Short time PM2.5 prediction model for Beijing-Tianjin-Hebei region based on Generalized Space Time Autoregressive (GSTAR)

Jinghao Hu¹, Shandong Wang¹* and Jian Mao²

¹ School of Earth Sciences and Engineering, Hohai University, Nanjing, Jiangsu, 210098, China
² School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing, Jiangsu, 210023, China

*Corresponding author’s e-mail: wsd@hhu.edu.cn (Shandong Wang)

Abstract. In recent years, with the rapid development of China's economy and urbanization, the irrational industrial structure, large energy consumption, and increased car ownership have made China's air pollution increasingly serious, especially in Beijing-Tianjin-Hebei, Yangtze River Delta, and the Pearl River Delta. As the primary pollutant affecting the air quality in the Beijing-Tianjin-Hebei region, the accurate prediction of PM2.5 concentration is of great significance for predicting heavy pollution weather, formulating the start-up mechanism of emergency plan and optimizing the production of enterprises. This study used the online hourly data of conventional pollutants from 80 national control stations in Beijing, Tianjin and Hebei from May 1, 2015 to May 31, 2019. Considering the influence of time and space, the GSTAR model of PM2.5 prediction is established. Comparing the prediction results of GSTAR model with ARMA and STAR, the validity of GSTAR model is verified by applying relevant accuracy evaluation indicators.

1. Introduction
Particulate matter refers to aerosol particles suspended in a solid or liquid state of gas. When the aerodynamic equivalent diameter of the particles is less than or equal to 2.5μm, it is defined as fine particles, ie PM2.5. Fine particulate matter has few components in the earth’s atmospheric composition, but it has caused great concern to the government and the public because of its serious harm to human health and the atmospheric environment. Epidemiological studies have confirmed that particulate matter can cause damage to the respiratory and cardiovascular systems, leading to lung cancer, cardiovascular disease, birth defects and even premature death[1]. In addition, PM2.5 will also have an impact on climate change, plant growth and ecosystems.

Currently, research methods for PM2.5 or other air pollutant predictions mainly include two categories[2]. Based on the physical model, a series of principles and mechanisms for the formation of haze are studied to carry out evolution and simulation to achieve the purpose of prediction. The other is based on numerical method, through the analysis of data, the corresponding mathematical model is established, and the fine prediction based on time and space is carried out. Compared with the physical model, the numerical value-based method has the characteristics of simplicity and high efficiency. Based on the statistics and analysis of historical data, the method obtains a reasonable prediction based on the current state. However, most PM2.5 concentration prediction models consider that spatial factors are
relatively small, and even neglect its role. The concentration of PM2.5 is often related to the spatial factors of its location. At the same time, there is a certain correlation between PM2.5 concentrations monitored by different monitoring base stations. Therefore, the prediction model's consideration of spatial factors is relatively necessary.

The classic method of simulating such spatially related data is the Vector Autoregressive (VAR) model. Although very flexible, the VAR model has too many unknown parameters and must be estimated only from a limited amount of data. The space-time autoregressive (STAR) model proposed by Cliff et al. in 1973 has been successfully applied in many fields of science, especially when there is a priori information about spatial dependence, but it requires the same parameter values for all locations. Borovkova et al. have introduced a more flexible class of models, the generalized space-time autoregressive model (GSTAR). Its model parameters allow each position to change. In this paper, the GSTAR model is used to establish the PM2.5 prediction model for 13 cities in the Beijing-Tianjin-Hebei region based on different spatial weight matrices, and the GSTAR model is compared with the ARMA model and the STAR model by using the accuracy evaluation index.

