

Project Report

Prediction and Fitting for PM2.5 data in China

Course: **Data Mining**

Name: **王 玫** **May Wang**

Student Number: **5120309691**

Class: **F1203026**

Phone Number: **18818212442**

E-mail: **Mary1994@sjtu.edu.cn**

2015 - 7 - 6

Content

Abstract	3
Chapter1 Introduction	3
1.1 Seasonal Characteristics and Spatial Distribution	3
1.2 Chemical Elements and Influential Factors	4
1.3 Modeling the PM2.5 data	4
1.3.1 Aided incremental support vector regression	4
1.3.2 Deep Recurrent Neural Network (DRNN)	4
1.3.3 Multiple Regression and BP Neural Network	5
1.4 Motivation	5
Chapter2 Related Work	6
2.1 Markov chain Monte Carlo (MCMC)	6
2.2 Particle Filters / SMC	6
2.3 Stochastic Volatility Model (SVM)	6
2.4 Self-Similarity	6
Chapter3 Fitting and Predicting	7
3.1 Rao-Blackwellization Method	7
3.2 Stochastic Volatility Model	9
Chapter4 Spatial Distinction	10
Chapter5 Temporal Fractal	11
Chapter6 Conclusion	12
ACKNOWLEDGMENTS	12
Reference	12

Abstract

Recently, the air pollution condition is mostly reflected by the PM2.5 value in China. How to find a proper model for PM2.5 data is very important to analysis the characteristics both temporally and spatially for each city. According to the PM2.5 data from five different cities in China these years, we found that it is perfectly suit for sequential Monte Carlo method, which is also known as particle filters and usually serves for economics and finance. Based on this method, we make a good fitting modeling for the training data and one-step prediction. In addition, in order to make a better prediction, we added the environmental factors with Neutral Network method. Another approach is to find the self-similarity from the data during different time slot and predict the probable following values. The result of our modeling and fitting methods is relatively better from the testing data.

Keywords – fine particulate matter (PM2.5); Modeling Fitting; Time Series Prediction; Rao-Blackwellization Method; stochastic volatility model(SVM); Markov chain Monte Carlo(MCMC); Sequential Monte Carlo (SMC); Neutral Network(NN); Self-Similarity

Chapter1 Introduction

Particulate matter (PM), also known as aerosol, is one of the major pollutants that affect air quality. Exposure to particular matter with aerodynamic diameters less than 2.5 μm (PM2.5) can cause lung and respiratory diseases and even premature deaths [1]. Some research studies have been done on analyzing the spatial distribution, temporal characteristics, chemical elements, measure methods, influential factors as well as the time series prediction and machine learning approaches. However, more research need to be done to refine the accuracy or reliability of the model.

1.1 Seasonal Characteristics and Spatial Distribution

In paper [2], it shows that the surface of PM2.5 was obtained by using Inverse distance weighting (IDW). Mean center, Standard deviational ellipse, and G index were used to describe discrete tendency and central tendency and spatial autocorrelations from macro and micro perspectives. The results showed seasonal change influences the change of mean center location in Henan province. East and most northeast regions in China have always had severe pollution. The distribution of PM2.5 is high cluster pattern on the whole. Local areas show low clusters in spring and summer. From the view of spatial autocorrelation, we should take regional cooperation to treat and control PM2.5 pollution

From [3] the author analyses the mass concentration variations characteristics of PM10 and PM2.5 in Guangzhou (China). They found the main pollution type and greater than PM2.5-10 in Guangzhou; seasonal variation of the ratios are shown as followed: Winter (0.80) = Autumn (0.80) > Spring (0.76) > Summer (0.62). Spatially, the maximum value of the ratio (0.85) occurs in South

(Panyu) of Guangzhou, followed by Center (0.76), North (Conghua, 0.75) and Northwest (Huadu, 0.72) of Guangzhou orderly.

