Multi-Scale Variation Prediction of $PM_{2.5}$ Concentration Based on a Monte Carlo Method

Chen Ding[1], Guizhi Wang[1, *] and Qi Liu[2]

Abstract: Haze concentration prediction, especially $PM_{2.5}$, has always been a significant focus of air quality research, which is necessary to start a deep study. Aimed at predicting the monthly average concentration of $PM_{2.5}$ in Beijing, a novel method based on Monte Carlo model is conducted. In order to fully exploit the value of $PM_{2.5}$ data, we take logarithmic processing of the original $PM_{2.5}$ data and propose two different scales of the daily concentration and the daily chain development speed of $PM_{2.5}$ respectively. The results show that these data are both approximately normal distribution. On the basis of the results, a Monte Carlo method can be applied to establish a probability model of normal distribution based on two different variables and random sampling numbers can also be generated by computer. Through a large number of simulation experiments, the average monthly concentration of $PM_{2.5}$ in Beijing and the general trend of $PM_{2.5}$ can be obtained. By comparing the errors between the real data and the predicted data, the Monte Carlo method is reliable in predicting the $PM_{2.5}$ monthly mean concentration in the area. This study also provides a feasible method that may be applied in other studies to predict other pollutants with large scale time series data.

Keywords: Monte Carlo method, random sampling, $PM_{2.5}$ concentration, chain development speed, trend prediction.

1 Introduction

With the rapid development of social economy and the improvement of people's living standards, a healthy and comfortable living environment has become the main pursuit goal of the public. However, air quality problems, especially haze problems, are the key factors that plague people's healthy life. According to satellite statistics, about 30% of the region, and nearly 800 million people in China suffer from smog damage like respiratory and cardiovascular diseases [Xie, Chen and Li (2014)], especially in Beijing and Shanghai. The frequency of fog in these regions is more than 50%, causing huge losses to the national economy [He, Wang, Wang et al. (2013)]. Therefore, predicting the concentration and trend of haze effectively, and rationally formulating prevention measures have far-reaching significance for people's health and social economic development.

Generally, the methods used to predict haze concentration at home and abroad mainly

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

[2] School of Computing, Edinburgh Napier University, Edinburgh, UK.

[*] Corresponding Author: Guizhi Wang. Email: wgz@nuist.edu.cn.

include meteorological methods [Saide, Carmichael, Spak et al. (2011)], statistical methods [Li, Bai, Shi et al. (2010)] and machine learning methods [Ordieres, Vergara, Capuz et al. (2005); Du, Lu and Dou (2017)]. In most of the previous studies, they mainly conducted short-term predictions of haze pollutant concentrations. Besides, some prediction models are a little complex and converge slowly, failing to mine potential information from data [Wang, Liu, Zhang et al. (2018)]. Therefore, motivated from these solutions, we propose an algorithm named Monte Carlo to predict the monthly average value of haze concentration. The Monte Carlo algorithm not only transforms some complex objects into the calculation of random numbers and digital features, but also is simple and easy to operate and has the characteristics of being able to visually explain the randomness of objects.

The rest of this paper is organized as follows. Section 2 mainly presents related work in haze prediction. Section 3 briefly introduces the related theories of Monte Carlo. Section 4 gives the modeling method of Monte Carlo. In Sections 5 and 6, the experimental results for prediction are reported. Finally, Section 7 concludes this paper.

2 Related work

In fact, the study of smog began very early. Limited by early detection equipment, Fuller et al. [Fuller, Carslaw and Lodge (2002)] designed an empirical model to predict the daily average concentrations of $PM_{2.5}$ and PM_{10} in London and the southeastern United Kingdom and also studied the relationship between various pollutants and $PM_{2.5}$ and PM_{10} and finally realized the daily average concentration of PM_{10}. Jian et al. [Jian, Zhao, Zhu et al. (2012)] proposed a statistical model, called Autoregressive Integrated Moving Average (ARIMA), whose results indicated that barometric pressure and wind velocity were anti-correlated and temperature and relative humidity were positively correlated with PM_{10} mass concentrations. Fu [Fu (2016)] gave an online update multivariate linear regression method and used meteorological elements as the influence factor of haze. Without the large amount of data, the model could be updated according to the results of the day, which improved the prediction accuracy. In recent years, some methods about machine learning and neural network have also been introduced for haze prediction by domestic and foreign scholars. Among them, Dong et al. [Dong, Yang, Kuang et al. (2009)] established a hidden Markov model to predict the high $PM_{2.5}$ concentration in haze weather and established the function mapping between parameters and variables. Then the trained hidden Markov model can better predict the high $PM_{2.5}$ concentration in the next 24 hours. Ai et al. [Ai and Shi (2015)] put forward a prediction model of BP artificial neural network and established a haze prediction system based on time series. The results showed that the model can accurately predict haze weather, but the convergence speed of the model is slow and easy to fall into local optimum situation. Ganesh et al. [Ganesh, Arulmozhivarman and Tatavarti (2018)] presented several ensemble models of neural network and regression predictors to predict the concentration of the $PM_{2.5}$ pollutant in Houston and New York on the basis of meteorological data and found that ensemble approach would perform promising prediction results compared to single model.

