東海大學

資訊工程學研究所

碩士論文

指導教授： 林正基博士
楊朝棟博士

以時間序列資料之模糊預測–以台中市空氣污染公開資料為例

The Prediction for Time-Series Data - a Case Study of Air Pollution Open Data at Taichung City

研究生:黃偉勝

中華民國一零六年七月

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 黃　偉　勝 所提之論文

以時間序列資料之預測-以台中市空氣汙染公開資料為例

經本委員會審查，符合碩士學位論文標準。

學位考試委員會
召　集　人 洪哲倫 簽章

委　　　員 呂峻益

姜自強

指 導 教 授 楊朝棟 簽章

指 導 教 授 林正基 簽章

中華民國　106 年　6　月　13　日

摘要

這個大數據時代,常會使用到統計或數學模型進行歷史數據的分析,並利用數據分析結果，預測未來的變化及趨勢。1965 年 Zadeh 的模糊理論出現，才能對模糊數加以分析及探討。目前，模糊理論已經被廣泛地運用在許多領域。近年來都市空氣品質日漸惡化，空氣汙染對人體健康的影響，空氣品質日漸受到重視，建立空氣汙染資料預測模式，提前讓民眾知道空污變化，採取適當策略以降低空氣污染對人體健康的負面衝擊，研究中選擇台中地區的測站為對象，挑選 2005 至 2015 年之間共 11 年的空污監測數據,使用逐步自回歸預測法建立台中市空氣污染的時間序列預測模型，模擬都市中空氣污染的變動趨勢。研究結果顯示，機率統計模式能有效預測空氣汙染資料的變化,在此模式的協助下,將可以提供空氣污染的變動趨勢,使民眾或政府提前做好預防措施，減少人體健康危害。

關鍵字:大數據、時間序列、空氣汙染、預測

Abstract

In this big data era, we always use the statistical or mathematical model for historical data analysis, and the use of data analysis results to predict future changes and trends. In 1965, Zadeh's fuzzy theory appeared in order to analyze and discuss the fuzzy number. At present, the fuzzy theory has been widely used in many fields. In recent years, the urban air quality is deteriorated, the impact of air pollution on human health and the air quality is increasingly valued. So to establish the air pollution data prediction model is in order to let people know the change of air pollution, and we will have time to take appropriate strategies to reduce the negative impact of air pollution on human health. In the research, I selected the total of eleven years of air pollution monitoring data between 2005 to 2015 from the Taichung air monitoring station. Using the method of autoregressive model to establish the time series prediction model of air pollution and to simulate the trend of air pollution in Taichung City. The result shows that probability statistics model can effectively predict the changes of air pollution from the data, and be able to provide changes in air pollution trends. Therefore, people or the government can do some precautionary measures in advance to reduce health hazards.

Keywords: Big Data, Time Series, Air Pollution, Prediction

致謝詞

時光飛逝轉眼間兩年的研究生涯就要進入尾聲了，經歷了大大小小的困難與挑戰，同時也學習成長了很多，在論文要完成之際，除了為自己感到開心之外更是從心底開始感到不捨即將離開東海大學了。我的論文能完成首先感謝我的指導教授林正基博士及楊朝棟博士，支持我想做的研究，並且提供所有我所需要的資源，在這兩年的期間教授的教導與督促甚至是老師在開會時常提到的態度，深刻地烙印在我心中，非常感謝老師兩年來的指導，經歷過後確實讓我學習成長了不少。

感謝我的口試委員洪哲倫及呂峻益教授特地撥空前來參加我的論文口試，委員們提出許多實用且寶貴的意見，使我的論文更加完整豐富。

特別感謝東海大學資管系姜自強博士，在論文找不到方向的時候，耐心且不厭其煩的教導，您就像一盞光明燈指引我前進的方向，讓我不再迷惘。

最後感謝實驗室的學長、同學及學弟們，在我研究生涯增添了豐富的色彩，在研究上彼此的激勵與共同成長，建立起的革命情感，我想那就是最美好的事情吧!

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

Due to the fast-paced nature of technology and rapid economic growth in Taiwan, our living habits change in many ways. Eventually resulting in different environmental problems. In 1980, the establishment of industrial parks were sprouting up all over the city, plus people drive the car or motorcycle instead of walking. Technology brings convenience, it also breeds pollution. The problem of air pollution is increasingly more serious in nowadays, and begun to harm the health of the human body.

This study will forecast the trend of air pollution data, let the people know the changes in air pollution ahead of time, take appropriate strategies to reduce the negative impact of air pollution on human health.

Time series predictions are often used in historical data, the most common methods such as autoregressive model, moving average model, autoregressive moving average model, which are based on continuous conceptual and assume that information is stable. In other words, these data are fixed over time. But usually the information is not so ideal, the time series of information will usually contain some uncertain changes, such as air pollution data.

This study is based on the study of air pollution data in Taichung, Taiwan, and establishes the time series forecasting model of air pollution to simulate the trend of air pollution in urban air. The simulation results can provide the public or the government to know the trend of air pollution ahead of time. People or the government can do some precautionary measures in advance to reduce health hazards.

1.2 Thesis Goal and Contributions

In this study, we selected 10 years of air pollution monitoring data from 2005 to 2015 at the Dali station, Zhongming station, Shairu station, Xitun station and Fengyuan station in Taichung area. Processing the data merging, data exploration, data analysis, use of autoregressive method to establish time series for the future data forecast, simulate the trend of air pollution in the city, finally we use MAPE (Mean absolute percentage error) to review the accuracy of the forecast.

1.3 Thesis Organization

The remainder of this thesis is structured as follows. In Chapter 2, we will describe the background information related to our work, including Big Data, Time Series, and Data Mining. In Chapter 3, Research Method, Method of Prediction, and Data Preprocessing are introduced. Chapter 4 shows the experimental results. And we summarized our work and the future work of this thesis in Chapter 5.

Chapter 2
Background Review and Related Work

2.1 BIG DATA

Big data [13] is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Almost 90% of the data of the world today was generated during the past two years. Moreover, approximately 90% of it is unstructured. Big Data is a way to collect a variety of information and data, and analyze the useful information to the user by the efficient data processing technology. The development of big data has four directions as shown in Figure 2.1:

• Volume: The amount of generated and stored data, the amount of data can easily reach tens of TB.

• Velocity: Refers to the speed of data processing.

• Variety: The types of big data source is very diverse, the easiest way is classified it structured and unstructured, structured data is mostly text, unstructured data are many be music, picture and video. These unstructured data cause difficulty on storage, mining and analyzing.

- Veracity: How to analyze and filter data error, data forgery and data anomaly, to prevent these “dirty data” damage data integrity and data accuracy.

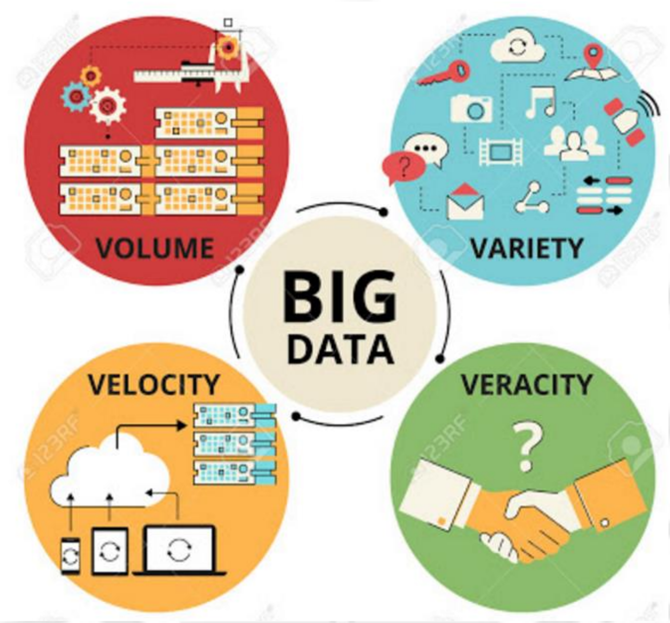


Figure 2.1: Big data-4V

2.2 Time Series

Time series data is based on regular time interval to do the continuous observation of the measured value. Time series analysis is to analyze the characteristics of the time series data has been occurred to predict the future value of the process. The analyze purpose of the time series is collected from the time series data and various type of models, such as trends, seasons, intervening events and other characteristics. Those models can summed up and estimated to reflect the historical data of the time series model.

