AIR CONDITION ANALYSIS AND FORECAST FOR NEW YORK CITY

STA2202 TIME SERIES ANALYSIS

ZEKUN LIU

Master of Computer Engineering

April 12, 2019

Summary

Daily air condition of 144 Cities in US is provided from the year of 2000 to 2016, in terms of the data of NO_2, O_3, SO_2 and CO. It records the mean value, maximum value of each single day and introduces a concept of AQI (Air Quality Index) separately for each air composition to better demonstrate the degree of air pollution. To narrow down the range research and summarize a more specific conclusion, this paper focuses on the air condition of SO_2 in New York City, to proceed time series analysis.

log transformation was first applied to the original data, since it was visualized as a multiplicative model, where the value of the datapoints is proportional to its past rather than a simple sum. Besides, the data of SO_2 demonstrates a stable pattern of seasonality. Therefore, in the process of model fitting, an extension of ARIMA (Autoregression Integrated Moving Average), SARIMA (Seasonal-ARIMA) was first performed, with the parameters determined by related ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function). Although it provides an acceptable result, it shows some limitations in processing large-period dataset and interpreting the seasonality. Consequently, SSA (Spectrum Singular Analysis), a methodology decompose time series with frequency estimation is utilized to further dealing with the seasonality and accurate forecast. Furthermore, both utilized models are compared with the fitting and forecasting result of the given data and concluded in terms of model adaptability and performance.

1.0 Introduction

Time series is a special form of dataset that, unlike those traditional datasets whose premise of performing data processing and model fitting is that, each observation in the dataset must guarantee identity and independent, or the result may be misleading. However, in the real world, there exist a type of dataset that the data changes with time index, namely, every datapoint indicate a dependence upon and the datapoint previous. Also, time index will be behaved as an insight variable that influences the pattern of the data. Therefore, new models are designed to perform such series analysis.

ARIMA [3] is the most popular and fundamental model that can well perform model fitting and forecasting. It analyzes the autocorrelation (ACF) and partial autocorrelation (PACF) between datapoints, interprets the patterns and fits the model. Basic ARMA model requires the stationarity of input data and AR (Autoregression) and MA (Moving average) regress the current datapoint and error separately with their previous terms. To enlarge the model, an integrated section is added to reduce non-stationary series by taking a sequence of differences. SARIMA is performed in this project, which is a modification version of ARIMA with a seasonal section, to further understand the seasonality of the proposed dataset.

However, SARIMA implies some limitations in this project. Therefore, SSA (Singular Spectrum Analysis) [4] is also applied to perform a more comprehensive decomposition of all possible periodicities of the given dataset and perform a more reliable prediction of the future data. The main idea of SSA is simple. It embeds the 1-D time series array to a matrix with a window length determined by the size of periodicity, and further modifies the resulting eigenvectors in the matrix to be fully orthogonal, namely, all separated frequencies should be separable by extended SSA methods. Then, reconstruct the matrix back to be several 1-d arrays to indicate trend, seasonality and residuals. Furthermore, such result decomposition and reconstruction will be applied for validation and forecast.

2.0 Problem Statement

Air condition has become a severe topic in recent years, since studies indicate multiple grave consequences caused by the increasing of air pollution. First and foremost, for example, human health will be affected due to the exposure of several particle pollutants and the content of air composition, such as respiratory diseases and cancer. In the long term, the rate of mortality will therefore increase [1]. Besides, studies show that air pollution may also bring an overall influence on global atmosphere, which may cause a potential climate change [2]. Therefore, understanding the change of air condition with time changes benefit a lot. As it can help to improve the analysis of some existed factors, such as some pollutant-related diseases, but also can potentially behave as an insight helper when forecast the possible trend and make future judgement.

Therefore, this paper takes interest in the daily data of SO_2 AQI in New York over 16 years from 2000 to 2016, understands the overall tendency of SO_2 in the past, as well as the possible trend in the future. To well perform the model, three-years' datapoints are removed to form a validation set for validating the model computed by training set. And the one with least validation error will be selected. Plus, during this prediction process, this project is also interested in a detailed seasonality analysis of different frequencies of the data. In the following sections, detailed implementation of both ARIMA and SSA in terms of data processing and data forecast will be introduced, followed by a comparison of the model reliability and prediction accuracy with the given dataset. Furthermore, conclusion will be made based on all the analysis.

3.0 Data Visualization

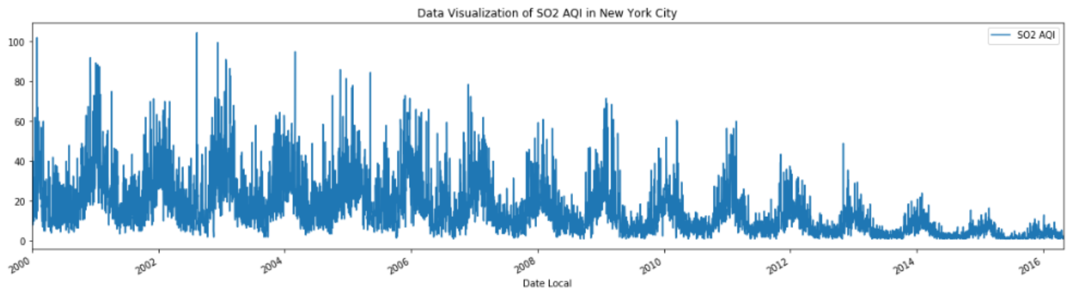
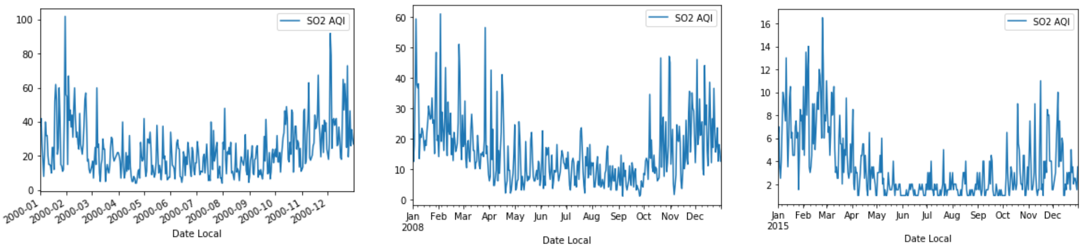


