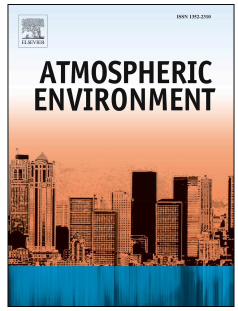
# Accepted Manuscript




Prediction of hourly $PM_{2.5}$ using a space-time support vector regression model

Wentao Yang, Min Deng, Feng Xu, Hang Wang

PII: S1352-2310(18)30153-5

DOI: 10.1016/j.atmosenv.2018.03.015

Reference: AEA 15885

To appear in: *Atmospheric Environment*

Received Date: 2 July 2017

Revised Date: 5 March 2018

Accepted Date: 6 March 2018

Please cite this article as: Yang, W., Deng, M., Xu, F., Wang, H., Prediction of hourly $PM_{2.5}$ using a space-time support vector regression model, *Atmospheric Environment* (2018), doi: 10.1016/j.atmosenv.2018.03.015.

# Prediction of hourly $PM_{2.5}$ using a space-time support vector regression model

**Abstract:** *Real-time air quality prediction has been an active field of research in atmospheric environmental science. The existing methods of machine learning are widely used to predict pollutant concentrations because of their enhanced ability to handle complex non-linear relationships. However, because pollutant concentration data, as typical geospatial data, also exhibit spatial heterogeneity and spatial dependence, they may violate the assumptions of independent and identically distributed random variables in most of the machine learning methods. As a result, a space-time support vector regression model is proposed to predict hourly $PM_{2.5}$ concentrations. First, to address spatial heterogeneity, spatial clustering is executed to divide the study area into several homogeneous or quasi-homogeneous subareas. To handle spatial dependence, a Gauss vector weight function is then developed to determine spatial autocorrelation variables as part of the input features. Finally, a local support vector regression model with spatial autocorrelation variables is established for each subarea. Experimental data on $PM_{2.5}$ concentrations in Beijing are used to verify whether the results of the proposed model are superior to those of other methods.*

**Keywords:** Real-time air quality prediction; spatial heterogeneity; spatial dependence; support vector regression; spatial clustering; Gauss vector weight function

## 1. Introduction

Epidemiologic studies have demonstrated that short-term (acute) exposure to air pollution can damage human health; of specific concern is particulate matter, which includes fine particulate matter ($PM_{2.5}$), that can accumulate in the respiratory system and directly increase the risk of death caused by lung cancer, cardiovascular disease, and pulmonary illness (Dominici et al., 2006; Diaz-Robles et al., 2015; Kloog et al., 2014; Di et al., 2017). Therefore, to protect the public from particulate matter or air pollution, real-time air quality prediction has been an active field of research in atmospheric environmental science.

Existing methods for real-time air quality prediction can be roughly classified into two categories: physically based methods and empirically based methods (Zhang et al., 2012). Physically based methods, also referred to as chemical transport models, aim to estimate air pollutants by using deterministic chemical transport models that encompass all major meteorological, physical, and chemical processes (Wayland et al., 2002). However, the performance of physically based approaches can be undermined by high uncertainty in the amount of emissions and the chemical reactions (i.e., reaction rates), which are presented at a fine space-time resolution. Comparatively, empirically based methods directly model the relationships between pollutant concentrations and relevant variables. Although empirically based methods cannot describe the pollution process, they are widely used to predict pollution because it these methods are easy to implement with suitable accuracy. Therefore, based on the in-depth development of geospatial data analysis in geographical information science (GIS), which provides an effective means to reveal the space-time distribution and evolution of air pollutant concentrations (Miller and Han, 2009), this paper focuses on empirically based methods.

Empirically based methods can be further grouped into two categories, namely, statistical methods and machine learning methods. Statistical methods generally assume that the data on air pollutant concentrations are generated by a given stochastic data model, and the stages of model building consist of model specification, coefficient estimation, model verification and statistical inference (Wasserman, 2004). Many statistical models, such as the multiple linear regression model (Abdul-Wahab et al., 2005; Ghazali et al., 2010), the land-use regression model (Hoek et al., 2008; Johnson et al., 2010; Wang et al., 2013), the geographically weighted regression model (Robinson et al., 2013), and the mixed-effect model (Lee et al., 2011; Kloog et al., 2014), have been adopted to predict air pollutant concentrations. Nevertheless, these specified models tend to oversimplify the complex non-linear relationships that exist between air pollutant concentrations and predictor variables.

Comparatively, machine learning methods have obvious advantages in handling complex non-linear relationships among environmental data. Machine learning methods mainly apply algorithmic models and treat the data mechanism as an unknown; additionally, the most commonly used models include artificial neural networks (ANNs) (Ordieres et al., 2005; Arhami et al., 2013), classification and regression trees (Brokamp et al., 2017), support vector regression (SVR) (Sánchez et al., 2011; Nieto et al., 2013), and hidden Markov models (Dong et al., 2009; Sun et al., 2013). Most of those machine learning methods are based on the assumptions of independent and identically distributed random variables (Pereira and Mello, 2011). However, data on air pollutants also exhibit the same characteristics as geospatial data (i.e., spatial heterogeneity and spatial dependence), which violates the assumptions of machine learning methods. Therefore, it is inappropriate to directly apply machine learning methods to model air pollutant data, and how to incorporate spatial heterogeneity and spatial dependence in the process of machine learning is an urgent problem that requires attention.

It has been shown that SVR outperforms other machine learning methods in predicting air quality because training for the SVR produces a global optimum (Sánchez et al., 2011; Nieto et al., 2013). As a result, this study aims to develop a space-time support vector regression (STSVR) model to predict hourly $PM_{2.5}$ concentrations. The STSVR model is developed by incorporating spatial dependence and spatial heterogeneity into the modelling process used by conventional support vector regression models.

## 2. Materials and Methods

### 2.1 Materials

#### 2.1.1 *Area description*

The study area was the city of Beijing, which is located in North China and is the capital of the People's Republic of China. The area has a monsoon-influenced humid continental climate, which is characterised by higher humidity in the summers and windier, colder, and drier winters. The daily average temperature in July is approximately 26.2°C, and in January, it is about -3.7°C. The annual precipitation is approximately 570 mm, with about three-fourths of the total precipitation falling between June and August. Annually, approximately 2,671 hours of bright sunshine is received, and monthly percent possible sunshine ranges from approximately 65% in July to approximately 47% in January and February.

