Beijing AQI Forecast

Xinnuo Lin, Harrison Zhu

June 7, 2018

Contents

1 Summary

Pollution is becoming a rising issue in the world today. In China, many of the 1.4 billion citizens live in Air Quality Index (AQI) levels (see [2]) significantly higher than the 'good' 0-50 interval recommended by the WHO. In particular, as the capital of the fast-growing nation, Beijing has long been suffering from the consequences of the nation's years of rapid industrialisation.

In this study, we used a pollution dataset collected online [3] concerning Aoti Centre in Beijing. We investigated the feasibility of building Autoregressive Integrated Moving Average (ARIMA) models to model the AQI. During the process, we considered a log transformation of our original data to stabilise the variance. We also considered stabilising the seasonal variance of our data by using a 'quasi-averaged' norm. As a result, we were able to conclude our study with an **ARIMA**(2,1,2) model on a log-transformed-stabilised data and obtain predictions that were then mapped back to the original series' scale. In particular, with 57 observations we obtained an error rate, that is defined by the number of points lying outside of the confidence bands, of 17.5%.

2 Initial data analysis and treatment

2.1 Description of the dataset

The dataset we used contains AQI at Aoti Zhongxin, Beijing, China between the 2nd of January 2015 to the 30th of January 2018. That is, 1125 data points. The method used to

calculate the AQI can be found here [2]. The data is subject to large fluctuations and the exact trend is difficult to spot by eye. This may be due to strong government efforts to crack down on pollution and the sporadic international events taking place in Beijing that require factories nearby to shutdown. There are also many missing values in this dataset, potentially due to sampling errors at the weather stations. The original time series is shown below (see Figure 1).

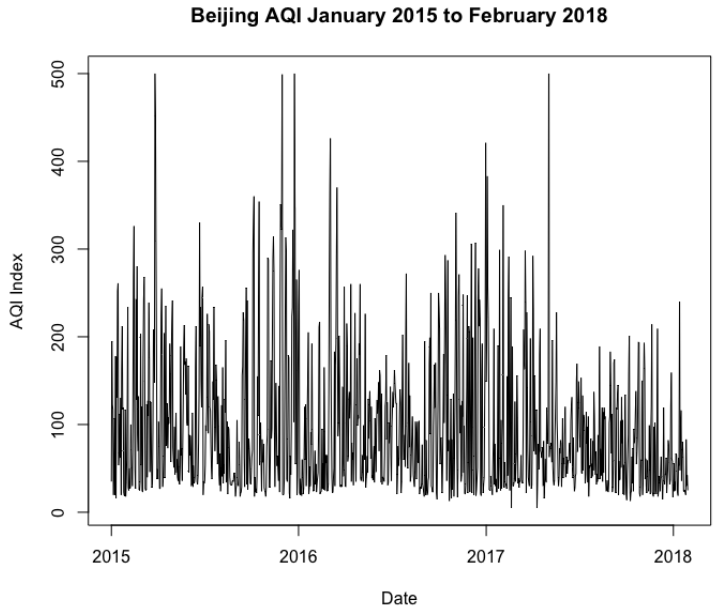

Figure 1: Plot of the original time series.

2.2 Data cleaning

There were less than 100 missing values in the original series and we sample the AQI at 12:00 pm, from the 2nd of January, 2015 to the 30th of January, 2018. We then imputed the dataset by taking the average between the 2 nearest times earlier and later than 12:00 for each day. For example, we would take the average of the AQI values at 11:00 and 13:00. This may have consequences on the accuracy of our models, but as there were less than 100 missing values compared to the size of the dataset, which contains 1000+ data points, this incurs only a minor penalty.

2.3 Step 1: Log transformation

Initially, we investigated the plot of the time series based on the processed data to uncover its structure. We can observe that the variance is time-dependent and there seems to be periodicity in the variance. Furthermore, looking at the ACF and PACF of the data (not shown, but the code can be supplied upon request), the autocorrelation is clearly large at lag 1. Hence, we can see that the data is not stationary.

In order to stabilize the variance, we first take the log transformation and denote the new time series by $LY_t := \log Y_t$. But it still looks like that the period of variance exists according to the time series plot of LY_t. Since the daily data indicates the air quality on the specific day, we assume that the periodicity of the variance should be 1 year. By separating $LY_t = \log Y_t$ into three parts of equal length of 365, denoted by I_1, I_2, I_3, the pattern of the variance (see Figure 2) in each part looks similar which somewhat supports our guess.

