Beijing PM2.5 Time Series Analysis and Prediction using Regression Models

Mengmeng Liu and Xin Cao

mliu301@gatech.edu and xincao@gatech.edu [1]

[1]Georgia Institute of Technology

May 2, 2016

1 Abstract

In this project, we analyzed the PM 2.5 data provided by U.S. Department of State Data Use Statement. We use three different methods to analyze the time series PM 2.5 observations in Beijing. They are Autoregressive Moving Average (ARMA), Neural Network (NN), and Support Vector Regression (SVR). The experiment results show that NN and SVR give more accurate result than ARMA. Besides, NN and SVR can correctly capture the fluctuation of PM 2.5 values in the next 400 days, but ARMA can only correctly predict the PM 2.5 values in the next 100 days. So if you want to predict the PM 2.5 in the short time range, then ARMA is the best model among these three models. However, NN and SVR will be the best model if you want to predict the PM 2.5 in longer time, such as prediction the PM 2.5 in next year.However, it is hard to say which model is the best, we need chose the appropriate model based on the data properties and objective of the analysis.

2 Introduction

The air pollution in modern cities is a severe problem which significantly affects humans life and health. PM 2.5 is a measurement a type of particulates or aerosol with a scale size less than 2.5 micrometers which usually suspends in the atmosphere. The majority of this aerosol consists of some chemicals such as organics, sulphate, amine, nitrate, black carbon and so on. The cause of PM 2.5 is very complex, such as protosomatic emission or production, and secondary emission or production. The protosomatic production includes but not limited to vehicles emissions, power plants emissions or even natural fires. The secondary production mainly comes from varieties of chemical reactions between different chemicals in the atmosphere, whose process is usually very complicated to investigate. Besides, the physical condition of atmosphere is also an important factor to affect PM 2.5, such as the temperature, pressure, humidity, wind orientation, wind speed. And therefore many atmospheric scientist developed some chemical model to explain and predict the influence of these chemical reactions on the production of PM 2.5.

However, even the models have been developed better and better, the prediction of PM 2.5 is still a hard problem. Because it is not only related to the chemical reactions, physical parameters, but might be also affected by some unknown minor factors, such as human local activities which are

hard to accurately measured and predicted. On the other hand, no matter which model of chemical or physical is used to study and predict PM 2.5, they more or less made some assumptions in the models because it would become extremely complicated and nearly impossible to solve without these assumptions. But these assumptions might also cause some unpredicted errors since the nonlinear behaviors could result in chaos. An alternative method is to use statistical models to investigate the variation of the PM 2.5, so that we do not need to consider the complex intermediate process.

3 Problem Statement

The objective of this project is to conduct the time series analysis of the PM 2.5 in Beijing from 2009 to 2015. Three statistical models are used and compared to see which model is good at long term prediction, which one is good for short term prediction. And finally we use the model to predict the future PM 2.5 values.

4 Data

The data used in this project is download from U.S. Department of State Data Use Statement. Its website is: `http://www.stateair.net/web/historical/1/1.html`

The data includes the hourly PM 2.5 observation data from 2009 to 2016 in Beijing. The scatter plot of PM 2.5 in each year are shown in 1 . There are some missing data in the data set, which are marked as -999. So we need clean the data before use them to train our model. And since the data we used covers many years, we aggregate the data into daily observations.

Since the data used in the time series analysis need to be continuous, we use the first order linear interpolation method to estimate the PM 2.5 observation values for missing data. Then we get a daily PM 2.5 observations without missing values. The total data set has 2484 data points, which cross almost 7 years time span from 2009 to 2015. The data is ordered by their time. We separate the whole data set into two parts: training data set and testing data set. The first 2200 data points are used to train the models, and the rest 284 data points are used to test or validate our models. and we use the same training and testing data set to train and test all the models we used in this project.

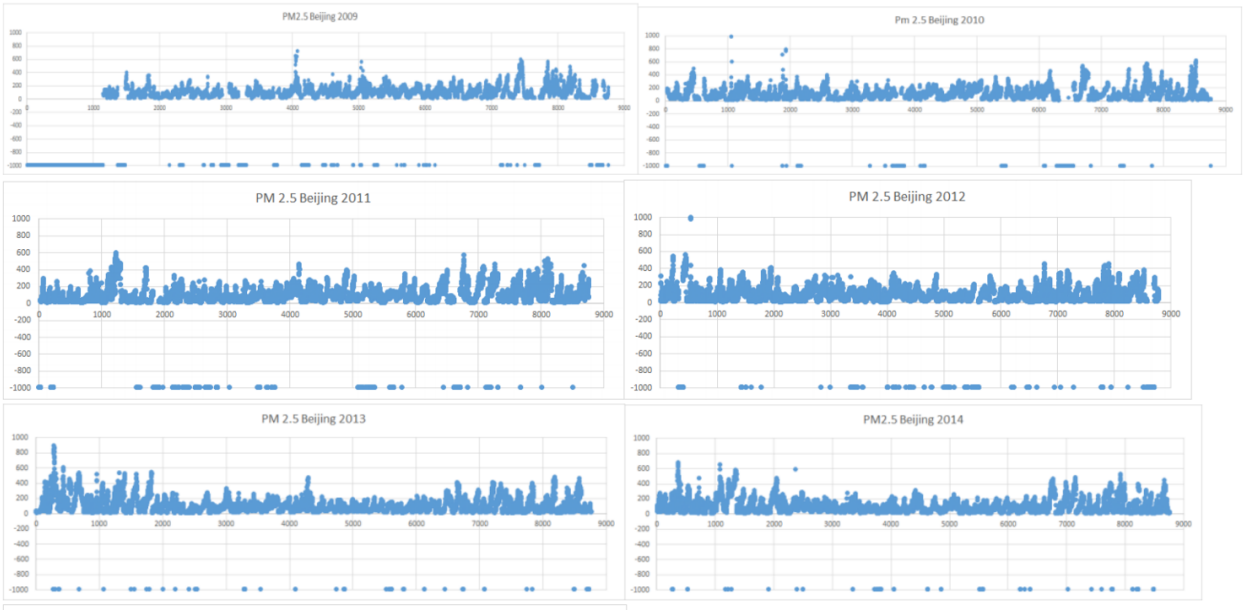

Figure 1: Original Data from

5 Methods

Three different statistical methods were used in the project. They are Autoregressive Moving Average (ARMA) Model, Neural Network (NN), and Support Vector Regression (SVR).The following sections provide more detail information to these models.

