the spatial heterogeneity of regressions. Under certain conditions, spatial heterogeneity can lead to spatial nonstationary. Lo [12] used GWR to estimate population size and found that the GWR model, which is local, can effectively deal with spatial nonstationary. The heterogeneity of a regression relationship can be described by allowing the parameters in the linear regression model to vary as unknown functions of the geographical location. The geographically and temporally weighted regression (GTWR) models are based on the GWR model. Spatial and temporal characteristics of data are used in GTWR models, which sets the stage for exploring the spatial nonstationary and temporal nonstationary of the regression.

The objectives of the present study are to analyse the influence of precipitation and temperature on air quality using a GTWR model approach and to examine the relationship between the indices and the climatic conditions in the study locations in China.

GTWR MODEL AND TWO-STEP ESTIMATOR

GTWR Model

Huang et al. [13] used spatial and temporal data and proposed the following GTWR model:

$$y_i = \beta_0\left(u_i, v_i, t_i\right) + \sum_{j=1}^{p} \beta_j\left(u_i, v_i, t_i\right) x_{ij} + \varepsilon_i, \qquad i = 1, 2, \cdots, n \tag{1}$$

where $\left(y_i; x_{i1}, x_{i2}, \cdots, x_{ip}\right)$ are observations of the response variable Y and explanatory variables $X_1, X_2, \cdots, X_p$ at location $\left(u_i, v_i, t_i\right)$ in the study region; $\varepsilon_i\left(i = 1, 2, \cdots, n\right)$ represents error terms with means equal to zero and a common variance σ^2; $\beta_j\left(u_i, v_i, t_i\right)\left(j = 0, 1, 2, \cdots, p\right)$ represents $p+1$ unknown functions of geographical locations and observation times.

Two-Step Estimator

For convenience, the regression model (1) is rewritten as follows:

$$y_i = \sum_{j=0}^{p} \beta_j \left(u_i, v_i, t_i\right) x_{ij} + \varepsilon_i, \qquad i = 1, 2, \cdots, n. \tag{2}$$

We set $x_{i0} \equiv 1$ to obtain the intercept term $\beta_0\left(u_i, v_i, t_i\right)$. Here, $\left(u_i, v_i, t_i\right)$ is any space-time cordinate point in ellipsoidal coordinates. Each regression coefficient function $\beta_j\left(u_i, v_i, t_i\right)\left(j = 0, 1, 2, \cdots, p\right)$ has continuous partial derivatives in the space position coordinates *u*, *v* and time coordinates *t* in the model (2). The term $\left(u_0, v_0, t_0\right)$ is any given point in the study area.

For each $j = 0, 1, 2, \cdots, p$, we use the Taylor formula in the neighborhood of $\left(u_0, v_0, t_0\right)$ and then have

$$\beta_j\left(u, v, t\right) \approx \beta_j\left(u_0, v_0, t_0\right) + \beta_j^{(u)}\left(u_0, v_0, t_0\right)\left(u - u_0\right) + \beta_j^{(v)}\left(u_0, v_0, t_0\right)\left(v - v_0\right)$$
$$+ \beta_j^{(t)}\left(u_0, v_0, t_0\right)\left(t - t_0\right).$$

At $\left(u_0, v_0, t_0\right)$, $\beta_j^{(u)}\left(u_0, v_0, t_0\right)$, $\beta_j^{(v)}\left(u_0, v_0, t_0\right)$ and $\beta_j^{(t)}\left(u_0, v_0, t_0\right)$ represent partial derivatives of $\beta_j\left(u, v, t\right)$ for *u*, *v*, *t* respectively. According to the local linear fitting method in the varying coefficient model, we can obtain an approximate expression of y_i, denoted as $\tilde{y}$, specific to the formula as

$$\tilde{y}_i = \sum_{j=0}^{p}\Big(\beta_j\left(u_0, v_0, t_0\right) + \beta_j^{(u)}\left(u_0, v_0, t_0\right)\left(u - u_0\right) + \beta_j^{(v)}\left(u_0, v_0, t_0\right)\left(v - v_0\right)$$
$$+ \beta_j^{(t)}\left(u_0, v_0, t_0\right)\left(t - t_0\right)\Big) x_{ij} + \varepsilon_i, \qquad i = 1, 2, \cdots, n.$$

Then the minimisation of the above formula can be expressed as

$$\sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2 w_i\left(u_0, v_0, t_0\right) \tag{3}$$

where $w_i\left(u_0, v_0, t_0\right) = K_h\left(d_{0i}\right) = \exp\left\{-d_{0i}^2 / h^2\right\}, i = 1, 2, \cdots n$, and $K(\cdot)$ is the kernel function; $d_{0i}, i = 1, 2, \cdots n$ are the distances between $\left(u_0, v_0, t_0\right)$ and $\left(u_i, v_i, t_i\right)$, and h is the bandwidth.

The weight matrix with Gauss kernel function and bandwidth h_1 can be expressed as

$$w_{h_1}\left(u_0, v_0, t_0\right) = diag\left[w_1\left(u_0, v_0, t_0\right), w_2\left(u_0, v_0, t_0\right), \cdots, w_n\left(u_0, v_0, t_0\right)\right],$$

$$Y = \left(y_1, y_2, \cdots y_n\right)^T,$$

$$X_1\left(u_0, v_0, t_0\right) = \begin{bmatrix} x_{10} & \dots & x_{1p} & x_{10}\left(u_1 - u_0\right) & \dots & x_{1p}\left(u_1 - u_0\right) \\ x_{20} & \dots & x_{2p} & x_{20}\left(u_2 - u_0\right) & \dots & x_{2p}\left(u_2 - u_0\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n0} & \dots & x_{np} & x_{n0}\left(u_n - u_0\right) & \dots & x_{np}\left(u_n - u_0\right) \end{bmatrix},$$

$$X_2\left(u_0, v_0, t_0\right) = \begin{bmatrix} x_{10}\left(v_1 - v_0\right) & \dots & x_{1p}\left(v_1 - v_0\right) & x_{10}\left(t_1 - t_0\right) & \dots & x_{1p}\left(t_1 - t_0\right) \\ x_{20}\left(v_2 - v_0\right) & \dots & x_{2p}\left(v_2 - v_0\right) & x_{20}\left(t_2 - t_0\right) & \dots & x_{2p}\left(t_2 - t_0\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n0}\left(v_n - v_0\right) & \dots & x_{np}\left(v_n - v_0\right) & x_{n0}\left(t_n - t_0\right) & \dots & x_{np}\left(t_n - t_0\right) \end{bmatrix}.$$

