PM$_{2.5}$ concentration prediction using times series based data mining

Jiaming Shen  
Shanghai Jiao Tong University, IEEE Honor Class  
Email: mickeysjm@gmail.com

Abstract

Prediction of particulate matter with diameter less than 2.5 $\mu$m (PM$_{2.5}$) is an important issue in control and reduction of pollutants in the air. In this report, I discuss three different models to deal with this PM$_{2.5}$ prediction problem, including basic Autoregressive Moving Average (ARMA) Model, Stochastic Volatility (SV) Model and Stock-Watson (SW) Model. Furthermore, I design an innovative model combining SW model with Time Series Neural Network. Finally, I conduct four experiments to compare and analyze the performances of above-mentioned models. The results show that this newly-designed model provide more accurate predictions of PM$_{2.5}$ concentration levels for the next six hours.

1 Introduction

Particulate matter (PM) is the term used for a mixture of solid particles and liquid droplets found in the air. In particular, fine particles with diameter less than 2.5 $\mu$m are called PM$_{2.5}$. Prediction of PM$_{2.5}$ is an important issue in control and reduction of pollutants in the air. Predictive models for PM$_{2.5}$ vary from extremely simple to extremely complex, but the ability to accurately forecast PM$_{2.5}$ concentration index remains elusive.

In this report, I discuss my effort to solve this problem using times series based data mining. I first adopt three distinctive models, including basic Autoregressive Moving Average (ARMA) Model, Stochastic Volatility (SV) Model and Stock-Watson (SW) Model. Furthermore, I design a new method combining SW model with Time Series Neural Network, adding external meteorological information such as temperature, pressure, humidity, windspeed and precipitation rate. Finally, I conduct four experiments to compare and analyze the performances of above-mentioned models, using the dataset provided by the U.S. Department of State. The results show that this newly-designed model provide more accurate predictions of PM$_{2.5}$ concentration levels for the next six hours.

2 Dataset

The dataset I use in this project is the U.S. Department of State air quality files [1], which contain PM$_{2.5}$ in concentration unit $\mu$g/m$^3$ from each post. Files include hourly data of following several attributes – Site, Parameter, Dates(LST), Year, Month, Day, Hour, Value, Unit, Duration, and QC Name. Site is the city where the measurements were taken. In this project, I only use four major cities in China, and their detail locations are shown in Table 1. Value is the measurement in concentration, and the value of missing data point is listed as -999. QC Name is the quality control status of the data; either valid or missing. Further explanations of this raw dataset file can be viewed in [2].

<table>
<thead>
<tr>
<th>Site</th>
<th>Location</th>
<th>Latitude and Longitude Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai</td>
<td>31.21, 121.44</td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td>39.95, 116.47</td>
<td></td>
</tr>
<tr>
<td>Guangzhou</td>
<td>23.12, 113.32</td>
<td></td>
</tr>
<tr>
<td>Chengdu</td>
<td>30.63, 104.07</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Geographic Coordinates for the locations of Four Major Cities

From the raw dataset files, I notice that these data have not been fully validated or quality assured, and thus they should be considered preliminary. Before conducting any further experiment, I first have a thorough look at the raw dataset, finding the number and percentage of missing data point. The missing data point means either the Value is -999 or the QC Name is labeled “Missing”. The results are shown in Table 2. As you can see, the percentage of missing data points of Guangzhou in Year 2014 is too high (above 5%), and thus I switch to the dataset in Year 2013. Except for Guangzhou, I use the latest dataset for analyzing all other major cities, namely using the dataset in Year 2014.

For the sake of temporal continuity, I did not simply delete those missing data. Instead, I use first-order interpolation method to replace them, and conduct simple statistical analysis in original data set. Results are shown in Figure 1 as
Table 2: Number of Missing Data Points of Four Major Cities in Three Different Years

<table>
<thead>
<tr>
<th>City</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai</td>
<td>279</td>
<td>173 (1.97%)</td>
<td>103 (1.17%)</td>
</tr>
<tr>
<td>Beijing</td>
<td>489</td>
<td>82 (0.94%)</td>
<td>99 (1.13%)</td>
</tr>
<tr>
<td>Chengdu</td>
<td>4370</td>
<td>1393 (15.9%)</td>
<td>285 (3.25%)</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>2188</td>
<td>371 (4.23%)</td>
<td>645 (7.36%)</td>
</tr>
</tbody>
</table>

Table 2: Number of Missing Data Points of Four Major Cities in Three Different Years

Well as Table 3. Obviously, Beijing has the worst air condition among all four major cities in China, which is consistence with our intuition.