In recent years, the study area has frequently suffered from severe air pollution. $PM_{2.5}$ has been shown to be the main air pollutant, and its concentrations are greatly influenced by emission sources (Zhang et al., 2015). Lv et al. (2016) reviewed recent studies that reported source apportionment results from 2000 to 2012 in Beijing. During this period, the annual average $PM_{2.5}$ concentrations gradually decreased. Summer is identified as the least polluted season, and winter is the most polluted season. The major compositions of $PM_{2.5}$ are sulphate, organic matter, nitrate and ammonium. It can also be found that vehicles, industry, dust, biomass burning, coal combustion and secondary products were major sources of $PM_{2.5}$. Two periods (i.e., before 2005 and after 2005) were further assessed to investigate differences between the source contributions. Specifically, the annual average contributions of vehicle exhaust increased from 6.8% before 2005 to 10.6% after 2005. The industrial contributions prior to and after 2005 were 6.9 and 15.5%, respectively. The contribution of dust was 13.3% before 2005 and 19.9% after 2005. Biomass burning also contributed less before 2005, with an annual average of 7.9%, than after 2005, when the annual average increased to 11.6%. The contributions of coal combustion were almost 15.0% in both periods.

2.1.2 *Data collection*

There are 35 air quality monitoring sites that record hourly average $PM_{2.5}$ concentrations. The tapered element oscillating microbalance method is used to measure $PM_{2.5}$ concentrations automatically. In addition, source apportionment by manual methods is applied to analyse pollutant components and to evaluate the accuracy of automatic monitoring. In general, the estimated error of automatic monitoring is less than 5%. The related information can be obtained from the official website of the Beijing Municipal Environmental Mentoring Center (http://www.bjmemc.com.cn). Fig. 1 shows the spatial distribution of these monitoring sites. In our experiment, air quality data were collected from these monitoring sites for the period between March and April 2014. Meanwhile, considering that meteorological elements are the main factors influencing changes in $PM_{2.5}$ concentrations, the meteorological elements for that same period were obtained from weather monitoring sites and selected as the predictor variables. The meteorological data utilised in the process of predicting the concentration of $PM_{2.5}$ are (1) surface temperature (ºC), (2) relative humidity (%), (3) wind force (level), (4) wind direction (angle), and (5) precipitation (mm).

[Insert Fig. 1 about here]

## 2.2 Methods

As discussed above, SVR provides a better learning strategy in modelling non-spatial data. However, spatial dependence and spatial heterogeneity make it necessary to extend SVR into the field of environmental or geospatial data analysis. Therefore, an STSVR model is developed in this paper by incorporating these spatial characteristics. The implementation of the STSVR model is shown in Fig. 2. First, spatial clustering analysis is used to address spatial heterogeneity of the space-time series of hourly $PM_{2.5}$ concentrations, and then spatial autocorrelation variables based on a Gauss vector weight function are identified to address spatial dependence of hourly $PM_{2.5}$ concentrations. Finally, a local SVR with spatial autocorrelation variables is employed to model hourly $PM_{2.5}$ concentrations.

[Insert Fig. 2 about here]

2.2.1 *Addressing spatial heterogeneity using spatial clustering analysis*

Spatial heterogeneity refers to the non-stationarity of the spatial processes generating the observed data (Jiang, 2014). Specifically, the statistical characteristics of $PM_{2.5}$ concentrations and the relationships between $PM_{2.5}$ and the associated factors may vary over space. To address spatial heterogeneity, it is common to build a local model, such as GWR and its variants, at each spatial location. However, this approach may be inappropriate for structured heterogeneity, which means that the model tends to be more dissimilar at locations that are farther apart. It is redundant to build a point-based model at each location, and thus, region-based models may be more suitable.

Spatial clustering algorithms can divide an entire study area into several homogeneous or quasi-homogeneous subareas; therefore, spatial clustering analysis is employed to group $PM_{2.5}$ data into several spatial clusters, and a local region-based model is built based on the results of spatial clustering. The existing methods for spatial clustering are mainly classified into five categories: hierarchical methods, partitioning methods, grid-based methods, density-based methods, and model-based methods (Liao, 2005). Most of these methods are presented using general-purpose clustering methods, which have a limited ability to recognise spatial patterns, including neighbours (Guo et al., 2003). To overcome this limitation, a model-based method, which is called a geographical self-organising map (GeoSOM) and considers geography's first law, is selected to find heterogeneous structures.

GeoSOM is developed by extending the conventional self-organising map algorithm to explicitly consider geographic information. In GeoSOM, first, the spatial coordinates of the objects are used as input vectors to search for the winning unit, which is called the geographical best match. Subsequently, the attribute values are used as input vectors, and only the units in the neighbourhood of the geographical best match are used to find the final best match in the output layer. Thus, both spatial proximity and attribute similarity within clusters can be guaranteed. In the process of spatial clustering analysis with GeoSOM, two crucial components need to be considered: the similarity measure and the cluster evaluation criteria. In this study, Euclidean distance is chosen as the similarity measure, and two types of clustering validity indices, namely, the DB index (Davies and Bouldin, 1979) and the Sil index (Rousseeuw, 1987), are used to select the number of clusters. A small DB index value or a larger Sil index value generally indicate better clustering results. The clusters that satisfy these two indices are chosen (Kryszczuk and Hurley, 2010).

2.2.2 *Addressing spatial dependence using* a *Gauss vector weight function*

Spatial dependence or spatial autocorrelation means that the $PM_{2.5}$ concentration at spatial $i$ and time $t$ not only depend on other associated factors but also depend on the previous concentrations at both that point and its neighbour (Tobler, 1970). Therefore, it is necessary to apply spatial autocorrelation variables as inputs in prediction models to handle spatial dependence. In the field of geospatial analysis, spatial autocorrelation variables are defined via a spatial weights matrix, $W$ ($n \times n$), which is the formal expression of spatial dependence between observation sites (Getis and Aldstadt, 2004).