3 Monte Carlo method

3.1 Basic principle

The Monte Carlo Method is called the statistical simulation method. It is based on the law of large numbers in probability statistics and uses a random number (pseudo-random number) for numerical calculation [Yi, Guan, Zhang et al. (2002)]. The basic idea is to establish a probability model related to the problem and let the parameters of the model be equal to the results of the problem, then calculate the statistical characteristics of parameters by a large number of repeated sampling tests on the model, finally obtain the approximate value of the problem.

Combined with the characteristics of the Monte Carlo algorithm, the purpose of our study is not to accurately gain the concentration of a certain day in the future, but to make the predicted value of haze concentration still consistent with the regularity of historical data and the distribution characteristics.

3.2 Monte Carlo method theory

Set the sample of random variables $X_1, X_2, \cdots, X_n$, its arithmetic mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

According to the law of large numbers:

$$\lim P\{|\bar{X} - E(X)| < \varepsilon\} = 1 \tag{2}$$

Then when n is large enough, $\bar{X}$ converges to $E(X)$ according to probability.

If random variable $X_1, X_2, \cdots, X_n$ independently identically distributed with non-zero finite variance σ^2 , from the central limit theorem, we have:

$$\lim_{N\to\infty} P\left(\frac{\sqrt{N}}{\sigma}|\bar{X}_N - E(X)| < x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt = \Phi(x) \tag{3}$$

when N is sufficiently large, there is an approximate expression as follows:

$$P\left(|\bar{X}_N - E(X)| < \frac{u_\alpha \sigma}{\sqrt{N}}\right) \approx \frac{1}{\sqrt{2\pi}}\int_{0}^{u_\alpha} e^{-t^2/2}dt \tag{4}$$

Generally, the error formula for the Monte Carlo method can be written as:

$$\varepsilon = \frac{u_\alpha \sigma}{\sqrt{N}} \tag{5}$$

Therefore, reducing the error of the Monte Carlo method can generally be achieved by reducing the sample standard deviation σ and increasing the sample size N. In the actual situation, the sample size is often certain, so the error can be decreased by reducing σ in most cases.

3.3 Monte Carlo method application

In general, as long as we can establish a suitable probability model for our problem, the Monte Carlo method can almost be applied to almost any field, including deterministic mathematical problems and stochastic problems.

For the deterministic mathematical problems, according to the Monte Carlo method, firstly, construct a probability model related to the solution, so that the solution is the probability distribution or mathematical expectation of the model we build, and then generate random numbers according to the known probability distribution. Perform multiple random samplings, and finally use the arithmetic mean value as the approximated estimate.

For the stochastic problems, the numerical simulation method is widely adopted, that is, a random sampling can be conducted to obtain the statistical characteristics of the parameters, according to the probability distribution of the actual problem, and finally we can obtain an approximate solution with a specific expected value. Since the haze concentration can be affected by different factors such as temperature, wind levels, humidity, CO concentration and so on, the prediction of haze concentration in this paper is a stochastic problem.

4 Monte Carlo method modeling

4.1 General steps of the Monte Carlo method

(1) Collect historical haze data and process the sample data;

(2) Establish a suitable probability model $Y = f(X)$ for the random variable X ;

(3) Use a random number generator to generate N random numbers that are uniformly distributed between 0 and 1;

(4) According to the generated random numbers and the established probability model, the sampling is repeatedly simulated N times in the sample data in order to obtain a large number of sampling function values $Y_1, Y_2, \cdots, Y_n$;

(5) Make use of N sampling function values to achieve the sample mean value, and then calculate the statistical characteristics of the object.

The flow chart of Monte Carlo simulation general steps is shown in Fig. 1:

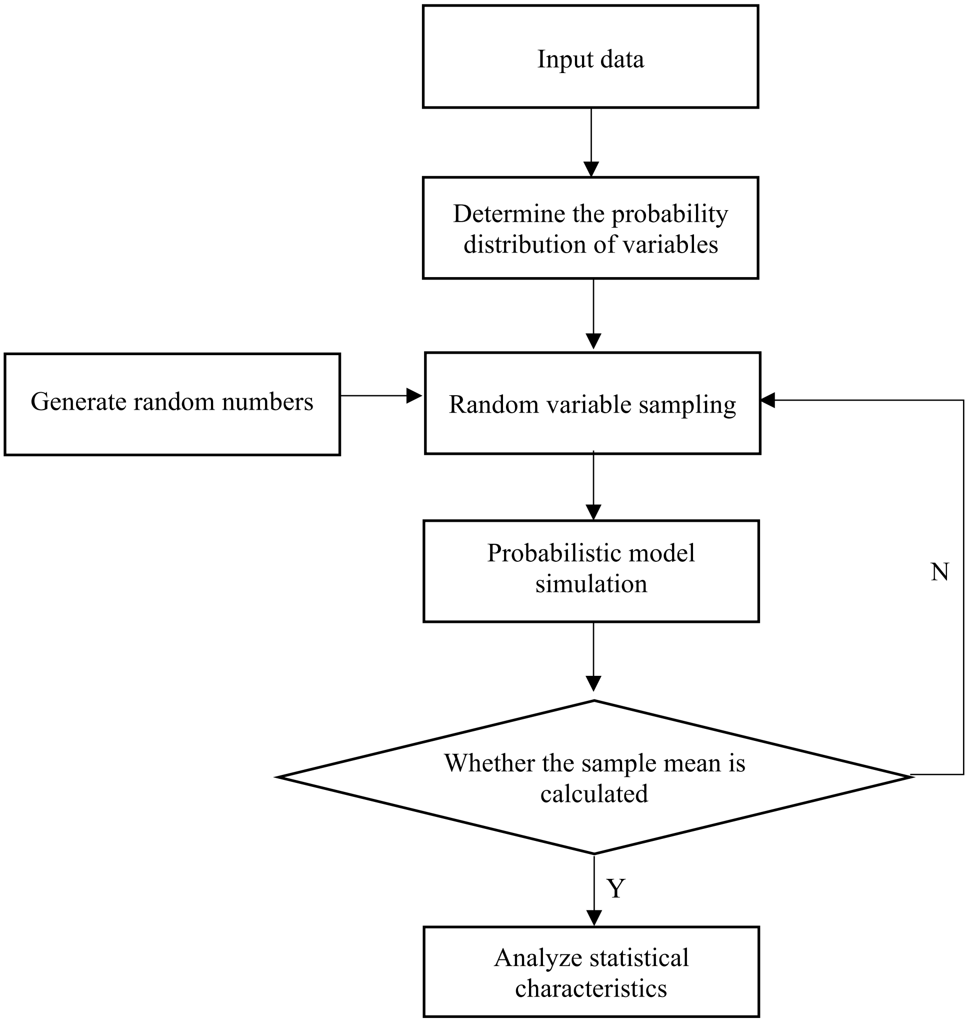
Input data
Determine the probability distribution of variables
Generate random numbers
Random variable sampling
Probabilistic model simulation
N
Whether the sample mean is calculated
Y
Analyze statistical characteristics

Figure 1: Monte Carlo simulation step flow chart

4.2 Probabilistic model establishment method

By consulting the literature, there are two commonly used methods for establishing a probabilistic model: The first one is the historical empirical method, that is to refer to the probabilistic model generally used by the predecessors when they review the related problems [Chi, Wang, Chen et al. (2015)]; The second is to use the statistical method to solve the probability model through historical data, usually including the parameter estimation method and the non-parametric estimation method. Given that there are few references on the probability model of haze concentration, it is quite difficult to achieve the probability model of the research object based on the experience of the predecessors and the non-parametric estimation does not need to rely on the prior knowledge of the relevant historical data. Therefore, this paper uses the kernel density estimation method in non-parametric estimation to estimate the probability density of the haze concentration data. The principle of kernel density estimation [Wu and Wang (1996)] is as follows:

Let $\mathrm{K}(\cdot)$ be a given Borel measurable function on R, $h_n > 0$ is a constant related to n, satisfying

$$\lim_{n\to\infty} h_n = 0 \tag{6}$$

Then, we have

$$\hat{f}_K(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) \tag{7}$$

where $\mathrm{K}(\cdot)$ is kernel function, h_n is bandwidth.

We can denote kernel density estimation $f(x)$ by $\hat{f}_K(x)$, that is, $\hat{f}_K(x)$ is kernel density estimation. Generally, we tend to select Gaussian kernel as our kernel function. The expression is as follows:

$$K(u) = \frac{1}{\sqrt{2\pi}} exp(-u^2/2) \tag{8}$$

Actually, under certain conditions of satisfying the kernel function, when the sample size n is large enough, the different choices of the kernel function are insensitive to the estimation results. The real impact on the performance of kernel estimation depends on the bandwidth h_n, because if h_n is too small, it will lead to the irregular shape of $\hat{f}_K(x)$ due to the increase of randomness. On the contrary, if h_n is too large, then $\hat{f}_K(x)$ will be overaveraged so that the properties of $f(x)$ cannot be well reflected.

According to the above principle, we use Python to draw a histogram of frequency distribution and perform kernel density estimation on the frequency histogram, and finally obtain a probability model of haze concentration.

4.3 Random number generation and sampling

There are many ways to generate random numbers, but most of them are pseudo-random numbers generated by mathematical recursive formulas. Although these pseudo-random numbers are not random numbers in the true sense, as long as their cycles are long enough, then they can reflect the randomness of sampling. In this way, we use the ‘uniform’ function in Python to generate a uniformly distributed random number r between 0 and 1, then divide the $\mathrm{PM}_{2.5}$ concentration interval into n parts, the probability corresponding to each interval i is p_i.

Set

$$p^k = \sum_{i=1}^{k} p_i, p_0 = 0 \tag{9}$$

then p^k represents the sum of the top k probability values.

Let

$$p_k = p^k - p^{k-1} \tag{10}$$

If the random number r falls on the interval $[p^k, p^{k-1})$, that means a $\mathrm{PM}_{2.5}$ concentration data is successfully extracted, and the corresponding probability is p_k. In this way, the random number r is matched with the $\mathrm{PM}_{2.5}$ concentration data one by one and we successfully complete one sampling.