Then

$$X\left(u_0, v_0, t_0\right) = \left[X_1\left(u_0, v_0, t_0\right), X_2\left(u_0, v_0, t_0\right)\right],$$

$$P(u_0,v_0,t_0)=\Big[\beta_0(u_0,v_0,t_0),\cdots,\beta_P(u_0,v_0,t_0),\beta_0^{(u)}(u_0,v_0,t_0),\cdots,\beta_P^{(u)}(u_0,v_0,t_0),$$
$$\beta_0^{(v)}(u_0,v_0,t_0),\cdots,\beta_P^{(v)}(u_0,v_0,t_0),\beta_0^{(t)}(u_0,v_0,t_0),\cdots,\beta_P^{(t)}(u_0,v_0,t_0)\Big]^T.$$

The solution of the least squares problem can be expressed in matrix form as

$$\hat{P}(u_0,v_0,t_0)=\left(\hat{\beta}(u_0,v_0,t_0)^T,\hat{\beta}^{(u)}(u_0,v_0,t_0)^T,\hat{\beta}^{(v)}(u_0,v_0,t_0)^T,\hat{\beta}_0^{(t)}(u_0,v_0,t_0)^T\right)$$
$$=\left[X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)X(u_0,v_0,t_0)\right]^{-1}X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)Y, \quad (4)$$

where

$$\hat{\beta}(u_0,v_0,t_0)=\left[\hat{\beta}_0(u_0,v_0,t_0),\hat{\beta}_1(u_0,v_0,t_0)\cdots,\hat{\beta}_P(u_0,v_0,t_0)\right]^T, \quad (5)$$

$$\hat{\beta}^{(u)}(u_0,v_0,t_0)=\left[\hat{\beta}_0^{(u)}(u_0,v_0,t_0),\hat{\beta}_1^{(u)}(u_0,v_0,t_0)\cdots,\hat{\beta}_P^{(u)}(u_0,v_0,t_0)\right]^T, \quad (6)$$

$$\hat{\beta}^{(v)}(u_0,v_0,t_0)=\left[\hat{\beta}_0^{(v)}(u_0,v_0,t_0),\hat{\beta}_1^{(v)}(u_0,v_0,t_0)\cdots,\hat{\beta}_P^{(v)}(u_0,v_0,t_0)\right]^T, \quad (7)$$

$$\hat{\beta}^{(t)}(u_0,v_0,t_0)=\left[\hat{\beta}_0^{(t)}(u_0,v_0,t_0),\hat{\beta}_1^{(t)}(u_0,v_0,t_0)\cdots,\hat{\beta}_P^{(t)}(u_0,v_0,t_0)\right]^T, \quad (8)$$

where (5) is the column vector determined by the estimated values of each regression coefficient function $\beta_j(u,v,t)(j=0,1,2,\cdots,p)$ at (u_0,v_0,t_0), and (6-8) are the column vectors determined by the estimated values of the partial derivatives about u,v and t respectively.

From (4) we can obtain

$$\hat{\beta}(u_0,v_0,t_0)=\left(I_{p+1},0_{p+1},0_{p+1},0_{p+1}\right)\left[X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)X(u_0,v_0,t_0)\right]^{-1}$$
$$X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)Y, \quad (9)$$

$$\hat{\beta}^{(u)}(u_0,v_0,t_0)=\left(0_{p+1},I_{p+1},0_{p+1},0_{p+1}\right)\left[X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)X(u_0,v_0,t_0)\right]^{-1}$$
$$X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)Y, \quad (10)$$

$$\hat{\beta}^{(v)}(u_0,v_0,t_0)=\left(0_{p+1},0_{p+1},I_{p+1},0_{p+1}\right)\left[X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)X(u_0,v_0,t_0)\right]^{-1}$$
$$X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)Y, \quad (11)$$

$$\hat{\beta}^{(t)}(u_0,v_0,t_0)=\left(0_{p+1},0_{p+1},0_{p+1},I_{p+1}\right)\left[X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)X(u_0,v_0,t_0)\right]^{-1}$$
$$X^T(u_0,v_0,t_0)W_{h_1}(u_0,v_0,t_0)Y, \quad (12)$$

where I_{p+1} and 0_{p+1} represent the $p+1$ order's unit matrix and $p+1$ order's zero matrix respectively. Let $(u_0,v_0,t_0)=(u_i,v_i,t_i)(i=1,2,\cdots,n)$; it is then easy to estimate the coefficient function at each observation position using (9):

$$\hat{\beta}(u_i,v_i,t_i)=\left(\hat{\beta}_0(u_i,v_i,t_i),\hat{\beta}_1(u_i,v_i,t_i)\cdots,\hat{\beta}_p(u_i,v_i,t_i)\right)^T$$
$$=\left(I_{p+1},0_{p+1},0_{p+1},0_{p+1}\right)\left[X^T(u_i,v_i,t_i)W_{h_1}(u_i,v_i,t_i)X(u_i,v_i,t_i)\right]^{-1}$$
$$X^T(u_i,v_i,t_i)W_{h_1}(u_i,v_i,t_i)Y, \qquad i=1,2,\cdots n. \quad (13)$$

Thus, the fitted value of dependent variables at (u_i,v_i,t_i) is

$$\begin{aligned} Y_i &= \sum_{j=0}^{p} \hat{\beta}_j\left(u_i, v_i, t_i\right) X_{ij} \\ &= x_i^T \hat{\beta}_j\left(u_i, v_i, t_i\right) \\ &= \left(X_i^T, 0_{1\times(p+1)}, 0_{1\times(p+1)}, 0_{1\times(p+1)}\right)\left[X^T\left(u_i, v_i, t_i\right) W_{h_1}\left(u_i, v_i, t_i\right) X\left(u_i, v_i, t_i\right)\right]^{-1} \\ & X^T\left(u_i, v_i, t_i\right) W_{h_1}\left(u_i, v_i, t_i\right) Y, \qquad i = 1, 2, \cdots n, \end{aligned} \tag{14}$$

where $X_i = \left(1, x_{i1}, \cdots, x_{ip}\right)$ are the column vectors composed by X_{0i} and the ith group observations of $X_1, X_2, \cdots, X_p$. Then the fitted values of the dependent variables Y at the observation positions are

$$\hat{Y} = \left(\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_n\right)^T = LY. \tag{15}$$

