JiaHong Yu

1/1/2016

**Research on Beijing ‘s PM 2.5 Prediction**

**Introduction:**

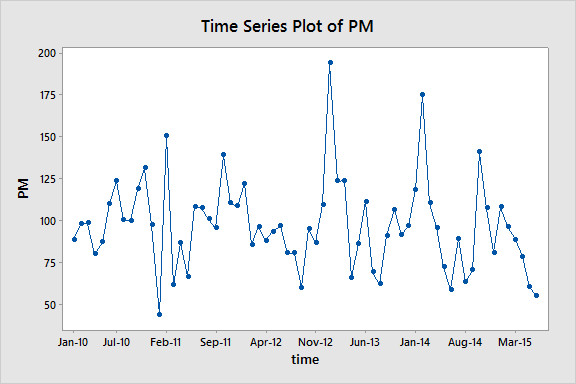
Beijing is the capital city of China which has severe air pollution. Being interested in how serious its pollution is and how the pollution level changes over time, I researched Beijing’s PM 2.5 hourly observations from Jan 1st, 2010 to Dec 31st, 2015 based on the data set ‘PM2.5 Data of Five Chinese Cities Data Set’ from UCI Machine Learning Repository. My research offers a solid overview of the air pollution in Beijing, which can help the Beijing government design its pollution control policies in 2016.

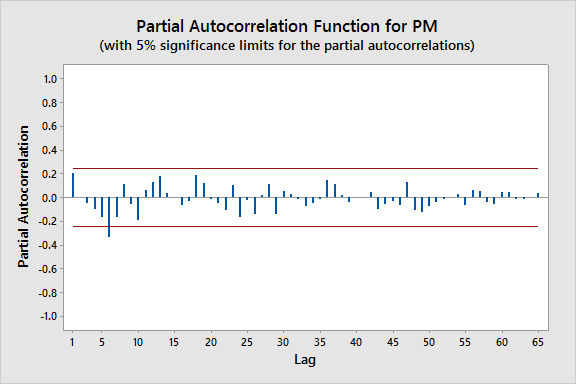
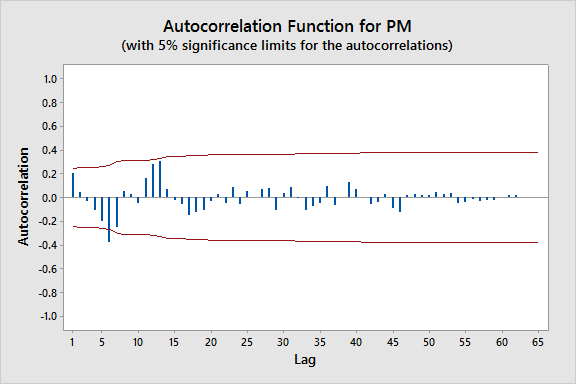
The original data set had many missing values named NA. To model time series I needed evenly spaced data. Therefore, a function called ‘na.ma’ in RStudio was used to fix these NAs. This function could replace the missing values based on a weighted exponential moving average. However, at some points in time, the wind direction information was missed and could not be fixed by ‘na.ma’ function because they were categorical values. Therefore, I manually fixed these 'NA's based on the wind directions around these points in time. In the end, all the NAs in this data set were fixed.

I made two subsets of the raw data—daily observations, and monthly observations. The daily observations were made by subsetting the observations at noon of each day; meanwhile, the monthly observations were obtained by calculating the average PM 2.5 of each month. For each subset, the last few records were held as the out-of-sample observations.

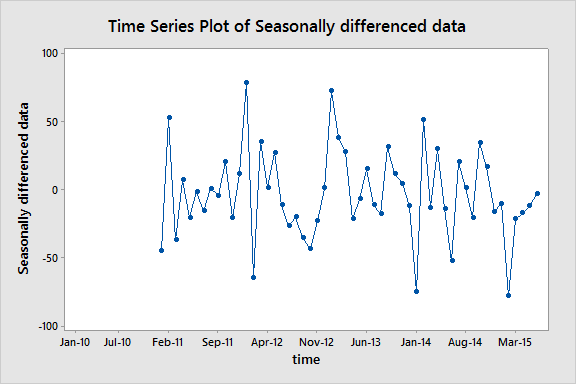
**Monthly Data Analysis:**

The monthly observations were first analyzed in Minitab 7. I loaded the monthly subset and drew its time series, ACF, and PACF plots. The plots are displayed below:

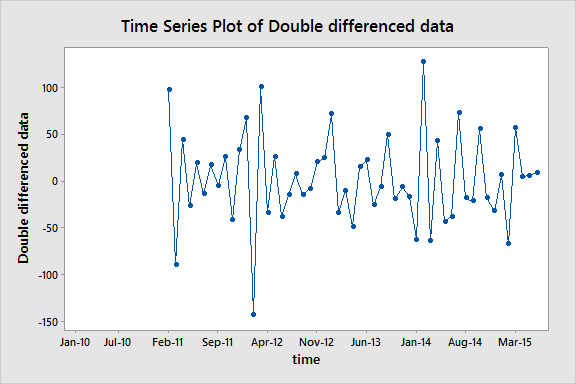


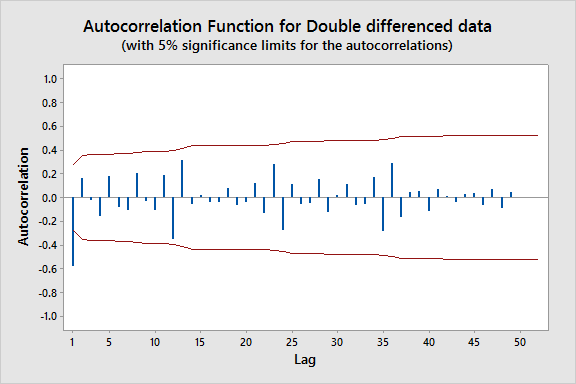


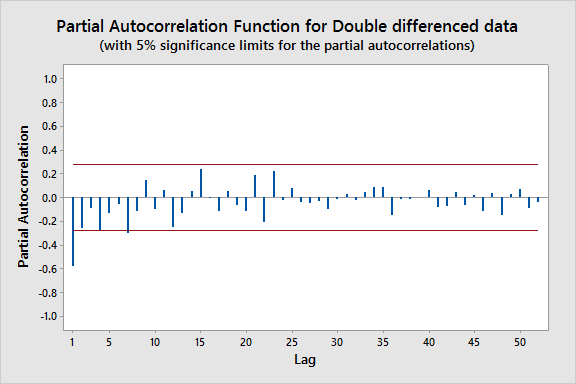
As we could see, there are apparent seasonal patterns in all above graphs: about every 12 months, Beijing’s PM 2.5 is strongly positive-correlated and about every six months, Beijing’s PM 2.5 is strongly negative-correlated. This phenomenon is because Beijing’s air pollution usually reaches its highest level in winter and reaches its lowest level in summer. To confirm the seasonal pattern, I made a time series plot for the seasonally differenced data:



As depicted above, this plot still does not have a stationary mean. Therefore, I applied the first difference to the seasonally differenced data:





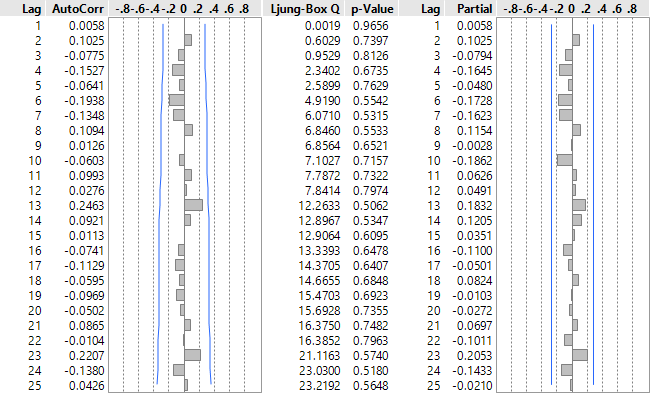


Now the above plot is stationary. There are also significant spikes at lag 1 in ACF and PACF plots while both graphs are sinusoidal. Though both plots have shown some seasonality as their autocorrelations exponentially decay at lag 12, lag 24, and lag 36, none of them show any significant spike at lag 12. Therefore, the best candidate model for the monthly PM 2.5 observations of Beijing is around Seasonal ARIMA (1,1,0) (0,1,0)12 or Seasonal ARIMA (0,1,1) (0,1,0)12.