1.2 Chemical Elements and Influential Factors

Paper [4] gives us an analysis of the chemical elements and that way it can lead to find the pollution resource. The paper found that the mass levels of PM2.5 sampled in urban and suburban site showed a seasonal variation, with much higher values in winter and spring, lower values in summer, and the lowest in autumn. Total elements in Shanghai PM2.5 increased gradually from summer to winter and kept in a relative higher level in spring. Enrichment factor (EF) analysis results demonstrated that K, Na, Ca, Mg, Al, Fe, Ba, Sr originated from natural resources, while As, Cu, Zn, Pb, Cd, Mn, Ni, Se emitted by anthropogenic resources. Rotated Component Matrix analysis testified that there exists a strong relationship among Ca, Fe, Mn, Ba, suggesting these elements originated from crustal resources.

The research shows that the PM2.5 index of yesterday, precipitation of yesterday, average temperature, average humidity, precipitation and wind velocity. [5] The weather condition and environmental factors always influence the PM2.5 value as well as big human events, like Olympic Games or APEC etc.

1.3 Modeling the PM2.5 data

Typically, there are two kinds of approaches to model the data. One is the comprehensive method, which formulates the regression function composed with the sub models of its main factors from high-level understanding, and then adjusts the parameters of the functions. The other method is the inverse way, which is firstly modeling the data directly and blindly, like black box, and after that try to give some explanations for each feature. As we are not able to know how the PM2.5 is affected exactly so far, usually we choose the second approach of them.

1.3.1 Aided incremental support vector regression

In paper [6] they formulate mathematically a spatial data aided incremental support vector regression (SaIncSVR) for spatio-temporal PM2.5 prediction. We conduct spatio-temporal PM2.5 prediction over 13 monitoring stations in Auckland New Zealand, and compare the proposed SaIncSVR with a pure temporal IncSVR prediction.

1.3.2 Deep Recurrent Neural Network (DRNN)

In paper [7], they introduce a Deep Recurrent Neural Network (DRNN) that is trained using a novel auto encoder pre-training method especially designed for the task of time series prediction.

The numerical experiments show that our proposed pre-training method is superior than a canonical and a state-of-the-art auto-encoder training method when applied to time series prediction. On the specific case of fine particulate matter (PM2.5) forecasting in Japan, the experiments confirm that when compared against the PM2.5 prediction system VENUS employed by the Japanese Government,.

1.3.3 Multiple Regression and BP Neural Network

The empirical study [5] shows that compared to regression prediction model, BP neural network prediction model obtains the nonlinear relation among the prediction factors after training. The predictive model based on BP neural network is of higher predictive precision and better prediction effect which can be used in kinds of prediction research and has good value to popularize.

1.4 Motivation

In order to make a better modeling for PM2.5 forecast prediction, we find a better approach as a combination of Sequential Monte Carlo (SMC) and Neural Network (NN), as well as Self-Similarity method, forming a more comprehensive model and getting better prediction result.

Chapter2 Related Work

From paper [8], we know that complex models often lead to integrals that cannot be solved analytically. And in the book [9], there is a model named Stochastic Volatility Model (SVM), which is a nonlinear state-space model. These models are usually used in economics and finance, but we found it also suit for PM2.5 data characteristics.

2.1 Markov chain Monte Carlo (MCMC)

It has created an increase in the popularity of Bayesian methods that utilize Markov chain Monte Carlo (MCMC) algorithms. Sequential Monte Carlo (SMC) methods are alternative simulation-based algorithms for solving analytically intractable integrals. [8]

2.2 Particle Filters / SMC

In these methods, a (partially) continuous probability distribution is approximated by a discrete distribution made of weighted draws termed particles. From one iteration of the algorithm to the next, particles are updated to approximate one distribution after another by changing the particle's location on the support of the distribution and their weights. SMC methods include the particle filter, which generalizes the Kalman filter and hidden Markov model (HMM) filter to nonlinear, non-Gaussian state space models.

2.3 Stochastic Volatility Model (SVM)

The stochastic volatility model can be considered as a generation of the celebrated Black-Scholes formula (Hull and White 1987). Shephard and Pitt (1977) suggested a way to use Gibbs sampler to obtain Bayes estimates. Liu and Sabatti (2000) developed an improved MCMC algorithm. Chen, Qin and Liu (2000) gave promising results for using a HMC-based algorithm to compute the Bayes estimates in a SV model. Details of the SV model can be found in book [9].