5.1 Autoregressive Moving Average Model

ARMA Model assumes that the value of $PM_{2.5}$ at time t, X_t, linearly depends on its previous p values, as shown in equation 1. Note that these p values does not necessarily be at continuous time. For example, the 1st one may be in time 10 seconds, then the 2nd value could be in time 30 seconds.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \epsilon_t \tag{1}$$

Where, Y_t is the data at time t, $Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ are the data points at previous time points before time t. $\alpha_0, \alpha_1, ...\alpha_p$ are their corresponding coefficients, ϵ_t is the noise term with the normal distribution.

From equation 1, we can see that ARMA model is actually a linear regression model based on time series analysis with single output, which is a one dimensional variable and with p input

variables or one variable with p dimensions. Physically speaking, if we assume the effect of PM 2.5 values lasts longer, we can use a larger p value, otherwise, we could use a smaller p value. For example, if we assume that the PM 2.5 value is only effected by the values of its previous day, we then make $p = 1$. After reading some papers of studying the PM air pollution, and trying multiple p values to train our model, we found that $p = 10$ is a reasonable number for our model. This means that PM 2.5 is a middle-long term pollution problem. The coefficients means the weight of each predictor, higher weight means lager impacts on current PM 2.5 values. Therefore, we think the data at previous day might have higer impact to the estimation of current day.

In order to better train the model, 10-fold cross validation was used in the model.

5.2 Neural Network

Originally, **Neural Network (NN)** were intended to model the learning and pattern recognition done by physiological neurons[4], figure 2 is an example of NN model structure, which only has one hidden layer with 4 nodes in the hidden layer. In the simplest NN (see equation 2), response variable Y is modeled as a linear combination of the derived features $Z_m = \sigma(\alpha_{0m} + \alpha_m^T X)$. In practice, sigmoid function (see equation) is usually be used as the activation function $\sigma(v)$ [4].

$$Y = \beta_0 + \sum_{m=1}^{M} (\beta_m \sigma(\alpha_{0m} + \alpha_m^T X)) \tag{2}$$

where,Y is the response variable, M is the number of nodes in hidden layer, $\sigma(v)$ is the activation function .

$$\sigma(v) = \frac{1}{1 + e^{-v}} \tag{3}$$

NN can model both the linear problem or non-linear problem, and can also predict continuous values after supervised training. However it generally over-parametrized and the optimization problem is unstable due to the random initial value. Besides, It is very susceptible to local minimum traps, thus is easy to over-fitting. It's best to scale all the inputs to have mean 0, and standard deviation 1.

In this project, we used back-propagation NN function provided by **MATLAB**.In the Neural Network model, we used 10-fold cross validation to tune the parameters such as learning rate, the

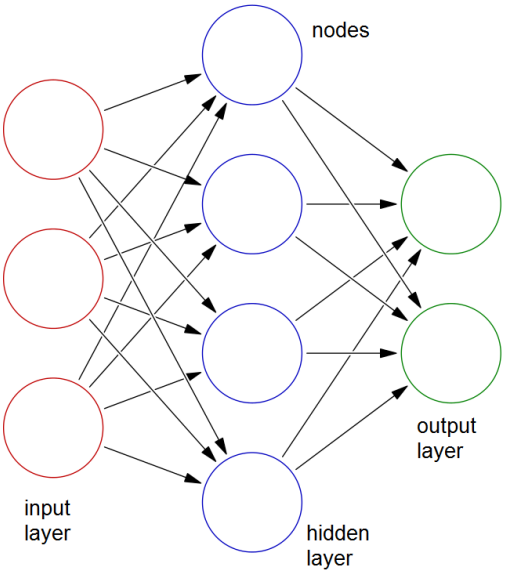

Figure 2: Basic Structure of Neural Network

option of transfer function. We choose these parameters in the model based on both the accuracy and efficiency. By trial and error, we found that the learning rate 0.1 is the best and the appropriate activation function is the commonly used sigmoid function. We set the error limit to be 0.00004, and iterate up to 100 iteration steps during the process of training models.

Before we trained the model, we first normalize the data into range of $0 \rightarrow 1$,and we reverse normalize the result and then get the final results after we get the result from our model. The results of NN are shown in figures in the results section. We use 10 inputs in Neural Network model, which means we assume that only the recent 10 PM 2.5 values will affect the observations of current value. The number we used in our model is reasonable from the aspect of real air pollution system evolution. Besides, high dimensional data can increase the possibility to avoid following into local optimum for our model. Because in high dimension, even if the model fall into a local optimum in a certain dimensional space, but the rest of dimension might still not an optimum, so it is hard to fall into a local optimum in all dimensions. For example, the saddle point is a local optimum in one dimension but not optimum in another dimension. Therefore, for high dimensional data, we can almost can get a global optimum! We also run the NN model many times in order to guarantee the final result is or close to the global optimum.

In order to better investigate the performance of NN, we train a couple of NN models with

different hidden layers and different nodes in each layer. Their results are shown in the result section.

5.3 Support Vector Regression

The original **Support Vector Machines (SVM)** algorithm was proposed by Vladimir N. Vapnik and Alexey Y. Chervonenkis in 1963[1] and originally intended for binary classification. SVM has become an important research topic in the pattern recognition field due to their unique advantages with respect to strong generalization ability, small samples, and the ability to process high-dimensional data[2,3]. SVM can be applied not only to classification problems but also to regression problems. and when SVM is used to regression problems, we call it **Support Vector Regression**(SVR). In this project, we use the SVR model provided by R in package **e1071**. And cross-validation was used when training our SVR model.

kernel is always used to mapping the samples into high dimensional feature space. In this project we use Radial Basis Function (RBF) kernel, because it implicitly projects the samples from the input space to a transformed space (i.e. feature space) with infinite dimensionality.