Set

$$A_i = \left[X^T\left(u_i, v_i, t_i\right) W_{h_1}\left(u_i, v_i, t_i\right) X\left(u_i, v_i, t_i\right)\right]^{-1}, \qquad i = 1, 2, \cdots n,$$

where

$$L = \begin{bmatrix} \left(X_1^T, 0_{1\times(p+1)}, 0_{1\times(p+1)}, 0_{1\times(p+1)}\right) A_1 X^T\left(u_1, v_1, t_1\right) W_{h_1}\left(u_1, v_1, t_1\right) \\ \left(X_2^T, 0_{1\times(p+1)}, 0_{1\times(p+1)}, 0_{1\times(p+1)}\right) A_2 X^T\left(u_2, v_2, t_2\right) W_{h_1}\left(u_2, v_2, t_2\right) \\ \vdots \\ \left(X_n^T, 0_{1\times(p+1)}, 0_{1\times(p+1)}, 0_{1\times(p+1)}\right) A_n X^T\left(u_n, v_n, t_n\right) W_{h_1}\left(u_n, v_n, t_n\right) \end{bmatrix}.$$

Equation (15) shows that the local linear estimation is a linear estimation of the L smooth matrix. Further, the residual vector of local linear estimation is

$$\hat{\varepsilon} = \left(\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n\right)^T = Y - \hat{Y} = \left(I - L\right) Y, \tag{16}$$

and the residual sum of squares is

$$RSS = \hat{\varepsilon}^T \hat{\varepsilon} = Y^T \left(I - L\right)^T \left(I - L\right) Y.$$

We obtain an estimation of the error variance $Var\left(\varepsilon_i\right) = \sigma^2$ as follows:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{tr\left(\left(I - L\right)^T \left(I - L\right)\right)} \\ &= \frac{Y^T \left(I - L\right)^T \left(I - L\right) Y}{tr\left(\left(I - L\right)^T \left(I - L\right)\right)}. \end{aligned} \tag{17}$$

The above method is termed the local linear estimation of the GTWR model. This estimation method not only provides the estimated values of the coefficient functions, but also generates estimated values of the partial derivatives of the coefficient functions about u, v and t. It is assumed that the functions $\beta_j\left(\cdot\right)$ possess equivalent degrees of smoothness and can be approximated equally well in the same interval. Optimal estimators of the smooth functions are obtained using the two-step estimator method if the functions possess different degrees of smoothness, where the degree of smoothness means the changed degree of coefficient function in the defined domain.

Assume that $\beta_p(u,v,t)$ is smoother than the other functions. The initial step involves obtaining an initial estimate of $\beta_0(u,v,t),\cdots,\beta_{p-1}(u,v,t)$. In this step the local linear estimation is used to generate the estimators of $\beta_0(u,v,t),\cdots,\beta_{p-1}(u,v,t)$, as shown in (13). We set

$$\hat{y}_i = y_i - \sum_{j=0}^{p-1} \hat{\beta}_j(u_i,v_i,t_i)x_{ij} = \beta_p(u_i,v_i,t_i)x_{ip} + \varepsilon_i, \qquad i = 1,2,\cdots,n.$$

In the second step a local least squares regression is fitted again by substituting the initial estimate in the local least squares problem. In this way a two-step estimator $\hat{\beta}_p(u,v,t)$ of $\beta_p(u,v,t)$ is obtained.

Specifically, we assume that $\beta_p(u,v,t)$ possesses a bounded second derivative so Taylor expansion can be used in the neighborhood of (u_0,v_0,t_0) as follows:

$$\beta_p(u,v,t) \approx$$
$$\beta_p(u_0,v_0,t_0) + \beta_p^{(u)}(u_0,v_0,t_0)(u-u_0) + \beta_p^{(v)}(u_0,v_0,t_0)(v-v_0) + \beta_p^{(t)}(u_0,v_0,t_0)(t-t_0).$$

The terms $\beta_p^{(u)}(u_0,v_0,t_0)$, $\beta_p^{(v)}(u_0,v_0,t_0)$ and $\beta_p^{(t)}(u_0,v_0,t_0)$ are partial derivatives of $\beta_p(u,v,t)$ about u,v and t at (u_0,v_0,t_0) respectively.

This naturally leads to the following weighted least squares problem:

$$\sum_{i=1}^{n}\Big\{\hat{y}_i - \Big[\beta_p(u_0,v_0,t_0) + \beta_p^{(u)}(u_0,v_0,t_0)(u_i-u_0) + \beta_p^{(v)}(u_0,v_0,t_0)(v_i-v_0)$$
$$+\beta_p^{(t)}(u_0,v_0,t_0)(t_i-t_0)\Big]x_{ip}\Big\}^2 w_{h_2}(u_0,v_0,t_0).$$

We let

$$G(u_0,v_0,t_0) = \begin{bmatrix} x_{1p} & x_{1p}(u_1-u_0) & x_{1p}(v_1-v_0) & x_{1p}(t_1-t_0) \\ x_{2p} & x_{2p}(u_2-u_0) & x_{2p}(v_2-v_0) & x_{2p}(t_2-t_0) \\ \vdots & \vdots & \cdots & \vdots \\ x_{np} & x_{np}(u_n-u_0) & x_{np}(v_2-v_0) & x_{np}(t_n-t_0) \end{bmatrix},$$

$e_{1,4} = (1,\ 0,\ 0,\ 0)^T$ and $\hat{Y} = \left(\hat{Y}_1,\hat{Y}_2,\cdots,\hat{Y}_n\right)^T$. Minimising the weighted least squares problem about $\beta_p(u,v,t)$, we obtain the two-step estimator $\hat{\beta}_p(u,v,t)$ of $\beta_p(u,v,t)$:

$$\hat{\beta}_p(u_0,v_0,t_0) = e_{1,4}^T\left[G^T(u_0,v_0,t_0)W_{h_2}(u_0,v_0,t_0)G(u_0,v_0,t_0)\right]^{-1}G^T(u_0,v_0,t_0)W_{h_2}(u_0,v_0,t_0)\hat{Y}, \tag{18}$$

where $W_{h_2}(u_0,v_0,t_0)$ is the weight matrix with the Gauss kernel function, and h_2 is a bandwidth in the second step.