2.4 Self-Similarity

A fractal is a natural phenomenon or a mathematical set that exhibits a repeating pattern that displays at every scale. Repeating pattern, in the other words, is self-similarity. Correlation Dimension changes sensitively with the change of attractor. Therefore, if a dynamic system works in different statuses, in other words, it has different movement trends, their attractors are different and their correlation dimensions are also different. Correlation dimension is a good way to analyze signals. But, except that, other fractal dimensions are useful in a lot of fields, like signal and image compression, computer vision and classification. In our work, we can use it in the PM2.5 data prediction

Chapter3 Fitting and Predicting

Some state space models have a special structure, where a subset of the state vector may be integrated out analytically. Analytical integration of part of the state vector reduces the Monte Carlo variation of the resulting estimator and will always improve its numerical efficiency.

3.1 Rao-Blackwellization Method

It is known as Rao-Blackwellization in the Monte Carlo literature because it is an implication of the Rao-Blackwell Theorem; e.g., see Robert and Casella (2004, p. 130). [8]

Algorithm 3 Rao-Blackwellized APF for the time-varying local level model

At $n = 0$, for $i = 1, \dots, N$

Draw $\mathbf{x}_{1,0}^{(i)}, P_{1,0|0}^{(i)} \sim p(\mathbf{x}_{1,0}), \mathbf{x}_{2,0}^{(i)} \sim p(\mathbf{x}_{2,0}), \mathbf{x}_{3,0}^{(i)} \sim p(\mathbf{x}_{3,0})$ and set $w_0^{(i)} = \frac{1}{N}$.

For $n = 1, \dots, T$:

(i) For $i = 1, \dots, N$, run the prediction step of the Kalman filter to obtain

the prediction errors and prediction error variances $\left\{ v_n^{(i)}, F_n^{(i)} \right\}_{i=1}^N$.

(ii) For $i = 1, \dots, N$, compute the importance weights $w_n^{(i)} \propto \mathcal{N}\left(v_n^{(i)}, F_n^{(i)}\right)$ and

normalize them: $\hat{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}$.

(iii) Resample N particles $\left\{ \mathbf{x}_{1,n-1|n-1}^{(i)}, P_{1,n-1|n-1}^{(i)}, \mathbf{x}_{2,n-1}^{(i)}, \mathbf{x}_{3,n-1}^{(i)} \right\}_{i=1}^N$ with probabilities

$\left\{ \hat{w}_n^{(i)} \right\}_{i=1}^N$ and for $i = 1, \dots, N$ set $w_n^{(i)} = \frac{1}{N}$.

(iv) For $i = 1, \dots, N$, draw $\mathbf{x}_{2,n}^{(i)} \sim \mathcal{N}(\mathbf{x}_{2,n-1}, 0.2)$ and $\mathbf{x}_{3,n}^{(i)} \sim \mathcal{N}(\mathbf{x}_{3,n-1}, 0.2)$ and

run the Kalman filter on each particle to obtain $\left\{ \mathbf{x}_{1,n|n}^{(i)}, P_{1,n|n}^{(i)} \right\}_{i=1}^N$.

$\mathbf{x}_n = (\mathbf{x}'_{1,n}, \mathbf{x}'_{2,n})'$: The state vector can be separated into two parts.

$p(\mathbf{x}_{1,n}, \mathbf{x}_{2,n} | \mathbf{y}_{1:n}; \theta) = p(\mathbf{x}_{1,n} | \mathbf{x}_{2,n}, \mathbf{y}_{1:n}; \theta) p(\mathbf{x}_{2,n} | \mathbf{y}_{1:n}; \theta)$: Marginal filtering distribution.

Particles are only simulated randomly from $p(\mathbf{x}_{2,n} | \mathbf{y}_{1:n}; \theta)$ while conditional on each individual

draw $\mathbf{x}_{2,n}^{(i)}$ the distribution $p(\mathbf{x}_{1,n} | \mathbf{x}_{2,n}^{(i)}, \mathbf{y}_{1:n}; \theta)$ can be evaluated analytically.