The RBF kernel on two samples x and x_i, represented as feature vectors in some input space, is defined by equation 4.

$$K(x, x_i) = exp(\frac{-||x - x_i||^2}{2\sigma^2} = exp(-\gamma||x - x_i||^2); \tag{4}$$

where $||x - x_i||^2$ is the distance between sample x and x_i, $\gamma = 1/(2\sigma^2)$ is the band-width of the kernel function.

In order to better train the model, 10-fold cross validation was used in the model. And we use grid-searching method to find the optimal parameters for SVR model.

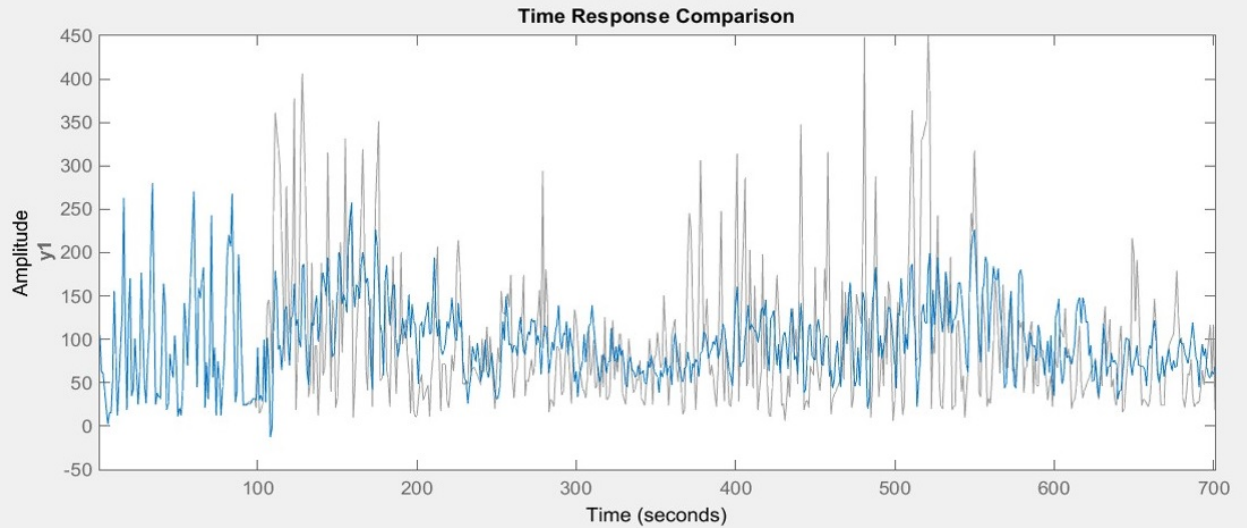

Figure 3: Training Data and its Prediction Result from ARMA Model

6 Results and Discussion

6.1 Autoregressive Moving Average Model

As Figure 3 and 4shows, we used the AR model to fit the time series of PM 2.5 values. From the result, we can see that the prediction after p time points start to become inaccurate and the error gradually increases. This is reasonable because our model assume that the PM 2.5 value is linear dependent on its previous p values, which could make a relative good approximation with p time points. However, with time going, the model's accuracy to predict the PM 2.5 values starts to decrease. This make sense, because the evolution of PM 2.5 is not actually a linear system, but a nonlinear system. Therefore, in order to better investigate PM 2.5 data, we need to use other models which can better describe a nonlinear system. An alternative way is to use nonlinear regression such as polynomial regression in the ARMA model instead of linear regression. For example, we may need use LASSO term plugged in the polynomial regression, in order to tune the parameters to avoid over-fitting, and we need to use cross validation to find out the best regularization/penalty term. Therefore, it is not easy or efficient to find the optimum polynomial order for this data set. However, NN and SVR can describe non-linear relationship in a much better way.And we will discuss their results in the following section.

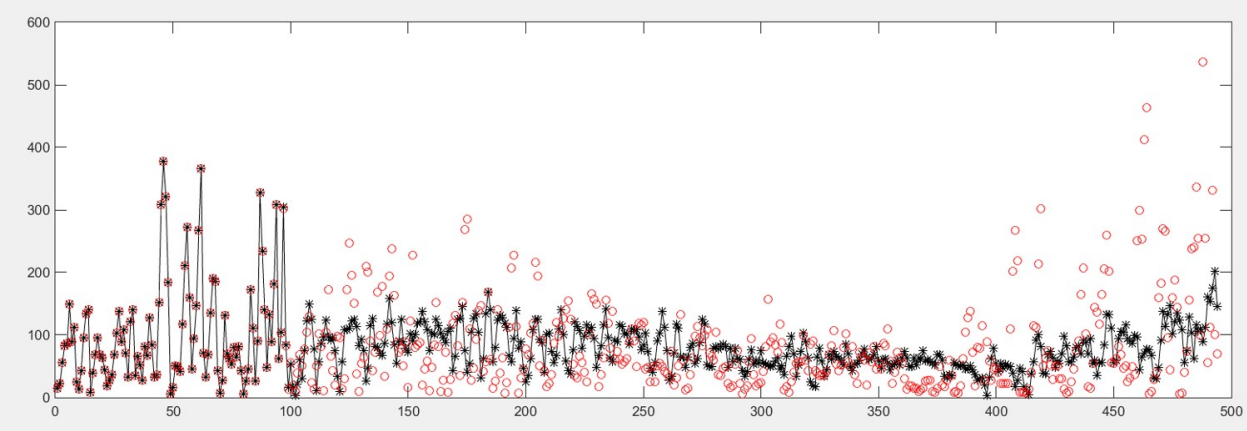

Figure 4: Testing Data and its Prediction Result from ARMA Model

6.2 Neural Network

The results of NN models with different hidden layers and different nodes are shown in Figure 5, Figure 6, Figure 7, and Figure 8. Their training and testing errors are listed in Table 1 and Table 2 respectively.