Suppose $n \times l$ samples $(\boldsymbol{x}_i(t),\ y_i(t))$ are observed at spatial location $i$ ($i$=1, …, $n$) and time $t$ ($t$=1, …, $l$), where $\boldsymbol{x}_i(t) \in R^m$ denotes the independent variables, and $y_i(t-1)$ denotes the predictive variable, i.e., the $PM_{2.5}$ concentrations in this study. Spatial autocorrelation variables can be defined using the following equation.

$$y_i^*(t-1) = \sum_{j=1}^{n} w(i,j) y_i(t-1), \tag{1}$$

where $w(i,j)$ (an element in $W$) represents the spatial weight between spatial locations $i$ and $j$. The general strategy for determining $W$ is based on spatial distance or spatial contiguity. The first assumes that the degree of correlation depends on the spatial distance, and the second determines the degree of correlation based on the spatial topology relationships. Both methods make the isotropic assumption, which assumes the effect from any direction can be regarded as equivalent.

However, the spreading process of air pollution is obviously anisotropic because air pollutants are usually transported based on the direction of the wind. The traditional strategy based on the isotropic assumption does not describe the spatial dependence of air pollutant concentrations. For example, in Fig. 3, if the wind direction is NE, the $PM_{2.5}$ concentrations at $p_0$ are directly affected by the concentrations at $p_1$, $p_2$, $p_3$, and $p_4$, and they may not be affected by the concentrations at $p_5$ and $p_6$. In addition, the affected degree is obviously negatively correlated with the angle $\theta$ and the distance (the angle is defined by the wind direction and the edge between two points, and the distance is defined by the spatial location of two points). Specifically, in the terms of $p_0$, although points $p_1$ and $p_2$ have the same angle, the affected degree of $p_1$ is higher than that of $p_2$ because $d_{01}$ is smaller than $d_{02}$. Likewise, the affected degree of $p_3$ is higher than that of $p_4$ because the NE$p_0p_3$ angle is smaller than the NE$p_0p_4$ angle, even though they are equally distant from $p_0$.

[Insert Fig. 3 about here]

A Gauss kernel function can only represent the dependence of spatial distance; it cannot describe the differences in direction. To simultaneously address anisotropy, the Gauss vector weight (GVW) is presented based on the Gauss kernel function. The GVW combines the direction and distance effects with the transport process of air pollutants, which can be described as

$$w_{ij}\left(d_{ij},\theta_{ij}(t)\middle|c\right)=\begin{cases} e^{-\frac{d_{ij}^2\sin\theta_{ij}(t)}{2c^2}} & if \quad 0^\circ\le\theta_{ij}(t)\le 90^\circ \\ 0 & if \quad 90^\circ<\theta_{ij}(t)\le 180^\circ \end{cases} \tag{2}$$

where $d_{ij}$ and $\theta_{ij}$ represent the distance variable and the angle variable, respectively. The distance can be calculated directly by spatial locations. Because the wind direction changes over time, the angle variable is a temporal variable that can be computed by the dynamic wind direction. It is noted that there is one bandwidth parameter, $c$, used, and it represents the trade-off between the direction and distance effects; this bandwidth parameter needs to be optimised.

2.2.3 *Modelling hourly $PM_{2.5}$ concentrations using STSVR*

A support vector machine (SVM) was developed to solve multi-dimensional function estimation problems using statistical learning theory (Vapnik, 2000). SVM can be divided into two main categories, namely, support vector classification and support vector regression. The former is used to address classification problems, and the latter is designed to handle problems associated with function approximation. Because the $PM_{2.5}$ concentrations are continuous values, predicting $PM_{2.5}$ concentration is a type of regression problem, and thus, SVR is suitable to model $PM_{2.5}$

concentrations. Conventional SVR methods are directly employed to model hourly $PM_{2.5}$ concentrations without considering spatial heterogeneity and spatial dependence. As discussed above, it is more suitable to build a spatially local model for each cluster or sub-area instead of a global model for the entire study area. Meanwhile, the spatial autocorrelation variables, $y_i{}^*(t-1)$, identified by the GVW function should be considered as the input variables.

Therefore, the STSVR model aims to find a series of local functions $f_{area(j)}(\boldsymbol{x}(t))$ ($j$=1, …, $k$) that can accurately predict the observation $y$ with the new input data $\boldsymbol{x}$ and spatial autocorrelation variables $y^*(t-1)$ at area $j$, and $k$ denotes the number of sub-areas obtained from spatial clustering analysis. Theoretically, a linear function $f_{area(j)}(\boldsymbol{x}(t))$ exists in the high dimensional feature space to formulate the non-linear relationship between the input data and the target data at sub-areas $j$ and $t$. The linear function is calculated using the following equation

$$f_{area(j)}(\boldsymbol{x}(t)) \ = \boldsymbol{w}^{\boldsymbol{T}}_{area(j)}\varphi([\boldsymbol{x}(t), y^*(t-1)]) + b_{area(j)}, \tag{3}$$

where the parameters $\boldsymbol{w}^{\boldsymbol{T}}_{area(j)}$ and $b_{area(j)}$ are the normal vector and the threshold at area $j$, respectively, and $\boldsymbol{x}^*(t) = [\boldsymbol{x}(t), y^*(t-1)]$ denotes the predictive variables. By solving the quadratic optimisation problem with inequality constraints, the STSVR model regression function can be obtained using the following equations

$$\boldsymbol{w}_{area(j)} = \sum_{t=1}^{l}\sum_{i=1}^{n(j)}(\beta^*_{i,area(j)}(t) - \beta_{i,area(j)}(t))\varphi(\boldsymbol{x}^*_i(t)) \tag{4}$$

$$f(\boldsymbol{x}^*) = \sum_{t=1}^{l}\sum_{i=1}^{n(j)}\left(\beta^*_{i,area(j)}(t) - \beta_{i,area(j)}(t)\right)K(\boldsymbol{x}^*_i(t), \boldsymbol{x}^*) + b_{area(j)} \tag{5}$$

where $\beta^*_i$ and $\beta_i$ represent the Lagrangian multipliers; $K(\boldsymbol{x}^*_i, \boldsymbol{x}^*)$ is called the kernel function, and any functions meeting Mercer's condition, such as Gaussian radial basis function, can be adopted as the kernel function, which can be defined as $exp(-0.5\|\boldsymbol{x}^*_i - \ \boldsymbol{x}^*\|^2/\sigma_{area(j)}{}^2)$ with a width of $\sigma_{area(j)}$.