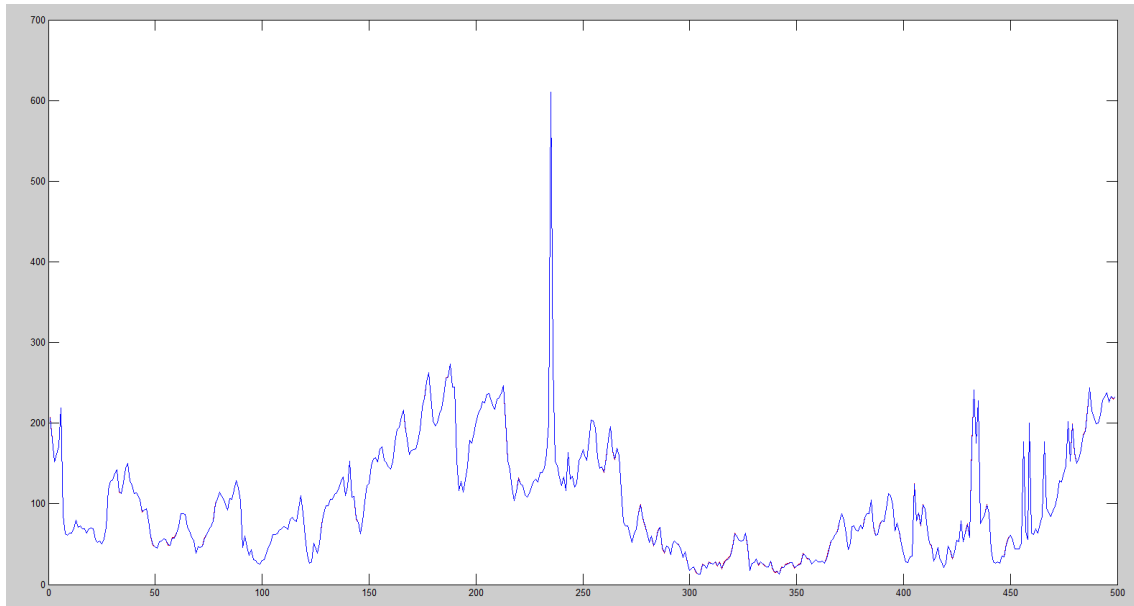
$$\mathbf{y}_n = Z_n(\mathbf{x}_{2,n}) \mathbf{x}_{1,n} + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, H_n(\mathbf{x}_{2,n})),$$

$$\mathbf{x}_{1,n} = T_n(\mathbf{x}_{2,n}) \mathbf{x}_{1,n-1} + \eta_n, \quad \eta_n \sim \mathcal{N}(0, Q_n(\mathbf{x}_{2,n})),$$