Figure 5 shows the results for a single layer NN model with 2 neural nodes. From this figure, we can see that the model output the training prediction with less magnitude compared with the original training data, so is the testing output.

Looking into its prediction for the testing data, we can see that this NN model fit the first 50 testing points well, but then its accuracy for predicting other values decreases. Especially, for the data points between 150 and 180 time unit (day) , the model did not capture trend of original data well. And after the 250 time units, there are certain shifts for the output prediction. However, the model can still describe the relative going up or dropping down during those period to some extent.

Figure 6 shows the result for NN model with 2 hidden layers, and 5 nodes in each layer . As Figure 6 shows,the prediction for training and testing data become slightly better comparing with the results of NN model with 1 hidden layer, and RMSE slightly decreases as well . In observation of the result, the trend of model output during 150 and 180 data points become slightly obvious. And the variation phase during the last 50 points becomes better, however, the amplitude becomes lower than the previous model, which might indicate the increasing hidden layers bring in a little over-fitting in the long term prediction.

Figure 7 shows the result for NN model with 3 hidden layers, and 5 nodes in each layer . As

Table 1: Training Errors of ANN using different hidden layers and neural nodes

methods	RMSE	SSE
1 layer 2 nodes	58.1665	0.0264
2 layer 5 nodes	57.2296	0.0260
3 layer 5 nodes	57.0184	0.0259
3 layer 10 nodes	56.2996	0.0256

figure 7 shows, the RMSE is very close to the previous case, so the better way to assess this model is compare the output and the original data from intuition viewpoint and based on the physical system analysis. We found that its performance is very close to the model with 2 hidden layers, and the model has weaker ability to predict the data with low value. We thought that it might be because the nodes in each layer is too small, since our data are ten dimensional vector variables. Therefore, we tried to increase the number of nodes in each layer.

figure 8 shows the result for NN model with 3 hidden layers, and 10 nodes in each layer . From the result, we can see that its RMSE is smaller than the previous one (see detail in Table 1 and Table 2), thus it is better than previous one. First of all, the model has the ability to predict both high and low range of PM 2.5 data for training data. Secondly, the lower panel of this figure shows that the model predict the testing data very well, not only in the beginning of time,but also in the middle and further time period, especially during 150 to 180 data points which shows the model can produce almost the same trend as that of the original data. We can also see the last 50 data points corresponds with the model prediction no matter the variation phase or the magnitude. This might be because increasing number of nodes in each layer has enough weight parameters to model the 10-days dynamic air pollution system. In conclusion, we figured out that the Neural Network model can predict PM 2.5 very good in a relative long period (about 280 days). And the model which has 3 hidden layers with 10 nodes in each layer is the best among the ones with different structures. We thought that although the increasing number of hidden layers and that of nodes in each layer might improve the accuracy of the prediction, the potential over-fitting problem might be brought in. Therefore, in practice, we usually try to use different parameters and structures to get a best model in order to balance the variance and bias.

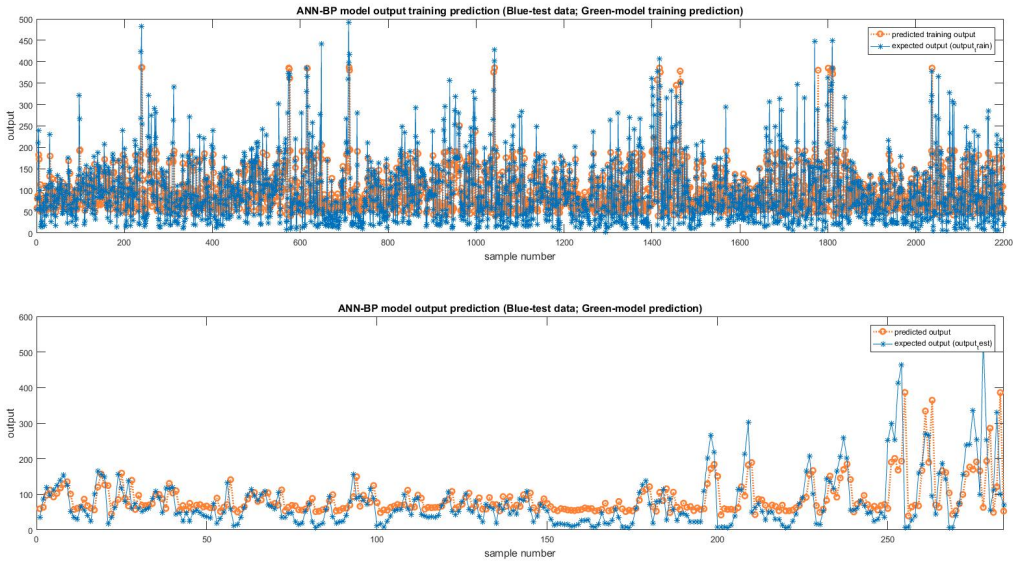

Figure 5: Prediction Results from Artificial Neural Network (1 hidden layer, 2 neural nodes)