Choosing Appropriate Bandwidth

In the process of calibrating the GTWR model, the model can be studied using cross-validation [14, 15]. Suppose the predicted value of y_i is denoted as a function of h. It can be written as $\hat{y}_{(i)}(h)$ in GTWR, so the sum of the squared error can be written as

$$CV(h_s,h_t) = \sum_{i=1}^{n}\left[y_i - \hat{y}_{(-i)}(h_s,h_t)\right]^2, \tag{19}$$

where $h_s = \sqrt{\frac{h^2}{\lambda}}$ and $h_t = \sqrt{\frac{h^2}{\mu}}$ are the space and time bandwidth parameters respectively; h_1 and h_2 can be calculated using cross-validation in the two-step estimator.

Then we analyse the rationality of the cross-validation method and note that

$$\begin{aligned} E(y_i - \hat{y}_{(-i)}(h_s,h_t))^2 &= E\left(Y_i - \hat{m}_{(-i)}(X_i)\right)^2 \\ &= E\left(Y_i - m\left(X_i\right) + m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)^2 \\ &= E\left(\varepsilon_i + m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)^2 \\ &= \sigma^2 + 2E\left(\varepsilon_i\left(m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)\right) + E\left(m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)^2, \end{aligned}$$

where $\hat{y}_{(-i)}\left(h_s,h_t\right) = \hat{m}_{(-i)}(X_i)$. Because ε_i and $m\left(X_i\right) - \hat{m}_{(-i)}(X_i)$ are conditionally independent, $E\left(\varepsilon_i\left(m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)\right)=0$ and we have

$$\begin{aligned} E(y_i - \hat{y}_{(-i)}(h_s,h_t))^2 &= \sigma^2 + E\left(m\left(X_i\right) - \hat{m}_{(-i)}(X_i)\right)^2 \\ &\approx \sigma^2 + E\left(m\left(X_i\right) - \hat{m}(X_i)\right)^2. \end{aligned}$$

Let $\left(Y_i - \hat{m}_{(-i)}(X_i)\right)^2$ be replaced with its mathematical expectation. Then

$$CV\left(h_s,h_t\right) \approx \sigma^2 + \frac{1}{n}E\left(m\left(X_i\right) - \hat{m}(X_i)\right)^2.$$

The second term is the average of the mean squared error of each point, so we can use the cross-validation method.

DATA AND MODELLING

Data and Pre-Analysis

Many factors such as precipitation, sunlight, temperature, atmospheric pressure, wind speed, relative humidity and pollutant emissions influence air quality. Generally, the AQI will be reduced with increases in temperature, atmospheric pressure and relative humidity when pollutant emissions remain constant. We have only considered the relationships among total monthly precipitation, mean monthly temperature and the mean monthly AQI.

To get a uniform distribution in the geographical position, we uniformly chose 67 cities for study (Table 1) to ensure the data can roundly reflect, if any, the variation of air quality with the geographical area in China. We collected 12 months (January-December 2013) of relevant data, viz. average monthly AQI, monthly precipitation, mean monthly temperature, latitude and longitude. The average monthly AQI data were obtained from the China National Environmental Monitoring Centre [16], and the monthly precipitation and mean monthly temperature data were obtained from the China Meteorological Data Sharing Service System [17]. We used Google to find the location of each city by latitude and longitude.

Table 1. Cities in China used as study sites

No.	Name	No.	Name	No.	Name	No.	Name
1	Xingtai	19	Shijiazhuang	35	Baoding	52	Handan
2	Tangshan	20	Beijing	36	Zhengzhou	53	Shenyang
3	Harbin	21	Changchun	37	Wuhan	54	Chengdu
4	Tianjin	22	Cangzhou	38	Nanjing	55	Changzhou
5	Shaoxing	21	Changsha	39	Taiyuan	56	Yancheng
6	Nantong	23	Huai'an	40	Jinhua	57	Dalian
7	Jinan	24	Lianyungang	41	Hohhot	58	Xuzhou
8	Hangzhou	25	Suzhou	42	Quzhou	59	Guiyang
9	Shanghai	26	Nanning	43	Zhongshan	60	Wenzhou
10	Dongguan	27	Guangzhou	44	Lishui	61	Zhuhai
11	Shenzhen	28	Zhaoqing	45	Huizhou	62	Xiamen
12	Suqian	29	Hengshui	46	Urumqi	63	Yinchuan
13	Haikou	30	Hefei	47	Ningbo	64	Taizhou
14	Fuzhou	31	Qinhuangdao	48	Taizhou	65	Zhoushan
15	Langfang	32	Nanchang	49	Chengde	66	Lanzhou
16	Xi'an	33	Zhangjiakou	50	Huzhou	67	Kunming
17	Qingdao	34	Xining	51	Lhasa		

In January the Chinese cities located on the south-eastern coast and most southern cities had lower average monthly AQI values, indicating good air quality (Figure 1). The average monthly AQI values of cities located in east-central and northern China were higher, indicting a poorer air quality. In January the air quality in Beijing was much worse than during other months (Figure 2).

The total monthly precipitation in most southern Chinese cities was relatively greater than that in inland cities as shown in Figure 3, where the box in the lower-left corner corresponds to the range of $\beta_1(u,v,t)$ estimation. If its absolute value is larger, the influence of x_{i1} on monthly precipitation is more important in the 67 cities. The rainfall in Hangzhou city was greatest in June and August as shown in Figure 4.

Among the 67 Chinese cities studied, the average monthly temperature increased from north to south and from inland to coastal areas (Figure 5). Figure 6 shows the average monthly temperature in Lanzhou city.