5 Monte Carlo simulation based on PM$_{2.5}$ concentration

5.1 Data preprocessing

This paper mainly studies the distribution of haze concentration in Beijing. Because the composition of haze is very complicated and PM$_{2.5}$ is the most important component of haze, we choose Beijing PM$_{2.5}$ concentration as the research object. According to the PM$_{2.5}$ daily concentration data (https://www.zq12369.com/) that can be collected, we select Beijing PM$_{2.5}$ daily concentration data from January 1, 2014 to March 31, 2018 as our training set, the daily concentration data from April 1 to June 30, 2018 as test set. However, there are a few missing data in the training set collected in this way. For time series data such as haze, the concentration of haze today tends to be closely related to that of yesterday and tomorrow. Therefore, the single missing value for the training data can be filled by selecting the average of the two-day observations before and after the missing value. Fig. 2 shows a line chart of Beijing PM$_{2.5}$ daily concentration data:

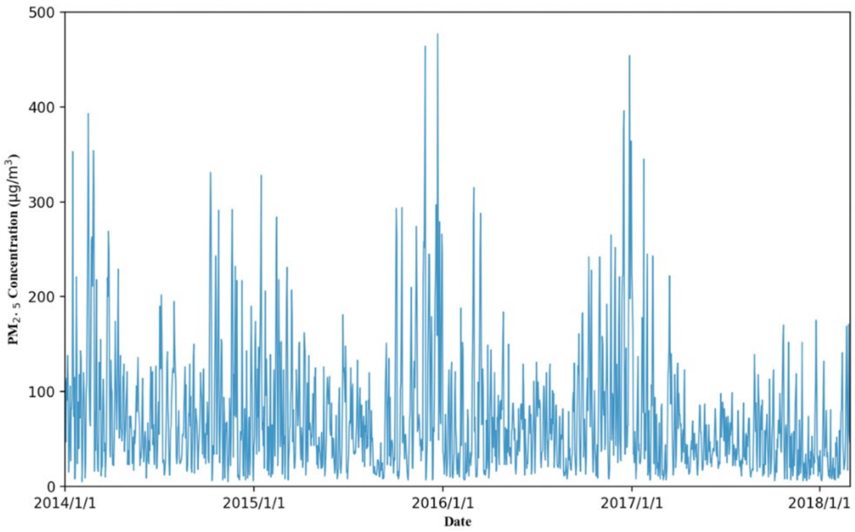

Figure 2: Beijing PM$_{2.5}$ daily concentration line chart

From this figure, we find that from January 2014 to the end of December 2016, there are many data with high peaks. Maybe people from all walks of life don't pay enough attention to the harm of haze and don't take effective preventive measures, which cause a large number of haze pollution in Beijing. In 2017, although Beijing PM$_{2.5}$ daily concentration peaks still have a slow upward trend, compared with previous years, the peak value has dropped sharply and the PM$_{2.5}$ concentration is basically stable between 0 and 200 $\mu g/m^3$. This phenomenon is closely related to the Chinese governments effective prevention measures against haze problems. The importance of the haze hazard has been continuously improved in all sectors of society and the frequency of high peaks of haze concentration has also been effectively reduced in the past year. From the overall perspective of historical data, we also find a rather regular phenomenon: Beijing PM$_{2.5}$ concentration shows a downward trend at the beginning of the year and gradually increased in the middle and late years, which has a certain reference value for predicting the concentration of $PM_{2.5}$.

Due to the large fluctuation of historical data, according to the error Eq. (5), in order to minimize the error caused by Monte Carlo prediction, we need to reduce the sample variance in the case of the same number of samples. Therefore, we can convert the historical data to logarithmic data and draw the result in Fig. 3.