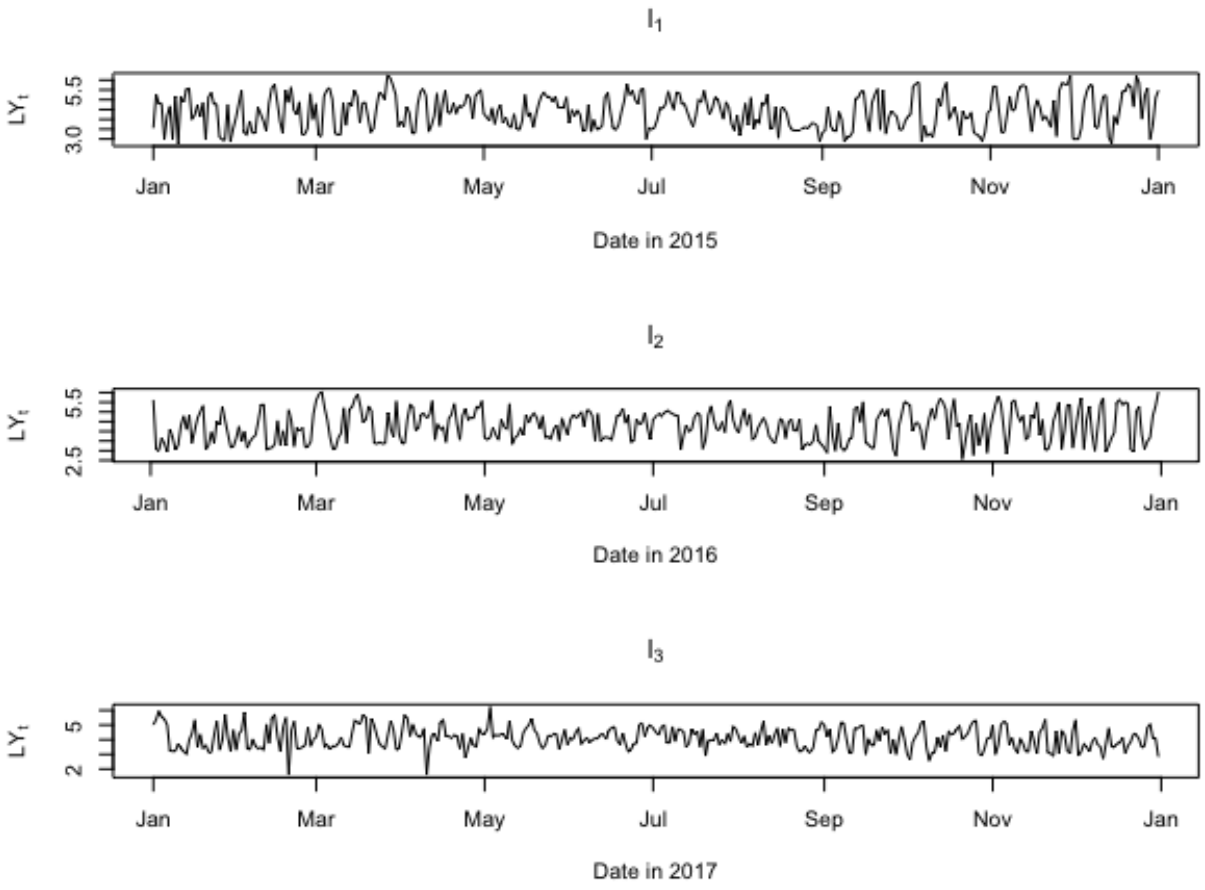

Figure 2: 3 equal parts of the log-transformed data LY_t: I_1, I_2, I_3.

2.4 Step 2: Quasi-averaged norm

To stabilise the time series for our analysis, we need to rescale series by dividing by the quasi-averaged norm $||\theta||_2/\sqrt{14}$, where $||\cdot||_2$ denotes the Euclidean norm.