Figure 1-1: The plot of the value of original SO_2 AQI in New York from 2000 to 2015

The proposed dataset is a collection of daily value of SO_2 AQI in New York, containing 5837 datapoints from January 1st of 2000 to April 30th of 2016. A general plot of such dataset based on time index is shown in Figure 1-1. It is obvious that, in the first place, the total data indicates gradually decreasing trend with time continuing. And it potentially implies a stable pattern of a yearly seasonality. This conjecture is further proved by Figure 1-2, that the plots of three random years of 2000, 2008 and 2015 demonstrate a similar pattern, namely, the value of SO_2 AQI in the beginning and at the end of the year is much higher than that in the middle of the year, although some tiny differences exist. Therefore, such dataset can be recognized as a typical form of time series, as a composition of trend, seasonality and residuals.



(a) SO_2 AQI data of 2000 *(b) SO_2 AQI data of 2008* *(c) SO_2 AQI data of 2015*

Figure 1-2: Three random yearly plots of SO_2 AQI.

When observing deeper into the amplitude of seasonality, it can be noticed that although the yearly data of SO_2 AQI demonstrates a seasonal pattern, the relative value over the given year is proportional to the value in the past. This characteristic implies a multiplicative model. Unlike addictive model, of which the data series is independently combined by the trend, seasonality, cyclic and residual, multiplicative model represents a product of those elements listed above. Such model cannot be directly performed by any

model, for it fails to perform decomposition. Therefore, log transformation should be introduced to the SO_2 AQI data, to convert multiplicative model to be addictive. Figure 1-3 shows the data after log transformation. And this data is finally utilized for analysis.

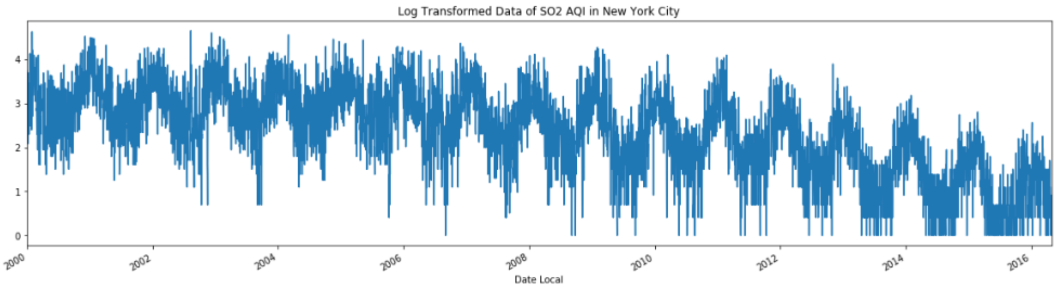


Figure 1-3: The plot of the value of original SO_2 AQI in New York from 2000 to 2015

4.0 Analysis with SARIMA

SARIMA is initially applied to the proposed dataset. Since the basic ARMA model only accepts stationary series as the input, the stationarity check and differencing are mandatory before determining the model. This also helps to identify the differencing orders of the finalized SARIMA model. After such parameters decided, the rest parameters of AR, MA, SAR (Seasonal-AR), SMA (Seasonal-MA) remain to be determined, with the help of Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots. When a range of these parameters decided, the finalized model with the least value of AIC (Akaike information criterion) will stand out as the best option. Furthermore, this model is utilized for forecasting. And a related forecasting error will be examined.

One limitation when applying ARIMA is that, the largest value of periodicity allowed is 200, while the period is 365 in this dataset. Therefore, the data is resampled to a monthly data as shown in figure 2-1.

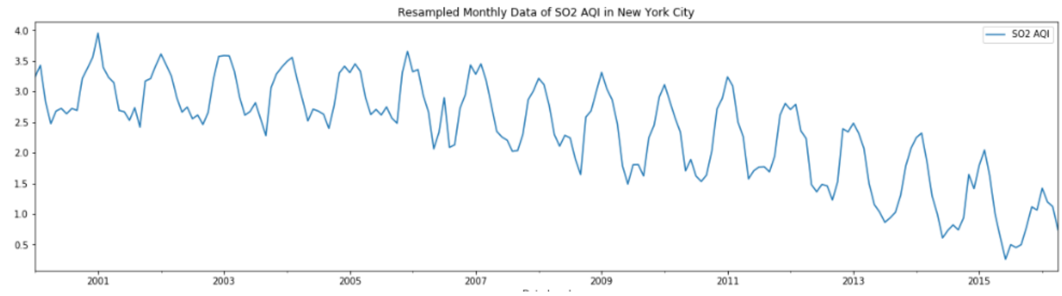


Figure 2-1: Resampled Monthly Data for SO_2 AQI

4.1 Stationarity and Differencing

ARIMA model requires stationary series as the input. Checking for the stationarity and reducing such non-stationary series to stationary is a must. In this section, a detailed check for stationarity and its related differencing will be introduced thoroughly.