Time series data according to observations can be divided into continuous or discrete, also called continuous time series and discrete time series. Time series are generally randomly distributed, and the future results of the sequence cannot be determined. The way which expressed in probability allocation, known as non-deterministic time series or stochastic time series. If the time series was change with a mathematical function, the prediction results are fixed, which called deterministic time series. Most of the time series data types are random, which can be divided into stationary, non-directional, trend, seasonal, and intervention events according to their sequence order and fluctuation.

2.2.1 Stationary Time Series

The observed values of the stationary time series are the same fixed level and same fixed regions, and this values do not change with time, as shown in Figure 2.2 If there is no special change or outlier, it is reasonable to deduce that the future observations of such time series are still at the same level and interval. Additionally, the predictive result can be improved by the dependency of successive observations.

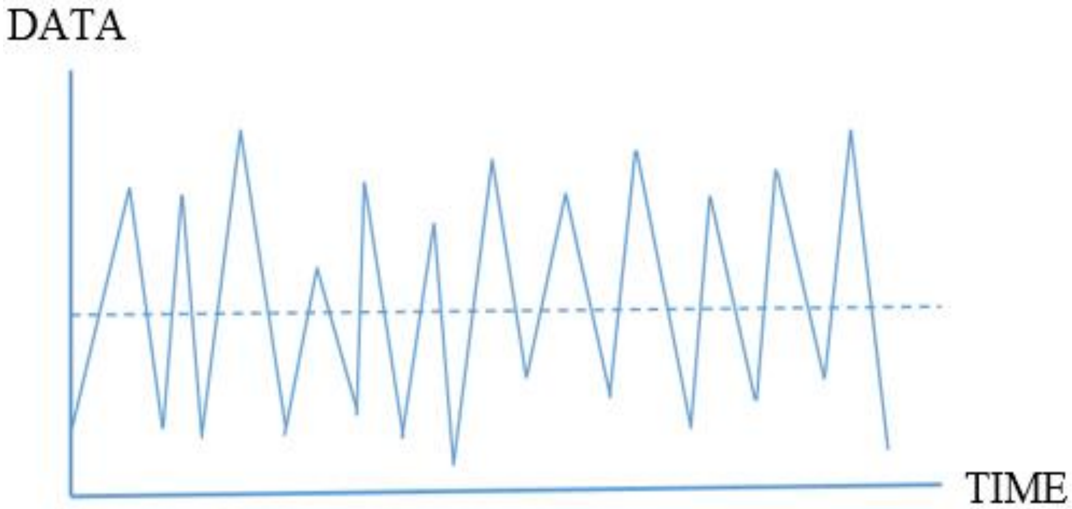


Figure 2.2: Stationary Time Series

2.2.2 Non-Stationary Time Series

The Non-Stationary time series will become fluctuating non-direction if it encounter the disturbed time series, as shown in Figure 2.3 External impact causes cumulative effects on the series, making the series unable to maintain a fixed level and difficult to estimate the predicted value.

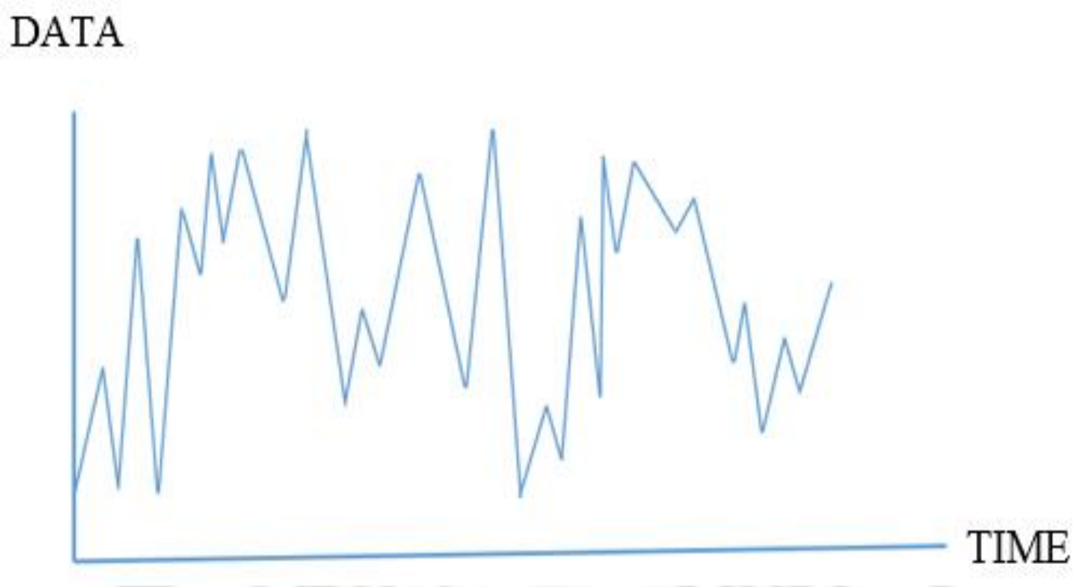


Figure 2.3: Non-Stationary Time Series

2.2.3 Trend Time Series

Trend time series is usually affected by long-term factors, leading to the average level of changes in the fixed trend. Each time points scattered variation is fixed, as shown in Figure 2.4 The average level of this series changes over time, so it can be assumed that this long-term factor will continuously effect the series in a fixed way，and finally conclude the predicted value.

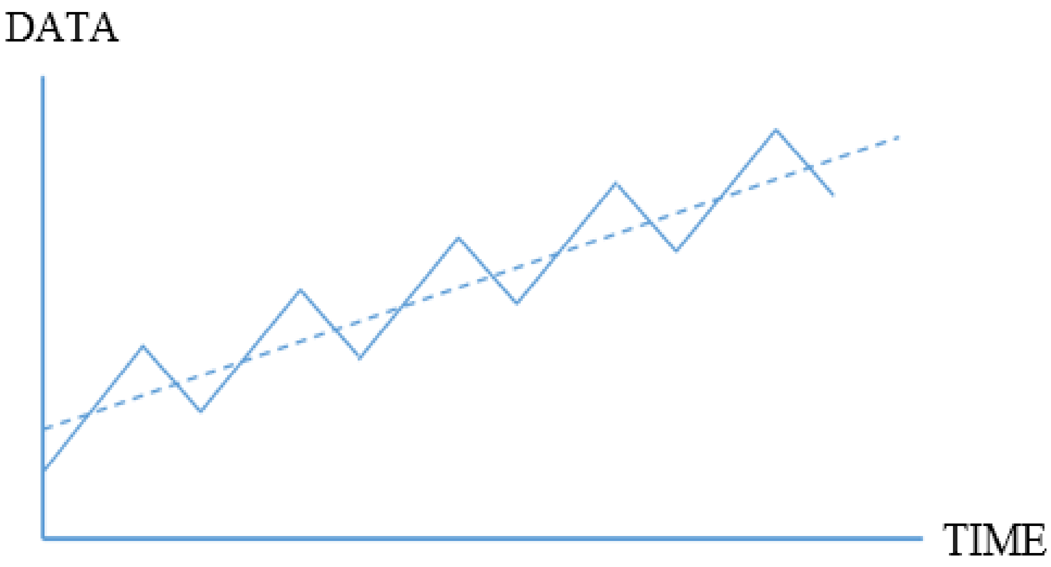


Figure 2.4: Trend Time Series

2.2.4 Season Time Series

Season time series can observe a similar fluctuation within a fixed time interval. As Figure 2.5 is a time series with both seasonal and trend factors. Since the average level of this type of sequence has a periodic change, it can be assumed that this period factor will continuously affect the series, and conclude the predicted value.