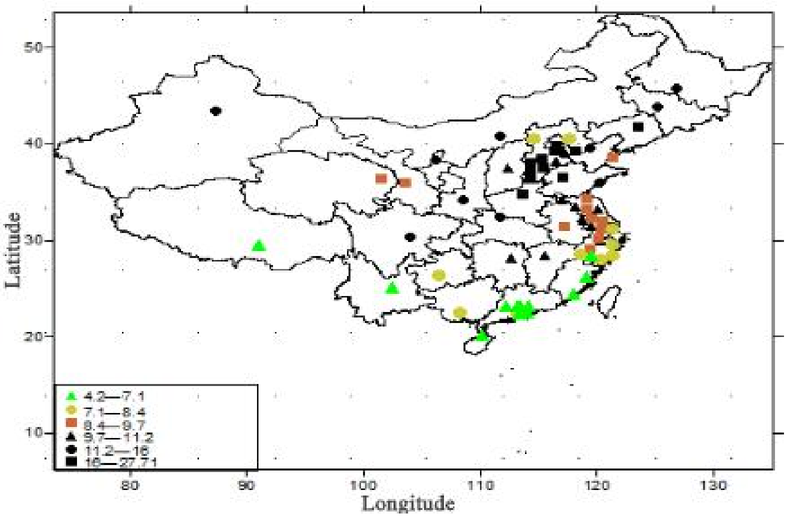

Figure 1. Average monthly AQI in 67 cities in January

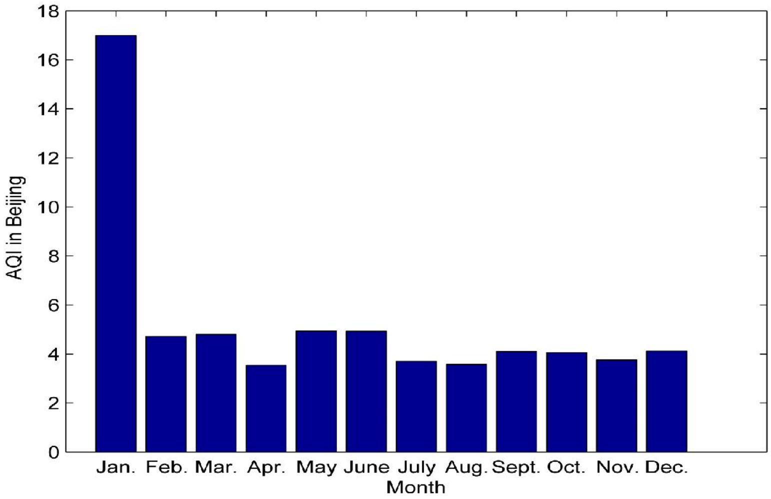

Figure 2. Average monthly air quality in Beijing city

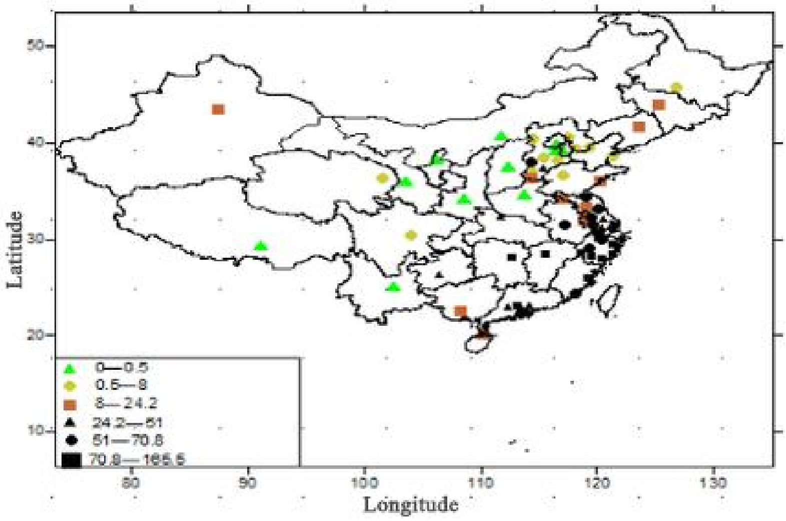

Figure 3. Total monthly precipitation (mm) in 67 cities in February

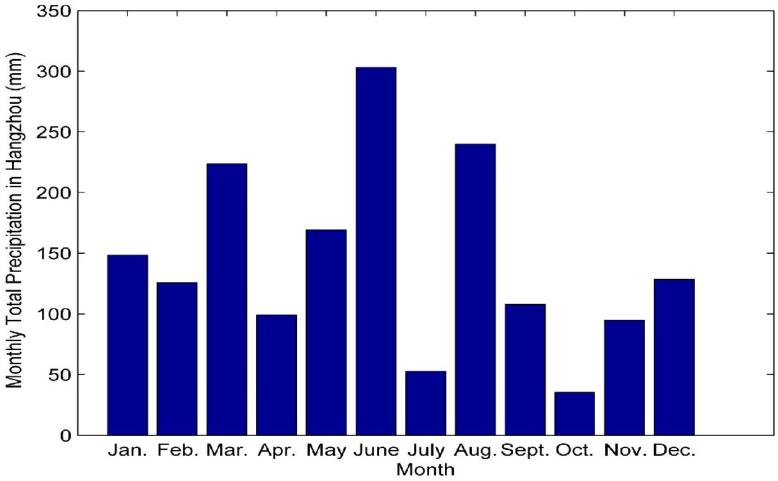

Figure 4. Total monthly precipitation (mm) in Hangzhou city

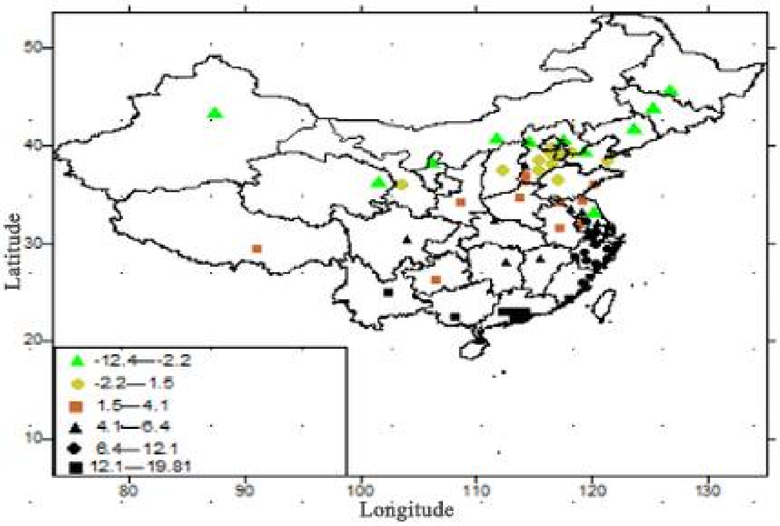

Figure 5. Average monthly temperature (°C) in 67 cities in February

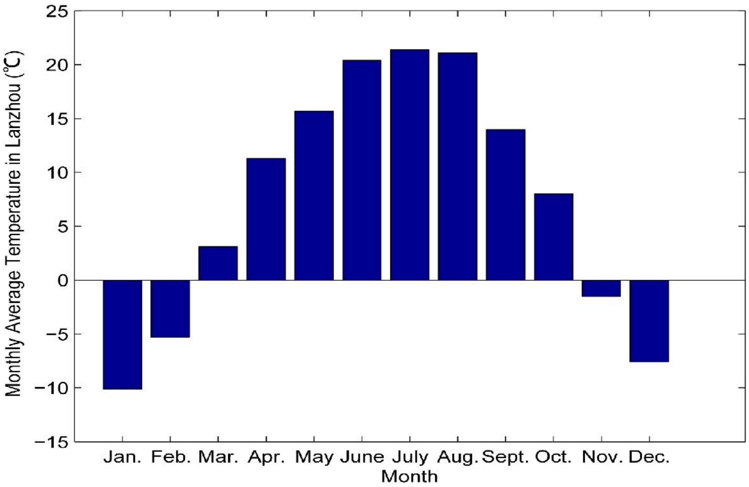

Figure 6. Average monthly temperature (°C) in Lanzhou city

Modelling

Let Y represent the average monthly AQI, X_1 the sum of monthly precipitation (mm) and X_2 the average monthly temperature (°C). We applied the GTWR model to the observational data of 12 months and 67 cities; the model can be written as

$$Y_i = \beta_0(u_i, v_i, t_i) + \beta_1(u_i, v_i, t_i) X_{i1} + \beta_2(u_i, v_i, t_i) X_{i2} + \varepsilon_i, \quad i = 1, 2, \cdots 67, \tag{20}$$

where $\beta_0(u_i, v_i, t_i)$ is the basis of average monthly AQI, $\beta_1(u_i, v_i, t_i)$ represents the average monthly rate of AQI associated with the sum of monthly precipitation, and $\beta_2(u_i, v_i, t_i)$ indicates the average monthly rate of AQI associated with the average monthly temperature.

SIMULATION AND ANALYSIS

The First Step

Using the significance test method of Cleveland et al. [18], Mei et al. [19] and Xuan et al. [20], we constructed the test statistic and obtained the global non-stationary-test p-value and the significance-test p-value of changes in each coefficient function in (19) through the third moment χ^2 approximation. As shown in Table 2, p is the global non-stationary-test p-value of the regression model, and p_0, p_1 and p_2 are the significance-test p-values of the regression cofficient functions β_0 , β_1 and β_2 respectively, which reflects significant changes in the regression coefficients using the local linear estimation method.

Table 2. Bandwidths and p-values obtained using local linear estimation

h_s	h_t	p	p_0	p_1	p_2
67.73 km	0.6801 month	1.31E-307	0.0000958	0.0002781	0.000233