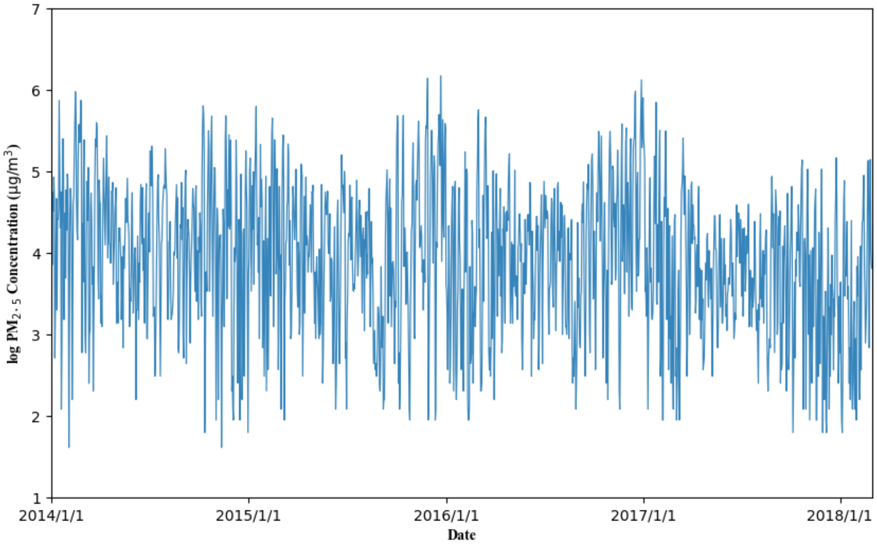

Figure 3: Beijing $PM_{2.5}$ daily logarithmic concentration line chart

According to Fig. 3, we could calculate the maximum $PM_{2.5}$ logarithmic concentration in Beijing as 6.17 $\mu g/m^3$, the minimum value is 1.61 $\mu g/m^3$, thus the range is 4.56 $\mu g/m^3$. Take 0.2 as the group distance, the total is divided into 24 groups, then the $PM_{2.5}$ logarithmic concentration interval can be divided. Then we can draw the histogram of $PM_{2.5}$ logarithmic concentration in Beijing, as is shown in Fig. 4.

According to the characteristics of the single peak and the symmetric distribution appearing in Fig. 4, we find that the $PM_{2.5}$ daily logarithmic concentration in Beijing is approximately normal distribution. Fig. 5 shows a comparison of the probability density function of the actual data with its normal distribution function with the same mean and standard deviation, and by comparing the theoretical probability with the actual data frequency value in Tab. 1, we can know that the degree of agreement between the two is higher. Meanwhile, $PM_{2.5}$ concentration is subject to short-term changes due to natural factors and human factors, so it is appropriate to use a normal distribution as the probability distribution of $PM_{2.5}$ daily logarithmic concentration in Beijing to a certain extent. Its mean value is 3.92 $\mu g/m^3$ and the standard deviation is 0.88 $\mu g/\mathrm{m}^3$. Further, it is found that a large amount of data is concentrated in the interval [3.6, 4.8]; that is, the actual value [36.6, 121.5], which shows that the $PM_{2.5}$ concentration in most of Beijing's time period is concentrated in the middle level, only a small amount of extreme values are in the picture.

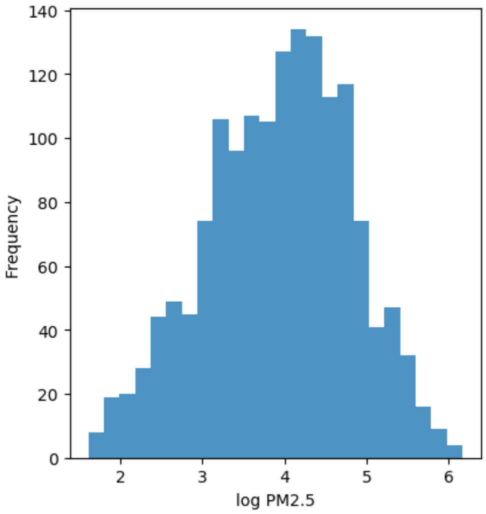

Figure 4: Beijing PM$_{2.5}$ daily logarithmic concentration histogram

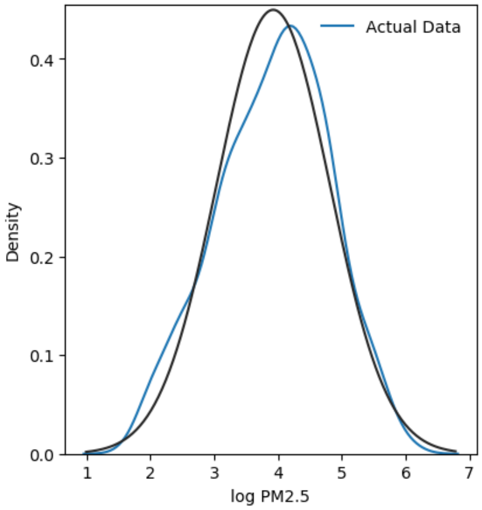

Figure 5: Comparison of PM$_{2.5}$ daily logarithmic concentration and normal distribution

Table 1: Normal distribution theory probability and actual frequency distribution table

Interval value	Theory probability	Actual frequency	Interval value	Theory probability	Actual frequency
1.8	0.46%	0.52%	4.2	9.38%	9.32%
2	0.74%	1.23%	4.4	8.67%	8.9%
2.2	1.22%	2.13%	4.6	7.96%	8.06%
2.4	1.89%	2.33%	4.8	7.89%	6.93%
2.6	2.8%	2.85%	5	5.56%	5.55%
2.8	3.91%	2.85%	5.2	3.43%	4.26%
3	5.19%	4.72%	5.4	2.91%	3.08%
3.2	6.49%	5.43%	5.6	2.39%	2.17%
3.4	7.72%	5.30%	5.8	1.1%	1.37%
3.6	8.68%	7.47%	6	0.58%	0.85%
3.8	9.26%	7.18%	6.2	0.19%	0.48%
4	9.32%	8.02%			

5.2 Random number generation and sampling

After data preprocessing, we can get a total of 1551 valid data from January 1, 2014 to March 31, 2018. In order to predict the monthly average haze concentration, we need to generate 30 pseudo-random numbers r_i $(1 \leq i \leq 30)$, which obeys uniform distribution. Then, the quantile p_k of the normal distribution with a mean value of 3.92 and a standard deviation of 0.88 corresponds to the random number r_i. Each r_i corresponds to a concentration value, so that the $PM_{2.5}$ daily logarithmic concentration is calculated in turn, that is, 30 samples are randomly extracted. As the program simulates a sufficient number of times, the sample data amount is gradually increasing. The mean of the simulation results gradually tends to a more stable value and this value is our predicted value.