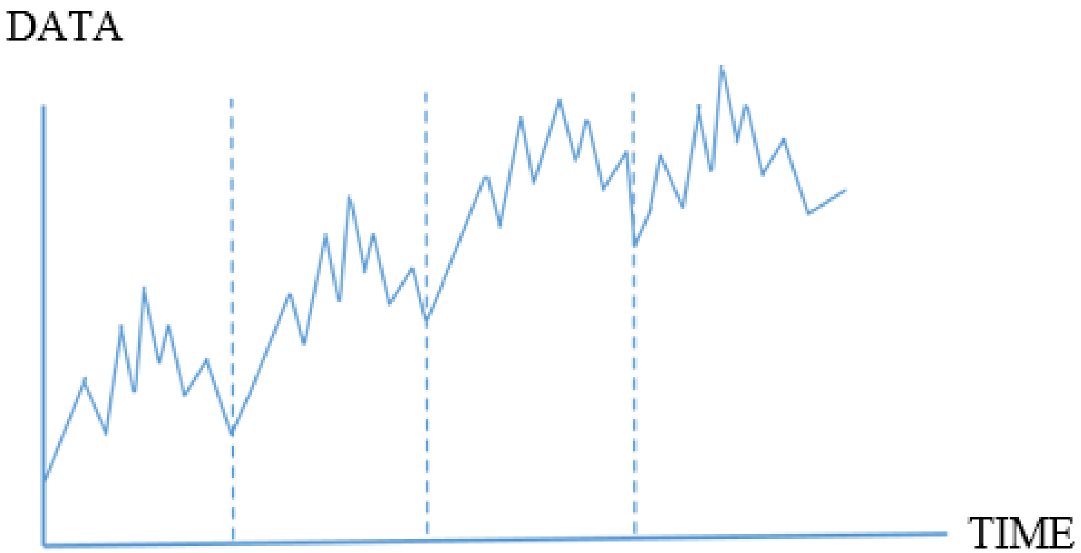


Figure 2.5: Season Time Series

2.2.5 Interventions Time Series

The interventions time series is disturbed by a single event, causing the minority observations in the series to behave differently from the other observations, as shown in Figure 2.6 Since the average level of this type of series does not change, and a single event is often unpredictable. Therefore, it can be assumed that this series will maintain the average level and change, and conclude the predicted value of series.

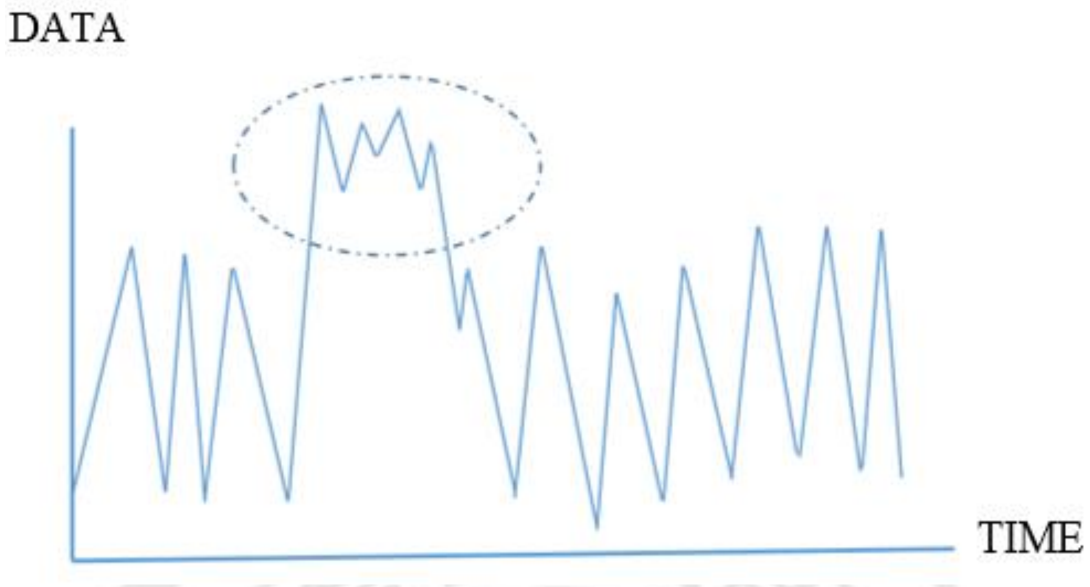
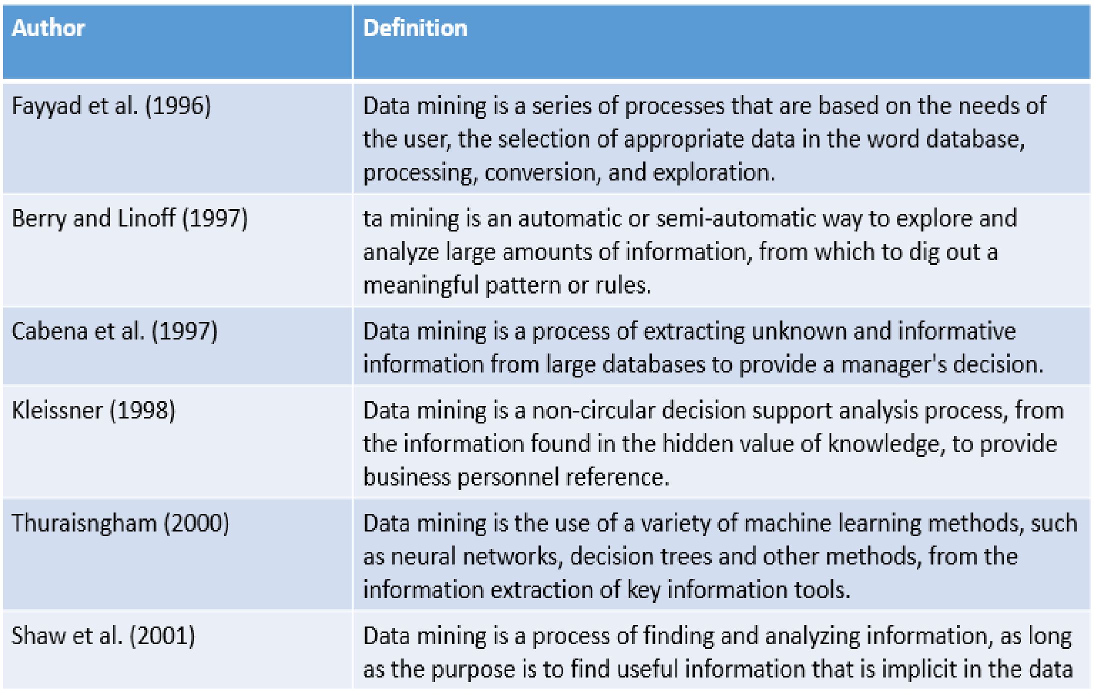


Figure 2.6: Interventions Time Series

2.3 Data Mining

Data mining is to analyze cumulative data and dig out a special data type or rules. Then through a series of data to sum up, organize and analyze the process, in order to obtain the most valuable information and knowledge.