In order to be as accurate as possible, we not only use data from the first 365 observations, but also the following 2 years of observations to do so, as there is a clear repetition in the size of the variance. To calculate the quasi-averaged norm of one particular day, the size of the sample data cannot be too big otherwise it cannot represent the variation of one exact

day. Thus, we only collect consecutively five values at a time from LY_t, from time $t, t+365$ and $t+365\times 2$, and use these values to obtain a quasi-averaged norm for these days. This pooling of 5 days is done because we only have data from the past three years data so it is not large enough to calculate the variance of one day accurately. Thus the quasi-averaged norms we need for time t, where $1 \leq t \leq 365$, we call $s_t^{(1)} \in \mathbb{R}^{365}$ are estimated as

$$s_t^{(1)} := \left\{ \begin{array}{ll} \sqrt{\frac{1}{14}\sum_{i=0}^{2}\sum_{j=0}^{4}(LY_{t+j+365i})^2}, & \text{if } t \bmod 5 = 1 \\ s_{t-4}^{(1)}, & \text{if } t \bmod 5 = 0 \\ s_{t+1-(t \bmod 5)}^{(1)}, & \text{otherwise.} \end{array} \right\} \tag{1}$$

As we are multiplying this value back later on when we model, it does not change the nature of our data.

3 Modelling

We considered the following 2 families of models:

1. ARIMA$(p, 1, q)$ with original series Y_t divided by the quasi-averaged norm: $Z_t = Y_t / s_t^{(1)}$.
2. ARIMA$(p, 1, q)$ with log-transformed series $\log LY_t$ divided by the quasi-averaged norm: $Z_t = \log LY_t / s_t^{(1)}$.

We can already tell by visualising the plots (see Figure 3) that the 2 differenced series from the 2 families are stationary. We gain further statistical confidence that the series are stationary after conducting KPSS tests at 5% level. We modelled using family 2 because by taking the log transform we were able to stabilise and reduce the scale of the variance ([1] , slide 28). To pick our best model, we considered the AIC, likelihood ratio statistics between nested models, significance of model parameters, but also model diagnostics such as the Ljung-Box white noise assessment, the Q-Q plot, the ACF and PACF of the modelled series and residuals, and finally parsimony.

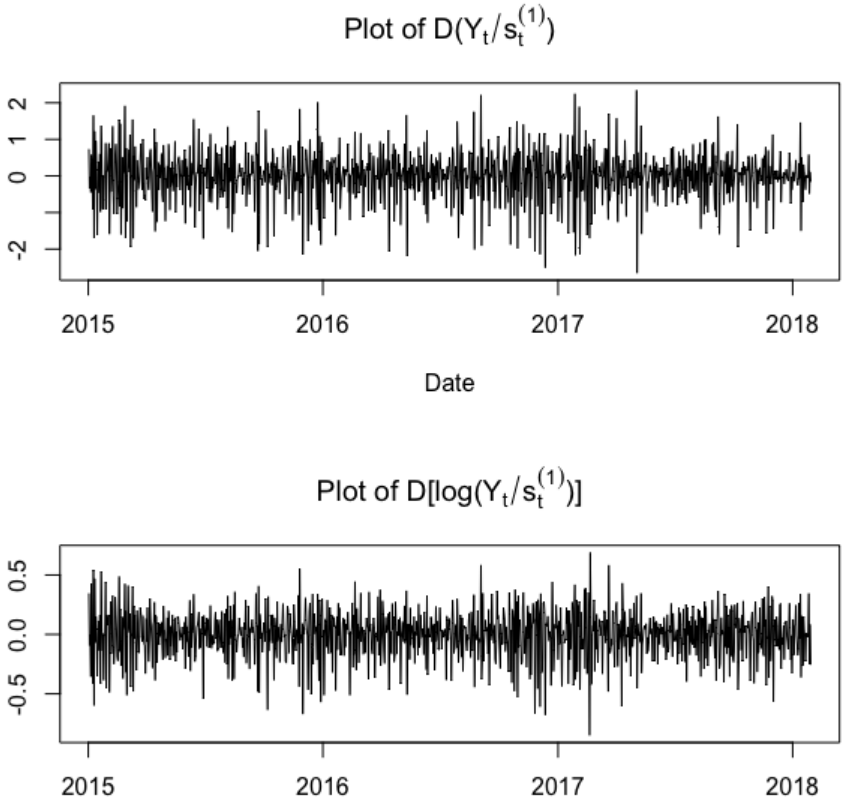

Figure 3: Plots differenced families 1 (top) and 2 (bottom) series.

3.1 Family 2

With the series $Z_t = \log LY_t/s_t^{(1)}$, we assess the feasibility of fitting an ARIMA(p, d, q) model. For this family, we first examined the ACF/PACF plots (see Figure 4) of the series $DZ_t = D\{\log Y_t/s_t^{(1)}\}$, where $D := I - B$ is the difference operator, and they suggest that we should opt for **ARIMA(0,1,3)** or alike models. The reasons are that the ACF **cuts off** at lag = 3 and the PACF **tails off**. For completeness, we stored the AICs for all ARIMA$(p, 1, q)$ with $0 \leq p, q \leq 5$ in Table 3.