4.1.1 Stationarity Check

As visualized in figure 2-1, the log transformed data indicates an apparent downward trend and a stable yearly seasonality, which is reasonable to be suspected as non-stationary. The following statistics tests further prove the same result. Adfuller (Augmented Dickey-Fuller) test and KPSS (Kwiatkowski–Phillips–Schmidt–Shin) test are both applied to check for an overall stationarity and the trend stationarity only. Both results are shown in figure 2-2. With the p-value greatly larger than 0.05 in Adfuller test and Statistic value larger than all critical values, the monthly data of SO_2 AQI in New York illustrate a non-stationary tendency.

```
ADF Statistic: 4.443174
p-value: 1.000000
Critical Values:
        1%: -3.467
        5%: -2.878
        10%: -2.575
```

(a) Adfuller Test Result

```
KPSS Statistic: 1.163341
p-value: 0.010000
Critical Values:
        10%: 0.347
        5%: 0.463
        2.5%: 0.574
        1%: 0.739
```

(b) KPSS Test Result

Figure 2-2: Statistics Test Result for stationarity

4.1.2 Differencing

Differencing is necessary for reducing a non-stationary series to stationary. Totally two types of differencing (non-seasonal and seasonal) can be applied, and the total order can not exceed 2. Based on such criteria, several differencing models are applied to the resampled data, and the one with least residual standard deviation will be selected. Table 1-1 listed the results and the model with non-seasonal differencing (0,1,0) x (0,0,0,0) indicates the best. However, since the data demonstrates an obvious seasonality, a seasonal differencing is necessary, to prevent the seasonality from eliminating in the long term. Therefore, both non-seasonal and seasonal differencing will be added to the ARIMA model.

Over-differencing should be prevented during this process. Partial autocorrelation plot (PACF) of differenced data may help for checking over-differencing. As shown in figure 1-2, the value of lag-1 in PACF is negatively smaller than -0.5, which implies a proper differencing, and the resulting series is reduced to be stationary.

ARIMA model (p, d, q) X (P, D, Q, S)	Residual Standard Deviation
(0,0,0) X (0,0,0,0) Original Data	0.594658
(0,1,0) X (0,0,0,0) One Order Non-Seasonal Difference	0.339629

(0,2,0) X (0,0,0,0) Two Order Non-Seasonal Difference	0.424147
(0,0,0) X (0,1,0,12) One Order Seasonal Difference	0.845484
(0,1,0) X (0,1,0,12) One Order Non-Seasonal and Seasonal Difference	0.397442

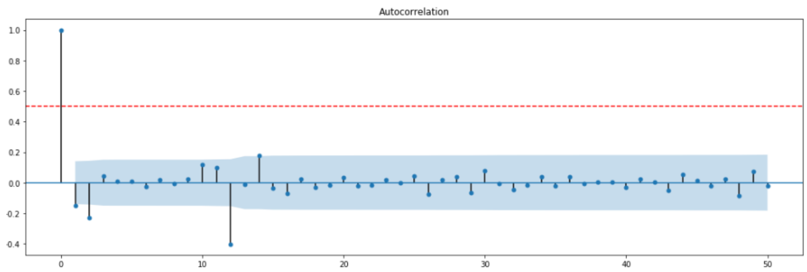
Table 1-1: Residual Standard Deviation of ARIMA Difference Model (0,0,0), (0,1,0), (0,2,0), (0,0,0) x (0,1,0,12), (0,1,0) x (0,1,0,12)

4.2 Parameters Determination

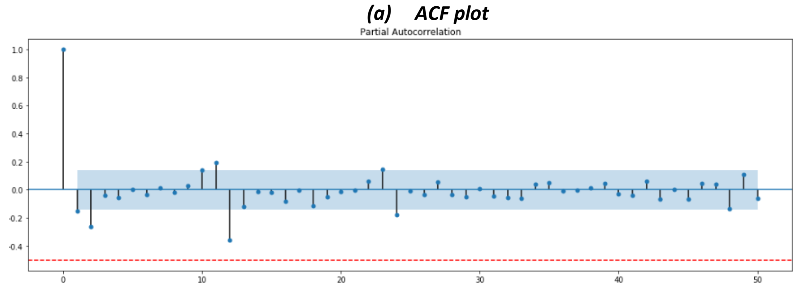
Totally 7 parameters in SARIMA model (p, d, q) x (P, D, Q, S) need to be determined. Two differencing orders d (non-seasonal), D (seasonal) have been decided from the previous section. Also, the parameter of monthly period is identified as 12. Before The rest of the parameters can be determined [5] as follows.

4.2.1 Non-Seasonal Parameters Determination

Typically, the order of AR and MA is easy to decide, by simply identifying the AR or MA signature from the ACF and PACF plot of the differenced series. AR signature is recognized as a sharp cutoff and/or a positive lag(k) of PACF while ACF decays more slowly, this situation also indicates under-differencing. While MA signature is opposite, by visualizing a sharp cutoff and/or a negative lag(k) of ACF and PACF decays more slowly. This in the contrast indicates a bit over-differencing. Based on the criteria above, figure 2-2 demonstrates the ACF and PACF of the differenced data series. It is observed that the rate of cutoff for both ACF and PACF are the same, but the value of the first two correlation are all negative, which can be reasonably suspect that the series is slightly over-differenced. Therefore, a MA (2) can be decided as a possible order of the finalized fitting model, while an AR model with order 1 or 0 can be a possible candidate.



(a) ACF plot



(b) PACF plot

Figure 2-2: ACF and PACF of Differenced SO_2 AQI

4.2.2 Seasonal Parameters Determination

The order of SAR (P) and SAM (Q) can also be determined from the ACF and PACF plot in figure 2-2. In ACF plot, the value of autocorrelation at every lag 12, where 12 is the period of the series, is always negative. Then SMA will be considered, this is also recognized as a typical behavior of seasonal differencing. Besides, SAR and SMA should not exist at the same time, as it may lead to over-fitting.

In conclusion, the finalized SARIMA model can be considered as (p,1,2) x (0,1,1,12). Such model will be applied to the preprocessed data to perform data fitting and forecast.