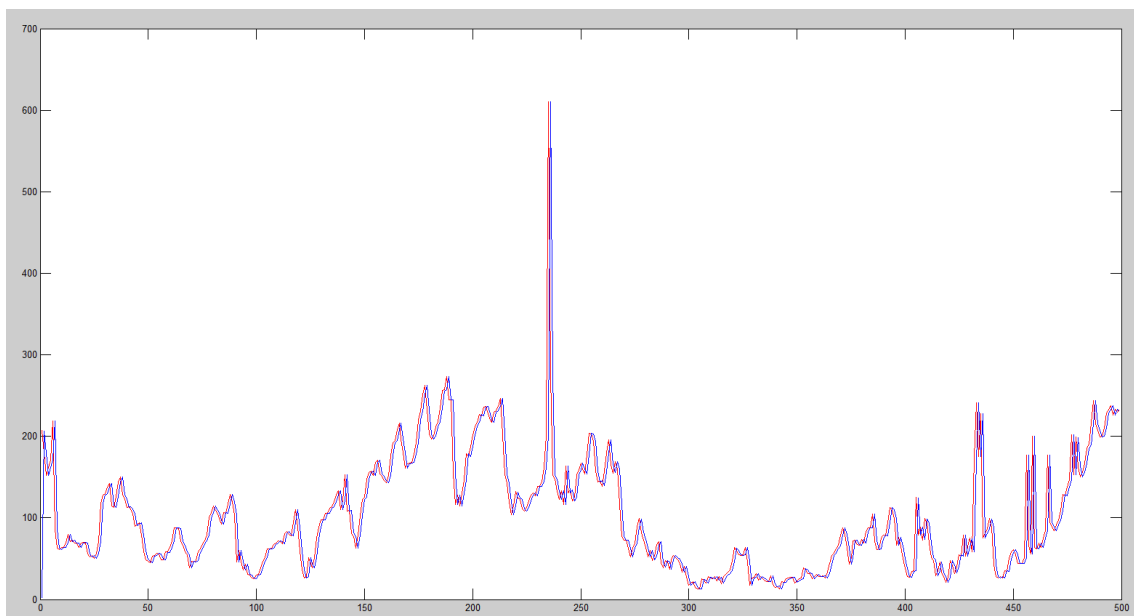
$$p_{ij} = p(\mathbf{x}_{2,n} = j | \mathbf{x}_{2,n-1} = i), \quad \mathbf{x}_{2,n} \in \{1, 2, \dots, k\},$$

Rao-Blackwellization is a linear, Gaussian state space model where the parameters in the state space matrices Z_n, T_n, Q_n, H_n depend upon the value of an additional discrete state variable $\mathbf{x}(2,n)$. The discrete state variables follow a first-order Markov process. These models are covered in Kim and Nelson (1999) and Frühwirth-Schnatter (2006).

We take the PM2.5 data of Beijing in spring, 2008 into the model. We can see that the fitting result is almost the same as the raw data. Here follows the fitting picture with Rao-Blackwellization method.



And also use this method, we can see the one-step prediction as following. The red curve is the true PM2.5 data and the blue one is the one-step prediction with Rao-Blackwellization method.



We can see from above that, the mechanism of Stock-Watson is that they only consider the influential factors of the nearest several preceding data and ignore the more previous influence. In this way, the prediction only for one step is typically similar to the previous one with normal noise containing the influence before.

3.2 Stochastic Volatility Model

SV model is a non-linear state model. There are three main parameters to be trained, as β , σ , ϕ . The β represents the log-variance, σ is the volatility of β , and ϕ is the persistence of the log-variance.

$$\Omega(\mathbf{x}) = \sum_{t=1}^T \left\{ \frac{\sigma}{x_t} + \frac{\sigma \beta_3 \text{exp}(x_t)}{\beta_3^2} + \frac{\sigma \alpha_3}{x_t^2 (1 - \phi_3)} \right\} + \sum_{t=1}^T \frac{\sigma \alpha_3}{(x_t + 1 - \phi_3) \beta_3}$$

ONCE THE BEST PARAMETER VALUES ARE FOUND, THE POSTERIOR LOG-VARIANCE IS

$$\Lambda = \frac{T + 10}{T} \left[0.2 + \beta_3^2 (1 - \phi_3) + \sum_{t=1}^T (x_t - \phi x_{t-1}) \beta_3 \right]$$

AND

$$\begin{aligned} \mu(\phi | \alpha_3, \mathbf{x}) &\propto \text{exp} \left\{ - \frac{\sigma \alpha_3}{\beta_3^2 (1 - \phi_3) + \sum_{t=1}^T (x_t - \phi x_{t-1}) \beta_3} \right\} (1 + \phi)_{10} (1 - \phi)^2 \\ \alpha_3 | \phi, \mathbf{x} &\sim \text{Gamma}(\mathcal{L}, \mathcal{L} + 10 \Lambda) \\ \beta_3 | \lambda, \mathbf{x} &\sim \text{Gamma}(\mathcal{L}, \mathcal{L} \sum_{t=1}^T \frac{\text{exp}(x_t)}{\beta_3^2}) \end{aligned}$$

The sampling algorithm:

Algorithm 1 Sequential Importance Sampling with Resampling (SISR)

At $n = 0$, for $i = 1, \dots, N$

Draw $\mathbf{x}_0^{(i)} \sim g_0(\mathbf{x}_0)$ and set $w_0^{(i)} = \frac{p(\mathbf{x}_0^{(i)})}{g_0(\mathbf{x}_0^{(i)})}$.