In addition, we can already conclude that the pure white noise and autoregressive models are not suitable models by looking at the high AIC values compared to all the other fitted models (see Table 4). In addition, by looking at their residual plots (not shown, but the code can be provided upon request), they do not satisfy the basic white noise assumptions, and therefore are not suitable models for our study.

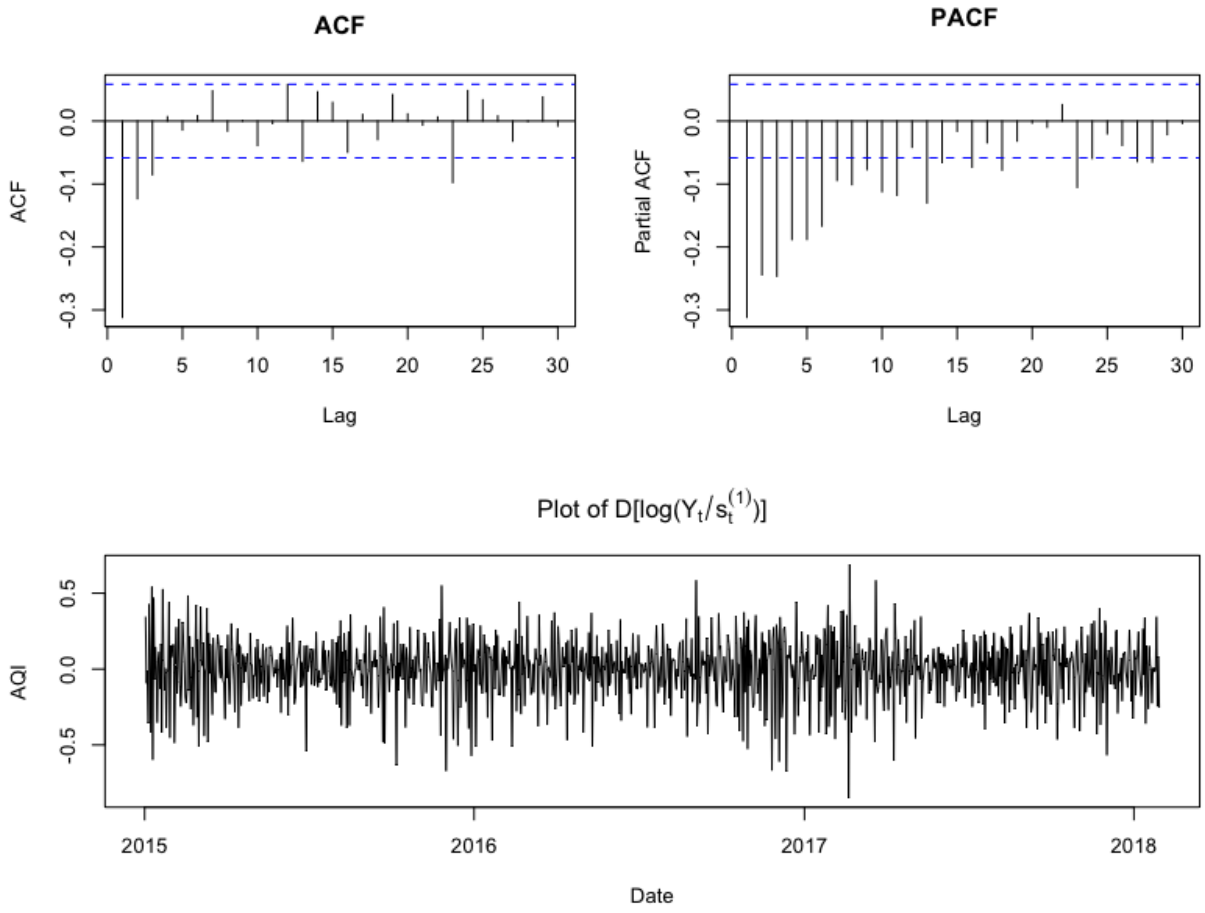

Figure 4: Initial plots of $\{DZ_t\}_t$, family 2.

3.2 ARIMA(0,1,q) Models

We first study the ARIMA(0,1,q) models as the correlograms suggest doing so. Looking at the log-likelihoods (see Table 1), we can see that the log-likelihood does not increase much when we go beyond order $q = 2$. Using log-likelihood ratio test with χ^2_1 sequentially, we can statistically check that perhaps simpler models can also be suitable.