These astonishing facts further strengthen the motivation of analyzing PM$_{2.5}$ concentration in China.

Figure 2: PM$_{2.5}$ of different status in four major cites of China

3 Task Goals

As I do not have a specific task goal in the original version of this project, I need to define them on my own. Actually, good task goals serve not only the purpose of conducting this project, but also decide whether this project is meaningful at all. Good task goals should be both intuitive and profound; they requires deep understandings of the whole problem and point out the direction I will be working on. In this project, I establish three different categories of task goals, namely PM$_{2.5}$ Trend Prediction, PM$_{2.5}$ Value Prediction, and PM$_{2.5}$ Hidden Factors Prediction. They will be discussed detailedly in following three subsections.

PM$_{2.5}$ Trend Prediction

Given a sequence of historical PM$_{2.5}$ data, we want to predict whether the following six hour’s PM$_{2.5}$ concentration indexes will increase, decrease or basically stay stable. This is call PM$_{2.5}$ Trend Prediction, which is the most basic task goal of our project.

PM$_{2.5}$ Value Prediction

Given a sequence of historical PM$_{2.5}$ data, we want to predict the exact value of PM$_{2.5}$ concentration indexes in next one or several time units. This is an advanced version of PM$_{2.5}$ Trend Prediction problem. Besides, if we only predict one time unit ahead, we call it One step ahead prediction, or if we want to predict several time units ahead, we call it Multiple steps ahead prediction. Obviously, the later one is a more challenging problem and require more complex model to solve it. Finally, we should pay attention to the usage of
“time unit”. The raw dataset provides hourly PM$_{2.5}$ concentration index, so intuitively the time unit could be “Hour”. We can see aggregate these values and change the “time unit” to be “Every Six Hour”, or simply “Day”. Here, in this project, I DO NOT change the time unit, and keep using “Hour” as the basic and single “time unit”, however, we can modify this in future works, for some other specific problems.

PM$_{2.5}$ Hidden Factors Prediction

Given a sequence of historical PM$_{2.5}$ data, we want to evaluate and predict the hidden factors which generate or influence the observable PM$_{2.5}$ concentration indexes. This is called PM$_{2.5}$ Hidden Factors Prediction, which is a higher level of prediction because it not only shows us the facts (observable PM$_{2.5}$ concentration indexes) on surface, but also analyzes the reasons in deeper level, i.e. hidden factors.

4 Experiment I: ARMA Model

The first model I used to achieve the task goals in previous section is Autoregressive Moving Average (ARMA) Model. The fundamental assumption of the model is that the value of the series at time $t$, $X_t$, linearly depends on its previous $p$ values (deterministic part) and on a random disturbance $\tilde{Z}_t$ (stochastic part). We can write

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \tilde{Z}_t,$$

where $\{\phi_1, \phi_2, \cdots, \phi_p\}$ are real constants called autoregressive (AR) coefficients. $\tilde{Z}_t$ is the disturbance at time $t$ and usually modeled as a zero-mean white noise process $\{Z_t\}$ with variance $\sigma^2$

$$\tilde{Z}_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q},$$

where $\{\theta_1, \theta_2, \cdots, \theta_q\}$ are called moving average (MA) coefficients. Combining the previous two equations, we can get

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

This defines a zero-mean autoregressive moving average (ARMA) process of orders $p$ and $q$, or ARMA($p, q$). When $q = 0$ only the AR part remains and ARMA model reduces to a pure autoregressive process of order $p$ denoted by AR($p$). Similarly, if $p = 0$, we obtain a pure moving average process of order $q$, MA($q$).