Data mining is to collect the new information, past unknown information and the model that can be explained from a great amount of data. Those information or models can be used for predicting future has not yet occurred. Moreover, data mining could be seen as a process that keep analyzing data for decision support. The definition of data mining as shown in Figure 2.7



Author	Definition
Fayyad et al. (1996)	Data mining is a series of processes that are based on the needs of the user, the selection of appropriate data in the word database, processing, conversion, and exploration.
Berry and Linoff (1997)	ta mining is an automatic or semi-automatic way to explore and analyze large amounts of information, from which to dig out a meaningful pattern or rules.
Cabena et al. (1997)	Data mining is a process of extracting unknown and informative information from large databases to provide a manager's decision.
Kleissner (1998)	Data mining is a non-circular decision support analysis process, from the information found in the hidden value of knowledge, to provide business personnel reference.
Thuraisngham (2000)	Data mining is the use of a variety of machine learning methods, such as neural networks, decision trees and other methods, from the information extraction of key information tools.
Shaw et al. (2001)	Data mining is a process of finding and analyzing information, as long as the purpose is to find useful information that is implicit in the data

Figure 2.7: The Definition of Data Mining

2.3.1 Data Mining Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

2.3.2 Data Mining Process

The knowledge discovery in databases (KDD) process is commonly defined with the stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data mining
(5) Interpretation/evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding
(2) Data Understanding
(3) Data Preparation
(4) Modeling
(5) Evaluation
(6) Deployment

or a simplified process such as (1) Pre-processing, (2) Data Mining, and (3) Results Validation.

2.3.2.1 Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

2.3.2.2 Data mining involves six common classes of tasks:

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.

- Summarization – providing a more compact representation of the data set, including visualization and report generation.

2.3.2.3 Results validation

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as overfitting, but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening.

Chapter 3
Research Method

In this study, we selected 10 years of air pollution monitoring data from 2005 to 2015 at the Dali station, Zhongming station, Shairu station, Xitun station and Fengyuan station in Taichung area. Processing the data merging, data exploration, data analysis, and finally the use of autoregressive method to establish time series for the future data forecast, simulate the trend of air pollution in the city. Finally, we use MAPE (Mean absolute percentage error) to review the accuracy of the forecast. The analysis flow chart of this paper as shown in Figure3.1.

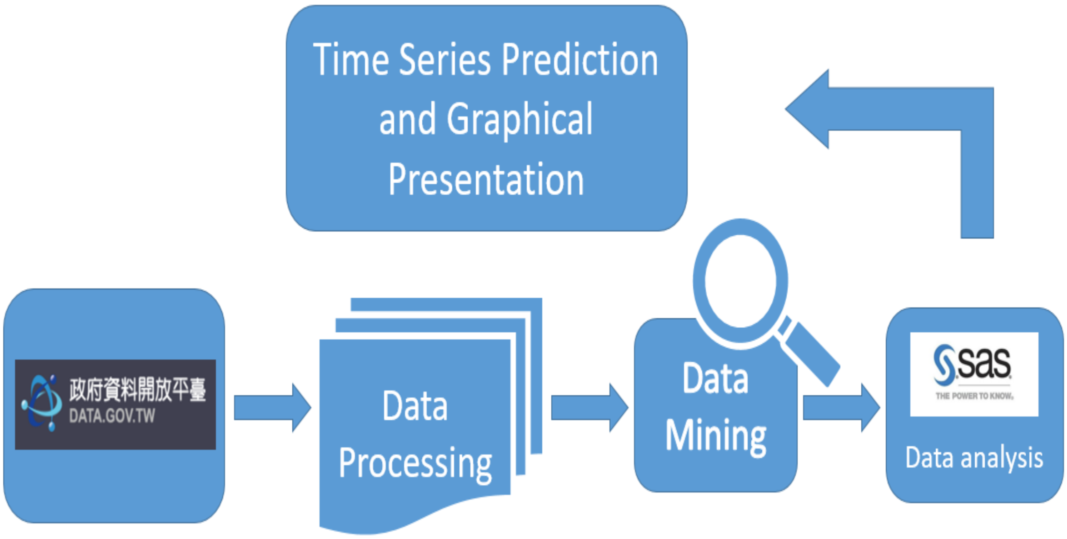


Figure 3.1: The Analysis Flow Chart

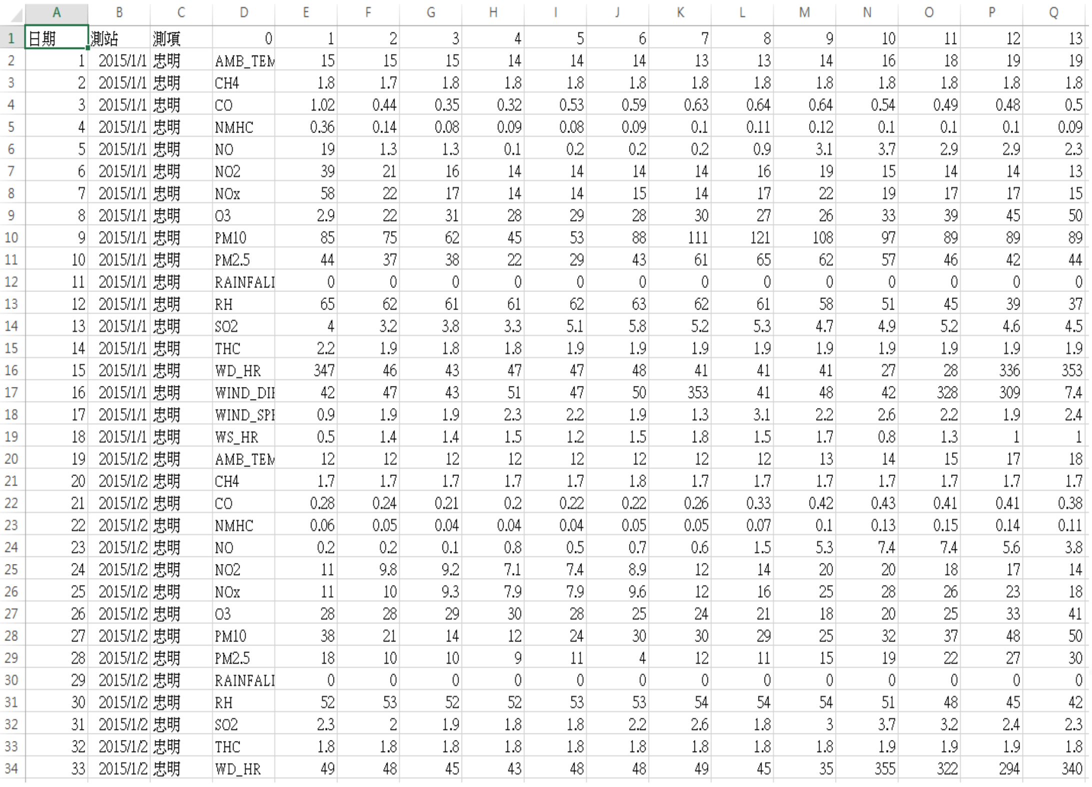
3.1 Research Instrument

This paper uses the tool software called SAS Enterprise Guide 7.1 workstation. The SAS Enterprise Guide's user interface is easy to use and with a powerful and versatile feature that allows users to import data in a variety of ways, such as Excel, Notepad, SAS data files, etc., Moreover, it can find hidden relationships between large data, and predict the future events. Among each analytical models, this software can be very convenient and easy to visualize and explain the results of the analysis or integration model.