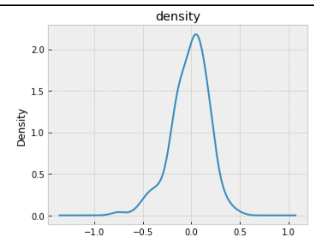
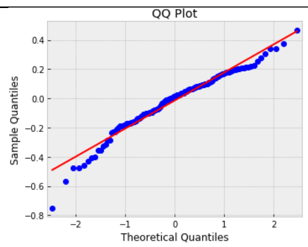
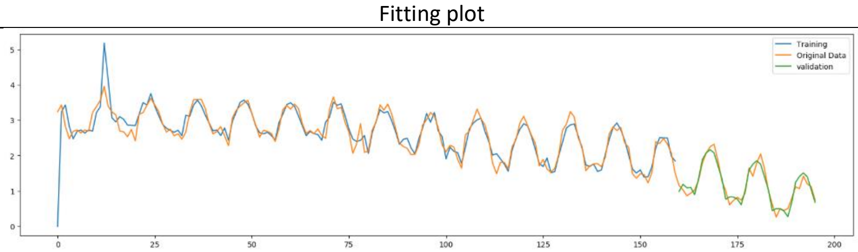
4.3 Model Fitting and Comparison

In this section, partially determined SARIMA model will be tested with the training set monthly data of SO_2 AQI in New York and make predictions. In this process, auto-arima performs to automatically generate the optimal model, with the range of each parameters of the SARIMA model required. AIC (Akaike information criterion) is behaved as the criteria for identifying the best fitting model. AIC relatively statistics the datapoints that fail to be fitted by the generated model. Therefore, least AIC value indicates a best fit.

Fitting results of auto-arima with different combinations of parameters are shown in figure 2-3, where the model of (1,1,2) x (0,1,1,12) has the lowest AIC and indicates the best model. However, adding both AR and MA will significantly complexify the model, which may potentially cause over-fitting. To further evaluate the fitness, first the training error is checked with normal distribution, as an ideal residual of time series should be a white noise. Also, validation error is computed of the model with and without the AR session. The fitting results are listed in table 1-2. See from the training and validation error, although model 2 with AR session illustrates the smaller training error, the validation error is the larger than the one without AR session, which indicates a slight over-fitting. Therefore, the optimal ARIMA model is indicated as (0,1,2) x (0,1,1,12).

```
Fit ARIMA: order=(0, 1, 2) seasonal_order=(0, 1, 1, 12); AIC=-56.668, BIC=-41.715, Fit time=1.985 seconds
Fit ARIMA: order=(1, 1, 2) seasonal_order=(0, 1, 1, 12); AIC=-65.399, BIC=-47.456, Fit time=2.105 seconds
Fit ARIMA: order=(1, 1, 2) seasonal_order=(0, 1, 0, 12); AIC=-15.300, BIC=-0.348, Fit time=0.824 seconds
Fit ARIMA: order=(2, 1, 2) seasonal_order=(0, 1, 1, 12); AIC=-64.422, BIC=-43.489, Fit time=2.403 seconds
Total fit time: 10.379 seconds
-65.39872804666608
```

Figure 2-3: Auto-ARIMA Fitting Result with Different ARIMA Model

<table>
<tr><td>ARIMA Model</td><td>Residual Density Plot</td><td>Residual qq-Plot</td><td>Training Error</td><td>Validation Error</td></tr>
<tr><td rowspan="3">Model 1 (0,1,2) x (0,1,1,12)</td><td>
</td><td>
</td><td>0.117</td><td>0.530</td></tr>
<tr><td colspan="4">Fitting plot</td></tr>
<tr><td colspan="4">
</td></tr>
<tr><td></td><td>Density Plot</td><td>qq-Plot</td><td>Training</td><td>Testing</td></tr>
</table>

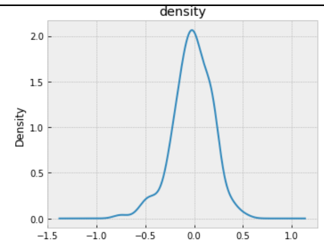
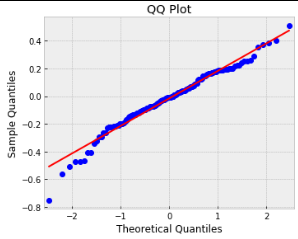
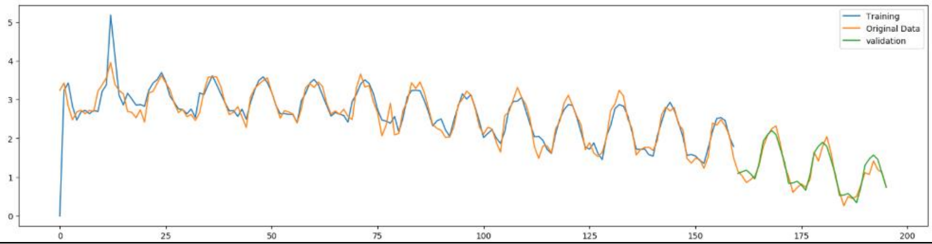
<table>
<tr><td></td><td></td><td></td><td>error</td><td>error</td></tr>
<tr><td rowspan="3">Model 2
(1,1,2) x
(0,1,1,12)</td><td></td><td></td><td>0.115</td><td>0.537</td></tr>
<tr><td colspan="4">Fitting Plot</td></tr>
<tr><td colspan="4"></td></tr>
</table>

Table 1-2: Fitting Results of Proposed Models

4.4 Summary

Based on the analysis above, the finalized model for monthly SO_2 AQI is determined as ARIMA (0, 1, 1) x (0, 1, 1, 12), with an acceptable fitting error. However, some drawbacks exist when applying ARIMA model. First, ARIMA model only accept a maximum period of 200, otherwise the computational cost is overly high. Fitting a such model with a seriously narrowed down model lacks accuracy and reliability. Besides, ARIMA model only generates a model from time domain, by analyzing the correlation of datapoints of different time index. However, it lacks the insight understanding of the detailed decomposition of the given data series, such as slow-vary trend and leading components of frequencies. Therefore, to take a more comprehensive look into the total data, singular spectrum analysis (SSA), a model-free technique for frequency decomposing, reconstruction and forecasting will be introduced in the following sections.