It can be found that there are two differences between STSVR and conventional SVR. First, spatial autocorrelation variables are included in the predictive variables in STSVR. Second, spatially local models need to be built for each sub-area in STSVR; in contrast, SVR aims to build a global model for the entire area. It is worth noting that the region is divided into several sub-areas to address spatial heterogeneity, but modelling spatial dependence is based on the data of the entire area, which means that the neighbours of spatial location $i$ not only include the elements of the sub-area of spatial location *i* but also the elements of other spatial areas.

## 3. Results

Experimental data were divided into two parts, one for modelling and the other (i.e., data from the last day) for prediction analysis. The latter data used for prediction analysis were regarded as unknown data and were not included in the building of the prediction model. First, GeoSOM was used to group $PM_{2.5}$ space-time series data into several clusters. It should be noted that the spatial variability of $PM_{2.5}$ concentrations can be identified from two aspects. The first is the result of spatial variability, which is directly analysed by $PM_{2.5}$ space-time series data; the second includes the causes of spatial variability, including meteorological elements, pollution source information, and topography. By contrast, it is easy to identify spatial variability based on $PM_{2.5}$ space-time series data because it does not consider multiple factors or the interactions among them. Specifically, we identified spatial clusters from the results of spatial variability, and $PM_{2.5}$ space-time series data were regarded as the input for spatial cluster analysis.

Two types of cluster evaluation indices, namely, the DB index and the Sil index, were employed to determine the optimal number of clusters. The values of the DB index and the Sil index with different numbers of clusters are shown in Fig. 4. The final number of clusters was chosen to be 14 because this number results in a relatively low DB index value and a relatively high Sil index value; the corresponding clustering results are shown in Fig. 5.

[Insert Fig. 4 about here]

[Insert Fig. 5 about here]

Through the partial autocorrelation function, the $PM_{2.5}$ concentration at time *t* is statistically relevant to the $PM_{2.5}$ concentrations at sites in the upwind direction at time *t*-1; hence, it is unnecessary to consider $PM_{2.5}$ concentrations in the upwind direction at times *t*-2, *t*-3, etc. in the input for STSVR. The bandwidth parameter *c* is set from 0, $0.1d_{\min}$, $0.2d_{\min}$, …, to $d_{\min}$, where $d_{\min}$ is the nearest-neighbour distance of spatial location *i*., The $d_{\min}$ changes over space because of the varying density of air quality monitoring stations. Considering that there is no structural method on how to efficiently set the STSVR parameters, we first fixed a bandwidth value, and other parameters were then determined by minimising the RMSEs; the smallest RMSEs for each bandwidth value are shown in Fig. 6. It can be found that the RMSEs decrease gradually with the increase in the bandwidth value, and finally, the RMSEs stabilise when the bandwidth value reaches $1.6d_{\min}$. Hence, the optimal value of the bandwidth parameter was selected to be $1.6d_{\min}$. After training the parameters, we can apply the STSVR model for predictive analysis.

[Insert Fig. 6 about here]

To demonstrate the effectiveness of the proposed STSVR model, other models, including the auto-regressive integrated moving average model with explanatory variables (ARIMAX) model, the global SVR model, and the space-time artificial neural networks (STANNs) model, were selected for comparative analysis. Meanwhile, ARIMAX can include covariates in ARIMA models, which can consider both temporal autocorrelation and other covariates (Brockwell and Davis, 1996). Global SVR aims to model all the data using a single model, but this approach cannot handle spatial heterogeneity. STANNs were initially presented by Cheng et al. (2009) to incorporate space-time autocorrelation into feedback ANNs. In addition, the result of the STANNs in each cluster (i.e., local STANNs) was used for comparative analysis. It is worth noting that because different models were constructed on the basis of different sizes of areas, there were obvious differences in the training sample sizes (listed in Table 1). The training sample sizes of local STANNs and STSVR were both $1440Ns(C)$, where $Ns(C)$ indicated the number of stations in subarea $C$ or cluster $C$ (listed in Table 2), and 1440 was derived from 24 samples a day within 60 days at each station. ARIMAX was used to analyse time series of a single station and then the training sample size was 1440×1. STANNs and Global SVR were constructed on the basis of the samples from the whole study area and then the training sample sizes were both 1440×35.

Two accuracy evaluation indices, i.e., a total accuracy index ( $p_t$ ) and a total absolute error index ( $e_t$ ), were used to quantitatively evaluate the predictive results of the different models. Their expressions are as follows

$$p_t = 1 - \frac{\sum_{i=1}^{n} |Prediction(i)_t - Observation(i)_t|}{\sum_{i}^{n} Observation(i)_t} \tag{6}$$

$$e_t = \frac{\sum_{i=1}^{n} |Prediction(i)_t - Observation(i)_t|}{n} \tag{7}$$

where $Prediction(i)_t$ and $Observation(i)_t$ represent the predicted value and the observation value at spatial locations *i* and *t*, respectively. The accuracies and the absolute errors of the next 1-6, 7-12 and 13-24 hours are listed in Table 1.

[Insert Table 1 about here]

Table 1 shows that the highest values of *p* for all methods occur at the next 1-6 hours, followed by the values at the next 7-12 hours; additionally, the values at the next 13-24 hours are usually the lowest. The total absolute errors at hours 1-6 hours lower than those during other periods, and the highest values occur after 13-24 hours. This means that the prediction accuracy tends to decrease as the prediction time increases; in other words, as the prediction time increases, the level of uncertainty increases. Further, according to the results of different methods, it can be found that the *p* from the STSVR model at the hours 1-6 and 6-12 are 0.720, 0.703, respectively, which are higher than the values from the other methods. The total absolute errors of the STSVR model at the first two periods are 22.96 and 31.99 ug/m$^3$, respectively, which are lower than the total absolute errors from the other methods. Therefore, it is demonstrated that the results of the proposed STSVR model are better than the results of the other methods.

Moreover, the results of four randomly selected stations (i.e., S1, S2, S20, and S31, whose locations are shown in Fig. 5) are shown in Fig. 7. In contrast, the curves from the STSVR model are closer to the actual change, which further verifies the effectiveness of the proposed method.

[Insert Fig. 7 about here]

Moreover, the accuracies of the STSVR model in different sub-areas are listed in Table 3. It is obvious that the accuracies dramatically change over space. The predicted results at C7 are superior to those at other clusters during the next 1-6 hours. However, the accuracy at C7 during the next 7-12 and 13-27 hours are not the highest. The predicted results at C1 are better than those at other clusters at hours 7-12, and the predicted results at C11 are better than those at other clusters at hours 13-24. Meanwhile, it was also found that the lowest accuracies at the next 1-6, 7-12, and 13-24 hours correspond to C13, C13, and C9, respectively.

