Air Quality Forecast Based on Influence of Meteorological Factors

Lauren Lingyun QU

August 14, 2024

I. Introduction

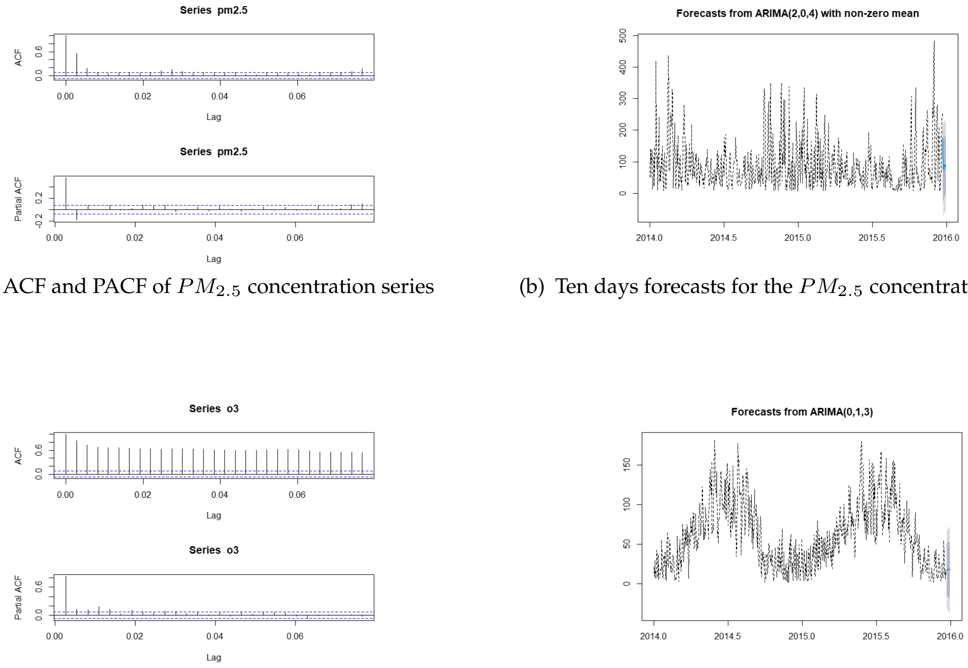
Nowadays, air pollution has become a serious problem. Beijing is positioned as top ranking polluted metropolitan cities in the world around 2019. Air quality forecasting is an effective way to protect public health by providing an early warning against harmful air pollutants. In this paper, Autoregressive Integrated Moving Average (ARIMA) and General Autoregressive Conditional Heteroskedasticity (GARCH) were used to investigate the cause of the air-pollution episodes. The data from January 2020 to December 2021 with 17520 observations were utilized in this study and are collected from the Bureau of Weather and Climate of China. The performance of the models was evaluated by RMSE value. Among the models, ARIMA with regression performed best for predicting $PM_{2.5}$ and O_3. The result suggests that unfavorable diffusion conditions (weak surface winds and high humidity) have induced heavy-haze pollution in the Beijing region over the two years.In order to examine the effect of emission reduction measures on pollutant concentrations, we must take into account the confounding effect of meteorological parameters, such as wind speed or air pressure. Here we explore the effect of meteorological variables on pollutant concentrations including $PM_{2.5}$ and O_3.

II. Hypotheses and Methodologies

A. Autoregressive Integrated Moving Average (ARIMA)

As we have mention in section II, ARMA model is an important model to deal with time series data which is stationary. If the data is non-stationary, we can extend the model to ARIMA model. Now we apply the ARIMA model to the our data. Firstly, we discuss the $PM_{2.5}$ concentration series. From the section II.C, we conclude that the series is stationary, so we can use the ARIMA(p,d=0,q) model, i.e. ARMA(p,q) model, without difference. Then, we need to determine p and q for the ARMA model. From the ACF plot and PACF plot 1(a), we can see the both the plots tail off. We can't directly choose the p and q from the plots. However, we can choose p and q with **AIC** criteria. We use the origin series except the last

10 days data as training data, and the last 10 days as testing data. After trying several different pairs of p and q on trainging data, we find that the ARMA model have a smaller AIC, 8112.98, when p is 2 and q is 4. It's important to note that we need to test whether the model is appropriate. Adopting the model ARMA(2,4) to the testing data, we get the predicted value and compute the $\textbf{RMSE} = 117.833$. And Figure 1(b) shows the observed data and predicted data and corresponding confidence interval with level 80% and 95%.