5.3 Simulation prediction

This paper use Python software programming and set the number of simulations to 10000 times in order to obtain the predicted mean of $PM_{2.5}$ concentration next month in Beijing. Take April as an example: we can predict the average concentration in April 2018 based on the historical data from January 1, 2014 to March 31, 2018. In the same way, we also predict the concentration in May and June. What needs to be changed is to expand the scope of our historical data from March 31 to April 30 and May 31, 2018. Finally, we can obtain predicted values from April to June. Unit: $\mu g/m^3$ (Tab. 2).

Table 2: Mean value and relative error of Beijing $PM_{2.5}$ monthly concentration from April to June 2018

Month	Actual mean value	Predictive mean value	Relative error	Relative mean error
April	65	68.07	4.7%	3.6%
May	57	58.88	3.3%	
June	50	51.39	2.8%	

From the data comparison in Tab. 2, the mean relative error of the predicted values from April to June 2018 is calculated to be 0.036. Therefore, the Monte Carlo method to predict the average monthly concentration of $PM_{2.5}$ by virtue of the $PM_{2.5}$ concentration value has a certain reliability. However, in Fig. 8, we find that the $PM_{2.5}$ daily concentration predicted value has a small fluctuation range and can’t roughly reflect the daily variation trend of Beijing $PM_{2.5}$ concentration, which indicates that the $PM_{2.5}$ concentration as a random variable cannot effectively simulate the change of $PM_{2.5}$ concentration in the region although the $PM_{2.5}$ concentration may be predicted well.

6 Monte Carlo simulation based on the chain development speed

In order to further study the change trend of $PM_{2.5}$ concentration in a month, it is not significant to rely solely on the concentration data as variables to simulate. Therefore, we consider the development speed of Beijing $PM_{2.5}$ logarithmic concentration value as a new random variable. The chain development speed is studied here. Assume that the chain development speed is recorded as S_t, then we have

$$S_t = \frac{D_t}{D_{t-1}}, t = 1,2,\cdots,\mathrm{n} \tag{11}$$

where D_t and D_{t-1} indicates the $\mathrm{PM}_{2.5}$ concentration levels for the t and $t-1$ phases respectively.

6.1 Data processing

According to the Eq. (11), the $\mathrm{PM}_{2.5}$ daily logarithmic concentration data need to be converted into the $\mathrm{PM}_{2.5}$ daily logarithmic concentration chain development speed. The total distance is 0.1, which is divided into 15 groups. In this way, we draw the $\mathrm{PM}_{2.5}$ daily concentration chain development speed histogram figure (Fig. 6). The histogram in Fig. 6 is still a single peak, and symmetrically distributed left and right, what's more, the distribution is still approximately normal distribution. In Tab. 3, we can also find that the theoretical probability is very close to the frequency of the actual data. Therefore, it is appropriate to use a normal distribution to simulate the probability distribution of the $\mathrm{PM}_{2.5}$ daily concentration chain development speed. Similarly, we figure out that the mean value of the sample data at this time is $\mu = 1.024$ and the standard deviation is $\sigma = 0.232$.

Table 3: Normal distribution theory probability and actual frequency

Interval value	Theory probability	Actual frequency	Interval value	Theory probability	Actual frequency
0.4	0.35%	0.25%	1.0	16.25%	18.08%
0.5	0.84%	0.94%	1.1	16.97%	23.32%
0.6	2.17%	2.18%	1.2	14.78%	15.02%
0.7	4.74%	4.11%	1.3	10.7%	8.04%
0.8	8.58%	7.36%	1.4	6.45%	4.55%
0.9	12.93%	10.54%			

By comparing the chain development speed of $\mathrm{PM}_{2.5}$ logarithmic concentration with the normal distribution chart of the same mean and variance, Fig. 7 can be obtained. We can see that the sample data have higher peaks than the normal distribution curve, and the rest are almost coincident, indicating that the $\mathrm{PM}_{2.5}$ concentration has a significant jump, but it has long-term stability as a whole. Among them, the jump performance in the $\mathrm{PM}_{2.5}$ concentration may appear large fluctuations in the short term, and the stability performance in the $\mathrm{PM}_{2.5}$ concentration is maintained at a low level in the long term.