2. Methods

2.1. STAR Model
A model that explicitly considers spatial dependence is called a spatiotemporal model. A class of such models - the spatiotemporal autoregressive model (STAR) and the spatiotemporal autoregressive moving average model (STARMA) were introduced by Cliff and Ord in the 1970s. Similar to the VAR model, the STAR model is characterized by a linear correlation of space and time. The essential difference from the VAR model is that the model construction in the STAR model imposes spatial dependence through the weight matrix. Spatial features such as distance between locations and adjacent locations are combined in such a weight matrix. The STAR model can be written as:

\[ Z(t) = \sum_{s=1}^{p} \sum_{k=0}^{\lambda_k} \Phi_{sk} W^{(k)} Z(t-s) + e(t) \]  

\[ \lambda_s \quad \text{: spatial lag} \]
\[ Z(t) \quad \text{: (N×1) of observation vector at time t} \]
\[ Z(t-s) \quad \text{: (N×1) of observation vector at time t-s} \]
\[ \Phi_{sk} \quad \text{: STAR parameters} \]
\[ W^{(k)} \quad \text{: weight matrix in spatial lag } k \]
\[ e(t) \quad \text{: the white noise with mean vector 0 and variance-covariance matrix } \sigma^2 I \]

The STARMA model is an extension of the STAR model, including an innovative moving average. Pfeifer and Deutsch (1980a, b, c) performed a comprehensive analysis of the STARMA model and described a three-stage modeling process in the spirit of Box and Jenkins (1976).

2.2. GSTAR Model
The generalized space-time autoregressive model (GSTAR) is a more flexible space-time model proposed by Borovkova et al. The GSTAR model is a specific form of the VAR (Vector Autoregressive) model. It reveals the linear dependence of space and time. The main difference between the GSTAR model and the VAR model is spatial dependence. In the GSTAR model, it is represented by a weight matrix. The STAR model also considers spatial dependencies, but it requires the same parameter values of all locations. The GSTAR model allows model parameters to vary from spatial location to location, and is more relevant than traditional STAR models.

The GSTAR model can be written as:

\[ Z(t) = \sum_{s=1}^{p} \sum_{k=0}^{\lambda_k} \Phi_{sk} W^{(k)} Z(t-s) + e(t) \]
\[ Z(t) = \sum_{k=1}^{p} [\Phi_{k0} + \Phi_{k1} W] Z(t - k) + e(t) \]

\[ Z(t) \]: (N×1) of observation vector at time t
\[ Z(t - k) \]: (N×1) of observation vector at time t-s
\[ \Phi_{k0} \]: the diagonal matrices with the diagonal elements as autoregressive and the space time for each location (\( \phi_{k01}, \ldots, \phi_{k0N} \))
\[ \Phi_{k1} \]: the diagonal matrices with the diagonal elements as autoregressive and the space time for each location (\( \phi_{k11}, \ldots, \phi_{k1N} \))
\[ W \]: weight matrix in spatial lag k
\[ e(t) \]: the white noise with mean vector 0 and

Ordinary Least Squares (OLS) is a general method for estimating generalized spatiotemporal autoregressive (GSTAR) parameters. But in some cases, the residuals of GSTAR are location-dependent. If OLS is applied to this case, the estimator is inefficient. In past studies, Kalman filters have been used for the estimation of ARIMA and STARMA model parameters. The results show that the estimation effect using the Kalman filter is significantly better than OLS. In this paper, the Kalman filter estimator is used to estimate the GSTAR parameters.

2.3. Determination of Space Weight
The selection and determination of the spatial weight matrix is a key issue in the establishment of the GSTAR model. There are many methods for defining the weight of the spatial weight matrix. But whether it is a simple variable definition or a complex function definition, its original intention is to truly and comprehensively reflect the spatial correlation among spatial geographic units. Borovkova gives the spatial weight matrices commonly used in the GSTAR model: binary weight matrix, weight matrix based on inverse of distance, and weight matrix based on cross-correlation reasoning.