(a) ACF and PACF of $PM_{2.5}$ concentration series

(b) Ten days forecasts for the $PM_{2.5}$ concentration series

(c) ACF and PACF of O_3 concentration series

(d) Presentation of data for each instance.

For O_3 in Figure 1(c), we know that first order of difference in enough to make the series stationary, so we can use the $\textbf{ARIMA}(\mathbf{p}, \mathbf{d} = \mathbf{1}, \mathbf{q})$ model to analyse the data. As dealing with $PM_{2.5}$ concentration series, we use **AIC** to choose p and q. Similarly, We use the origin series except the last 10 days data as training data, and the last 10 days as testing data. ARMA model have a smaller AIC, 8112.98, when p is 0 and q is 3. Adopting the model ARIMA(0,1,3) to the testing data, $\textbf{RMSE} = \mathbf{11.192}$. And Figure 1(d) shows the observed data and predicted data and corresponding confidence interval with level 80% and 95%.

For PM2.5 concentration, we first employ all 7 weather factors with the data of 2 years except last 10 days to construct a linear model.

(1)

$$\begin{aligned}\mu_{1t} = &\beta_{01} + \beta_{11}TEMP + \beta_{21}PRES + \beta_{31}RAIN + \beta_{41}HUMI + \beta_{51}DEWP + \beta_{61}WS + \\ &\beta_{71}I(WD = NE) + \beta_{81}I(WD = NW) + \beta_{91}(WD = SE) + \beta_{101}I(WD = SW)\end{aligned}$$

As VIF of each variable is over 10, the collinearity exists among the variables used in the model 1. The model is over-fitted so we need to remove some redundant variables out of the model, which can be done by way of step regression method with AIC criteria.

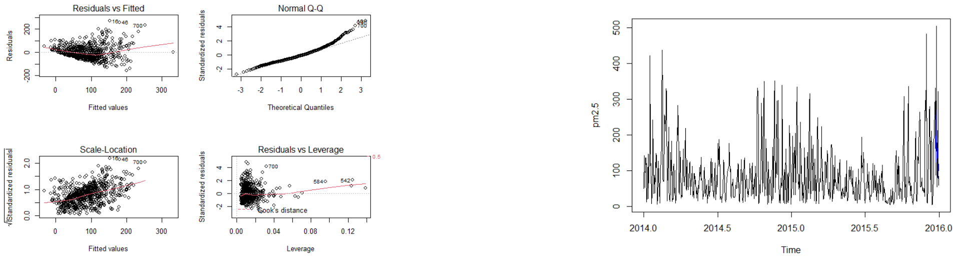
(2)

$$\begin{aligned}\mu_{1t} =& \beta_{01} + \beta_{21} PRES + \beta_{31} RAIN + \beta_{41} HUMI + \beta_{51} DEWP + \beta_{61} WS + \\ & \beta_{71} I(WD = NE) + \beta_{81} I(WD = NW) + \beta_{91}(WD = SE) + \beta_{101} I(WD = SW)\end{aligned}$$

The same as just now, we compute the VIF of each coefficients in the model 2 and the results are as follow. Now the VIF of all variables are below , which means that the there is weak collinearity existing among the variables used in the model 2, and The estimated model is:

(3)

$$\begin{aligned}\mu_{1t} =& 3133.4058 - 2.9591 PRES - 60.7432 RAIN + 3.1725 HUMI - 5.2055 DEWP \\ & - 12.1075 WS - 178.8003 I(WD = NE) - 199.4114 I(WD = NW) \\ & - 170.0278(WD = SE) - 174.9247 I(WD = SW)\end{aligned}$$



(e) Residual plots $PM_{2.5}$

(f) Predicted values of the $PM_{2.5}$ concentration of last 10 days

From first column of figure $1(e)$, we suspect that the errors are heterogeneous and auto-related. And both the p-values from Breusch-Pagan test and Durbin Watson test are below 0.05, which confirm our hypothesis.