Here, the kernel function is the Gaussian kernel function and all bandwidths are determined by the cross-validation method described in 'Choosing Appropriate Bandwidth' section. In Table 2 the significance test results show that the global non-stationary-test p-values of the regression model and the significance-test p-values of the regression coefficient functions are very small (approximately 0). This indicates significant differences in the monthly average AQI, the monthly precipitation sum and the monthly average temperature in the 67 cities. In other words, the monthly precipitation sum and monthly average temperature have a significant influence on the monthly average AQI.

Using the spatio-temporal data set and the results calculated by SAS software, we drew by means of Surfer (a software made by Golden Software) the β_0, β_1 and β_2 distribution maps of 67 cities in February, which are shown in Figures 7-9 respectively. In Figure 7 the values of β_0 less than 6 are mainly from southern locations and south-eastern coastal areas. These cities generally have sufficient rainfall and higher average temperatures or abundant forest cover. Thus, the distribution of the benchmark monthly average air index β_0 is lowest on the southern coast, south-eastern coast and south-western border, and gradually increases from these areas to northern China.

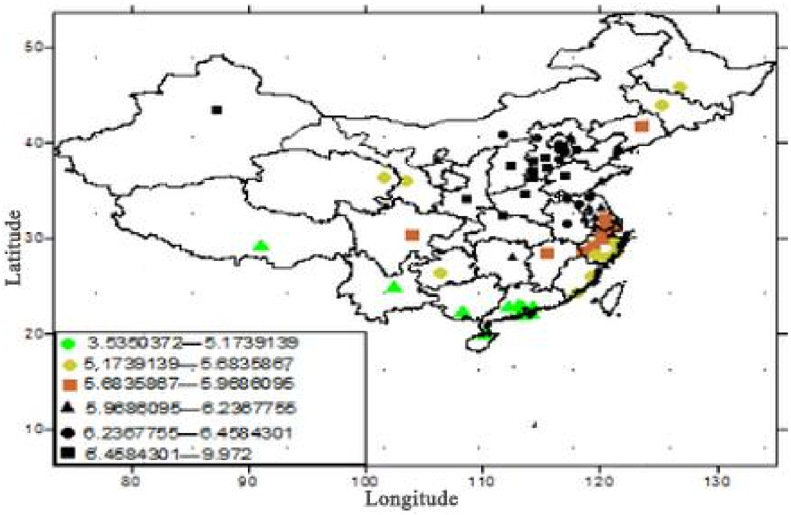

Figure 7. The β_0 distribution of 67 cities in February

Figure 8 shows that the β_1 values of the 67 cities in February are mostly negative. Only seven cities (Nanning, Hohhot, Lanzhou, Xining, Lhasa, Kunming and Yinchuan) have positive values. Either the wind has a greater impact on these cities (such as Lanzhou) or they have significant forest coverage (such as Nanning). Although we did not consider the impact of factors such as wind on the AQI, the indices of the seven cities result in positive β_1 values. Precipitation clearly has the largest impact on cities in north-eastern China including Jiangsu and Zhejiang provinces. It has less impact on the Yangtze River cities and most of the southern cities. This is mainly because industry is relatively developed and precipitation is relatively low in northern China and the north-eastern cities. On the whole, cities with adequate monthly rainfall along the Yangtze River and most of the southern cities have significantly high technological and industrial development with better and stable air quality, although considerable seasonal variation can increase the influence of precipitation. Therefore, it has less impact on AQI of the cities along the Yangtze River cities and most of the southern cities.