3.2 Data preparation

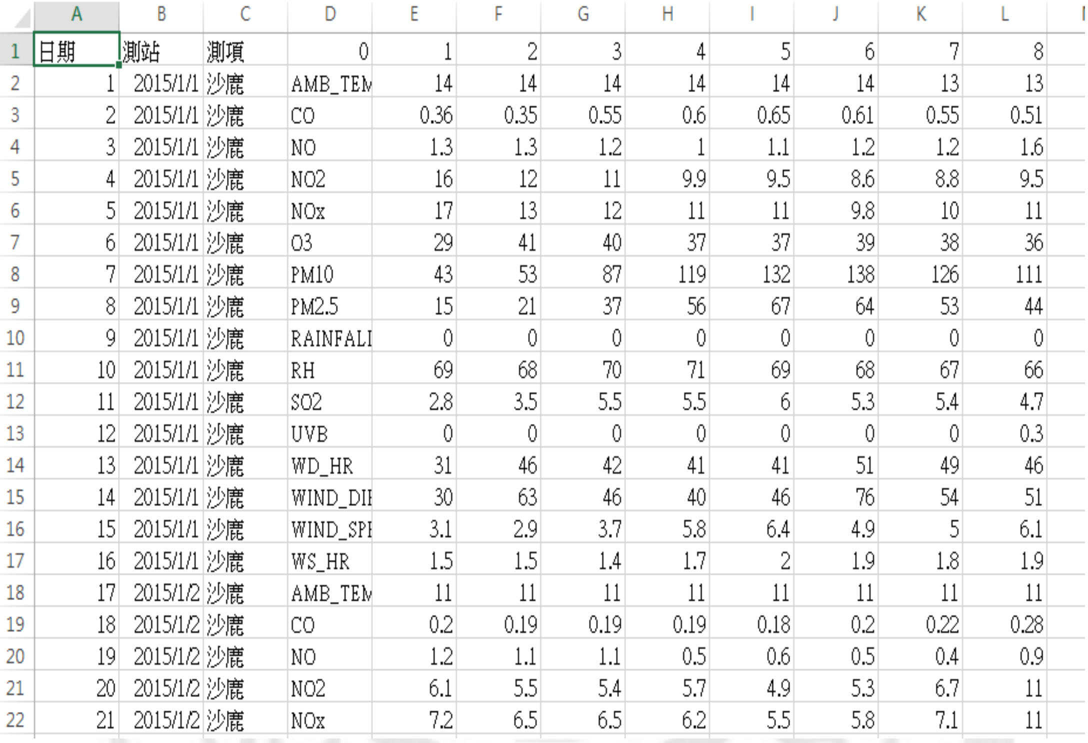
The source of this paper is from the opening platform of government. Collected the air pollution-related information between the five stations from the Taichung City in 2005 to 2015. There are more than 5 million data in this 11 years, including ten measure items such as temperature, CO , NO, NO2, NOx, O3, PM10, PM2.5, RAINFALL and SO2. The government open platform of the original data as shown in Table 3.2 and 3.3.

Table 3.1: The Government Open Platform of the Original Data (Zhongming Station)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	日期	測站	測項	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1	2015/1/1	忠明	AMB_TEM	15	15	15	14	14	14	13	13	14	16	18	19	19
3	2	2015/1/1	忠明	CH4	1.8	1.7	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
4	3	2015/1/1	忠明	CO	1.02	0.44	0.35	0.32	0.53	0.59	0.63	0.64	0.64	0.54	0.49	0.48	0.5
5	4	2015/1/1	忠明	NMHC	0.36	0.14	0.08	0.09	0.08	0.09	0.1	0.11	0.12	0.1	0.1	0.1	0.09
6	5	2015/1/1	忠明	NO	19	1.3	1.3	0.1	0.2	0.2	0.2	0.9	3.1	3.7	2.9	2.9	2.3
7	6	2015/1/1	忠明	NO2	39	21	16	14	14	14	14	16	19	15	14	14	13
8	7	2015/1/1	忠明	NOx	58	22	17	14	14	15	14	17	22	19	17	17	15
9	8	2015/1/1	忠明	O3	2.9	22	31	28	29	28	30	27	26	33	39	45	50
10	9	2015/1/1	忠明	PM10	85	75	62	45	53	88	111	121	108	97	89	89	89
11	10	2015/1/1	忠明	PM2.5	44	37	38	22	29	43	61	65	62	57	46	42	44
12	11	2015/1/1	忠明	RAINFALI	0	0	0	0	0	0	0	0	0	0	0	0	0
13	12	2015/1/1	忠明	RH	65	62	61	61	62	63	62	61	58	51	45	39	37
14	13	2015/1/1	忠明	SO2	4	3.2	3.8	3.3	5.1	5.8	5.2	5.3	4.7	4.9	5.2	4.6	4.5
15	14	2015/1/1	忠明	THC	2.2	1.9	1.8	1.8	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9
16	15	2015/1/1	忠明	WD_HR	347	46	43	47	47	48	41	41	41	27	28	336	353
17	16	2015/1/1	忠明	WIND_DIF	42	47	43	51	47	50	353	41	48	42	328	309	7.4
18	17	2015/1/1	忠明	WIND_SPI	0.9	1.9	1.9	2.3	2.2	1.9	1.3	3.1	2.2	2.6	2.2	1.9	2.4
19	18	2015/1/1	忠明	WS_HR	0.5	1.4	1.4	1.5	1.2	1.5	1.8	1.5	1.7	0.8	1.3	1	1
20	19	2015/1/2	忠明	AMB_TEM	12	12	12	12	12	12	12	12	13	14	15	17	18
21	20	2015/1/2	忠明	CH4	1.7	1.7	1.7	1.7	1.7	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.7
22	21	2015/1/2	忠明	CO	0.28	0.24	0.21	0.2	0.22	0.22	0.26	0.33	0.42	0.43	0.41	0.41	0.38
23	22	2015/1/2	忠明	NMHC	0.06	0.05	0.04	0.04	0.04	0.05	0.05	0.07	0.1	0.13	0.15	0.14	0.11
24	23	2015/1/2	忠明	NO	0.2	0.2	0.1	0.8	0.5	0.7	0.6	1.5	5.3	7.4	7.4	5.6	3.8
25	24	2015/1/2	忠明	NO2	11	9.8	9.2	7.1	7.4	8.9	12	14	20	20	18	17	14
26	25	2015/1/2	忠明	NOx	11	10	9.3	7.9	7.9	9.6	12	16	25	28	26	23	18
27	26	2015/1/2	忠明	O3	28	28	29	30	28	25	24	21	18	20	25	33	41
28	27	2015/1/2	忠明	PM10	38	21	14	12	24	30	30	29	25	32	37	48	50
29	28	2015/1/2	忠明	PM2.5	18	10	10	9	11	4	12	11	15	19	22	27	30
30	29	2015/1/2	忠明	RAINFALI	0	0	0	0	0	0	0	0	0	0	0	0	0
31	30	2015/1/2	忠明	RH	52	53	52	52	53	53	54	54	54	51	48	45	42
32	31	2015/1/2	忠明	SO2	2.3	2	1.9	1.8	1.8	2.2	2.6	1.8	3	3.7	3.2	2.4	2.3
33	32	2015/1/2	忠明	THC	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.9	1.9	1.9	1.8
34	33	2015/1/2	忠明	WD_HR	49	48	45	43	48	48	49	45	35	355	322	294	340

Table 3.2: The Government Open Platform of the Original Data (Shairu station)