Now we analyse $\hat{Y_{1t}}$ in the same way as processing x_{1t}. To begin with, we test whether the series $\hat{Y_{1t}}$ is stationary. The p-value from the ADF test is below 0.01, so we can conclude that this series is stationary. From the acf and pacf plot, we choose the order as $p = 1, d = 0, q = 0$. Actually it is also a AR(1) model. From the figure $2(b)$, we reckon that the model is proper. Adopting the model to the last 10 days' data, we can get the predicted value of $\hat{Y_{1t}}$ of the last 10 days, i.e. t is from 2015-12-23-2015-12-31. With $\hat{\mu_{1t}}$ and $\hat{Y_{1t}}$, where t is from 2015-12-23-2015-12-31, we can get $\hat{x_{1t}}$, the predicted $PM_{2.5}$ concentration of the last 10 days. And the **RMSE** is 97.194.

For O_3 concentration, from the scatter plot of all 7 meteorological variables, we include all of them in the linear model. The variables temperature and dew point temperature are included in quadratic form and the others are in linear form.

(4)

$$\begin{aligned}\mu_{2t} =& \beta_{02} + \beta_{12} TEMP^2 + \beta_{22} PRES + \beta_{32} RAIN + \beta_{42} HUMI + \beta_{52} DEWP^2 + \beta_{62} WS + \\ & \beta_{72} I(WD = NE) + \beta_{82} I(WD = NW) + \beta_{92} (WD = SE) + \beta_{102} I(WD = SW)\end{aligned}$$

From the VIF of each coefficients in this model, which are all below 10, there is weak collinearity existing among the variables used in the model 5. The estimated model is:

(5)

$$\begin{aligned}\mu_{2t} =& 68.20.7 + 0.1108 TEMP^2 - 0.0556 PRES + 8.9147 RAIN - 0.2861 HUMI \\ & + 0.0235 DEWP^2 + 12.6430 WS - 0.9834 I(WD = NE) - 10.3988 I(WD = NW) \\ & + 6.0290 (WD = SE) + 8.0613 I(WD = SW)\end{aligned}$$

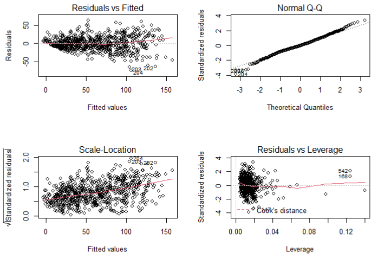
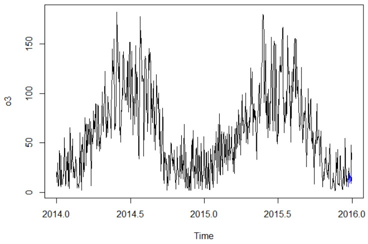


Figure 2. : Residual plots O_3

() Residual plots O_3

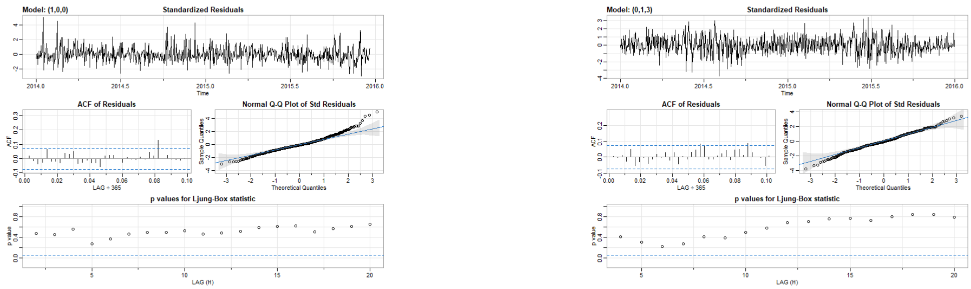


(a) Predicted values of the O_3 concentration of last 10 day

At the same time, with the meteorological factors of last 10 days, we can predict the $\hat{\mu_{1t}}, t$.