5.0 Analysis with SSA

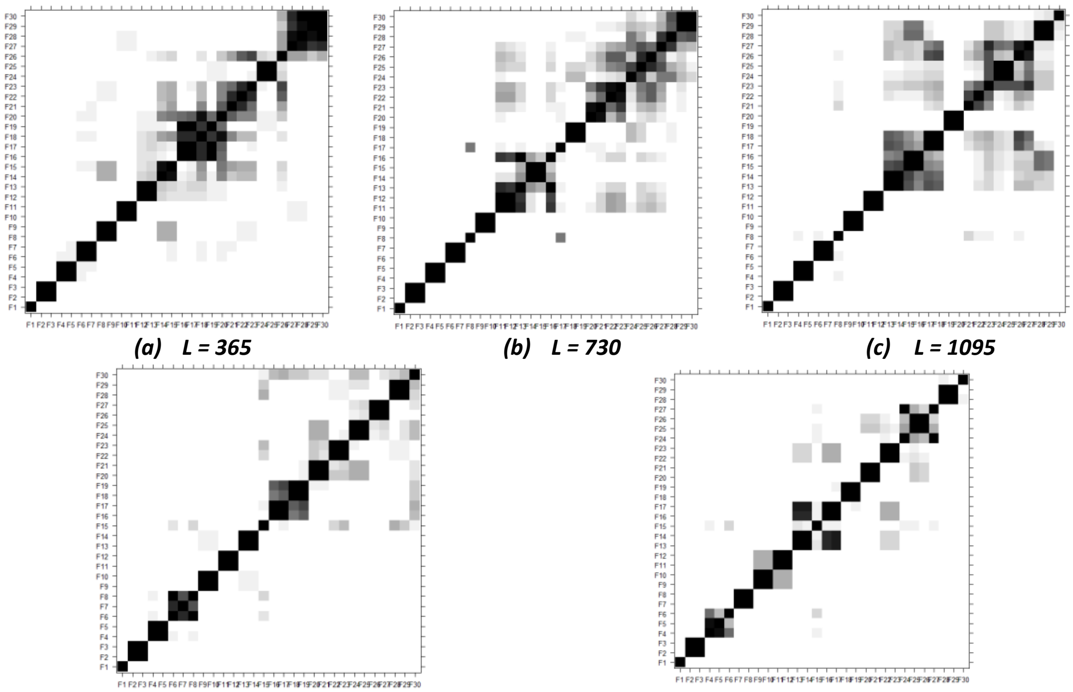
Singular Spectrum Analysis (SSA) and its extensions are typical in such area. Totally four steps are gone through when performing this model. First, for univariant analysis, the given 1-d time series will be embedded into a Hankel matrix L X K, where L is denoted as the window length that needs to be determined for optimal decomposition. Then in the decomposition step, the computed matrix will be converted into a sum of eigenvectors of Rank 1. With a proper decomposition, these eigenvectors can be grouped to be trend, seasonality and residuals. Several modifications will be applied in this procedure to obtain the optimal groups. Then the finalized groups will be utilized in the reconstruction and forecast step to make further uses. Divided training and validation set will also apply in this section, with training set utilized as input data. The detailed implementation of SSA with such data series will be introduced in the following section.

5.1 Data Embedded and Separability

Although SSA is known as a model-free methodology that requires no premise such as stationarity, or any parameters and allows any data input, window length L is of great priority when performing data embedding. It not only decides the size of the resulting Hankel matrices, also, a well-determined window length may help to provide a good separability. Separability [6] is an important notion in understanding and performing SSA, as it signifies the extraction of leading frequency components with residuals. Long-term time series always illustrate a weak separability, which means after decomposition there may exist a proper choice of grouping, while strong separability means any choices of grouping can bring a success reconstruction. Therefore, strong separability is always required in decomposition step. Several extensions of basic SSA are designed for further decomposition in a nested way for advanced decomposition.

5.1.1 Determination of Window Length

Usually, the window length should be smaller than half of the length of the data, and be a multiplication of the period, if a seasonal data series provided. W-correlation matrix may help to decide optimal window length as it can detects weak separability and group identification. The window length providing a better separability will be selected. Figure 3-1 lists the w-correlation plots corresponding to different window length. Of all these plots window length of 365 and 1825 both indicates a bad separability as they fail to separate leading components of trend from noises. The rest both show an effective separability, while the window length of 1460 indicates the best, as it separates the most leading components among all the window length.



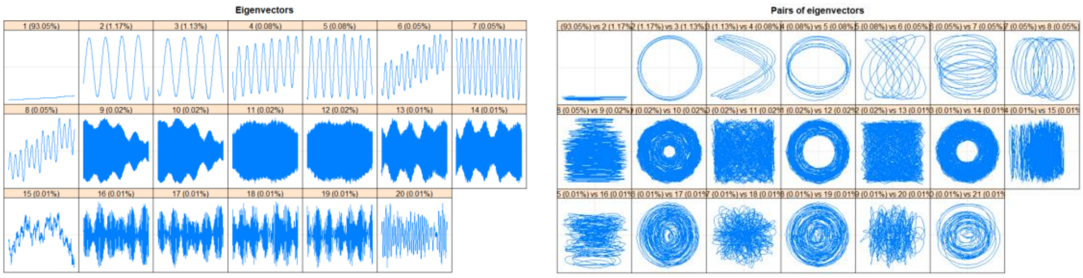
(a) L = 365 ***(b) L = 730*** ***(c) L = 1095***

(d) L = 1460 ***(e) L = 1825***

Figure 3-1: W-correlation of Elementary Components of different Window Length (L)

5.2 Decomposition and Reconstruction

Basic SSA utilized conventional SVD to perform decomposition. Sometimes Truncated SVD may applied as a shorter version to automatically generate a subset of eigenvectors and neglect the rest. This may significantly reduce computation and space cost. However, this only benefits short term series. As for long term series with large window size, Truncated SVD may lose some information by dropping the eigenvectors and may not fully represent the given series. Therefore, a full version SSA is applied for preliminary decomposition. The resulting eigenvectors are visualized in figure 3-2.