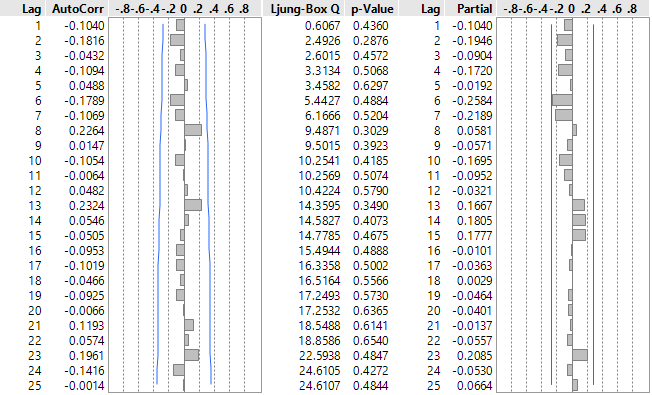
The Minitab worksheet was saved as a CSV file and opened in JMP. Because the monthly subset was estimated by the mean value of each month’s PM 2.5 observations, I had to drop the other variables that may have affected Beijing’s PM 2.5, like wind direction, as they were vectors and could not be averaged. Only a time series ARIMA could be made for the monthly PM 2.5 observations.

After multiple attempts, the two best ARIMA models from JMP were Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept and Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept. Their residuals’ ACF and PACF plots are shown below:

1. Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept



b. Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept.



Both the ACF and PACF plots are very clean because no autocorrelation coefficient exceeds their threshold. This phenomenon means, there is no autocorrelation in residuals under both models.

1. Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept

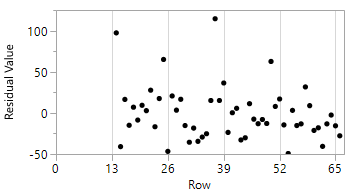


b. Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept

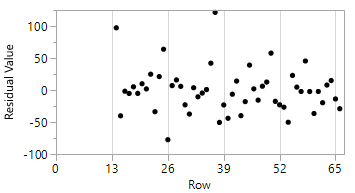


Both models are approximately significant as the p-values of their ARIMA coefficients are smaller or equal to 0.05.

1. Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept



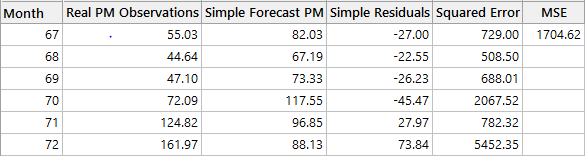
1. Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept



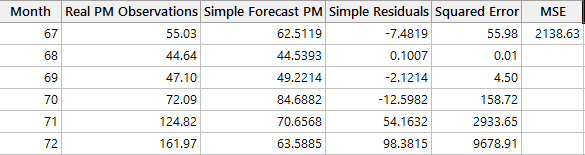
The residual plot is not very perfect as there are some outliers on the top of both models’ residual plots. This phenomenon is most likely because the Chinese Spring Festival at the beginning of each year always involves a lot of fireworks.

All above analysis proved that Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept and Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept fit the monthly PM 2.5 observations of Beijing very well. However, I needed to compare my model’s forecasts with the real out-of-sample observations in the last half year of 2015(6 months).

Seasonal ARIMA(0,1,1)(0,1,1)12:



Seasonal ARIMA(0,1,1)(1,1,0)12:



We could see, Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept’s MSE(1704.62) is much lower than Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept’s(2138.63). Therefore, Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept is the best available model to the forecast monthly average PM 2.5 of Beijing.