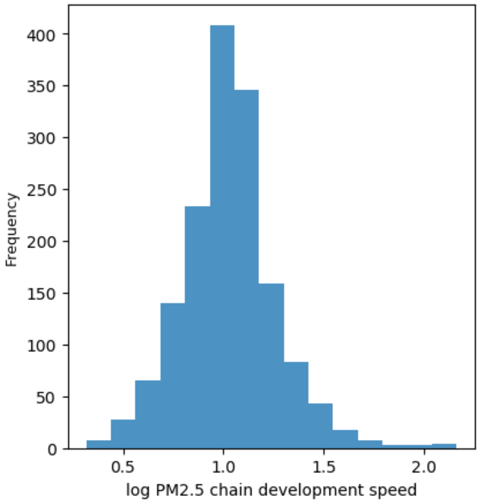

Figure 6: Beijing PM2.5 daily logarithmic concentration chain development speed histogram

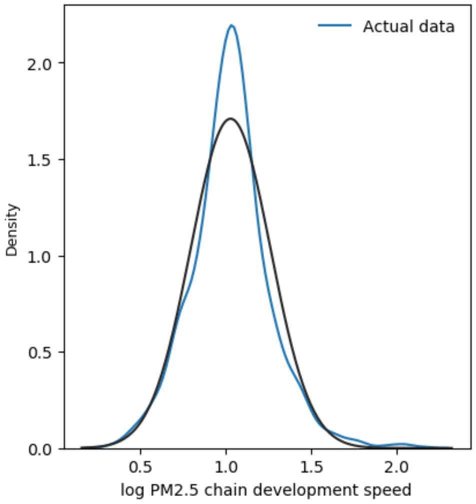

Figure 7: Comparison of PM2.5 daily logarithmic concentration chain development speed and normal distribution

In the above two figures, based on the mean value Zof the sample data $\mu = 1.024 > 1$, we can know that the chain development speed of PM2.5 daily logarithmic concentration is mainly concentrated in the interval $[1, 1.2]$, which shows that the daily concentration of PM2.5 in Beijing has a slight increase compared with the previous period. The PM2.5 concentration in Beijing has had a slow growth trend over the past period of time.

6.2 Random number generation and sampling

According to the probability distribution of the chain development speed of PM2.5 daily logarithmic concentration (subject to the mean distribution of $\mu = 1.024$, standard deviation $\sigma = 0.232$),

$$P(s \geq s_d) = \frac{1}{\sqrt{2\pi}\sigma}\int_{s_d}^{+\infty} e^{-\frac{(s-\mu)^2}{2\sigma^2}} dx \tag{12}$$

According to the Eq. (12), a random number r_i $(1 \leq i \leq 30)$ satisfying $(0, 1)$ uniform distribution can be generated by computer, then the quantile function of the normal distribution p_k corresponds to the random number r_i. In addition, we also need to calculate the chain development speed of PM2.5 daily logarithmic concentration s_d, then repeat the simulation N times, we can achieve the mean value of the chain development speed of the PM2.5 daily logarithmic concentration at different times in a month. Here take data in June as an example and $N = 10000$. Tab. 4 gives the partial chain development speed predicted values.

Table 4: PM2.5 daily logarithmic concentration chain development speed predicted value

Date	Predicted value	Date	Predicted value
6-1	1.039	6-17	1.000
6-3	1.018	6-19	1.007
6-5	1.002	6-21	1.025
6-7	1.019	6-23	1.036
6-9	1.006	6-25	0.992
6-11	0.985	6-27	1.031
6-13	1.015	6-29	1.017
6-15	1.031		

6.3 Simulation results

Assume that the logarithm of the PM2.5 concentration at the current time is D_0, the calculation formula of the haze concentration in the t phase from the current time is as follows:

$$D_t = D_0 \prod_{t=1}^{n} H_t, t = 1,2,\cdots,30 \tag{13}$$

Let the logarithm of PM2.5 concentration on May 31, 2018 be the starting value D_0. According to Eq. (13) and the data in Tab. 4, the PM2.5 daily logarithmic concentration in the next month can be calculated, then we can obtain the PM2.5 daily concentration predicted value in Tab. 5 by indexing it.

Table 5: PM2.5 daily concentration predicted value

Date	Predicted value	Date	Predicted value
6-1	41.23	6-17	56.96
6-3	44.52	6-19	58.71
6-5	45.90	6-21	65.78
6-7	48.65	6-23	75.52
6-9	50.00	6-25	80.69
6-11	46.72	6-27	80.20
6-13	49.83	6-29	75.27
6-15	56.99		

Based on the above predicted values and combined with Fig. 8, the average monthly concentration in June is 53.66 $\mu g/m^3$ and the relative error is 18.12%. Compared with the predicted value of 51.39 $\mu g/m^3$ calculated directly from PM2.5 concentration, we find that although the second method is slightly higher than before in the concentration prediction, the concentration of PM2.5 in Beijing shows a slow growth phenomenon in the period of June, and it is better to match the trend of the actual curve. Therefore, it is considered that the chain development speed of PM2.5 daily logarithmic concentration can be used as a random object to more comprehensively evaluate the $PM_{2.5}$ concentration change.