	A	B	C	D	E	F	G	H	I	J	K	L
1	日期	測站	測項	0	1	2	3	4	5	6	7	8
2	1	2015/1/1	沙鹿	AMB_TEM	14	14	14	14	14	14	13	13
3	2	2015/1/1	沙鹿	CO	0.36	0.35	0.55	0.6	0.65	0.61	0.55	0.51
4	3	2015/1/1	沙鹿	NO	1.3	1.3	1.2	1	1.1	1.2	1.2	1.6
5	4	2015/1/1	沙鹿	NO2	16	12	11	9.9	9.5	8.6	8.8	9.5
6	5	2015/1/1	沙鹿	NOx	17	13	12	11	11	9.8	10	11
7	6	2015/1/1	沙鹿	O3	29	41	40	37	37	39	38	36
8	7	2015/1/1	沙鹿	PM10	43	53	87	119	132	138	126	111
9	8	2015/1/1	沙鹿	PM2.5	15	21	37	56	67	64	53	44
10	9	2015/1/1	沙鹿	RAINFALI	0	0	0	0	0	0	0	0
11	10	2015/1/1	沙鹿	RH	69	68	70	71	69	68	67	66
12	11	2015/1/1	沙鹿	SO2	2.8	3.5	5.5	5.5	6	5.3	5.4	4.7
13	12	2015/1/1	沙鹿	UVB	0	0	0	0	0	0	0	0.3
14	13	2015/1/1	沙鹿	WD_HR	31	46	42	41	41	51	49	46
15	14	2015/1/1	沙鹿	WIND_DII	30	63	46	40	46	76	54	51
16	15	2015/1/1	沙鹿	WIND_SPI	3.1	2.9	3.7	5.8	6.4	4.9	5	6.1
17	16	2015/1/1	沙鹿	WS_HR	1.5	1.5	1.4	1.7	2	1.9	1.8	1.9
18	17	2015/1/2	沙鹿	AMB_TEM	11	11	11	11	11	11	11	11
19	18	2015/1/2	沙鹿	CO	0.2	0.19	0.19	0.19	0.18	0.2	0.22	0.28
20	19	2015/1/2	沙鹿	NO	1.2	1.1	1.1	0.5	0.6	0.5	0.4	0.9
21	20	2015/1/2	沙鹿	NO2	6.1	5.5	5.4	5.7	4.9	5.3	6.7	11
22	21	2015/1/2	沙鹿	NOx	7.2	6.5	6.5	6.2	5.5	5.8	7.1	11

3.3Data Preprocessing

Data preprocessing includes data cleaning, data integration, data transformation, and data reduction (Peng & Chien, 2003). The preprocessing flow chart of this paper as shown in Figure3.2 and the data after processing as shown in Table 3.3, 3.4 and 3.5.

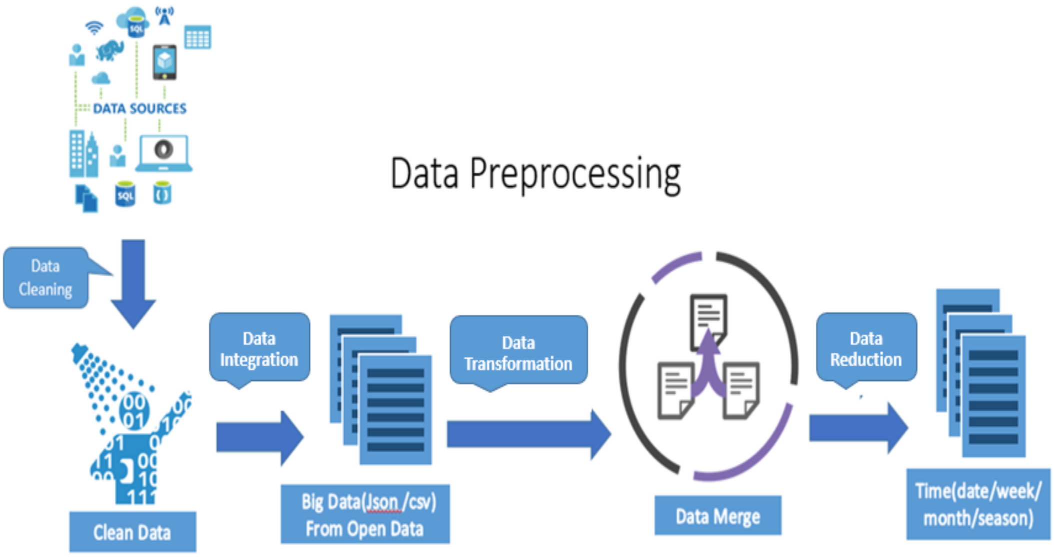


Figure 3.2: The Preprocessing Flow Chart

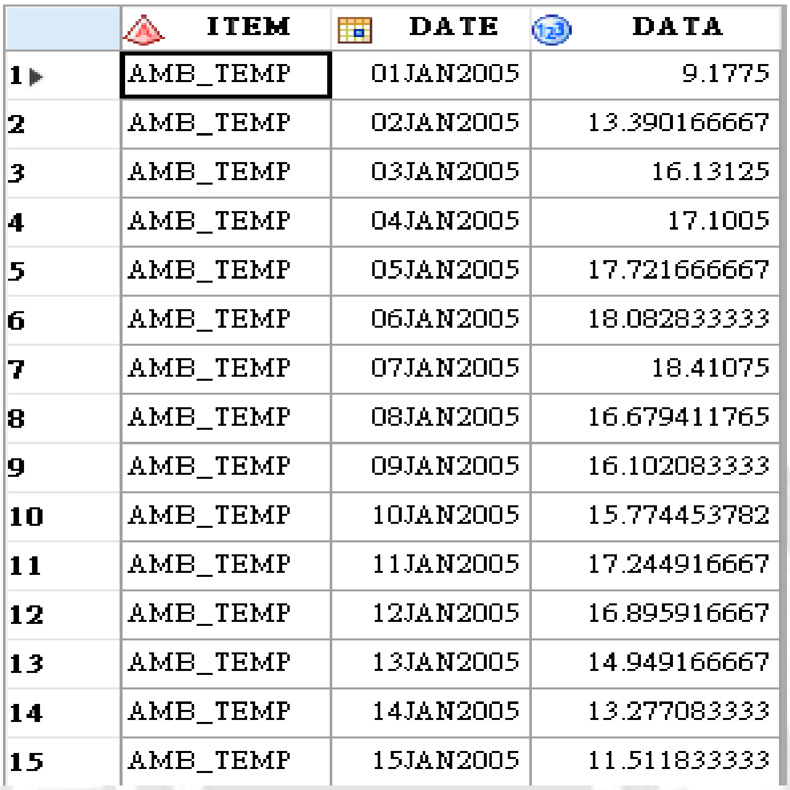
- Data cleaning: Handle the missing values, smoothing the messy data, finding the outliers, and correcting the inconsistency of the data. This paper replaces the outliers and the missing values with averages.

- Data integration: Combining data from different sources than stored in the same data file. Different sources of information may be recognized as mistaken information due to attribute definition or unit definition Therefore, it must be re-view and put the same information together. Due to the public information of items on the annual number are different , we must sort out the annual common items, and then delete the non-common items, than those data could be integrated into same data file, this paper is the use of Visual Basic for Applications (VBA) integration is complete.

- Data transformation: Transforming data into a suitable form of mining. As the government open platform is the original source of csv file, and csv file cannot store more than 5 million copies of data. Therefore, this paper uses the SAS Enterprise Guide to convert data into SAS file.