[Insert Table 2 about here]

## 4. Conclusions and future work

The paper develops an extended support vector regression model to predict hourly $PM_{2.5}$ concentrations. Spatial heterogeneity and spatial autocorrelation are incorporated into the modelling process of SVR. First, spatial clustering is executed to address spatial heterogeneity by dividing the study area into several sub-areas. Using a novel Gauss vector weight function approach, spatial autocorrelation variables are determined and selected as a part of the input features. Finally, the traditional algorithm of SVR is adopted to map the relationships of each local sub-area. The experiment data on $PM_{2.5}$ concentrations in Beijing are used to verify that the proposed method is superior to comparative methods, and the proposed method had high prediction accuracy and reliability. The main reason for this is that STSVR can address spatial heterogeneity, spatial autocorrelation, non-linearity, and external regressors simultaneously; in contrast, the comparative methods (i.e., ARIMAX, global SVR, STANNs, and local STANNs) address only some of these characteristics. The performance comparisons of the different methods are shown in Table 3.

[Insert Table 3 about here]

In fact, spatial heterogeneity can be classified into spatial local heterogeneity and spatial stratified heterogeneity (Wang et al., 2016). STSVR can mainly be used to address spatial stratified heterogeneity, which means that the relationships between air pollutant concentrations and other relevant variables change over spatial areas, but they are uniform within the same area. However, spatial local heterogeneity refers to relationships that change across spatial locations. That is, STSVR cannot address spatial local heterogeneity well. Meanwhile, we make an implicit assumption that the relationships between $PM_{2.5}$ concentrations and meteorological elements satisfy stationary conditions, which means the relationships will not change over time. Then, under this condition, we can employ statistical models to make predictions. It is obvious that the assumption will not be valid if the time span is long. For at least one year of data, it may be necessary to construct seasonal models or dynamic models, and an available strategy is to split the data into a short time-span series. In addition, STSVR cannot accurately predict abnormal or outlier patterns, such as pollution episodes (Zhang et al., 2012). However, abnormal patterns occur at low frequencies during this time period, and they are considered to result from an unknown or novel mechanism (Jiang et al., 2003). However, only a single model is used to fit all the samples of a sub-area. Because of the relatively small number of abnormal samples, this single model cannot describe the novel mechanism implicit in the abnormal patterns. Hence, it is difficult for STSVR to predict the abnormal patterns well.

Future studies should focus on improving the following aspects of the STSVR model: (1) explore the space-time clustering method to correctly identify space-time heterogeneity and not just spatial heterogeneity; and (2) develop a hybrid method to address extreme concentrations to solve the problems commonly encountered in empirically based approaches. Moreover, only five meteorological parameters were selected to predict the $PM_{2.5}$ concentrations, and this may explain why none of the accuracies of the included methods were very high. In the future, if we can collect other statistical data, including environmental data and human activity data, these additional features can be added to improve prediction results.

## References:


Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Modell. Softw. 20(10), 1263-1271.

Arhami, M., Kamali, N., Rajabi, M.M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ. Sci. Pollut. R. 20(7), 4777-4789.

Brockwell, P.J., Davis, R.A., 1996. Introduction to time series and forecasting. Springer-Verlag, New York.

Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos. Environ. 151, 1-11.

Cheng, T., Wang, J.Q., 2009. Accommodating spatial associations in DRNN for space–time analysis. Comput. Environ. Urban. 33(6), 409-418.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE T. Pattern Anal. 1(2), 224-227.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., et al., 2017. Air pollution and mortality in the Medicare population. New. Engl. J. Med. 376(26), 2513.

Diaz-Robles, L., Cortés, S., Vergara-Fernández, A., Ortega, J.C.,2015. Short-term health effects of particulate matter: a comparison between wood smoke and multi-source polluted urban areas in Chile. Aerosol. Air. Qual. Res. 15(1), 306-318.

Dominici, F., Peng, R.D., Bell, M.L., Pham, L., Mcdermott, A., Zeger, S.L., et al., 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Jama-J. Am. Med. Assoc. 295(10), 1127.

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D., 2009. $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. Expert. Syst. Appl. 36(5), 9046-9055.

Getis, A., Aldstadt, J., 2004. Constructing the spatial weights matrix using a local statistic. Geogr. Anal. 36(2), 147-163.

Ghazali, N.A., Ramli, N.A., Yahaya, A.S., Yusof, N.F.F.M., Sansuddin, N., Al Madhoun, W.A., 2010. Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. Environ. Monit. Assess. 165(1), 475-489.

Guo, D., Peuquet, D.J., Gahegan, M., 2003. ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. GeoInformatica. 7(3), 229-253.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42(33), 7561-7578.

Jiang, B., 2014. Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. Geojournal. 80(1), 1-13.

Jiang, S.Y., Li, Q. H., Li, K.L., Wang, H., Meng, Z.L., 2003. GLOF: a new approach for mining local outlier. In IEEE Proceeding of 2003 International Conference on Machine Learning and Cybernetics, 1, 157-162.

Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S., Özkaynak, H., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmos. Environ. 44(30), 3660-3668.

Kloog, I., Chudnovsky, A.A., Just, A.C., Nordio, F., Koutrakis, P., Coull, B.A., et al., 2014. A new hybrid spatio-temporal model for estimating daily multi-year $PM_{2.5}$ concentrations across northeastern USA using high resolution aerosol optical depth data. Atmos. Environ. 95(1), 581-590.

Kryszczuk, K., Hurley, P., 2010. Estimation of the number of clusters using multiple clustering validity indices. In Multiple Classifier Systems. Springer Berlin Heidelberg, 114-123.

Lee, H.J., Liu, Y., Coull, B. A., Schwartz, J., Koutrakis, P., 2011. A novel calibration approach of Modis AOD data to predict $PM_{2.5}$ concentrations. Atmos. Chem. Phys. 11(11), 9769-9795.

Liao, T.W., 2005. Clustering of time series data—a survey. Pattern Recogn. 38(11), 1857-1874.

