Introduction

Air quality has been taken seriously all the time since it is closely related to everyone’s life no matter where you are. PM2.5, known as microscopic solid or liquid matter suspended in Earth's atmosphere, is almost produced by human activity like burning of fossil fuels. It may also cause cardiopulmonary disease, cancer of the trachea and lung in human. Study on PM2.5 has been becoming more and more essential.

This project is about analysis on PM2.5, and the important factors that might affect the air quality in Beijing. We would like to figure out what important factors significantly influenced PM2.5 and how to predict PM2.5. Our goal is fitting appropriate time series models to explore the association between PM2.5 concentration and any other weather factor such as temperature, humidity, and wind speed. We will also use the model to predict the PM2.5 a few steps into future to verify the validity of our model.

Data

Our data is on UCI Machine Learning Repository named ‘PM2.5 Data of Five Chinese Cities DataSet'. It includes the hourly data of PM2.5 and some other weather information in Beijing, Shanghai, Guangzhou, Chengdu, and Shenyang from January 1st

, 2010 to December 31st, 2015.

Respond variable: PM\_US.post (PM2.5 provided by US Embassy)

Predictor variables: TEMP(temperature),  HUMI(humidity), PRES(pressure), Iws(Cumulated wind speed), DEWP(Dew Point)

Data cleaning

For our project, we just focus on Beijing. We pick the maximum PM2.5 in PM\_US.post as our daily data and use na.ma function to replace the missing value of PM2.5 by simple moving average. Moreover, we found that there exist plentiful missing data of humidity and pressure from November 7th, 2015 to November 21st, 2015. Thus we fill these data by wunderground website using its daily average humidity and pressure.

ARIMA model

ARIMA models combine differencing with auto regression and moving average model together. In our assumption, ARIMA model would help us to achieve the best forecasting model since the same stationarity and invertibility conditions will be used for autoregressive and moving average models apply to the ARIMA model(p,d,q) when we are using differencing to stabilize the variance of our time series. The main process we are used here is: Try arima model with or without different groups of predictors at the beginning, then check their residual by ACF and PACF plotting to see whether or not look like white noise. Also observing the prediction by the model, comparing to the original data and compute the mean squared error. The best model would be selected from these comprehensive comparison.

For fitting ARIMA model without any regressors. We first try to use auto.arima() function in R that could help us to find appropriate values for the order and degree automatically. It uses Hyndman and Khandakar algorithm to minimize AIC, AICc, and BIC to obtain the best p, d, q. The output of auto. arima function is ARIMA(0,1,3) which is same as the result we got from the best.arima function(in tutorial7). With p=0, we just having integrated moving average model:IMA(1,3). According to the output from R we got the coefficient: -0.4474, -0.3142, and -0.1827. This is an ARIMA(0,1,3) or IMA(1,3) model (see attached table in Appendix):

After adding temperature and cumulated wind speed as regressor we still get ARIMA(0,1,3), but the acf and pacf of the residual does not seem improved there still have plentiful significant spike. Although the prediction is still not performing well, it became more sensitive to the decreasing of PM2.5 and predicted some decreasing point. In order to discuss our prediction convenient, we cut out the part of test set data(see attached Graph: prediction with temp&iws in Appendix). This is an ARIMA(0,1,3) or IMA(1,3) with 2 regressors temperature(x1) and wind speed(x2) (see attached table in Appendix):

Next, we add one more predictor humidity and get out new order ARIMA(1,2,2) this time. For this model we noticed that most of the residuals show that the model has captured the patterns in the data quite well, although there still exist a little autocorrelations in the residual according to the graph(see attached Graph: ACF of ARIMA model with temp&humi&iws in Appendix) which means we still need to improve. The result of ACF and PACF have almost achieved our expectation which is approaching white noise model. The final prediction also improved a lot(see attached Graph: prediction with temp&humi&iws in Appendix), though the increasing tendency being not obvious is a fly in the ointment. This is an ARIMA(1,2,2) with three regressors temperature(x1), humidity(x2), and wind speed(x3) model:

With predictor temperature, cumulated wind speed, humidity, and pressure, we get a better model. The residual graph got closer to white noise and the final prediction also appears a little increasing tendency as what we expect. Furthermore, our test set is nearly all in the prediction intervals (see attached Graph: prediction with temp&humi&iws&press in Appendix). This is an ARIMA(1,2,2) with four regressors temperature(x1), humidity(x2), wind speed(x3), and pressure(x4) model (see attached table in Appendix):

When we fitted model with all predictor temperature, cumulated wind speed, humidity, pressure and dew point, we did not get a more ideal outcome. And we compared the MSE of this model to the model without dew point variable, the mean squared error was getting bigger. Thus we comprehended that dew point is not relate to PM2.5 and we should not use it to predict PM2.5 concentration.

We also tried to test Seasonal ARIMA models, but the arima() function will only allow a seasonal period up to m=350 which our data have already exceed. Thus we decide to create the harmonic functions and then fit the model. However the output was disappointing, it only caught some micro-trend with a decreasing trend on the whole. The mean squared error is also the biggest in all ARIMA model we tried.

After comparing the ACF plot of the residuals, mean squared error and the prediction performance, we are able to conclude that ARIMA(1,2,2) with four regressors temperature, humidity, wind speed, and pressure model would be the best model we select to predict the PM2.5.