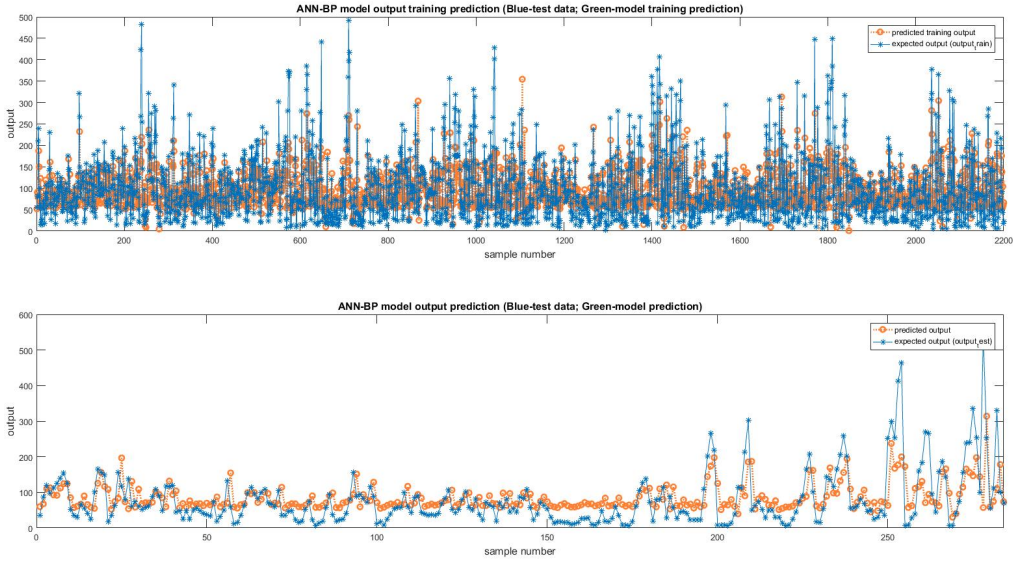

Figure 6: Prediction Results from Artificial Neural Network (2 hidden layer, 5 neural nodes)

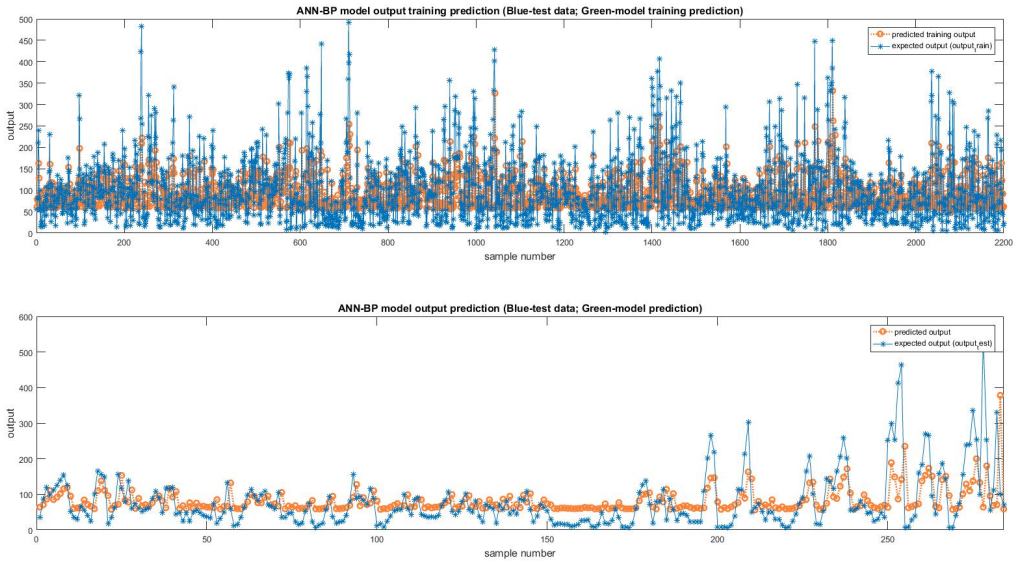

Figure 7: Prediction Results from Artificial Neural Network (3 hidden layer, 5 neural nodes)

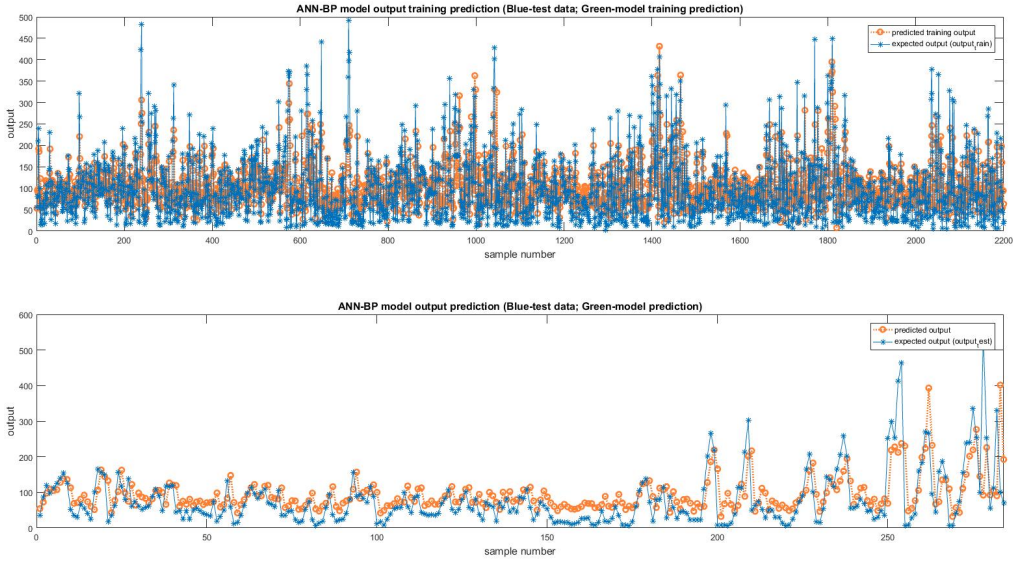

Figure 8: Prediction Results from Artificial Neural Network (3 hidden layer, 10 neural nodes)

Table 2: Testing Errors of ANN using different hidden layers and neural nodes

methods	RMSE	SSE
1 layer 2 nodes	58.3932	0.2056
2 layer 5 nodes	62.1253	0.2188
3 layer 5 nodes	58.6473	0.2065
3 layer 10 nodes	61.8876	0.2179

Table 3: Training and Testing Errors of SVM

	RMSE	SSE
training	58.7716	0.0267
testing	60.1504	0.2118

6.3 Support Vector Regression

The prediction results of SVR are shown in Figure 9. From Figure 9, we can see that the prediction for testing data in the short or middle period is very good, but there is still a slight phase shift in the last 50 days. From table 3, we can see that the RMSE of training data of SVR is higher than that of NN, but RMSE of testing data is lower than the Neural Network. This may indicate that NN is a little over-fitting.But it is hard to say SVR is better than NN. Because the results of NN is sensitive to initial value, its number of hidden layers, number of neural nodes in each layer and other parameters.