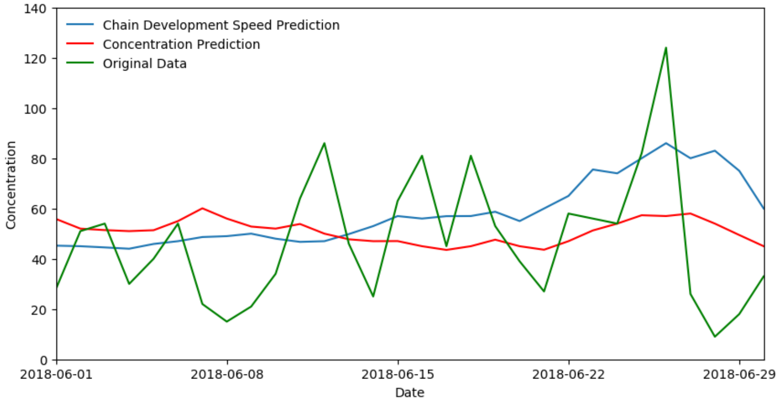

Figure 8: Comparison of two simulation methods

7 Conclusion

In order to solve the prediction of $PM_{2.5}$ concentration in Beijing, this paper adopts the Monte Carlo method, which is simple and easy to implement, to establish a normal distribution probability model based on $PM_{2.5}$ logarithmic concentration data and the chain development speed of $PM_{2.5}$ logarithmic concentration as two different random variables. Through a large number of random samplings, the mean value of $PM_{2.5}$ monthly concentration can be calculated, and the validity of the model is verified by comparing the relative error between the predicted value and the true value. The results show that based on historical effective data, Monte Carlo method can be used to predict the monthly average value of $PM_{2.5}$ concentration and the trend of $PM_{2.5}$ concentration change with certain reliability. Therefore, combined with Monte Carlo's broad applicability, as long as we can obtain valid data of $PM_{2.5}$ (also can be PM_{10}, SO_2 concentration, etc.) and calculate its corresponding probability model, then we can reasonably predict the mean and trend of $PM_{2.5}$ monthly concentration in this area.

Acknowledgements: This work was supported by Major Program of the National Social Science Fund of China [Grant No. 17ZDA092].

Declaration of interests: The authors declare no conflicts of interest.

References

Ai, H. F.; Shi, Y. (2015): Study on prediction of haze based on bp neural network. *Computer Simulation*, vol. 32, pp. 402-405.

Chi, D. C.; Wang, Z. H.; Chen, T. T.; Xu, X. J. et al. (2015): Application of monte carlo and arima models for precipitation forecast. *Journal of Shenyang Agricultural University*, vol. 46, pp. 187-191.

Du, E. L.; Lu, X. L.; Dou, L. L. (2017): Application of support vector machine in forecast of haze weather. *Inner Mongolia Science Technology & Economy*, vol. 2017, no. 17, pp. 57-58.

Dong, M.; Yang, D.; Kuang, Y.; He, D.; Erdal, S. et al. (2009): $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications*, vol. 36, no. 5, pp. 9046-9055.

Fu, Q. R. (2016): Research on haze prediction based on multivariate linear regression. *Computer Science*, vol. 43, pp. 526-528.

Fuller, G. W.; Carslaw, D. C.; Lodge, H. W. (2002): An empirical approach for the prediction of daily mean PM_{10} concentrations. *Atmospheric Environment*, vol. 36, no. 9, pp. 1431-1441.

Ganesh, S. S.; Arulmozhivarman, P.; Tatavarti, V. S. N. R. (2018): Prediction of $PM_{2.5}$ using an ensemble of artificial neural networks and regression models. *Journal of Ambient Intelligence & Humanized Computing*, vol. 2018, pp. 1-11.

He, H.; Wang, X. M.; Wang, Y. S.; Wang, Z. F.; Liu, J. G. et al. (2013): Formation mechanism and control strategies of haze in china. *Bulletin of Chinese Academy of Sciences*, vol. 3, no. 3, pp. 344-352.

Jian, L.; Zhao, Y.; Zhu, Y. P.; Zhang, M. B.; Bertolatti, D. (2012): An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science of the Total Environment*, vol. 426, pp. 336-345.

Li, W. F.; Bai, Z. P.; Shi, J. W.; Liu, A. X. (2010): Pollution characteristics and sources of fine particulate matter in ambient air in Tianjin city. *Research of Environmental Sciences*, vol. 23, no. 4, pp. 394-400.

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). *Environmental Modelling & Software*, vol. 20, no. 5, pp. 547-559.

Saide, P. E.; Carmichael, G. R.; Spak, S. N.; Gallardo, L.; Osses, A. E. et al. (2011): Forecasting urban PM_{10} and $PM_{2.5}$ pollution episodes in very stable nocturnal conditions and complex terrain using WRF-Chem CO tracer model. *Atmospheric Environment*, vol. 45, no. 16, pp. 2769-2780.

Xie, Y. B.; Chen, J.; Li, W. (2014): An assessment of $PM_{2.5}$ related health risks and impaired values of Beijing residents in a consecutive high-level exposure during heavy haze days. *Environmental Science*, vol. 35, pp. 1-8.

Yi, Z. Q.; Guan, J. F.; Zhang, X. H. (2002): The Monte Carlo Method and its application. *Physics and Engineering*, vol. 12, no. 3, pp. 45-49.

Wang, B. J.; Liu, P. Z.; Zhang, C.; Wang, J. M.; Chen, W. J. et al. (2018): Research on hybrid model of garlic short-term price forecasting based on big data. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 283-296.

Wu, X. Z.; Wang, Z. J. (1996): *Nonparametric Statistical Method*. Higher Education Press, China.