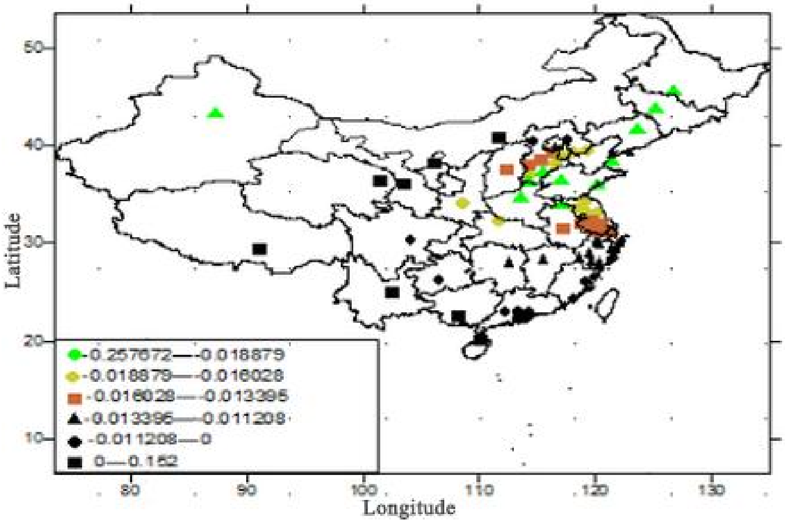

Figure 8. The β_1 distribution of 67 cities in February

Figure 9 shows the February distribution of β_2 in 67 Chinese cities. Cities where the average monthly temperature has the largest impact on the monthly AQI are in the eastern region, Yangtze River region, Tibet and Urumqi regions (temperature changes are more obvious in these cities), followed by those in the eastern coast and the south-eastern coastal areas. Cities in the southern area and Sichuan province have less impact on the average monthly air quality, mainly because the area temperature changes are more moderate. Northern and north-eastern cities are greatly influenced by rainfall, and the average monthly temperature of these cities has little impact on the average monthly AQI (greater than -0.079), but the index value remains negative. The β_2 values for Lanzhou, Hohhot, Xining and Yinchuan are positive, indicating that in these cities the average monthly humidity is not the major factor affecting the AQI. We found that the impacts of the total monthly precipitation and the monthly average temperatures for these four cities are small and these are not the main factors. Wind as well as other factors might be important, but they are not part of this study.

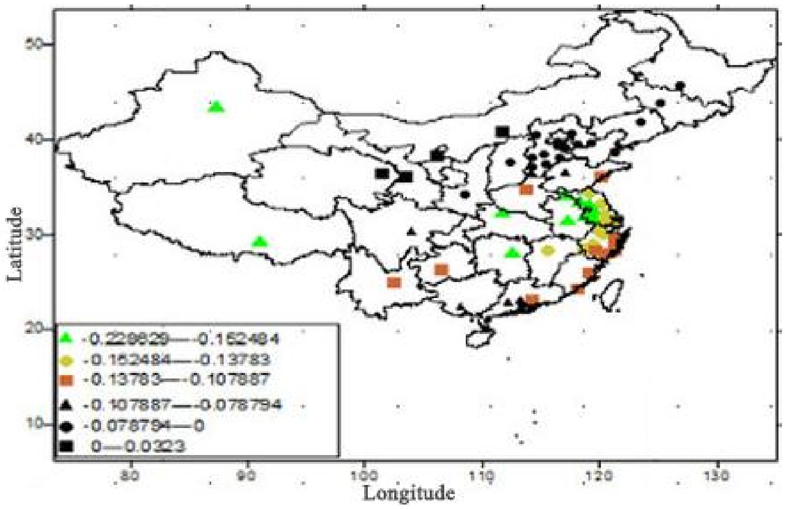

Figure 9. The β_2 distribution of 67 cities in February

The Second Step

The second step of the two-step estimator selects the optimal smooth degree of variable X_1, which is the sum of monthly precipitation, based on the first step, which uses the local linear estimation. From the two-step estimator, the significance-test p-value of the regression coefficient function β_1 is 6.702E-10, as shown in Table 3. Table 3 also states the significance-test p-value of the regression coefficient function β_1 based on the first-step estimator to be 0.0002781, which is greater than the significance-test p-value from the second-step estimation. Thus, the two-step estimator is a better estimator of the coefficient function β_1.

Table 3. Bandwidths and p-values based on two-step estimator

	h_s	h_t	p	p_0	p_1	p_2
The first step	67.73 km	0.6801 month	1.31E-307	0.0000958	0.0002781	0.000233
The second step	65.09 km	0.6396 month	0		6.702E-10	

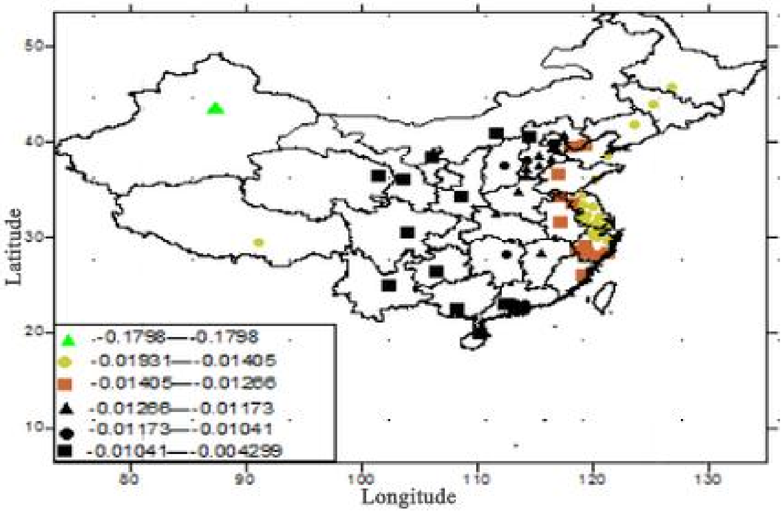

Figure 10. The β_1 distribution of 67 cities in February based on two-step estimator

Figure 10 shows that the February β_1 values of 67 cities obtained by the two-step estimation are all negative. Urumqi has the largest precipitation impact with a 0.1798 rate of change. When the monthly precipitation increases by one unit, the monthly average AQI decreases by 0.1798, while the average temperature remains unchanged. The possible reason that Urumqi has the largest February precipitation effect among the 67 cities is that it generates significant atmospheric pollution from coal burning. Inversion weather and the atmospheric stability of Urumqi in February are not conducive to the diffusion and dilution of polluted air [21]. However, precipitation has a large impact on the AQI of north-eastern, northern, eastern and most southern coastal cities, including Tibet. The rate of change is between -0.01931 and -0.01041. The effect of rainfall on the AQI of Xi'an, Lanzhou, Nanning and other mid-western and south-western cities is relatively small. Local factors may affect the AQI. For example, Nanning is greatly influenced by a ridge of high pressure in February [22]. Lanzhou is influenced by valley terrain conditions; temperature inversion conditions are stable and deep in February, and the wind has a relatively high impact on the AQI [23].