From first column of figure 2, we suspect that the errors are heterogeneous and auto-related. And both the p-values from Breusch-Pagan test and Durbin Watson test are below 0.05, which confirm our hypothesis. So we need to analyse series $\hat{Y_{2t}}$

Now We analyse $\hat{Y_{2t}}$ in the same way as processing x_{2t}. To begin with, we test whether the series $\hat{Y_{2t}}$ is stationary. The p-value from the ADF test is below 0.01, so we can conclude that this series is stationary. From the acf and pacf plot, we choose the order as $p = 1, d = 0, q = 0$. Actually it is also a AR(1) model. From the figure $2(c)$, we reckon that the model is proper. Adopting the model to the last 10 days' data, we can get the predicted value of $\hat{Y_{2t}}$ of the last 10 days.



(b) Sample ACF and PACF of the Y_{1t} concentration. Lag is in terms of days

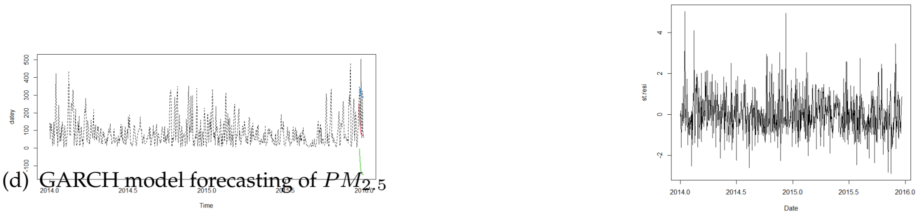
(c) Sample ACF and PACF of the Y_{2t} concentration. Lag is in terms of days

B. General Autoregressive Conditional Heteroskedasticity (GARCH)

GARCH without regression

As ARMA models without regression of weather variables were used to model the conditional mean of a process when the conditional variance was constant and we notice these time series of $PM_{2.5}$ and O_3 are highly volatile periods tend to be clustered together. so with fitted ARIMA model of $PM_{2.5}$ and O_3, we think that use ARCH or GARCH model to develop to model changes in volatility.

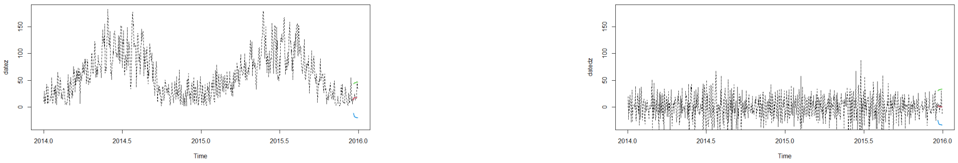
Then we use this ARMA-GARCH model to forecast concentration of $PM_{2.5}$ in the last ten days and plot as Figure 2(d). In this model, we compare prediction with true data and get RMSE=177.26 and mean bias=125.19 with these ten predicted values. Therefore, we can conclude that the analysis of $PM_{2.5}$ sequences using the ARMA-GARCH model is better than using ARMA model alone. However, just using ARMA-GARCH model for time series of $PM_{2.5}$ does not yield an accurate prediction value.



(d) GARCH model forecasting of $PM_{2.5}$

(e) Residuals of GARCH model on regression of $PM_{2.5}$

When it comes to time series of O_3, according to ARIMA model section, we know that the stationary of O_3 is not satisfied and first order difference of O_3 satisfies stationary. So we fit an ARMA(0,3)-GARCH(2,3) for the first order difference of O_3. Other steps is the same as forecasting the concentration in last ten days of 2015. The result of prediction of first order difference of O_3 is plotted as Figure 2(g). And use such forecast we could also get the prediction of O_3 and it is shown in Figure 2(d). through this model to predict the concentration of O_3, we calculate RMSE=11.35 and mean bias=21.26. Obviously, the prediction for ozone is better than that for $PM_{2.5}$. This result is consistent with ARIMA model.