Lv, B., Zhang, B., Bai, Y., 2016. A systematic analysis of $PM_{2.5}$, in Beijing and its sources from 2000 to 2012. Atmos. Environ. 124, 98-108.

Miller, H.J., Han, J.W., 2009. Geographic data mining and knowledge discovery. New York: CRC Press.

Nieto, P.G., Combarro, E.F., Del Coz Díaz, J.J., Montañés, E., 2013. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. Appl. Math. Comput. 219(17), 8923-8937.

Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., 2005. Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Modell. Softw. 20(5), 547-559.

Pereira, C.M.M., Mello R.F.D., 2011. Learning process behavior for fault detection. Int. J. Artif. Intell. T. 20(05), 969-980.

Robinson, D.P., Lloyd, C.D., McKinley, J.M., 2013. Increasing the accuracy of nitrogen dioxide ($NO_2$) pollution mapping using geographically weighted regression (GWR) and geostatistics. Int. J. Appl. Earth. Obs. 21, 374-383.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20(20), 53-65.

Sánchez, A.S., Nieto, P.G., Fernández, P.R., Del Coz Díaz, J.J., Iglesias-Rodríguez, F.J., 2011. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Math. Comput. Model. 54(5), 1453-1466.

Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W., Liu, S., 2013. Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in northern California. Sci. Total. Environ. 443(3), 93-103.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 234-240.

Vapnik, V., 2000. The nature of statistical learning theory. Springer Science & Business Media.

Wang, J.F., Zhang T.L, Fu, B.J., 2016. A measure of spatial stratified heterogeneity. Ecol. Indic. 67, 250-256.

Wang, R., Henderson, S.B., Sbihi, H., Allen, R.W., Brauer, M., 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmos. Environ. 64(1), 312-319.

Wasserman, L., 2004. All of statistics. Springer New York.

Wayland, R.A., White, J.E., Dye, T.S., 2002. Communicating real-time and forecasted air quality to the public. Environ. Manage. 30(6), 28-36.

Zhang, Y. L., Cao, F., 2015. Fine particulate matter ($PM_{2.5}$) in china at a city level. Sci. Rep. 5, 14884.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: History, techniques, and current status. Atmos. Environ. 60(32), 632-655.

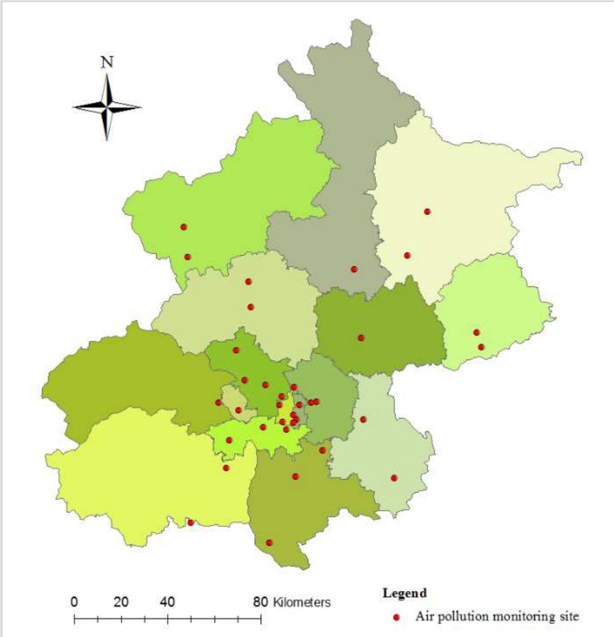
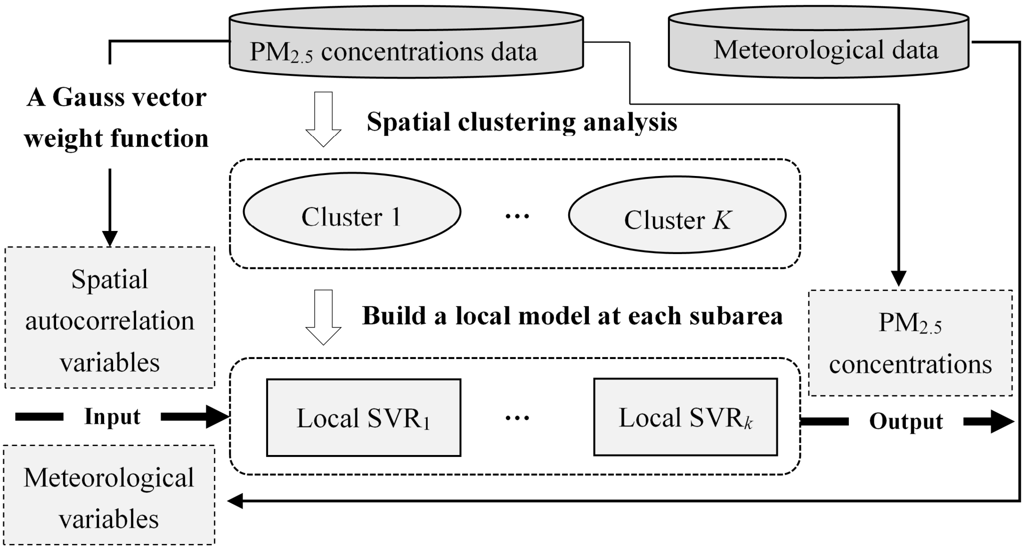