The explicit model equations of Seasonal ARIMA (0,1,1)(0,1,1)12 are given below:





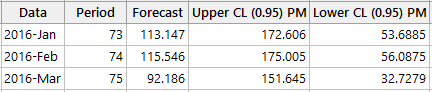
Therefore, the forecast equations for the next three periods are as follows:

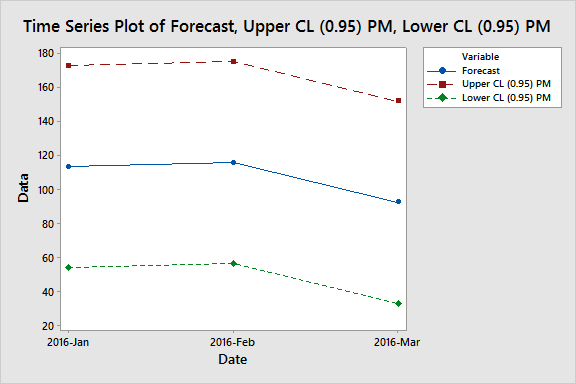






Based on the chosen model, I combined the in-sample and out-of-sample monthly observations and forecasted the average PM 2.5 value of the next year’s first three months (2016-Jan, 2016-Feb, 2016-Mar) with a 95% confidence interval:

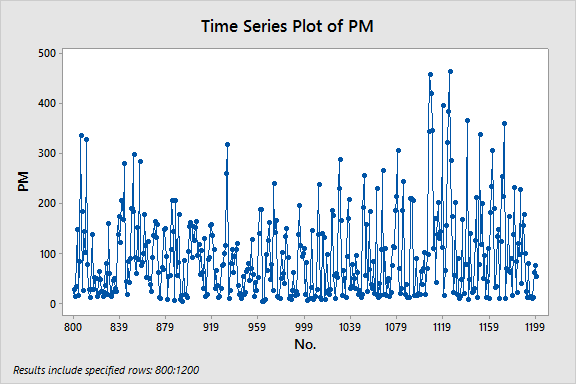
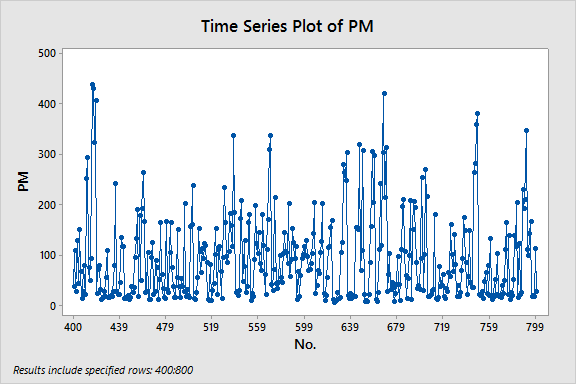
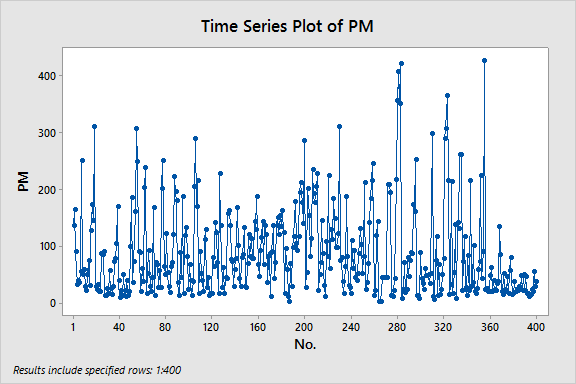
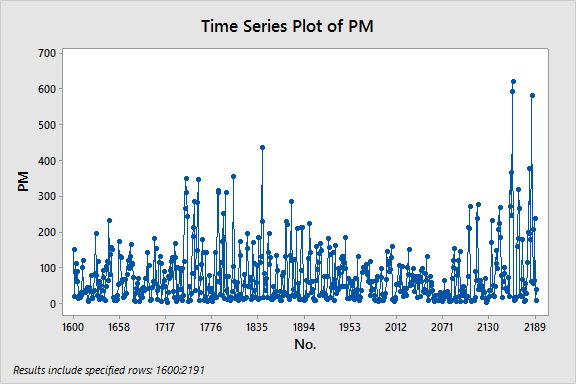
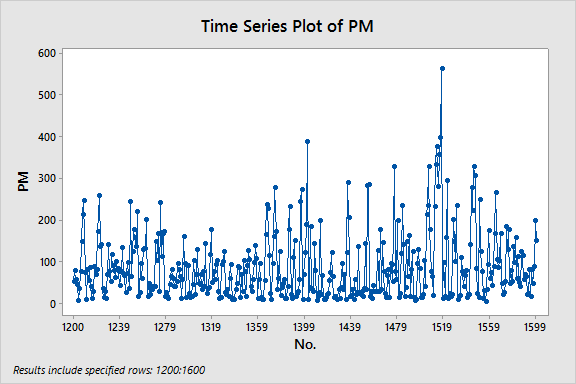




From the time plot, Beijing’s monthly PM 2.5 value would decrease in the next 3 months.