Therefore, combing the results from NN and SVR together may give a better result. And we will do experiment to verify this in the future.

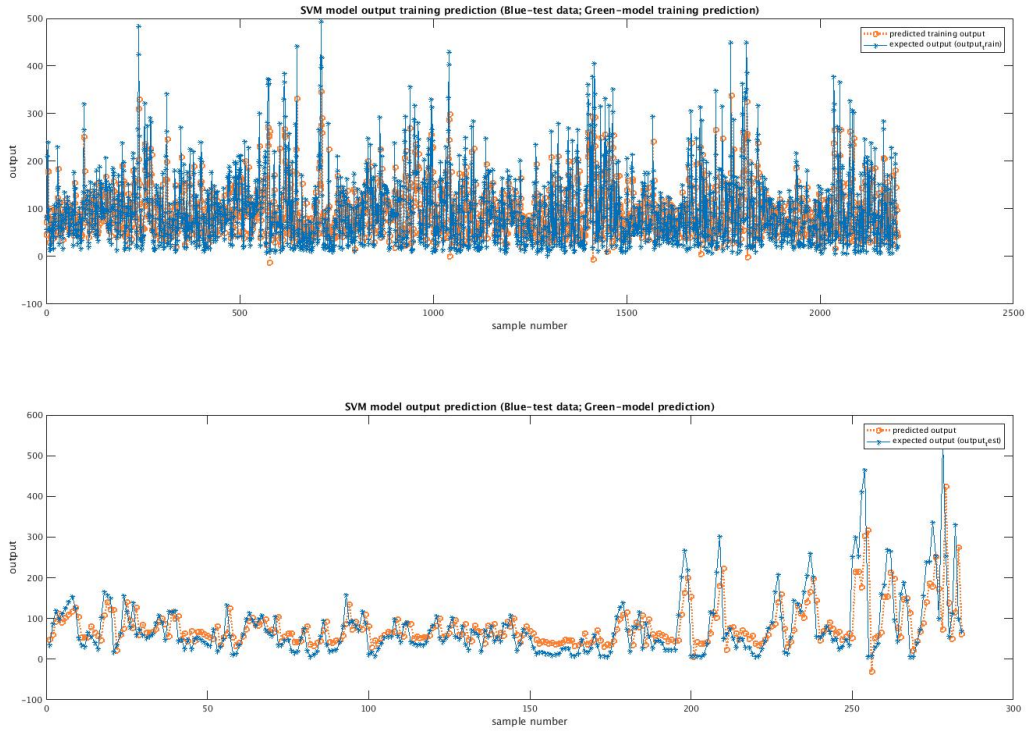

Figure 9: Prediction Results from SVR

7 Future Works

In the future, we will test our model use other data, and try to use resemble models to see whether it will give a better result or not. For example, combing the NN and SVR model together, or combing these three models used in the project together. we also want to analyze the models' sensitiveness to time scales. We will train our model at different time scales, including hourly, monthly, seasonally, and yearly. And then we analyze how our prediction result will change with time scales.

8 Conclusion

In this project, we analyze the PM 2.5 data in Beijing from 2009 to 2015 by ARMA, NN, and SVM. We discussed the characteristics of each model and tuned the parameters for each model, in order to get the best validation/testing results. We found that ARMA model has good performance to predict short term time series variation, but lose its predictive ability in a long term variation. This is because ARMA assume the relationship between the predictor and response variables is linear, but PM 2.5 is not linearly with the time in the long term. NN and SVM is good at long term prediction. For NN, its number of hidden layers and number of nodes in each layer could affect the performance and efficiency for validation or testing the data.

It is hard to say which model is best for this data set, because we found that NN got better training error, while SVR got better testing error.

From this project, we also found that cross validation is very useful for parameters tuning. In practice of computational statistics, we not only need know how to choose the appropriate model and how to tune the parameters, but we also need analyze the data in its specific domain. Such as reading relative papers, study its background and physical meaning, and so on. This can really help us better understand the data and the potential prediction results.

Reference

1. Wikipedia, "Support vector machine". Wikimedia Foundation, Inc., 19 December 2015 (Accessed 3 March 2016) [https://en.wikipedia.org/wiki/Support_vector_machine].

2. Huang C., Davis L. S., and Townshend J. R. G., "An Assessment of Support Vector Machines for Land Cover Classification". International Journal of Remote Sensing, 23(4): 725-749 (2002).

3. Luo J., Zhou C., Leung Y. and Ma J, "Support Vector Machine for Spatial Feature Extraction and Classification of Remotely Sensed Imagery". Journal of Remote Sensing, 6(1): 50-55 (2002).

4. Lectures and Notes of ISYE 7406.

5. Simon N. Wood, "Generalized additive models with integrated smoothness estimation". inside-R, (Accessed 3 March 2016) [`http://www.inside-r.org/r-doc/mgcv/gam`]

6. Wikipedia, "Regression analysis". Wikimedia Foundation, Inc., 18 February 2016 (Accessed 3 March 2016) [`https://en.wikipedia.org/wiki/Regression_analysis`].