For $n = 1, \dots, T$:

(i) For $i = 1, \dots, N$ draw $\mathbf{x}_n^{(i)} \sim g_n(\mathbf{x}_n | \mathbf{x}_{n-1}^{(i)}, \mathbf{y}_n; \psi)$ and compute

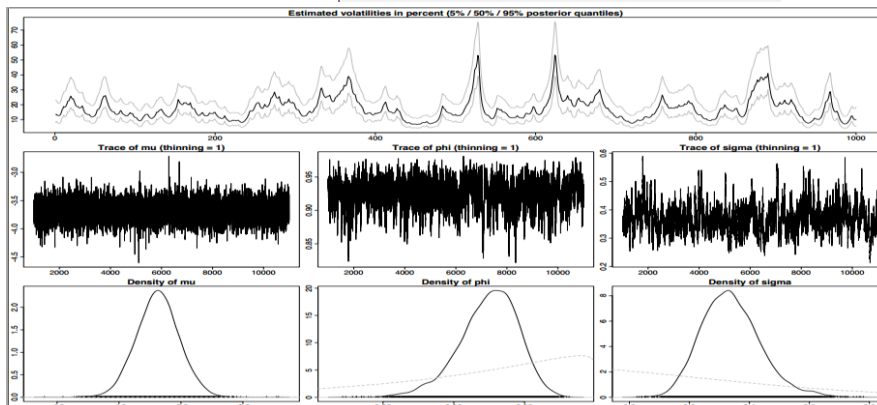
the importance weights $w_n^{(i)} \propto w_{n-1}^{(i)} \frac{p(\mathbf{y}_n | \mathbf{x}_n^{(i)}; \theta) p(\mathbf{x}_n^{(i)} | \mathbf{x}_{n-1}^{(i)}; \theta)}{g_n(\mathbf{x}_n^{(i)} | \mathbf{x}_{n-1}^{(i)}; \psi)}$.

(ii) For $i = 1, \dots, N$ normalize the importance weights: $\hat{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}$.

(iii) Resample N particles with probabilities $\{\hat{w}_n^{(i)}\}_{i=1}^N$ and for $i = 1, \dots, N$ set $w_n^{(i)} = \frac{1}{N}$.

After translating the training data to satisfy the SV model, we can get the result of it on PM2.5 data. The posterior distribution of three parameters in the aggregated results are like this:

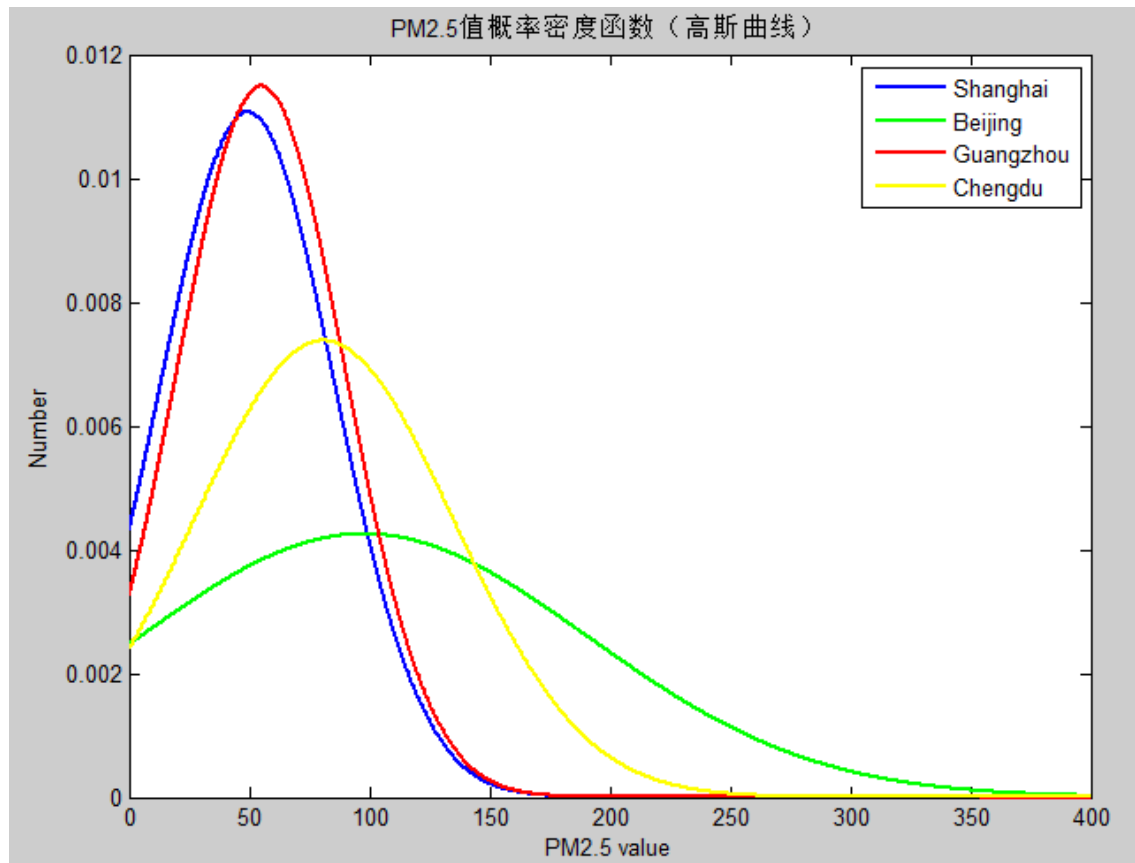
Parameter	Mean	Standard Error
β	-3.69	0.173
σ	0.37	0.0049
ϕ	0.92	0.021



Chapter4 Spatial Distinction

Intuitively we know that the average air quality in different city is different from each other, so we can make some simple analysis of the PM2.5 data, like mean value, variance, Gaussian distribution, etc. to represent their PM2.5 characteristics.

Here following the fitting Gaussian distributions of the four cities.



And the parameters of each city:

CITY	MEAN	STANDARD DEVIATION	MEDIAN
Shanghai	49.54	36.03	40
Beijing	97.95	93.83	72
Guangzhou	55.15	34.72	49
Chengdu	81.06	54.09	67

We can easily distinguish the cities from the different distribution of their PM2.5 dataset. However the distribution of Shanghai and Chengdu are relatively similar to each other, so that we can make other correlation information to find more characteristics to distinguish them.

Chapter5 Temporal Fractal

The time series: $\mathbf{x} = \{x_1, x_2, x_3, \dots, x_n\}$.

$$R(t) = \sum_{j=1}^{j=n} x(j)x(j-t).$$

Autocorrelation Method:

$$\mathbf{X}_i = [x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(m-1)\tau}], \quad i = 1, 2, \dots, N$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+\tau} & \dots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \dots & x_{2+(m-1)\tau} \\ \vdots & \vdots & & \vdots \\ x_N & x_{N+\tau} & \dots & x_{N+(m-1)\tau} \end{bmatrix},$$

Where m is embedding dimension, N is the number of vectors in the new phase space, and

$$N = n - (m - 1)\tau$$

The Euclidean Distance of all the points in X :