(a) Vector Plot for Eigenvectors *(b) Pair Plot between two Eigenvectors*

Figure 3-2: Visualization of Decomposed Eigenvectors

It can be noticed that the first 8 eigenvectors show a weak separability that needs to be modified, where an extension of SSA should be considered. Such extension can be only performed in a nested way of those weakly-separated components. Iterative Oblique SSA and DeriveSSA is two main method for advanced decomposition. Iterative OSSA focuses on separating those sine waves with similar frequencies, which is rare in a real time series, while DeriveSSA focuses on the sine waves with similar amplitude. Therefore, the latter one will be utilized in this situation.

Weakly-separable section after preliminary decomposition can be identified as follows. Noticed from figure 3-1(d) the eigenvector (ET) after 15 may formed by noises, the ET1-15 can be regarded as the combination of trend and oscillations. Figure 3-2(a) further proves this observation, that ET1-8 and ET15 and an obvious trend and ET9-14 may illustrate some seasonality. Therefore, the ET1-15 may be further decomposed. The resulting eigenvectors may show in figure 3-3. The resulting eigenvectors can be grouped together and reconstructed as in figure 3-4 shows.

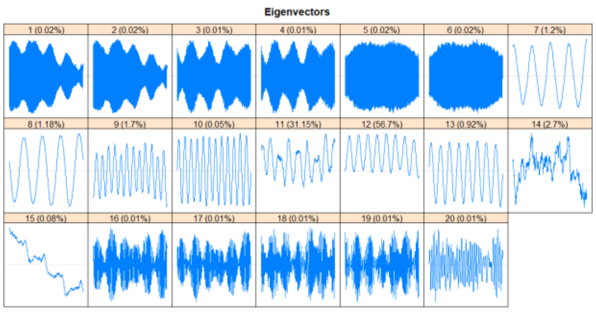


Figure 3-3: Eigenvectors after DeriveSSA

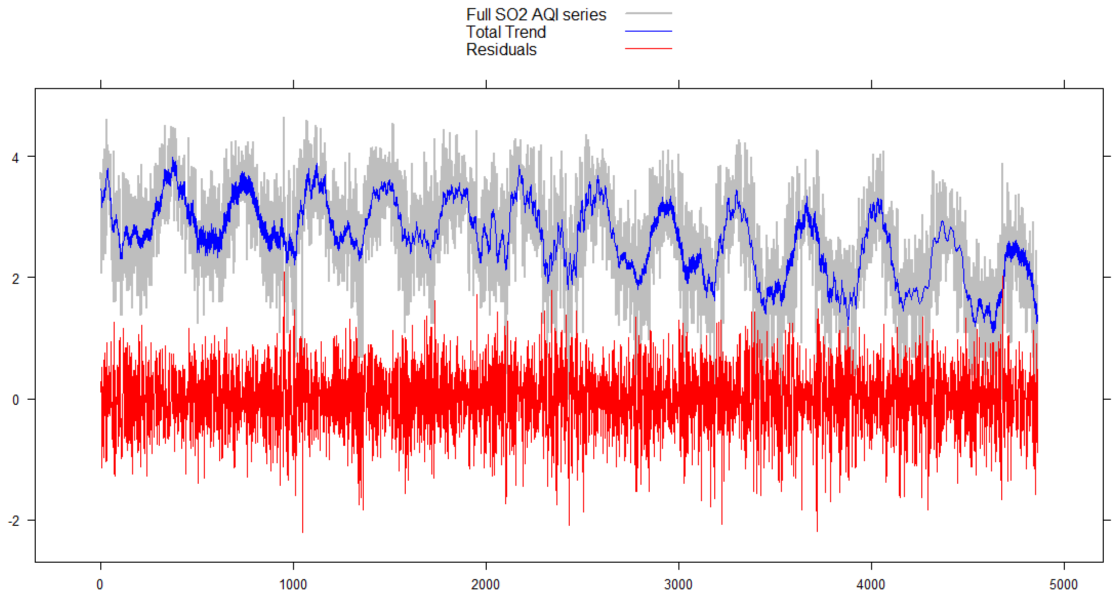
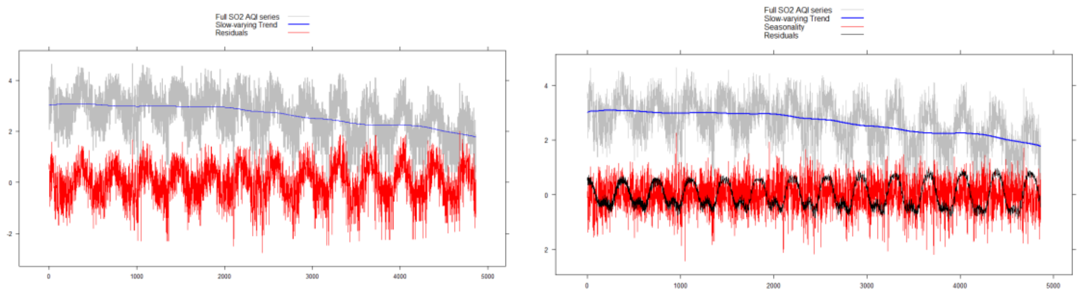


Figure 3-3: Decomposition of Total Trend and Residuals of SO_2 AQI of New York

5.2.1 Decomposition of Trend and Seasonality

The extracted trend shown in figure 3-3 is a mixture of show-vary trend and seasonality. To further separate these component, sequential SSA will be performed. After basic SSA the leading component can be extracted as figure 3-2 (a) shows, the first slow-varying eigenvector contributes the most and can be grouped and reconstructed as the trend as shown in figure 3-4(a). The rest eigenvectors are the mixture of seasonality and noises. The mixed component will then be regarded as a new input for extracting seasonality. The procedure of second-step for seasonality extraction is the same as the one introduced before, by applying basic SSA with a window length of 1701, further nested decomposing the weakly-separable eigenvectors and reconstructing back to the 1d-array. The resulting extraction can be fully viewed in figure 3-4(b).