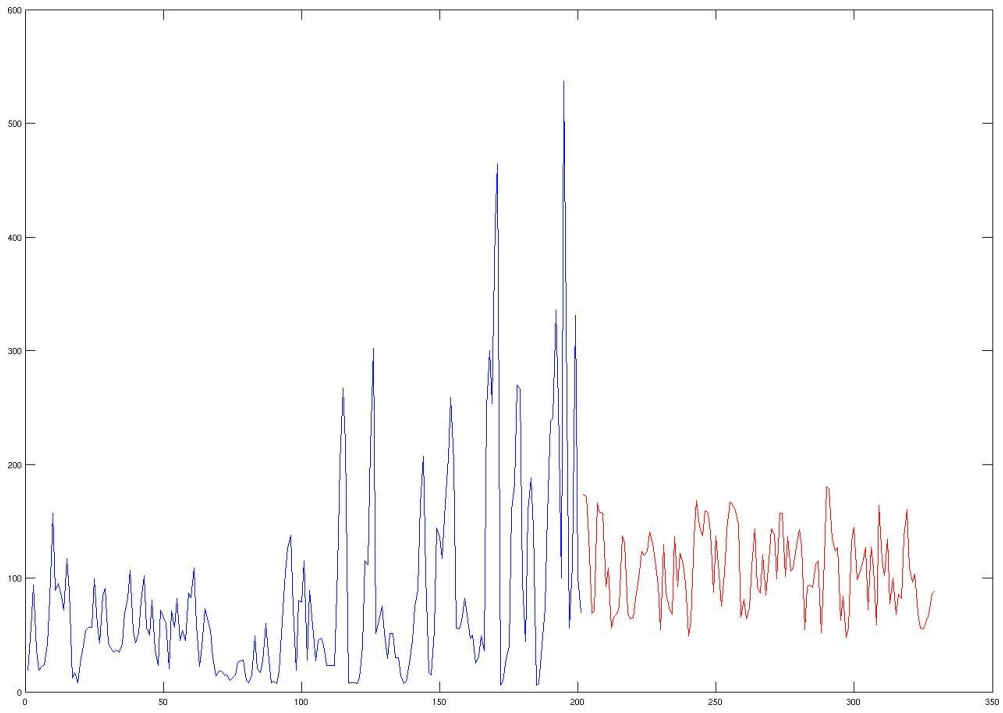

Figure 10: Training and Testing Error of One-layer Neural Network

Appendix

8.1 NN Results

We also used NN in predict PM 2.5 in Beijing in the current week, and compared with the current PM 2.5 data from Beijing and observed the air pollution locally in Beijing (one member of our group went there) from both data learning side and intuitive observation viewpoint side. Figure 10 shows the prediction result for the PM 2.5 in the last 200 days of 2015 (in blue line) and the first 128 days prediction of 2016 (in red line).

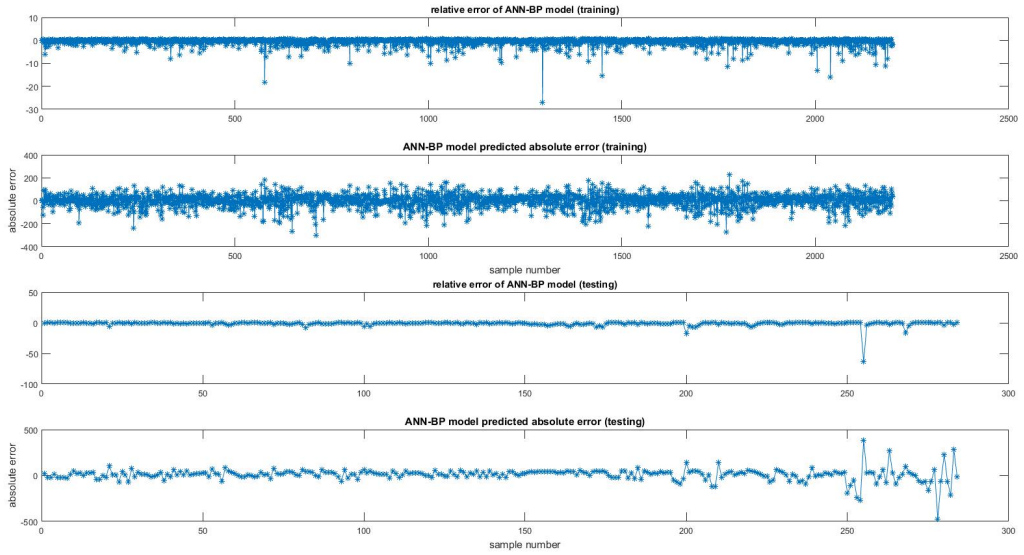

Figure 11: Training and Testing Error of One-layer Neural Network

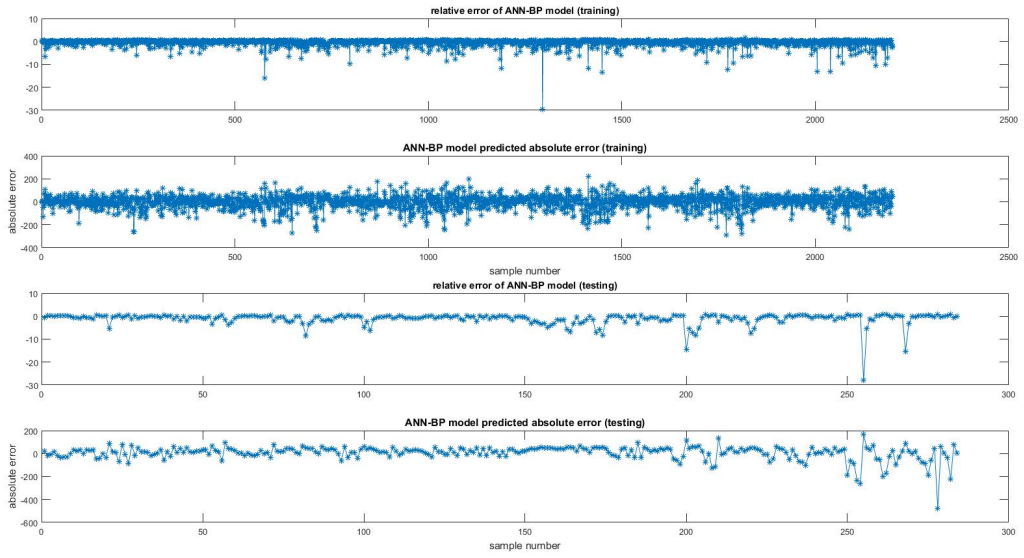

Figure 12: Training and Testing Error of Two layers Neural Network with 5 nodes in each layer