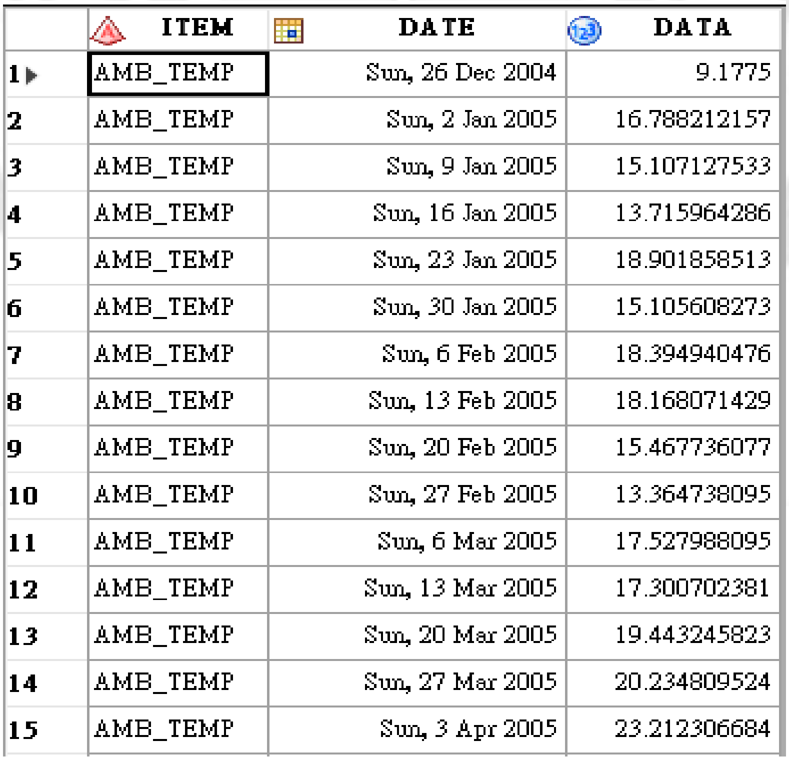
- Data reduction: Data will affect the establishment of mining model, in general, high-dimensional data calculation is more complex, and need to spend more time. So the analyst must determine whether the need for the data reduction in order to reduce the data dimension. However, this data must retain the integrity of the information as much as possible. The public information provided by the government is based on an hourly basis. This paper transforms it into daily, weekly, monthly and quarterly units for analysis and forecast.

Table 3.3: Data after Processing (Based on days)



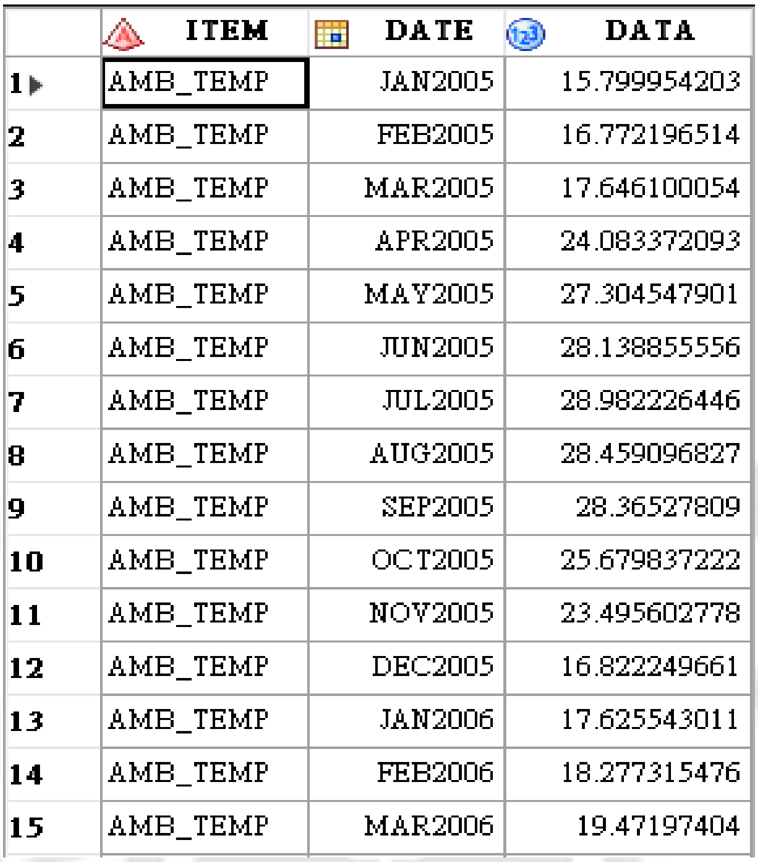
	ITEM	DATE	DATA
1 ▸	AMB_TEMP	01JAN2005	9.1775
2	AMB_TEMP	02JAN2005	13.390166667
3	AMB_TEMP	03JAN2005	16.13125
4	AMB_TEMP	04JAN2005	17.1005
5	AMB_TEMP	05JAN2005	17.721666667
6	AMB_TEMP	06JAN2005	18.082833333
7	AMB_TEMP	07JAN2005	18.41075
8	AMB_TEMP	08JAN2005	16.679411765
9	AMB_TEMP	09JAN2005	16.102083333
10	AMB_TEMP	10JAN2005	15.774453782
11	AMB_TEMP	11JAN2005	17.244916667
12	AMB_TEMP	12JAN2005	16.895916667
13	AMB_TEMP	13JAN2005	14.949166667
14	AMB_TEMP	14JAN2005	13.277083333
15	AMB_TEMP	15JAN2005	11.511833333

Table 3.4: Data after Processing (Based on Weeks)



	ITEM	DATE	DATA
1 ▸	AMB_TEMP	Sun, 26 Dec 2004	9.1775
2	AMB_TEMP	Sun, 2 Jan 2005	16.788212157
3	AMB_TEMP	Sun, 9 Jan 2005	15.107127533
4	AMB_TEMP	Sun, 16 Jan 2005	13.715964286
5	AMB_TEMP	Sun, 23 Jan 2005	18.901858513
6	AMB_TEMP	Sun, 30 Jan 2005	15.105608273
7	AMB_TEMP	Sun, 6 Feb 2005	18.394940476
8	AMB_TEMP	Sun, 13 Feb 2005	18.168071429
9	AMB_TEMP	Sun, 20 Feb 2005	15.467736077
10	AMB_TEMP	Sun, 27 Feb 2005	13.364738095
11	AMB_TEMP	Sun, 6 Mar 2005	17.527988095
12	AMB_TEMP	Sun, 13 Mar 2005	17.300702381
13	AMB_TEMP	Sun, 20 Mar 2005	19.443245823
14	AMB_TEMP	Sun, 27 Mar 2005	20.234809524
15	AMB_TEMP	Sun, 3 Apr 2005	23.212306684

Table 3.5: Data after Processing (Based on Months)



	ITEM	DATE	DATA
1 ▶	AMB_TEMP	JAN2005	15.799954203
2	AMB_TEMP	FEB2005	16.772196514
3	AMB_TEMP	MAR2005	17.646100054
4	AMB_TEMP	APR2005	24.083372093
5	AMB_TEMP	MAY2005	27.304547901
6	AMB_TEMP	JUN2005	28.138855556
7	AMB_TEMP	JUL2005	28.982226446
8	AMB_TEMP	AUG2005	28.459096827
9	AMB_TEMP	SEP2005	28.36527809
10	AMB_TEMP	OCT2005	25.679837222
11	AMB_TEMP	NOV2005	23.495602778
12	AMB_TEMP	DEC2005	16.822249661
13	AMB_TEMP	JAN2006	17.625543011
14	AMB_TEMP	FEB2006	18.277315476
15	AMB_TEMP	MAR2006	19.47197404

3.4 Autoregressive

In statistics and signal processing, an autoregressive (AR) model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term); thus the model is in the form of a stochastic difference equation.