CONCLUSIONS

Our analysis has demonstrated that spatio-temporal heterogeneity prevails in real air quality data, and these data change over both time and space in the study areas. The GTWR approach can simultaneously deal with both spatial and temporal heteroscedasticity and produces good results in handling the relationships among precipitation, temperature and air quality. This study shows that cities in northern China generally have the highest air quality. There are ten cities with the worst air quality; seven are located in Hebei province. These data are consistent with our simulation results. The monthly average AQI varies with the mean monthly temperature in each city. When the monthly precipitation increases by one unit, the northern Chinese cities, eastern coastal cities, north-eastern cities and Urumqi have maximum rates of change. The simulation results are consistent with the results found in real life and those from pre-analysis. The GTWR model is useful for simulation and analysis of the distribution characteristics of the Chinese AQI.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their valuable suggestions. This work was supported by National Natural Science Foundation of China (11261031).

REFERENCES

1. C. A. Pope, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito and G. D. Thurston, “Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution”, *J. Am. Med. Assoc.*, **2002**, *287*, 1132-1141.
2. S. J. Santosa, T. Okuda and S Tanaka, “Air pollution and urban air quality management in Indonesia”, *CLEAN Soil Air Water*, **2008**, *36*, 466-475.
3. L. Sichletidis, I. Tsiotsios, A. Gavriilidis, D. Chloros, I. Kottakis, E. Daskalopoulou and T. Konstantinidis, “Prevalence of chronic obstructive pulmonary disease and rhinitis in northern Greece”, *Respiration*, **2005**, *72*, 270-277.
4. K. N. Grigoropoulos, P. T. Nastos, G. Ferentinos, A. Gialouris, T. Vassiliou, J. Mavroidakos, D. Avgeri, V. Kalabokis and D. Saratsiotis, “Spatial distribution of PM1 and sinus arrhythmias in Athens, Greece”, *Fresenius Environ. Bull.*, **2008**, *17*, 1426-1431.
5. K. N. Grigoropoulos, P. T. Nastos and G. Ferentinos, “Spatial distribution of PM1 and PM10 during Saharan dust episodes in Athens, Greece”, *Adv. Sci. Res.*, **2009**, *3*, 59-62.
6. P. T. Nastos, A. G. Paliatsos, M. B. Anthracopoulos, E. S. Roma and K. N. Priftis, “Outdoor particulate matter and childhood asthma admissions in Athens, Greece: A time-series study”, *Environ. Health*, **2010**, *9*, 45, doi: 10.1186/1476-069X-9-45.
7. M. Xiang, Y. Han, Z. Deng, “Spatial-temporal distribution characteristic of Chinese cities air pollution in 2007”, *Admin. Tech. Environ. Monit.*, **2009,** *21*, 33-36.
8. J. Zhang, J. Sun, G. Wang, L. An and W. Wang, “Relation between the spatial-temporal distribution characteristics of air quality index and meteorological conditions in Beijing”, *Meteorol. Environ. Sci.*, **2014**, *37*, 33-39.
9. A. Poupkou, P. Nastos, D. Melas, C. Zerefos, “Climatology of discomfort index and air quality index in a large urban mediterranean agglomeration”, *Water Air Soil Pollut.*, **2011**, *222*, 163-183.
10. P. Kassomenos, A. N. Skouloudis, S. Lykoudis and H. A. Flocas, “Air-quality indicators for uniform indexing of atmospheric pollution over large metropolitan areas”, *Atmos. Environ.*, **1999**, *33*, 1861-1879.
11. X. Li, “Air quality forecasting based on GAB and fuzzy BP neural network”, *J. Huazhong Univ. Sci. Technol. Nat. Sci. Edn.*, **2013**, *41*, 63-69.
12. C. P. Lo, “Population estimation using geographically weighted regression”, *GISci. Remote Sens.*, **2008**, *45*, 131-148.
13. B. Huang, B. Wu and M. Barry, “Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices”, *Int. J. Geogr. Inf. Sci.*, **2010**, *24*, 383-401.
14. J. Fan, I. Gijbels, T.-C. Hu and L.-S. Huang, “A study of variable bandwidth selection for local polynomial regression”, *Stat. Sinica*, **1998**, *6*, 113-127.
15. W. Zhang and S.-Y. Lee, “Variable bandwidth selection in varying-coefficient models”, *J. Multivariate Anal.*, **2000**, *74*, 116-134.
16. China National Environmental Monitoring Centre, “Data of average monthly air quality index”, **2014**, http://www.cnemc.cn/ (Accessed: April 2016).

17. China Meteorological Data Sharing Service System, "Data of monthly total precipitation and mean monthly temperature", **2014**, http://cdc.nmic.cn/home.do (Accessed: April 2016)
18. W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting", *J. Am. Stat. Assoc.*, **1988**, *83*, 596-610.
19. C. Mei and N. Wang, "Recent Regression Analysis Methods", Science Press, Beijing, **2012**, pp. 160-220.
20. H. Xuan, S. Li and Y. Zhang, "Influence analysis of geographically and temporally weighted regression model", *J. Lanzhou Univ. Technol.*, **2013**, *39*, 135-138.
21. X. Liu, Y. Zhong, H. Qing and Z. Guo, "The variety characteristics and influencing factors of air quality in Urumqi and its surrounding cities", *Desert Oasis Meteorol.*, **2010**, *4*, 12-17.
22. H. Xie, "Analysis on weather factors affecting on atmosphere quality in Nanning city", *Yunnan Environ. Sci.*, **2005**, *24*, 28-31.
23. Y. Yu, D. S. Xia, L. H. Chen, N. Liu, J. B. Chen and Y. H. Gao, "Analysis of particulate pollution characteristics and its causes in Lanzhou, Northwest China", *Chinese J. Environ. Sci.*, **2010**, *31*, 22-28 (in Chinese).