$$\varepsilon_{ij}(m) = \|\mathbf{X}_i - \mathbf{X}_j\|.$$

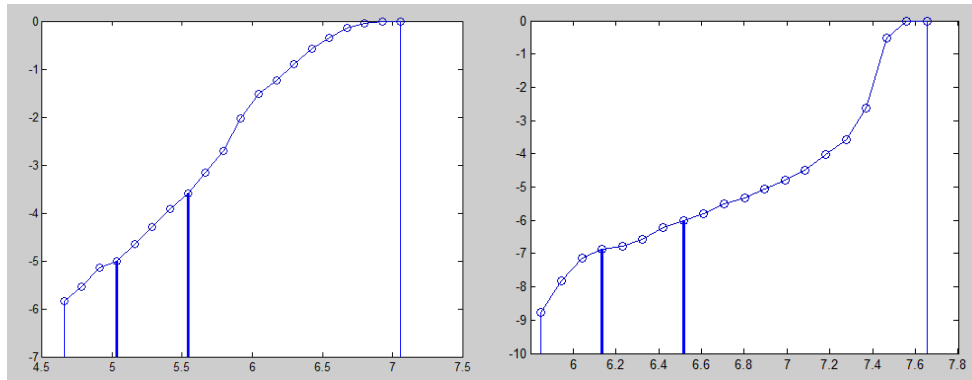
$$C_m(r) = \frac{1}{N(N-1)} \sum_{i \neq j} [H(r - \varepsilon_{ij}(m))],$$

correlation integral:

$$H(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}.$$

Heaviside Function:

$$\text{correlation dimension: } D = \frac{d \ln C_m(r)}{d \ln r}.$$



N=500

N=5000

Maybe $D=-6$ is the time saturation correlation dimension method for Beijing.

Chapter6 Conclusion

The project is enlightened by our Prof. Yuan and Dr. Wang, we found that the PM2.5 dataset is perfectly suit in the SMC model. After a good talk we started this research.

In our work, we use several methods to find a suit model and try to make a more reliable prediction. We use SMC model to estimate the parameters of the model, and use Rao-Blackwellized method to fit the PM2.5 dataset and one-step prediction. In addition, we also analysis the spatial distribution as well as the time correlation. What's more, we also consider the weather conditions and add them into the model by Neural Network, which is proved to improve the predicting accuracy.

And still there are a lot of promising works we can do in the future. As we have considered at the beginning, the noise can changed as Levy Process and take more influential factors into consideration, like vehicle flowrate etc. Furthermore, some big events also affect the air condition as well, which could also be considered and learned later.

ACKNOWLEDGMENTS

Thanks for Prof. Yuan and TA Yanbo giving us sighted guidance to go through this project. Jiaming Shen and I are so interested in the prediction of PM2.5 that we hope to do this research continually this summer. Hopefully make up a paper in the end.

Reference

- [1] He X, Xue Y, Li Y, et al. Air quality analysis based on PM 2.5 distribution over China[C]//Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, 2012: 2494-2497.
- [2] Yang Y, Yang K, Zhu Y, et al. The seasonal characteristics and spatial distribution of PM 2.5 in China[C]//Geoinformatics (GeoInformatics), 2014 22nd International Conference on. IEEE, 2014: 1-4.
- [3] Liu R, Fan F. Mass concentration variations characteristics of PM 10 and PM 2.5 in Guangzhou (China)[C]//Earth Observation and Remote Sensing Applications (EORSA), 2014 3rd International Workshop on. IEEE, 2014: 111-115.
- [4] Senlin L, Xiaohui C, Minghong W, et al. Chemical elements and their source apportionment

of fine particulates (PM_{2.5}) in Shanghai atmosphere[C]//Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on. IEEE, 2008: 3758-3761.

[5] Chen Y, Qin H, Zhou Z. A comparative study on multi-regression analysis and BP neural network of PM_{2.5} index[C]//Natural Computation (ICNC), 2014 10th International Conference on. IEEE, 2014: 155-159.

[6] Song L, Pang S, Longley I, et al. Spatio-temporal PM_{2.5} prediction by spatial data aided incremental support vector regression[C]//Neural Networks (IJCNN), 2014 International Joint Conference on. IEEE, 2014: 623-630.

[7] Ong B T, Sugiura K, Zettsu K. Dynamic pre-training of Deep Recurrent Neural Networks for predicting environmental monitoring data[C]//Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014: 760-765.

[8] Creal D. A survey of sequential Monte Carlo methods for economics and finance[J]. *Econometric Reviews*, 2012, 31(3): 245-296.

[9] Liu J S. Monte Carlo strategies in scientific computing[M]. Springer Science & Business Media, 2008.

[10] 分形理论及其应用[M]. 科学出版社, 2011.