**Daily Data Analysis:**

I then started to study the forecast model for the daily PM 2.5 observations of Beijing. First, Minitab 7 was used to make the time series plot:

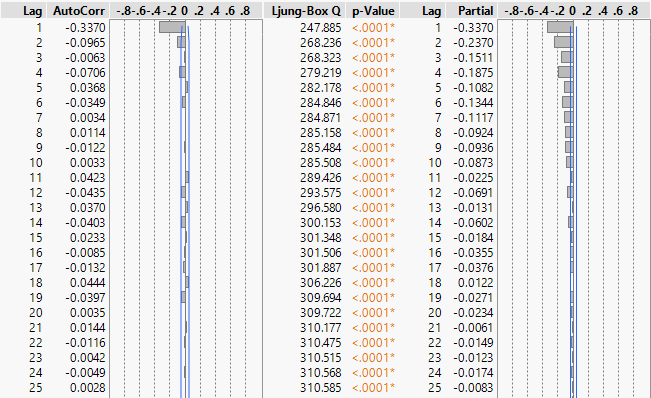
From the time series plot, there is no clear trend or seasonality of PM2.5 observation data. Thus, I used R to analyze this data set.

ARIMA models only include information from the past observations of a series. To make predictions for the future PM 2.5, Some other useful information like Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), Cumulative wind speed (m/s), and hourly precipitation (mm) are needed in the data set.

According to the KPSS test in R, PM 2.5 and Cumulative wind speed are not stationary. The p-values of each variable are given below:

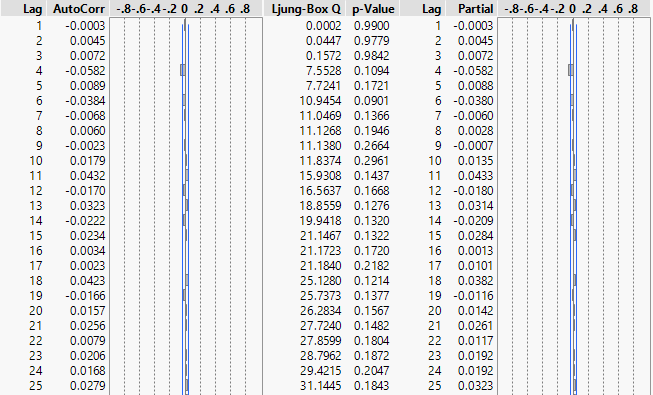
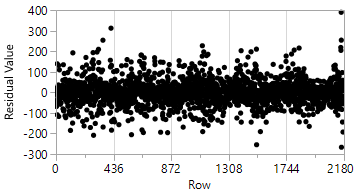
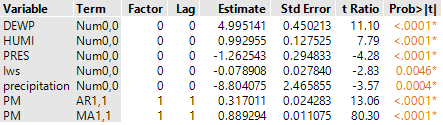
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PM | DEWP | HUMI | PRES | TEMP | Iws | precipitation |
| 0.04155 | 0.1 | 0.1 | 0.1 | 0.1 | 0.01093 | 0.1 |

Therefore, I needed to difference the data first and then applied the ordinary regression to the differenced variables in JMP. The ACF and PACF plots of the regression residuals are given below:

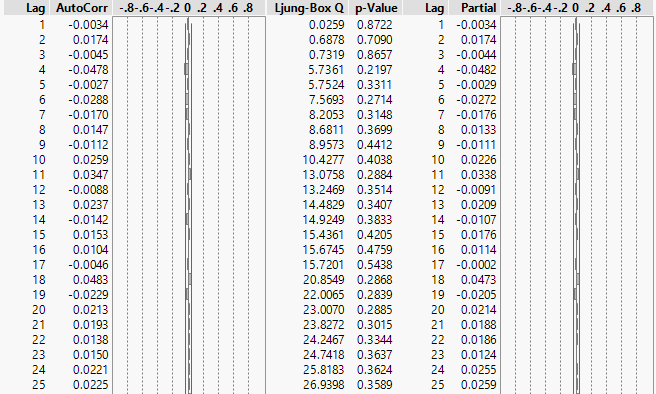
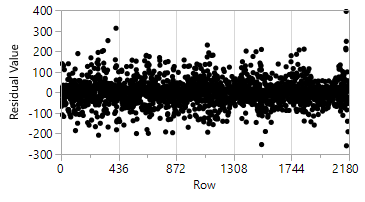
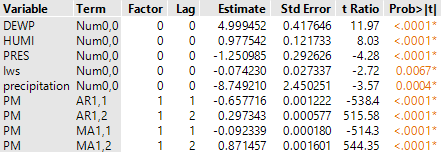


Obviously, the residuals are strongly autocorrelated. Therefore, it was necessary to fit autocorrelated regression noise instead with ARMA models. Based on the ACF and PACF plots of the ordinary regression residuals, multiple different models were tried in JMP and the two best candidates were: ARMA(1, 1) and ARMA (2, 2):

**ARMA(1, 1)**

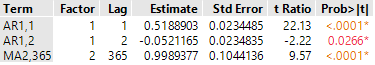


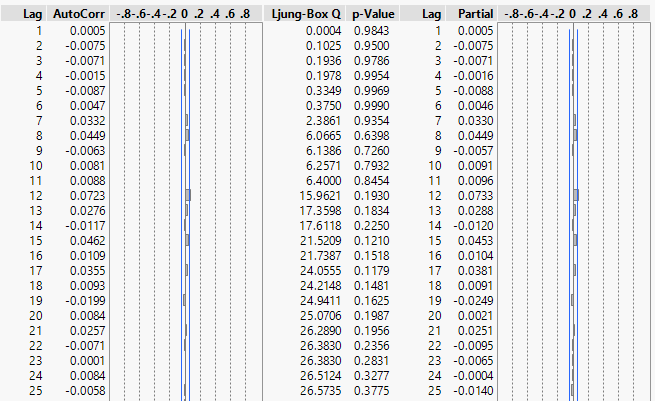
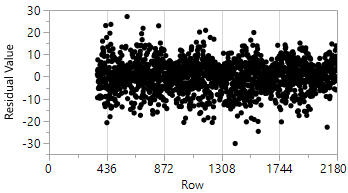
**ARMA(2, 2)**



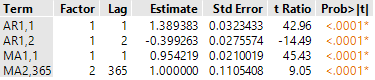
As seen from the results, all residuals are clean and all model parameters are significant. To choose a better model, I forecasted the out-of-sample data (12/22/2015 - 12/31/2015) and calculated the out-of-sample MSE. An ARIMA model had to be fitted for each predictor in JMP:

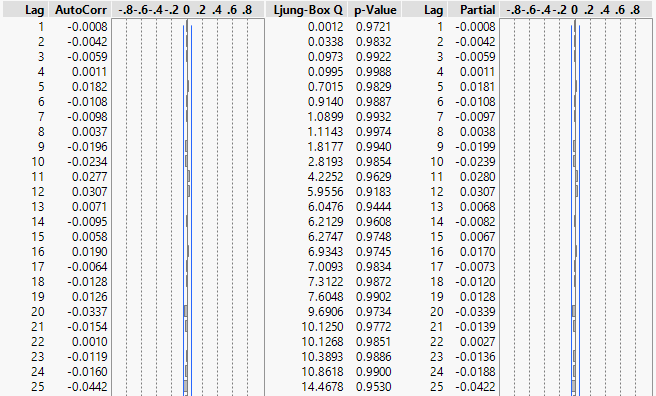
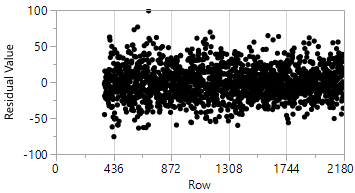
**Dew Point:** Seasonal ARIMA(2,0,0)(0,1,1)365 without intercept