Before I apply this model to the PM$_{2.5}$ values in Shanghai of the first quarter in Year 2014. I first smooth the raw data using moving average of every six hours, see Figure 3.

After that, I run the Time Series Toolbox in Mathematica [3] and work on the model selection part using AIC as criterion. Results are shown in the table below:

You can see that the top four candidate models have very similar performs in terms of AIC, and later consideration, I pick ARMA(2,1) process and train a model based on it. Results are shown in Table 5. Finally, I use this process to estimate 3 days ahead PM$_{2.5}$ concentration index for the purpose of trend prediction and exhibit the results in Figure 4.
The second characteristic of SV model is the existence of static parameters, namely $\mu$, $\phi$ and $\sigma_{\eta}$. This is very different from the Stock-Watson Model discussed in next section. In order to make inference on these parameters, I resort to the full Bayesian framework and utilizes Markov Chain Monte Carlo (MCMC) samplers.

The level $\mu \in \mathbb{R}$ is equipped with the usual normal prior $\mu \sim N(b_{\mu}, B_{\mu})$. I set $b_{\mu} = 0$ and $B_{\mu} = 100$ for non-informative prior. For the persistence parameter $\phi \in (-1, 1)$, I choose $(\phi + 1)/2 \sim \mathcal{B}(a_{\phi}, b_{\phi})$, implying

$$p(\phi) = \frac{1}{2 \mathcal{B}(a_{\phi}, b_{\phi})} \left( 1 + \phi \right)^{a_{\phi} - 1} \left( 1 - \phi \right)^{b_{\phi} - 1}$$

where $\mathcal{B}(a_{\phi}, b_{\phi})$ denotes the beta function, and I choose the hyperparameters $a_{\phi} = 20, b_{\phi} = 1.1$. Finally, for the volatility of log-variance $\sigma_{\eta} \in \mathbb{R}^{+}$, we choose $\sigma_{\eta}^{2} \sim \mathcal{B}_{\eta} \times \chi^{2} = \mathcal{G}(1/2, 1/2 \mathcal{B}_{\eta})$. The hyperparameter $\mathcal{B}_{\eta}$ is chosen to be 0.1, which is of minor influence in empirical applications.

The aggregated results including posterior distribution of three parameters and the predicted volatility are shown in Table 6 and Figure 5.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>sd</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>-3.69</td>
<td>0.173</td>
<td>-3.973</td>
<td>-3.41</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.92</td>
<td>0.021</td>
<td>0.888</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma_{\eta}$</td>
<td>0.37</td>
<td>0.049</td>
<td>0.088</td>
<td>0.21</td>
</tr>
</tbody>
</table>

![Figure 5: Result of SV model on PM$_{2.5}$ dataset](image)

Finally, we should notice that all hyperparameters, estimated latent states and predicted volatility are in the “Hidden State Space”. Thus, SV model provides an excellent tool to conduct hidden factors prediction, however, it’s not suitable for making PM$_{2.5}$ Trend or Value Prediction. In future work, I think I could use SV model to extract some latent features of PM$_{2.5}$ concentration index. For example, I can use the predicted volatility as the measure of “aggressiveness” of one-step or multiple-steps prediction.
6 Experiment III: Stock-Watson Model

Stock-Watson (SW) Model is recently proposed in [7], and its brief explanation can be found in Section 2.5.7 of [5]. It is a time-varying random walk plus local level model.

\[
\pi_n = x_{1,n} + \varepsilon_n, \quad \varepsilon_n \sim N(0, \exp(x_{2,n}))
\]

\[
x_{1,n+1} = x_{1,n} + \eta_n, \quad \eta_n \sim N(0, \exp(x_{3,n}))
\]

\[
x_{2,n+1} = x_{2,n} + \omega_{1,n}, \quad \omega_{1,n} \sim N(0, 0.2)
\]

\[
x_{3,n+1} = x_{3,n} + \omega_{2,n}, \quad \omega_{2,n} \sim N(0, 0.2)
\]

where \(x_{1,n}\) in this project is the unobserved time-varying mean of \(\text{PM}_{2.5}\), and \(x_{i,n}\) for \(i = 2, 3\) are unobserved log-variances. Conditional on the log-variance \(x_{2,n}\) and \(x_{3,n}\), the remaining model is a linear, Gaussian state space model, and thus it can then be calculated exactly by the Kalman Filter.