**Fig. 1.** Spatial distribution of air quality monitoring stations in Beijing

450




451

452 **Fig. 2.** The general framework of STSVR

453

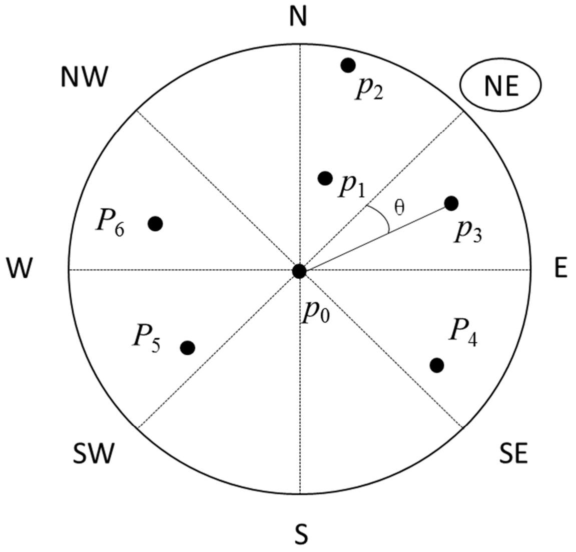



**Fig. 3**. An illustration of the spatial dependence of air pollutant concentrations

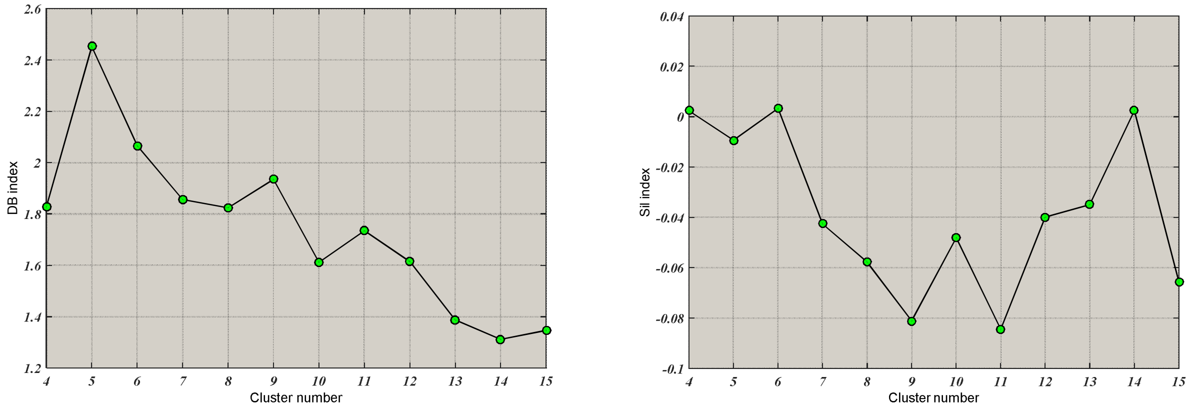
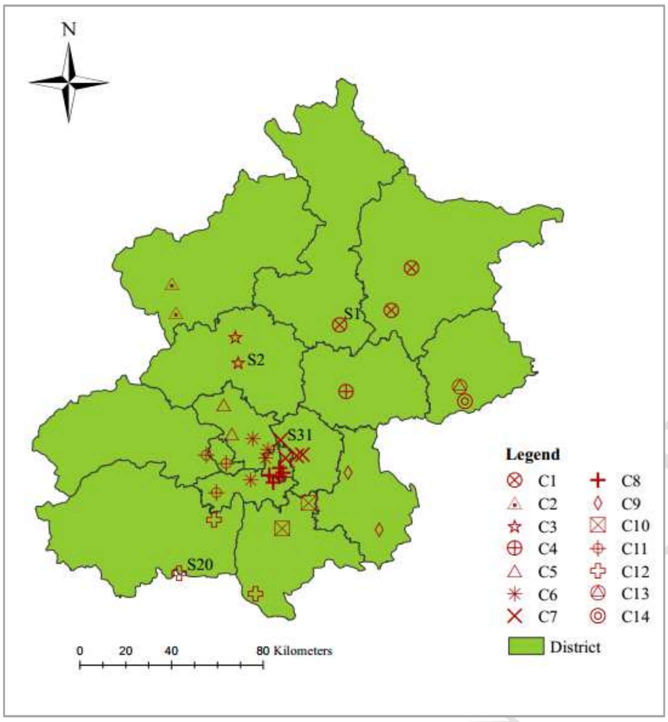