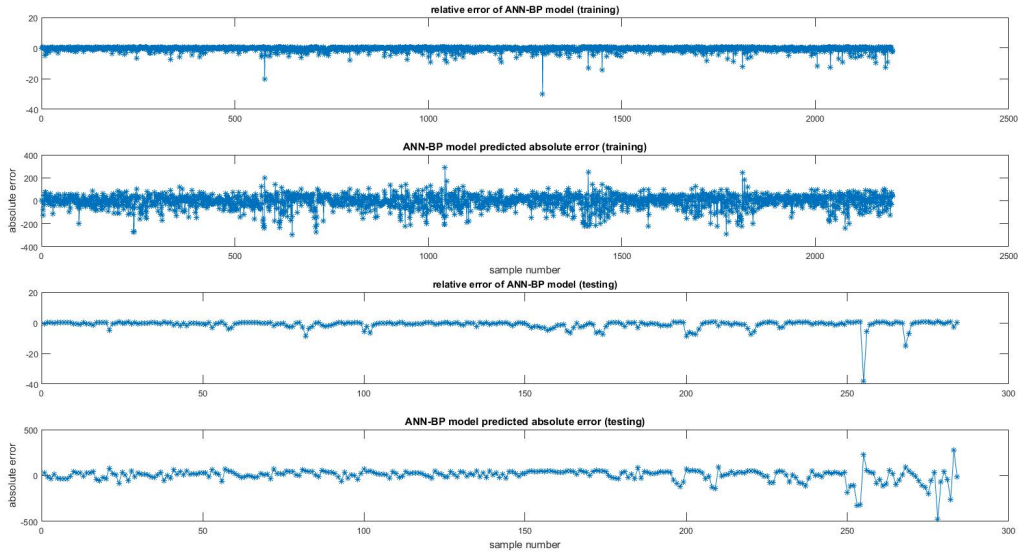

Figure 13: Training and Testing Error of Three layers Neural Network with 5 nodes in each layer

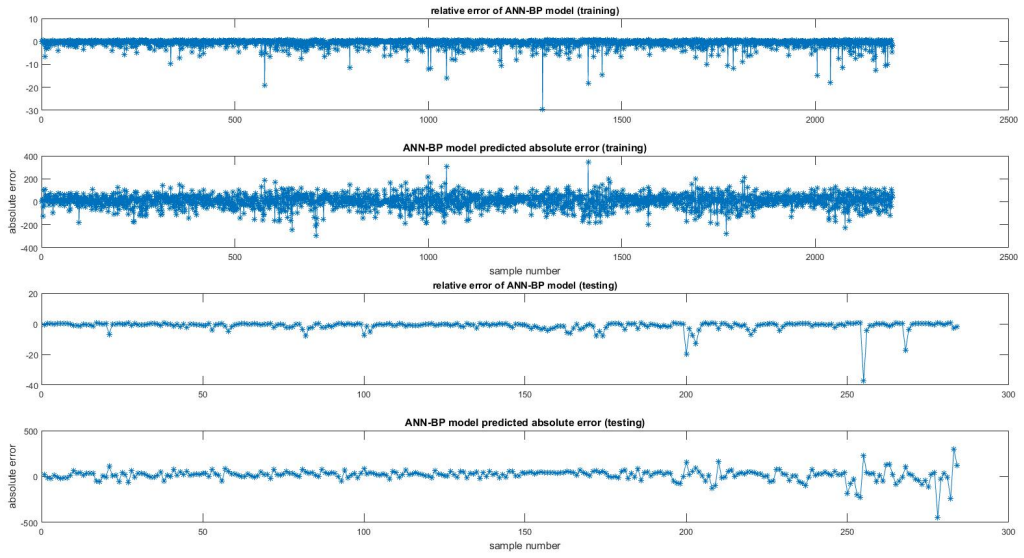

Figure 14: Training and Testing Error of Two layers Neural Network with 5 nodes in each layer

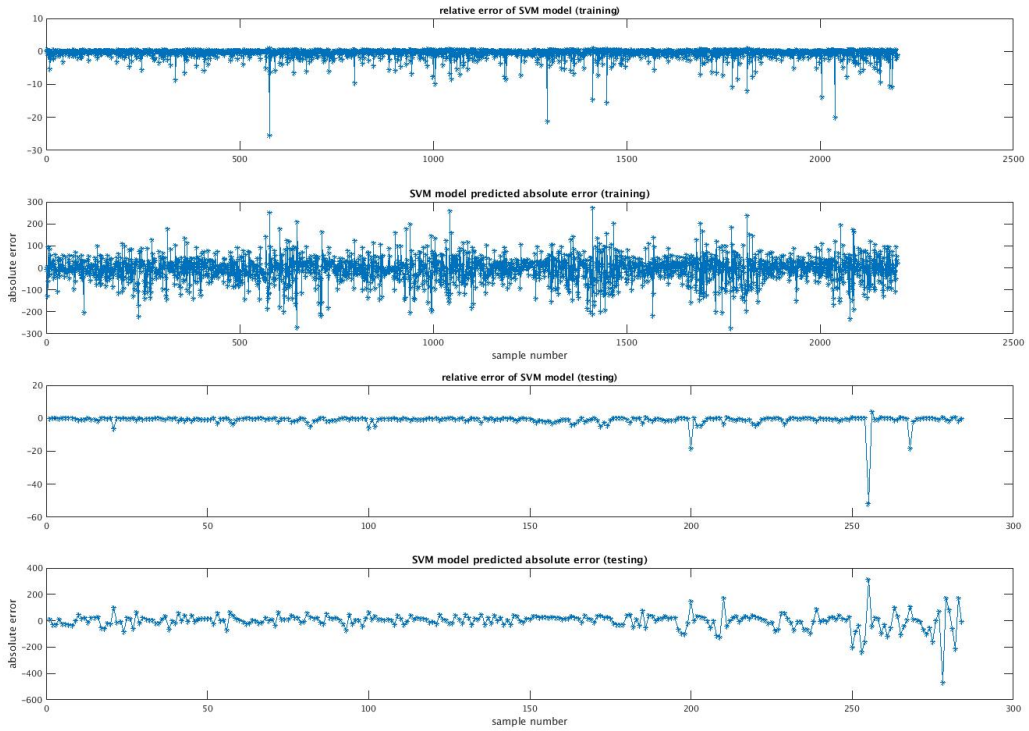

Figure 15: Training and Testing Error of Two layers Neural Network with 5 nodes in each layer