	MA0	MA1	MA2	MA3	MA4	MA5
Log-likelihood	215.86	423.29	466.38	467.45	470.84	471.90
AIC	-429.71	-842.58	-926.76	-926.90	-931.69	-931.80

Table 1: Log-likelihood values for ARIMA(0,1,q) models with q from 0 to 5.

Fitting an **ARIMA**$(0, 1, 2)$ as suggested by the correlograms and a low AIC value of -926.76 (see Table 3), we obtain the model:

$$DZ_t = \epsilon_t - 0.72\epsilon_{t-1} - 0.27\epsilon_{t-2}$$

where $\epsilon_t \overset{i.i.d.}{\sim} N(0, 0.0255)$. A quick sanity check on the roots of the AR and MA polynomials reveals that we do in fact have causal and invertible processes as the roots are outside the unit disk $\mathbb{D} \subset \mathbb{C}$. Using a likelihood ratio test again with ARIMA(0,1,3), we obtain a p-value of 0.14, indicating that we should keep the simpler model, ARIMA(0,1,2). Indeed, by looking at the confidence intervals of the coefficients of ARIMA(0,1,3), it produced a coefficient such that the 95% interval of that coefficient contains 0. Hence, we may have overparameterisation in the moving average polynomial.

However, after examining the diagnostic plots (Figure 5, Q-Q plot, Ljung-Box assessment and cumulative periodogram), we can see that there are issues with normality of the residuals and the noise being white noise. In particular, most of the points plotted on the **Ljung-Box** plot seem to be below the 5% bar, indicating severe violations of the residuals being white noise assumption. Despite this, the **cumulative periodogram** provides a visual test to assess whether the residuals are white noise or not ([1], slide 95; possible since our transformed series is stationary). As we can see, the points clearly lie within the confidence bands, indicating that statistically we are confident that we have white noise residuals. The **Q-Q plot** also sees most of the points lying near the 45 degree line, thereby supporting normality in the residuals.

As a result, there is contradiction of our assumptions of the residuals being white noise and so we should see if adding autoregressive elements to our model will be more suitable.

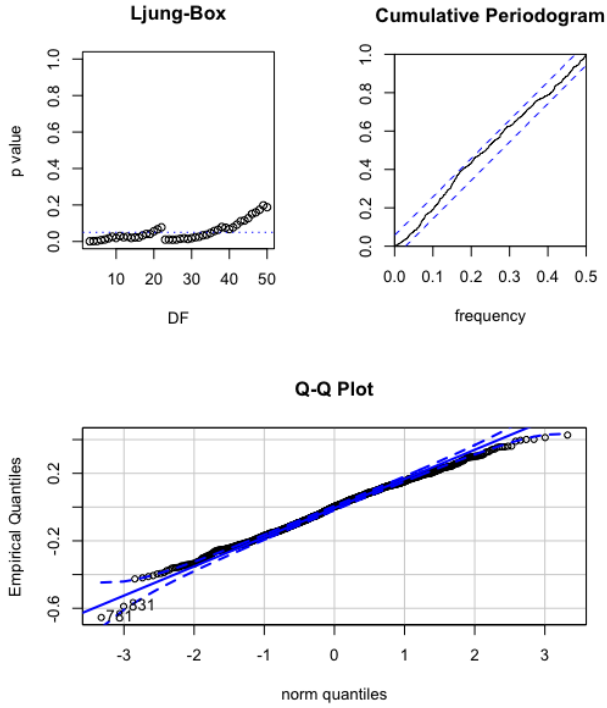

Figure 5: Diagnostic for ARIMA(0,1,2) of $\{DZ_t\}_t$, family 2.

3.3 ARIMA($p, 1, q$) Models

We begin by calculating the log-likelihoods of several ARIMA($p, 1, q$) models (Table 2). We see that there is not much increase in log-likelihood as we increase p or q beyond $(p, q) = (2, 2)$. Again, using likelihood ratio tests we have statistical backing for choosing $p, q \approx 3$. As a small remark, notice that the logliklihood has somehow decreased when we changed from ARIMA(2,1,2) to ARIMA(2,1,3). This may be because the 2 models are so similar that the maximum likelihood procedure in R has given us a very similar, albeit smaller, likelihood value. Therefore, it is immediate that with a likelihood ratio statistic we would prefer the simpler model ARIMA(0,1,2).