3. Case Study
3.1. Data description

![Figure 1. Research area](image)
The Beijing-Tianjin-Hebei region is located between 36°01′~ 42°37′N, 113°04′~ 119°53′E, east of Bohai Sea, west of Taihang Mountain, north of Yanshan Mountain, northwest high and southeast low. The area is mainly including Beijing, Tianjin and Hebei Province. The Beijing-Tianjin-Hebei region is China's political, economic and cultural center. The rapid development of the regional economy, combined with the topographical conditions on three sides, makes it a stable geography and meteorological environment with high air pollution and relatively dense urban air quality monitoring points. It is one of the typical urban agglomeration areas. According to statistics, 11 out of 13 cities rank among the top 20 most polluted countries in the country. The air quality in some cities and the number of pollution days accounted for 40% of the total number of days in the year. In 2016, the average number of air days in 13 cities in the Beijing-Tianjin-Hebei region was only 56.8%. Among the exceeded days, the number of days with PM2.5 and PM10 as the primary pollutants was the highest. The short-term PM2.5 forecast for the Beijing-Tianjin-Hebei region is typical. In this paper, the PM2.5 hour concentration model of the region from May 1, 2019 to May 30, 2019 was used to predict the PM2.5 concentration on May 31, 2019, and compare it with the actual concentration. The figure below shows PM2.5 hour concentration monitoring data from May 1 to May 30, 2019 in Beijing-Tianjin-Hebei region.

![PM2.5 Concentration in Beijing-Tianjin-Hebei Region](image)

**3.2. The GSTAR Model for PM2.5 Prediction**

From the analysis, the space-time autocorrelation coefficient is decremented and smeared, and the partial correlation coefficient tends to zero after the 1st and 1st order spatial delays and the 1st time delay, then it can be preliminarily judged that the sample is the 1st time delay and the 1st order space. In the delayed autocorrelation process, the selected model is the space-time autocorrelation model GSTAR(1), and its specific expression is:

\[ Z(t) = [\phi_{10} + \phi_{11} W] Z(t-1) + e(t) \]

**3.3. GSTAR Model with Binary Weight Matrix**

A binary weight matrix is established and the Kalman filter is used for parameter estimation. The GSTAR model parameters of 13 cities are shown in the following table:

| City          | Phi10  | Std.Error | P.Value | Phi11  | Std.Error | P.Value |
|---------------|--------|-----------|---------|--------|-----------|---------|
| Beijing       | 0.24479 | 0.15101   | 0.03208 | 0.28302 | 0.16187   | 0.01775 |
| Tianjin       | 0.08534 | 0.27225   | 0.00202 | 0.40213 | 0.14344   | 0.01693 |
| Baoding       | 0.13277 | 0.33535   | 0.03413 | 0.56129 | 0.33396   | 0.0174  |
| Tangshan      | 0.64291 | 0.21722   | 0.03396 | 0.37456 | 0.12436   | 0.03441 |
| Langfang      | 0.21459 | 0.265     | 0.02022 | 0.12312 | 0.32657   | 0.00762 |
| Shijiazhuang  | 0.13444 | 0.27383   | 0.00727 | 0.29863 | 0.29877   | 0.01335 |
The table is the estimated value of the equation parameters and the significance test is performed. The P value indicates the significance level of the t test. It can be seen that the P value is less than 0.05. The assumption that the coefficient is irrelevant to the dependent variable indicates that the independent variable can explain the dependent variable.

3.4. GSTAR Model with Inverse Distance Weight Matrix
An inverse distance weight matrix is established and the Kalman filter is used for parameter estimation. The GSTAR model parameters of 13 cities are shown in the following table:

| City        | Phi10  | Std.Error | P.Value | Phi11  | Std.Error | P.Value |
|-------------|--------|-----------|---------|--------|-----------|---------|
| Beijing     | 0.38532| 0.21587   | 0.02478 | 0.19657| 0.20888   | 0.02364 |
| Tianjin     | 0.20778| 0.33315   | 0.01826 | 0.32707| 0.2201    | 0.02059 |
| Baoding     | 0.35518| 0.27309   | 0.03296 | 0.47289| 0.1425    | 0.00867 |
| Tangshan    | 0.10455| 0.1642    | 0.00652 | 0.213  | 0.26217   | 0.00688 |
| Langfang    | 0.5888 | 0.25159   | 0.00761 | 0.56606| 0.29969   | 0.02022 |
| Shijiazhuang| 0.55501| 0.12469   | 0.03512 | 0.59234| 0.16838   | 0.01303 |
| Handan      | 0.30185| 0.12148   | 0.03469 | 0.15531| 0.3116    | 0.0239  |
| Qinhuangdao | 0.29264| 0.18885   | 0.01214 | 0.57517| 0.15276   | 0.01038 |
| Zhangjiakou | 0.44358| 0.13446   | 0.01482 | 0.53903| 0.33456   | 0.02385 |
| Chengde     | 0.44461| 0.12096   | 0.03178 | 0.59663| 0.12326   | 0.02153 |
| Cangzhou    | 0.55385| 0.25348   | 0.01936 | 0.36476| 0.15352   | 0.0317  |
| Xingtai     | 0.11842| 0.33883   | 0.029   | 0.47497| 0.11964   | 0.03448 |
| Hengshui    | 0.29072| 0.22871   | 0.02638 | 0.10706| 0.14946   | 0.01236 |