(f) GARCH model forecasting of O_3

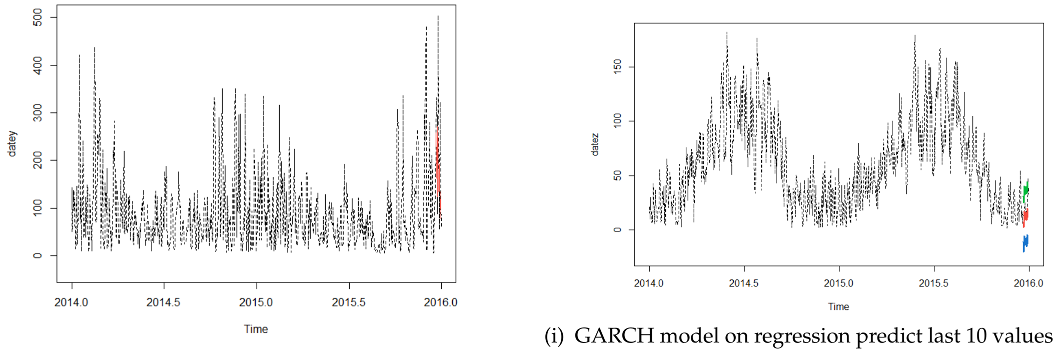
(g) GARCH model forecasting of first order difference of O_3

GARCH WITH REGRESSION OF WEATHER VARIABLE

Then we consider to use the residuals of regression of $PM_{2.5}$ and O_3 on weather variables to fit a GARCH model as the p-value of ARCH-test on residuals of ARIMA model are both less than 0.05.(0.01561 of $PM_{2.5}$ and 2.415×10^{-10} of O_3). So after fitting a linear regression model on meteorologrical variables, it is still reasonable to use a GARCH model.

For $PM_{2.5}$, with a AR(1) model fit in last section, we use a AR(1)-GARCH(1,1) choosed by AIC. The residuals of AR(1)-GARCH(1,1) plot as Figure 2(e). The Arch test of these residuals is 0.9837, so after fitting a GARCH model, there is no arch affect on residuals. And the ACF and PACF of residuals and squares are plot as Figure **??** . So we deem that it is suitable for residuals of regression to fit AR(1)-GARCH(1,1). Then we use this model to predict the concentration of $PM_{2.5}$ in last 10 days of 2015. We add the values estimated by regression on result of GARCH model as the forecasting value and we plot as Figure 2(h). The RMSE is 97.02685 and bias is 76.001, both are less than GARCH model without regression on meteorological factors.

For O_3, with a MA(1) model in last section, we use a MA(1)-GARCH(1,1) with a lowest AIC score. The same as $PM_{2.5}$, use it to predict concentration of O_3 in last 10 days of 2015 as shown in Figure 2(i).The RMSE is 13.2318 and bias is 10.1425 is less than GARCH fitted on raw data of O_3.



(h) GARCH model on regression predict last 10 values of $PM_{2.5}$

(i) GARCH model on regression predict last 10 values of O_3

Figure 2. : Forecasting.

III. Discussion

This analysis is still not perfect because of following constraints. Firstly, the data is only from one resource. Although the Bureau of Weather and Climate is an authoritative institution, if data from multiple resource is available, we may acquire a more accurate result.

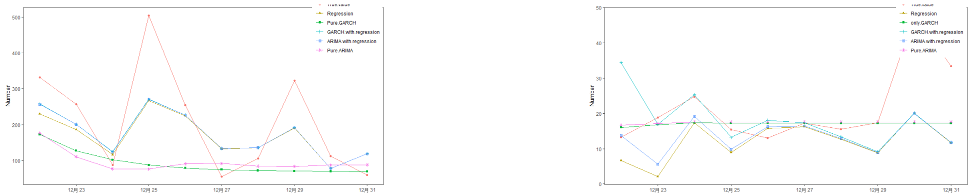
Second, this analysis assumed that the data is without seasonality. However, the concentration of particles can fluctuate during one year. This may lead to inaccuracy in the research.

Lastly, the research may expand the time range and take into account of some special events in China. For example, when there are important meetings, some vehicles are forbidden to travel. Change in number of cars will change the pollutant in the air, and may introduce changes to the weather quality. Therefore, the research may step further by eliminating the special days and analysis within the normal days.