There are two appealing characters of this model. First, the introduction of \(x_{1,n}\) enables the observation \(\pi_n\) to possess non-zero mean. Second, there are no static parameters that need to be estimated due to the special structure of this model. We can fit this model using Rao-Blackwellized particle filter technique. I plot the one step ahead prediction in Figure 6.

When viewed in high level, this model seems to fit the trend and values of \(\text{PM}_{2.5}\) concentration index. However, there are several major drawbacks. First, this model requires the true observation every time when it makes further prediction, and thus it can only make accurate one-step ahead prediction. Second, it presents sort of time-delay when making prediction. Actually, if we observed this results in detail, the average difference between true value and predicted value is 12.7463, which is not good enough. Despite of all these problems, Stock-Watson model is currently the most promising off-the-shelf method and serves as the foundation of an extended model discussed in the following section.

7 Experiment IV: Stock-Watson Model with External Information

All above-mentioned models have a common problem, i.e., they failed to take consideration of external information. When talking about \(\text{PM}_{2.5}\), people normally think of the weather data. Based on this intuition, I combine the prediction output by SW model with the external weather information. To be more specific, every time we get a output from SW model, instead of using this value as the prediction directly, I view it as a part of input features which also contains the weather information such as temperature, pressure, humidity, windspeed and precipitation rate. Our goal is to output the final prediction based on the input feature vector. This is a regression problem! In order to solve this problem, I adopt a time series neural network with the architecture shown in Figure 7.

Using the first 500 hourly \(\text{PM}_{2.5}\) concentration indexes of Shanghai in Year 2012, I train this neural network with 10 hidden units and 2 units time delays. In addition, I adopt the Bayesian Regularization in the training process to deal with overfitting problem. Results are shown in Figure 8 and Figure 9.

As you can see, the error in most of time stamps belongs to range \([-7, +8]\). The MSE reduces to approximately 50,
when using this method.

Finally, I will compare this method with another two methods through solving multi-step prediction problems. I train the models with first 500 hours day points and evaluate its performance using six steps ahead prediction. As mentioned below, the original version SW model is not able to make multi-step prediction, so I modify it by letting the next “true” observation equal to the last prediction. Results are shown in Figure 10.

You can see the true value of PM$_{2.5}$ is decreasing, while the multiple steps ahead predictions using only SW model remain basically the same. This is because SW model adjust its predictions by comparing them with the true observation values, however, in order to achieve multi-step prediction, we let the “true” observation to be same with last prediction and thus SW model loses its ability to recover from bad prediction. The modified version of SW model overcomes this problem by using external information, and achieves a better performance.

8 Conclusion & Future Works

There are several places I can further improve in future. First, I should clean my original dataset more carefully. In this project, I simply adopt the first-order interpolation method which only works when the number of missing data in a row is not very high. However, after a more detailed observation of raw dataset, I find sometimes there will be over 30 continuous missing data. This phenomenon requires more careful data pre-processing.

Second, I should put more effects on the latent variables. These hidden factors not only serves as a way to improve the prediction accuracy, which is basically this report focused on, but also should be considered as the fundamental reasons leading to the PM$_{2.5}$ concentration index trend. As mentioned in “Experiment II: Stochastic Volatility Model”, I could use the SV model as a method to extract hidden features of observable PM$_{2.5}$ values.

Finally, I should develop my own model to make PM$_{2.5}$ concentration index prediction, based on this problem’s own characteristics. In this report, I briefly talk about one simple modification of Stock-Watson Model and show its corresponding performance improvement. Through further analyze on this problem and more discussion with TAs, I will figure out a more suitable model to deal with it.

9 Reference

[1]. U.S. Department of State.
[2]. Air Quality Data Files.
[3]. Mathematica Time Series Toolbox
[8]. Matlab Neural Network Toolbox