(a) Decomposition of Slow-Varying Trend

(b) Full Decomposition of Slow-Varying Trend, Seasonality and Residual

Figure 3-4: Performance of Sequential SSA for Trend, Seasonality and Residual Decomposition

5.3 Forecast

After all components for training set have been extracted, the corresponding SSA model is prepared for validation. There are mainly two methods for forecasting: one is recurrent forecasting, which is a continuation of linear recurrent relation (LRR). Namely, there exist a series of coefficients that the extracted tendency can be represented by a sum of a series of sine waves. In the contrast, the other forecasting method, vector forecasting, achieves prediction in a verse way. It adds columns to the end of embedded matrix to make predictions. Good separability achieved in embedded step may lead to similar result with both methods, while such result may vary with an approximate separability. Both forecasting methods are applied to the validation set as shown in figure 3-6, as well as the total trend extracted from the whole data series.

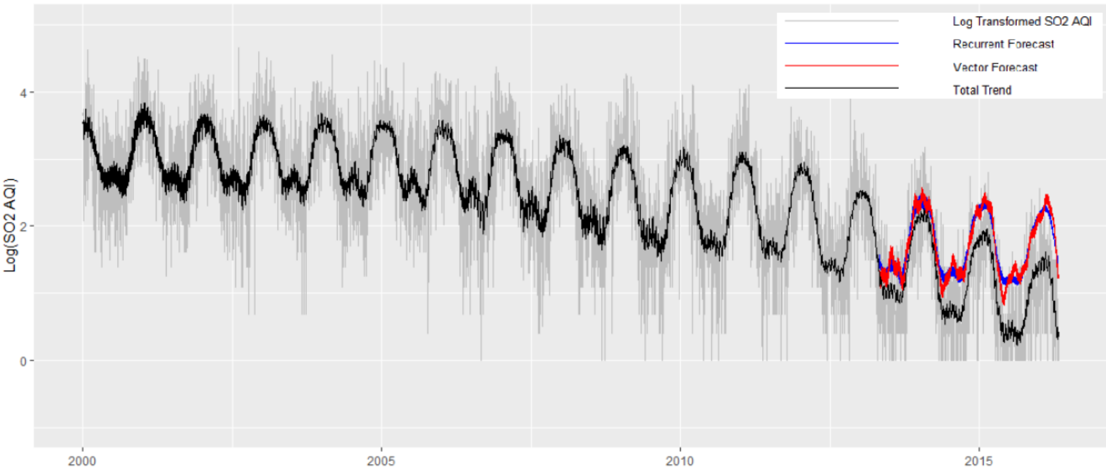


Figure 3-6: Validation Prediction of Last 3 Years and Original Transformed Data

As figure 3-6 shows, the prediction performed by recurrent forecast and vector forecast demonstrate a similar result, which implies a good separability by determining the embedded window length and group in reconstruction step. However, they both show an apparent difference with the original extracted trend, as the overall predicted tendency is higher than the reality. This mainly due to the lack of comprehensive. Since the decomposed SSA model is generated only by the whole training set with one determined window length and one set of leading components, it fails to collect a detailed possible change of trend of several subseries, and therefore the resulting validation prediction may only reflect a partial possible trend rather than a comprehensive one. Hence, a more complex function is applied to the data, for computing the optimal parameters and obtaining accurate model.

5.3.1 Optimal Forecast

In this section, an optimal forecast function will be introduced and applied to the log transformed data for better prediction. Same as before, the training set contains the data for the first 13 years while 3-year's data is left for validation. The main modification is, instead of generating the with only one training set, a sliding parameter is added to the original model, so that the training set will be further divided into several subseries with the window sliding, for an optimal model computation. Besides, a series of different window length value and number of leading components will be simultaneously checked with one certain training subseries. Finally, the combination of window length and number of components that achieves least validation mean square error will be chosen as the optimal SSA model and perform the accurate prediction.

By tuning the window length from 1*365 to 6*365, number of components from 1 to 100, the same optimal result computed with sliding parameter set as 2 and 3. Such result shows that, with an optimal window length of 4*365, with the first 17 leading eigenvectors, the validation error may achieve 0.544 as its least. Figure 3-7 demonstrates the validation predictions as well as a future prediction for two years after 2016. The fitting result of optimal model well interprets the log transformed data and predict the value with little difference, while figure 3-8 demonstrates the results for optimal model with original data series.