	MA0	MA1	MA2	MA3	MA4
AR0	215.86	423.29	466.38	467.45	470.84
AR1	273.52	464.66	467.02	468.66	473.58
AR2	308.18	468.50	472.89	472.74	473.64

Table 2: **Loglikelihood** values for ARIMA($p, 1, q$) models.

Assessing the AIC for a collection of models (Table 3), we see that in fact **ARIMA(2,1,2)** seems to statistically fit well, with an AIC of -935.79. That is:

$$DZ_t - 0.98B[DZ_t] + 0.27B^2[DZ_t] = \epsilon_t - 1.69\epsilon_{t-1} + 0.69\epsilon_{t-2} \tag{2}$$

with $\epsilon_t \overset{i.i.d.}{\sim} N(0, 0.025)$. Again, a quick sanity check on the roots of the AR and MA polynomials reveals that we do in fact have causal and invertible processes as the roots are outside the unit disk $\mathbb{D} \subset \mathbb{C}$. In addition, none of the 95% confidence intervals of the coefficients of the model contain 0, and hence indicating that they are indeed significant at 5% level.

This seems like a good model in fact and is confirmed when we examined its diagnostic plots (Figure 6), whereby we see that now the Ljung-Box assessment seems to be look better than the one for ARIMA(0,1,2). Finally, a check on the ACF/PACF (see Figure 7) of the residuals reveal that they mostly lie below the bars, meaning that there is almost no correlation between the residuals.

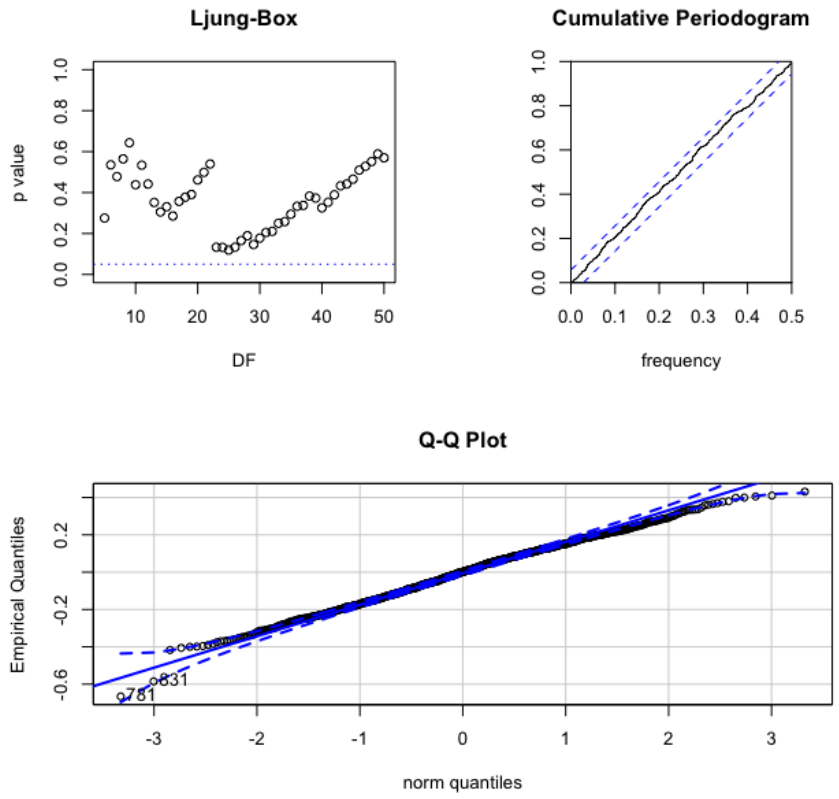

Figure 6: Diagnostic for ARIMA(2,1,2) of $\{DZ_t\}_t$, family 2.

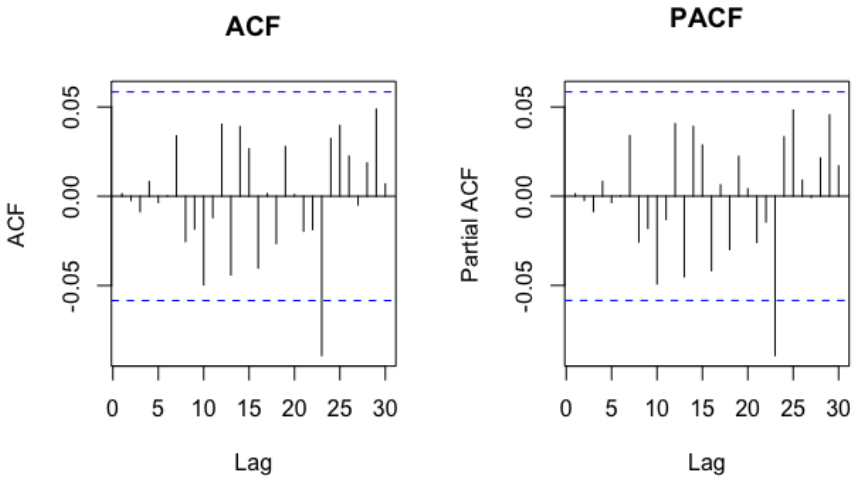

Figure 7: ACF and PACF for the residuals of ARIMA(2,1,2) of $\{DZ_t\}_t$, family 2.