The P value is less than 0.05. The assumption that the coefficient is irrelevant to the dependent variable indicates that the independent variable can explain the dependent variable.

3.5. Diagnostic Checking Model
The model is tested. The partial correlation and autocorrelation of the residuals are significantly 0, indicating that there is no significant autocorrelation between the model residuals in time and space. The residual sequence is close to the random error, that is, the selected GSTAR model. The experimental data can be well explained, and the model has a good fitting effect.

3.6. The Accuracy of Forecasting GSTAR Model
This paper predicts the real value of PM2.5 concentration in the air, which is a regression problem. Therefore, the evaluation index uses the root mean square error (RMSE) and the average absolute error (MAE).
Table 3. Accuracy index

| Description            | Equation |
|------------------------|----------|
| MAE                    | \( \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \) |
| RMSE                   | \( \text{RMSE} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \) |

The GSTAR-Binary and GSTAR-Distance constructed in this paper are compared with the traditional ARMA and STAR models by evaluation indicators RMSE and MAE. The RMSE and MAE of the GSTAR-Distance model are 0.5722 and 2.0311, respectively. Compared with other models, the accuracy of the model is greatly improved. In summary, the improved GSTAR-Distance model established in this paper has a better prediction effect on PM2.5 concentration.

Table 4. Accuracy comparison

|                | RMSE    | MAE    |
|----------------|---------|--------|
| ARMA           | 0.7943  | 2.6353 |
| STAR           | 0.6321  | 2.4515 |
| GSTAR- Binary  | 0.6032  | 2.2572 |
| GSTAR- Distance| 0.5722  | 2.0311 |

4. Conclusion

Using the monitoring data to predict and analyze the urban air quality effectively can not only provide a good basis for people to travel, but also provide solid and powerful assistance for further taking corresponding measures to control pollution. Thus, the PM2.5 concentration prediction model is established to obtain better generalization ability and faster convergence speed. The accurate prediction of air quality such as PM2.5 concentration in Beijing-Tianjin-Hebei region provides a theoretical basis for the catastrophic impact of air pollution. It is of great significance for the integration of meteorological disciplines and information disciplines, and has broad application prospects.

References

[1] A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5[J]. Jun Ma, Yuexiong Ding, Jack C.P. Cheng, Feifeng Jiang, Zhiwei Wan. Journal of Cleaner Production.
[2] Recursive neural network model for analysis and forecast of PM10 and PM2.5[J]. Fabio Biancofiore, Marcella Busilacchio, Marco Verdecchia. Atmospheric Pollution Research.
[3] Calibration of VAR models with overlapping data[J]. Ralph Frankland, Andrew D. Smith, James Sharpe, Rishi Bhatia, Stuart Jarvis, Parit Jakhria, Gaurang Mehta. British Actuarial Journal.
[4] Weighted least absolute deviations estimation for periodic ARMA models[J]. Baoguo Pan, Min Chen, Yan Wang. Acta Mathematica Sinica, English Series. 2015(8)
[5] Martin R L, Oeppen J E. (1975) The identification of regional forecasting models using space-time correlation functions Transactions of the Institute of British Geographers, 66: 95-118.
[6] S. Borovkova, H. P. Lopuhaa and B. N. Ruchjana. Consistency and Asymptotic Normality of Least Squares Estimator in Generalized STAR Models. Statistica Neerlandica 62, nr.4, pp. 482-508 (2008).