**Fig. 4.** The DB index and SI index varied with different cluster numbers

460




461

462 **Fig. 5.** The clustering results from the $PM_{2.5}$ concentration data

463

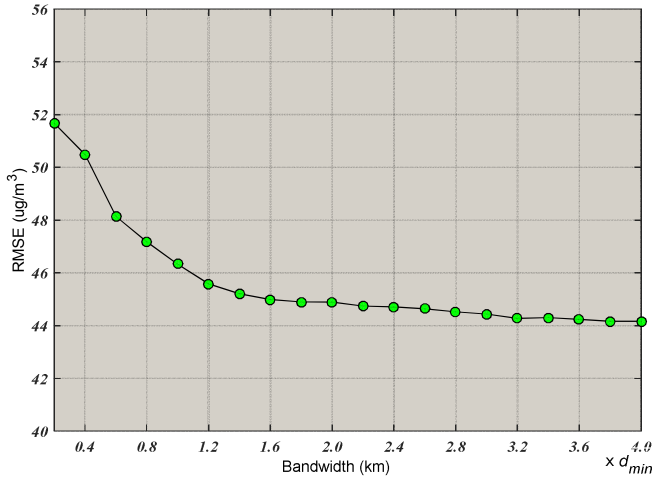



**Fig. 6.** The RMSEs of different bandwidth parameters

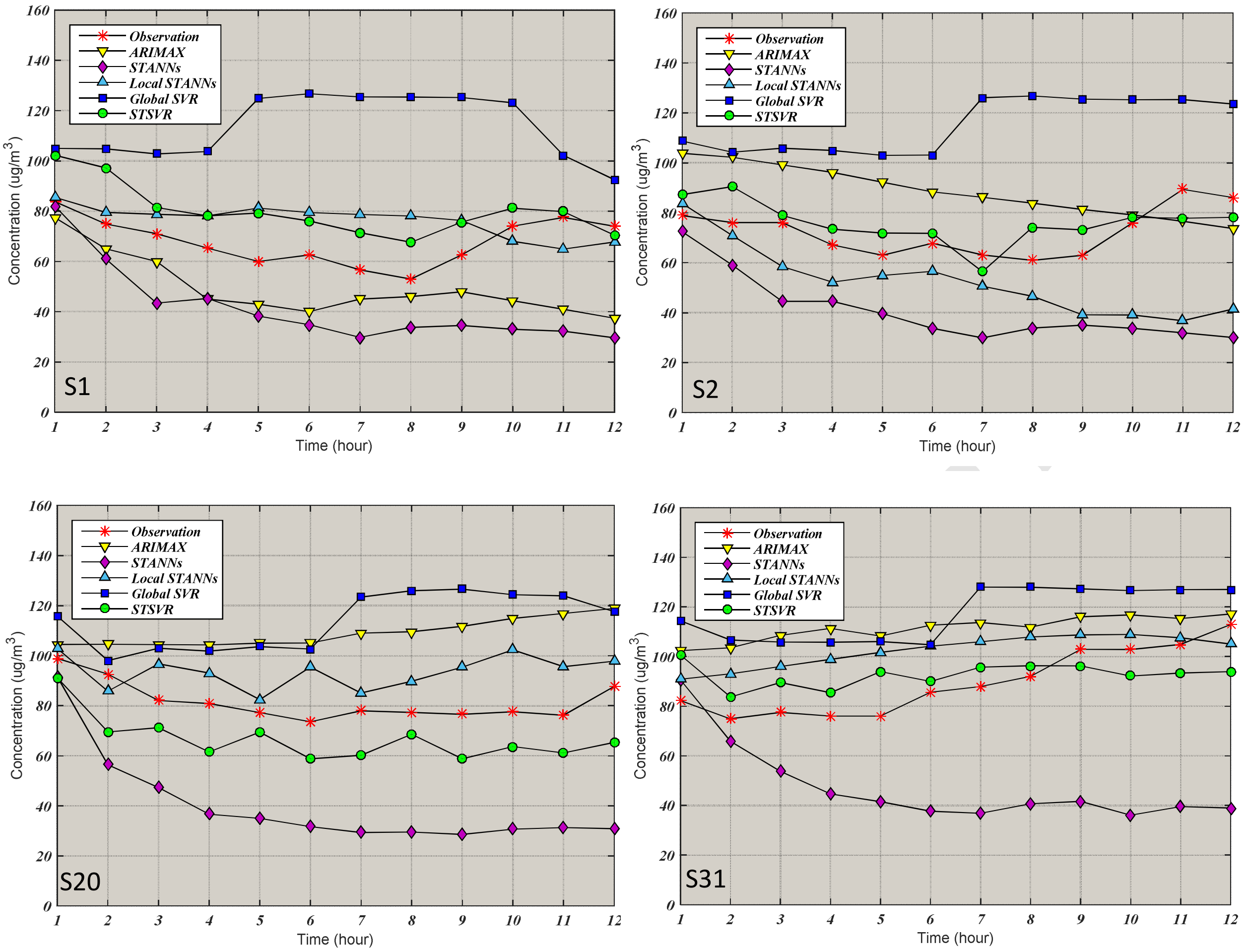



**Fig. 7.** The curves of different methods for the following 1-12 hours

**Table 1** Comparison of different methods applied to the entire area

| Time | | 1-6 hour | | 7-12 hour | | 13-24 hour | |
|---|---|---|---|---|---|---|---|
| Methods | Training sample sizes | ***p*** | ***e*** (ug/m$^3$) | ***p*** | ***e*** (ug/m$^3$) | ***p*** | ***e*** (ug/m$^3$) |
| ARIMAX | 1440×1 | 0.681 | 26.64 | 0.667 | 34.20 | 0.444 | 72.09 |
| STANNs | 1440×35 | 0.625 | 33.09 | 0.395 | 63.49 | 0.409 | 76.88 |
| Local STANNs | 1440×$Ns(C)$ | 0.731 | 21.61 | 0.687 | 32.18 | 0.437 | 72.91 |
| Global SVM | 1440×35 | 0.691 | 24.63 | 0.675 | 29.95 | 0.667 | 44.51 |
| STSVR | 1440×$Ns(C)$ | 0.769 | 19.76 | 0.703 | 31.81 | 0.594 | 53.79 |

**Table 2** Comparisons among different clusters or sub-areas using the STSVR model

| Time | | 1-6 hour | | 7-12 hour | | 13-24 hour | |
|---|---|---|---|---|---|---|---|
| Clusters | *Ns*(*C*) | ***p*** | ***e*** ($ug/m^3$) | ***p*** | ***e*** ($ug/m^3$) | ***p*** | ***e*** ($ug/m^3$) |
| C1 | 3 | 0.832 | 11.79 | 0.871 | 8.93 | 0.746 | 29.78 |
| C2 | 2 | 0.781 | 18.46 | 0.710 | 24.73 | 0.490 | 57.36 |
| C3 | 2 | 0.860 | 13.21 | 0.750 | 23.88 | 0.568 | 53.93 |
| C4 | 1 | 0.757 | 21.54 | 0.872 | 11.63 | 0.638 | 41.60 |
| C5 | 2 | 0.727 | 36.32 | 0.523 | 70.15 | 0.424 | 92.17 |
| C6 | 4 | 0.828 | 14.08 | 0.658 | 33.11 | 0.480 | 71.89 |
| C7 | 4 | 0.873 | 10.86 | 0.531 | 55.52 | 0.430 | 80.44 |
| C8 | 5 | 0.825 | 14.55 | 0.840 | 17.88 | 0.714 | 39.22 |
| C9 | 2 | 0.759 | 23.45 | 0.573 | 55.37 | 0.420 | 75.62 |
| C10 | 2 | 0.732 | 22.47 | 0.770 | 25.51 | 0.665 | 44.29 |
| C11 | 3 | 0.564 | 24.91 | 0.688 | 26.86 | 0.817 | 21.58 |
| C12 | 3 | 0.742 | 24.68 | 0.667 | 46.74 | 0.615 | 55.76 |
| C13 | 1 | 0.433 | 32.14 | 0.347 | 51.52 | 0.655 | 42.65 |
| C14 | 1 | 0.691 | 23.38 | 0.686 | 23.35 | 0.709 | 31.23 |

479 **Table 3** The performance comparison of different methods

| Performance / Models | Heterogeneity | Autocorrelation （anisotropy） | Non-linearity | External regressors |
|---|---|---|---|---|
| ARIMAX | √ | × | × | √ |
| STANNs | × | × | √ | × |
| Local STANNs | √ | × | √ | × |
| Global SVR | × | × | √ | √ |
| STSVR | √ | √ | √ | √ |

The symbol √ (×) denotes that the model can (or cannot) address the corresponding characteristic. It is worth noting that the third column (i.e., autocorrelation) indicates whether the model can address anisotropy of the spreading process of air pollution.

480

481

# Highlights:

- Spatial clustering analysis was used to handle spatial heterogeneity among $PM_{2.5}$ data.
- A Gauss vector weight was presented to define spatial autocorrelation variables as so to accord with the transport process of air pollutants
- An extended support vector regression model was constructed by considering the spatial characteristics of the air pollutant data, namely spatial dependence and spatial heterogeneity.