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	-429.71	-842.58	**-926.76**	-926.90	-931.69	-931.80
AR1	-543.03	-923.32	-926.03	-927.33	-935.28	-933.31
AR2	-610.37	-929.00	**-935.79**	-933.28	-933.29	-931.41
AR3	-679.63	-934.55	-933.82	-931.93	-931.27	-929.45
AR4	-718.16	-933.00	-930.58	-936.70	-939.68	-939.36

Table 3: AIC values for ARIMA models for family 2. The column number indicates the order of the moving average component and the row number the order of the autoregressive component.

To conclude, we also tested using the likelihood ratio test between ARIMA(2,1,2) and ARIMA(0,1,2) and obtained a p-value of 0.0015, indicating a high-level of significance in adopting the former model. Thus it seems statistically plausible to use **ARIMA(2,1,2)** as our best model from family 2.

4 Forecasting

Taking ARIMA(2,1,2), we adopted the integrated model prediction method described in [1] (slides 205 - 213). In practice, we applied the `forecast` function from the library `Forecast` to an `Arima` model object. We then multiplied back the manually-estimated variances and obtained our mean and 95% confidence intervals.

Formally, from the `forecast` function we obtained $\{\tilde{Z}_t = \log \tilde{Y}_{n+h}/s^{(1)}_{29+h}\}_{h=1\to 59}$, as we note that 1 February is the 30th day in the week. Hence the reconstruction back to the original scale is:

$$\tilde{Y}_{n+h} = e^{\tilde{Z}_{n+h}} s^{(1)}_{29+h},$$

for $h = 1 \to 59$. This series $\{\tilde{Y}_{n+h}\}_{h=1\to 59}$ is our **prediction**, and we do this for our upper 95% confidence bands and esimation of the mean.

We predicted $h = 59$ steps ahead and overlaid the actual data from **1 February to 31 March 2018** on top of our predictions (see Figure 8).