For $PM_{2.5}$, ARIMA and GARCH model with linear regression on meteorolograical variables could reduce the RMSE of forecasting compared with only use, so we think that there is correlation of meteorolograical variables and concentration of pollution and it is useful to take meteorolograical variables in to account to predict the concentration of $PM_{2.5}$. However, it is not true for O_3, it is possible that the linear model is not suitable or the correlation between meteorological variables and concentration of O_3 is not significant.

With all ten values pridicted by these method and we plot them in Figure 3(a) and Figure 3(b). When we considered the prediction of $PM_{2.5}$, with a lower RMSE of ARIMA and GARCH model with linear regression of meteorological variables, they also show well the volatility of $PM_{2.5}$ concentration, such volatility is not shown in other methods. So for $PM_{2.5}$, ARIMA with regression methods is simpler than GARCH and its RMSE is just 0.1 more than GARCH.

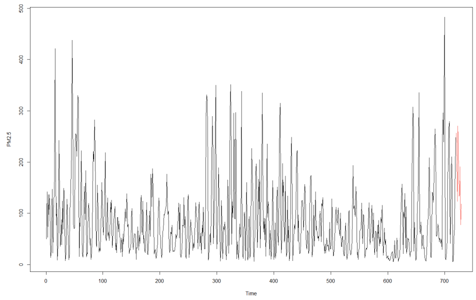
When it comes to O_3, it is different from situation of $PM_{2.5}$. ARIMA or GARCH straightly on concentration of O_3 has a lowest RSME as 11.192. However, it is so stable that not reflect the volatility of O_3 concentration. So with a little higher but better reflection of volatility, we still choose ARIMA with regression methods which is the same as in situation of $PM_{2.5}$.



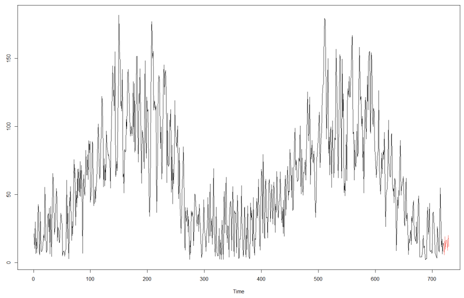
(a) All method forecasting concentration of $PM_{2.5}$ in last 10 days of 2015

(b) All method forecasting concentration of O_3 in last 10 days of 2015

The prediction of concentration of these two pollutent is shown in table 1. Including other values, we plot Figures as Figure 3(c) and Figure 3(d). The red lines show the value we forecasting for concentration of two pollutants.



(c) Forecasting of $PM_{2.5}$



(d) Presentation of data for each instance.

Table 1—: Forecasting of last 10 days

DATE	12/22	12/23	12/24	12/25	12/26
$PM_{2.5}$	256.773	200.299	123.528	270.443	226.581
O_3	13.728	5.602	19.125	9.8	16.228
DATE	12/27	12/28	12/29	12/30	12/31
$PM_{2.5}$	133.336	135.842	191.371	77.932	118.587
O_3	16.425	12.864	8.924	19.987	11.753

IV. Conclusion

In this paper, we construct several models to explore the effect of meteorological variables, on pollutant concentrations including $PM_{2.5}$ and O_3. ARIMA with regression performs relatively perfect. Regression on meteorological factors controls the confounding effect of weather and ARIMA model accounts for the latent pattern of $PM_{2.5}$ and O_3 series.

The characteristics of the surface meteorological variables during moderate and severe haze pollution episodes in the Beijing region are as follows: weaker pressure, higher temperature, particularly in the plains region, higher relative humidity, weak winds and lower visibility. Orographic wind convergence zones resulted in the pollution accumulation in the piedmont plain and restrained the diffusion of pollutions; as a result, severe regional haze pollution developed. Recirculation and regional transport, along with the poorest diffusion conditions and favorable secondary transformation conditions under high emissions and the hygroscopic growth of aerosols, led to the explosive growth and the highest hourly average concentration of PM2.5 in Beijing. Considering that decreasing wind speeds and weakened southerly winds resulted in more stable atmospheric conditions and weaker dispersion abilities, an effort should be made to control emissions and prevent the air polluting episode.