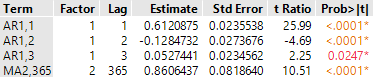


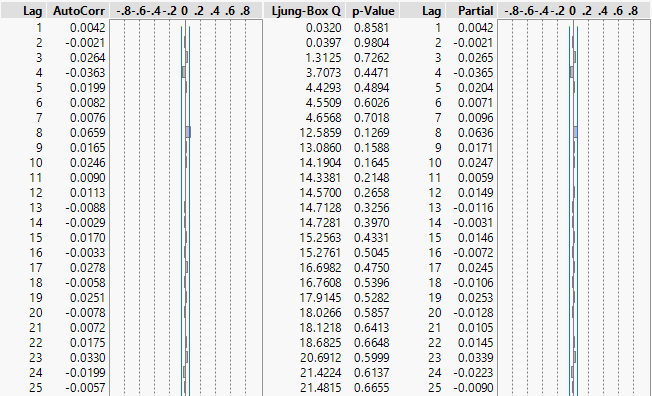
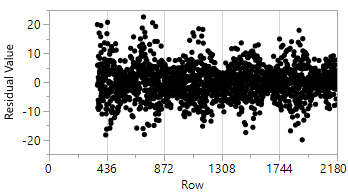
**Humidity:** Seasonal ARIMA(2,0,1)(0,1,1)365 without intercept



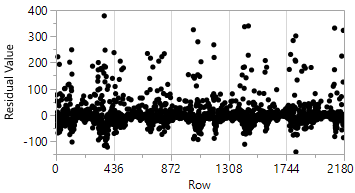
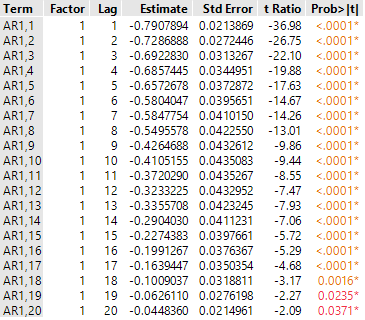


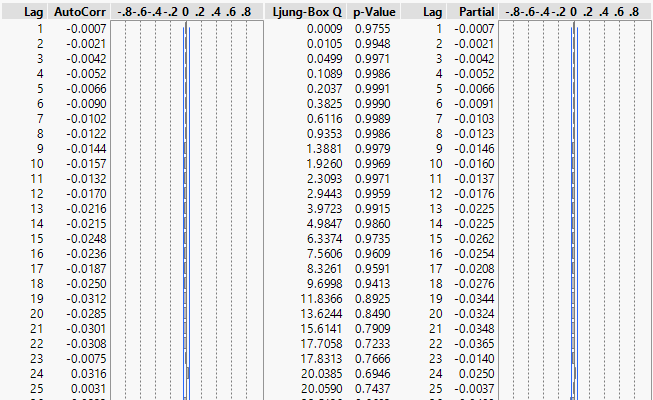
**Pressure:** Seasonal ARIMA(3,0,0)(0,1,1)365 without intercept



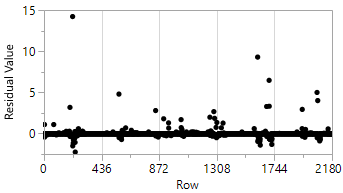
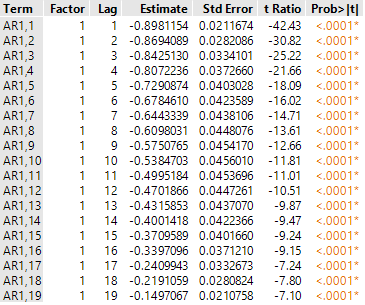


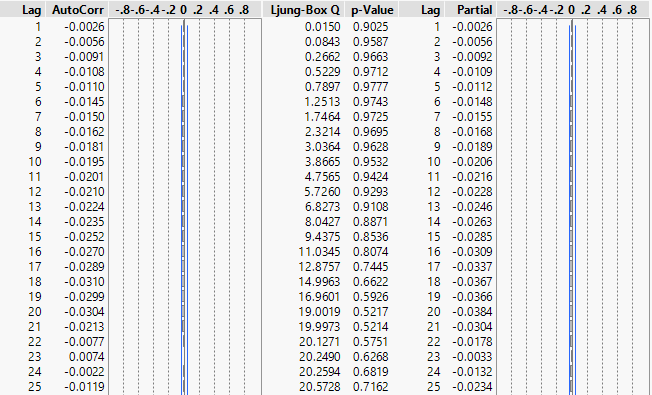
**Cumulated wind speed:** ARIMA(20,1,0) without intercept