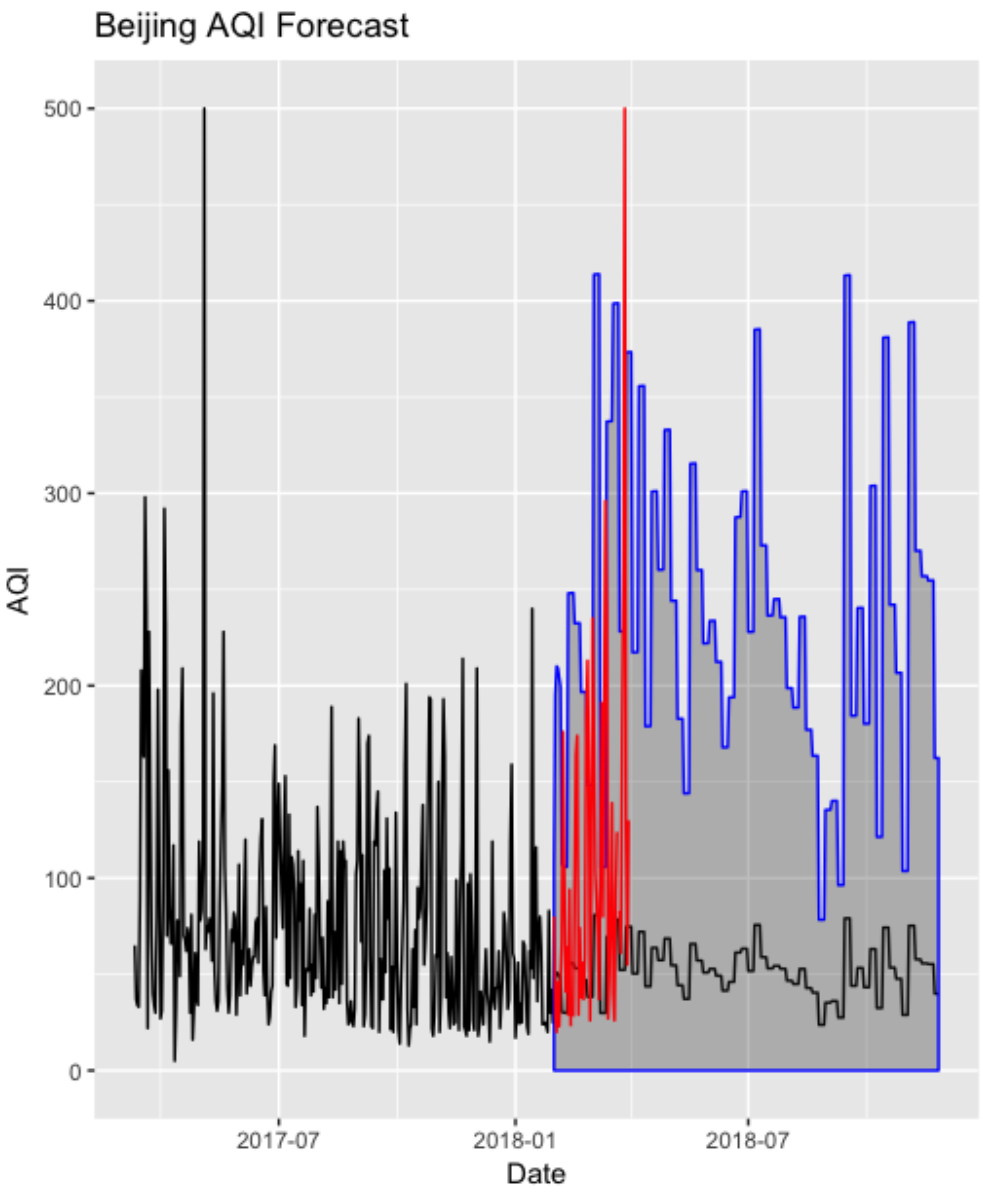

Figure 8: Plot of data and 300-steps-ahead prediction (black) overlaid with the actual data (59 days, red). The blue bands indicate the confidence intervals. The model used here is ARIMA(2,1,2) for $\log Y_t / s_t^{(1)}$.

Finally, we note that there were 2 missing values in our actual data (1 February to 31 March) and so we have 57 actual observations. There were 10 observations that fell outside of the confidence region, indicating an error rate of 17.5%. Clearly, the confidence intervals are massive, but this is because our data is very noisy and that we used an integrated model (taking a difference of $d = 1$). In general, weather forecasting is difficult to predict and so it is not unusual to have such large confidence bands.

5 Conclusion

In general, our dataset is very noisy and has a lot of irregular patterns due to the rapid changes going on in China. Nevertheless, we were still able to capture the important features of this dataset.

We analysed the AQI time series and chose ARIMA(2,1,2) to be the best fitted model for the processed data LY_t divided by the quasi-averaged norm. In fact, there are some other ARIMA models that have approximately equal AIC value and log-likelihood. For those that have more parameters, it's more appropriate to choose the simpler model. Compared to the other simpler models, the Q-Q plot of the residuals for ARIMA(2,1,2) has a more significant straight line at angle 45°. Thus,we consider ARIMA(2,1,2) as the best model given by (3). The fit is confirmed by the fact that the actual values in specific months are mostly contained in the prediction confidence intervals produced by ARIMA(2,1,2).

During the testing process, since the variance varies over time, we considered using the ARMA-GARCH model to the data but the model fits are not good as the residuals are not distributed as white noise (confirmed by the Q-Q and Ljung-Box plots). Perhaps a more complicated distribution like student-t or skewed student-t distribution could also produce good models. However, results from these models tend to be difficult to be reconstructed back to the original scale, especially when we have processed our original time series. It would thus be interesting in the future to investigate this further.

Finally, the best results that we produced were after we stabilised the variability in the data by dividing the log-transformed series LY_t by its quasi-averaged norm. We didn't take into account of any time-dependence of the variance into and assumed that the variance for each day is constant without any noise. Therefore further investigations could be carried out to present a more rigorous approach to stabilising the variability.

References

[1] Thibaud, E., (2018).Time Series. *EPFL*.

[2] Wikipedia. Computing the AQI, Air quality index, `https://en.wikipedia.org/wiki/Air_quality_index#Computing_the_AQI`.

[3] Wang, Xiaolei, (2018). Data of Historical Air Quality in the Country, `http://beijingair.sinaapp.com`.