Journal of Information and Computing Science
J. Info. Comput. Sci., Vol. 18, No. 2, pp. 117-128

DOI: 10.4208/JICS-2023-008
December 2023

Prediction of $PM_{2.5}$ Concentration in Beijing Based on Bayesian Hierarchical Autoregressive Spatio-Temporal Model

Jing Wang[1] and Chunzheng Cao[1,*]

[1] *School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Abstract. Here, a hierarchical autoregressive spatio-temporal model under the Bayesian framework is proposed to address the simultaneous multi-site $PM_{2.5}$ prediction. The true daily average concentration of $PM_{2.5}$ is regarded as a potential spatio-temporal process, then the temporal correlation is described by the first-order autoregressive process and the spatial correlation is captured based on the Matérn process, which greatly improves the efficiency in dimension reduction and synchronous prediction. In addition, meteorological factors such as daily maximum temperature, relative humidity and wind speed are used as explanatory variables to improve the prediction accuracy. The combination of Bayesian method and MCMC can realize parameter estimation and prediction process due to the model's hierarchical structure. The empirical analysis of daily $PM_{2.5}$ concentration in Beijing shows that the proposed model has good interpolation or prediction performance in both spatial and temporal dimensions.

AMS subject classifications: 62C10, 60J10

Key words: Bayesian method, Hierarchical model, Autoregressive, Spatio-temporal model, $PM_{2.5}$ prediction, Markov Chain Monte Carlo (MCMC).

1 Introduction

As one of the main air pollutants, $PM_{2.5}$, due to its small particle size, can be directly inhaled by the human body, and has a long residence time in the atmosphere and a long transportation distance, so it has a great impact on human health and atmospheric environmental quality. Medical studies have shown that too high concentration of $PM_{2.5}$

Translated from *Journal of Nanjing University of Information Science & Technology*, 2023, 15(1): 34-41.

[*] Corresponding author. *Email addresses:* **jwang_jane@163.com** (J. Wang), **caochunzheng@163.com** (C. Cao).

will not only lead to an increase in the incidence and mortality of cardiopulmonary diseases [1], but also affect the cardiovascular system, nervous system and immune system of the human body [2-3], and even have toxic effects on genetic materials at different levels such as chromosomes and DNA, causing cancer and birth defects [4-5].

Research on $PM_{2.5}$ includes data collection methods, mechanisms, causes and influencing factors [6-7]. From a statistical point of view, $PM_{2.5}$ concentration in a region over a period of time is regarded as a typical spatio-temporal data set, and relevant research focuses on spatial interpolation and short-term or long-term prediction in time. The space-time Kriging method [8-9] is a popular method for spatial interpolation of $PM_{2.5}$, which can realize linear and unbiased optimal estimation of unobserved locations based on the spatio-temporal position relationship and spatio-temporal variation characteristics of spatio-temporal data, while the prediction of $PM_{2.5}$ in time dimension can be made using mechanism analysis or statistical modeling methods. Mechanism analysis methods mainly model the physicochemical processes of the generation, conversion and diffusion of air pollutants, such as CMAQ model [10]. The statistical modeling method is to capture the characteristics of the data to obtain the change rule of pollutant concentration, including Multivariable Linear Regression (MLR) [11], Generalized Additive Model (GAM) [12-13], as well as various extension models of statistical learning models such as BP neural network [14] and Long Short-Term Memory (LSTM) [15-16]. Compared with mechanism analysis method, statistical method relies less on pollution source data, transmission mode and physical mechanism, and focuses more on the law of data itself. Quantitative analysis has more advantages in accuracy, and is a powerful tool for processing complex data.

In recent years, many studies have focused on the spatio-temporal characteristics and statistical inference of $PM_{2.5}$ concentration. For example, Cheam et al. [17] applied EM algorithm to the inference of parametric spatiotemporal mixed model to cluster air quality data. Based on the semi-parametric spatiotemporal model, Clifford et al. [18] use Gaussian Markov random field to approximate the spatial random effect and non-parametric time trend, and make Bayesian inference to predict the concentration of atmospheric particulate matter. These studies focus more on the flexibility of models and calculations and do not take into account the meteorological variables that play an important role in triggering air pollution. Some studies have also developed spatio-temporal models containing meteorological variables and applied them to spatio-temporal prediction [19-20]. For example, Wan et al. [21] conducted a comprehensive study on $PM_{2.5}$ concentration in Beijing by establishing a fine parametric statistical model, analyzed the spatio-temporal dependent structure of $PM_{2.5}$ concentration and made a prediction. However, when dealing with large-scale data, especially multi-site synchronous prediction, such spatio-temporal models will face excessive computational complexity.

In this paper, a Bayesian Hierarchical Autoregression (BHAR) spatio-temporal model was established for the average daily $PM_{2.5}$ concentration of 35 air quality monitoring points in Beijing, based on the Bayesian framework, stratified model theory