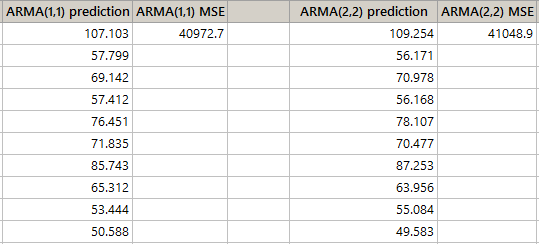


**Precipitation:** ARIMA(19,1,0) without intercept:



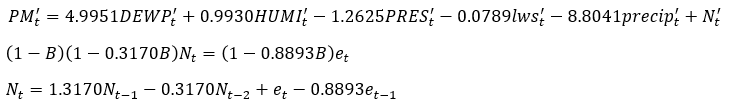


Based on these ARIMA models, both the related out-of-sample predictors and the out-of-sample PM 2.5 could be forecasted. I then compared these predictions with the real out-of-sample observations to derive the MSE:

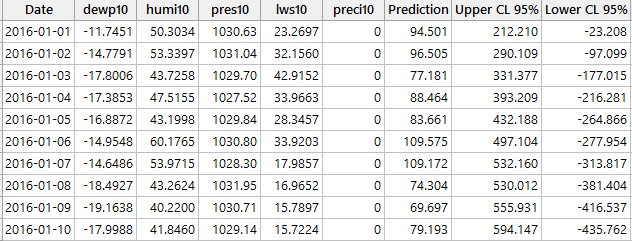


ARMA (1,1) has a smaller MSE than ARMA (2, 2), and ARMA (1,1) is less complicated than ARMA (2,2). Thus, ARIMA (1,1,1) was picked as the best model for the autocorrelated regression noise.

The explicit model equation is given below:



After validating these models, I forecasted the predictors of the next ten days. Then, I predicted the future ten days’ PM2.5 values based on these assumptive predictors.



Since PM 2.5 cannot be negative, the lower 95% confidence bound was set as 0. The time series plot for the PM 2.5 predictions in the next ten days（2016-01-01 to 2016-01-10） is shown below:



**Conclusion:**

Considering even the best model for daily forecast has an MSE larger than 40,000, it is challenging to predict the daily PM 2.5 in the future accurately. The later the predicted date is, the further the upper 95% confidence limit expands. Therefore, it is impossible to forecast Beijing’s long-term daily air pollution level, as the 95% confidence interval will only become larger and larger. However, in the short term, the daily predictions are still worthwhile.

On the other hand, because the monthly forecast model focuses on the average PM 2.5 of each month themselves and has excluded all independent variables, it has a comparatively smaller MSE (1704.62). Therefore, the predictions of the monthly average PM 2.5 are much more precise than the daily PM 2.5 forecasts, in both the short and long term.

According to the daily forecasts, Beijing's PM 2.5 would most likely be around 100 in the first ten days of 2016. People shall stay indoors if possible and wear masks while working outdoors.

However, as the monthly forecasts show, Beijing’s air quality will likely improve after February 2016 as March’s average PM 2.5 will drop about twenty percent compared to February’s average PM 2.5. Thus, I suggest Beijing residents practice outdoor activities in Spring 2016.

Based on my daily forecast model, precipitation has the most negative impact on PM 2.5 compared to the other predictors. This relationship is reasonable because rain can lower PM 2.5 by taking away the particulates in the air. Therefore, I recommend the Beijing government to create artificial rainfall when Beijing’s air pollution reaches a hazardous level. However, because artificial rain may have a potential negative impact on the macro environment of China, the most critical objective of Beijing is to set an environmentally friendly development scheme instead of the extensive growth it adopts now. It will surely bring temporary pain, but in the long term, it is the only way to fix Beijing’s horrible air pollution completely.

**Reference:**

S. X. Chen (2016). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities]. Beijing, China: Peking University, Guanghua School of Management, Center for Statistical Science.