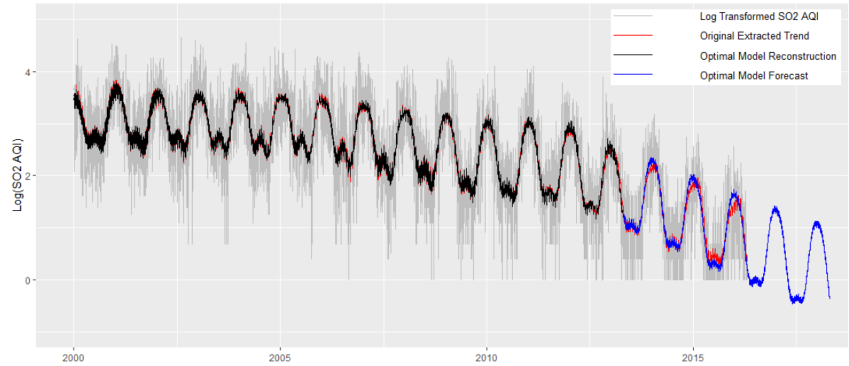


Figure 3-7: Plot of Transformed Data, Original Trend, Reconstruction and Forecast Based on Optimal Model

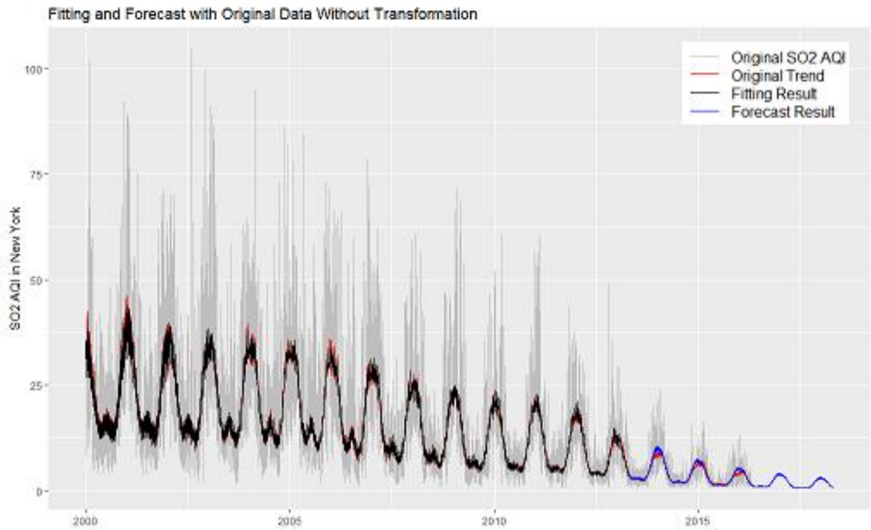


Figure 3-8: Plot of Original Data, Original Trend, Reconstruction and Forecast Based on Optimal Model

6.0 Data Analysis

From both models of ARIMA and SSA, an obvious decreasing trend can be visualized for the value of SO_2 AQI, especially after the year of 2012. And from SSA, with all components extracted as shown in figure 3-3, a comprehensive decomposition of the change of SO_2 AQI in New York from 2000 to 2013 can be fully understood. It demonstrates a yearly pattern that the SO_2 AQI value is much higher than usual in the beginning and end of the year, while the midyear may reach a slight increment. Besides, several insight frequencies can be viewed by ESPERIT. The result is shown in figure 3-5, that the series also indicates a half-year and quarterly-year seasonal pattern.

period	rate	Mod	Arg	Re	Im
364.975	0.000116	1.00012	0.02	0.99997	0.01722
-364.975	0.000116	1.00012	-0.02	0.99997	-0.01722
18.965	0.000093	1.00009	0.33	0.94570	0.32531
-18.965	0.000093	1.00009	-0.33	0.94570	-0.32531
121.693	0.000060	1.00006	0.05	0.99873	0.05161
-121.693	0.000060	1.00006	-0.05	0.99873	-0.05161
182.695	-0.000016	0.99998	0.03	0.99939	0.03438
-182.695	-0.000016	0.99998	-0.03	0.99939	-0.03438
6.856	-0.000055	0.99995	0.92	0.60857	0.79343
-6.856	-0.000055	0.99995	-0.92	0.60857	-0.79343
11.157	-0.000064	0.99994	0.56	0.84552	0.53382
-11.157	-0.000064	0.99994	-0.56	0.84552	-0.53382
20.507	-0.000172	0.99983	0.31	0.95326	0.30157
-20.507	-0.000172	0.99983	-0.31	0.95326	-0.30157
6.991	-0.000534	0.99947	0.90	0.62230	0.78209
-6.991	-0.000534	0.99947	-0.90	0.62230	-0.78209

Figure 3-5: Frequency Estimation of Extracted Seasonality

7.0 Model Comparison

Two typical models are performed for the analysis of the daily data of SO_2 AQI in 16 years. Both models achieve an accurate fitting and forecasting for the validation set, and they both succeed in reflecting the total trend of the data input. Also, some differences occur. First, ARIMA, as a mature and hands-on time series model, it is easy to apply and determining the optimal parameters for the fitting model. Also, it works well with the data with apparent linear relations. However, one limitation is that it fails to accept data input with large frequencies, which is common in real world in recent year. Therefore, some other methodologies may help for such dataset. For SSA, it well decomposes the input data with different frequencies, providing a deeper understanding for the data itself. And it is more complex than ARIMA as several extensions are designed for SSA, for dealing with time series with different characteristics.

Reference

[1] Zmirou, Denis, Joel Schwartz, Marc Saez, Antonella Zanobetti, Bogdan Wojtyniak, Giotta Touloumi, Claudia Spix et al. "Time-series analysis of air pollution and cause-specific mortality." Epidemiology (1998): 495-503.

[2] Seinfeld, John H., and Spyros N. Pandis. Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, 2016.

[3] Brockwell, Peter J., Richard A. Davis, and Matthew V. Calder. Introduction to time series and forecasting. Vol. 2. New York: springer, 2002.

[4] Golyandina, Nina, Anton Korobeynikov, and Anatoly Zhigljavsky. Singular Spectrum Analysis with R. Springer, 2018.

[5] Robert Nau. "Statistical forecasting: notes on regression and time series analysis." People.duke.edu. http://people.duke.edu/~rnau/411arim3.html (accessed March 30, 2019)

[6] Golyandina, Nina, and Alex Shlemov. "Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series." arXiv preprint arXiv:1308.4022 (2013).