Together with the moving-average (MA) model, it is a special case and key component of the more general ARMA and ARIMA models of time series, which have a more complicated stochastic structure; it is also a special case of the vector autoregressive model (VAR), which consists of a system of more than one interlocking stochastic difference equation in more than one evolving random variable.

Contrary to the moving-average model, the autoregressive model is not always stationary as it may contain a unit root. In this study, we will use the autoregressive method to analyze and predict the air pollution data in Taichung City.

3.5 Mean Absolute Percentage Error

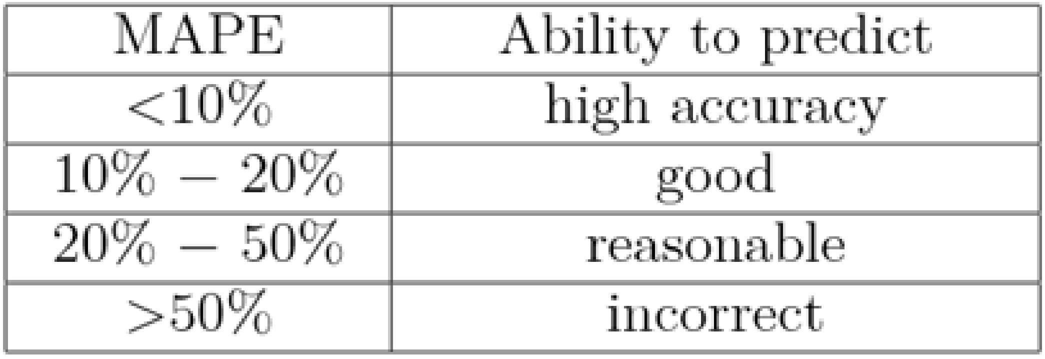
When the correctness of the model is compared, the MAPE (Mean Absolute Percentage Error) can be used as the criterion of evaluation. [27]. When the value is smaller, the difference between the predicted value and the observed value is smaller, the result is also better. The calculation method can be expressed as follows:

$$MAPE = \frac{1}{n}\sum\left|\frac{\mathrm{y} - y^t}{\mathrm{y}}\right|$$

y：Actual Value y^t：Forecast Value n ：Forecast Value Number

The Standards of predictive accuracy for MAPE as shown in Table3.3.

Table 3.6: Standards of predictive accuracy for MAPE.



MAPE	Ability to predict
<10%	high accuracy
10% – 20%	good
20% – 50%	reasonable
>50%	incorrect

Chapter 4
Experimental Results

This paper was based on data preprocessing, data merging, data exploration, data analysis, and finally using autoregressive method to obtain the drawing of the forecast value about air pollution data in Taichung City.

Figures 4.1, 4.2 and 4.3 are time-series predictions based on days, with temperature, CO and O3 as examples, black lines are actual values, and blue lines are predictions. The predicted and actual values of the known data are quite close to each other, as shown in Figures 4.1, 4.2 and 4.3. However, the prediction of the unknown value is very inaccurate.

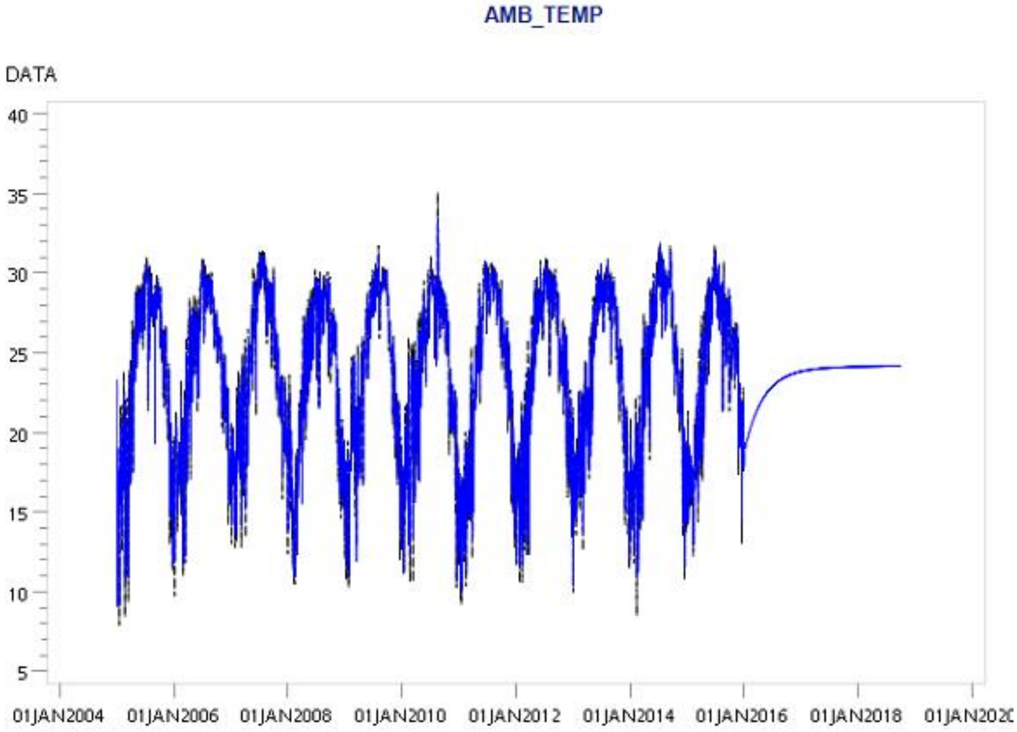


Figure 4.1: Time-Series Predictions Based On Days (Temperature)

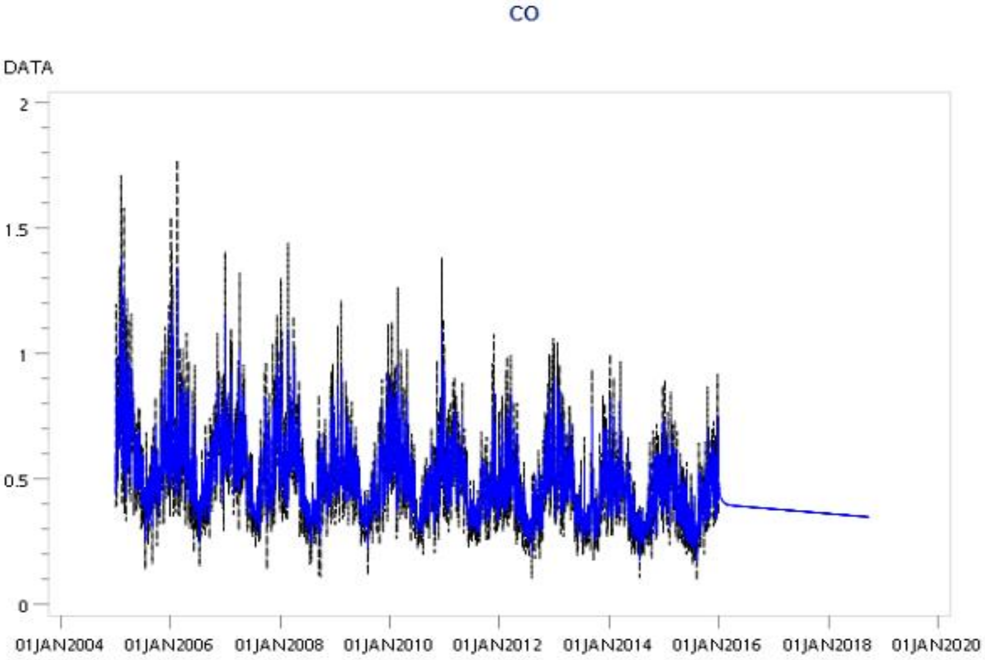


Figure 4.2: Time-Series Predictions Based on Days (CO)

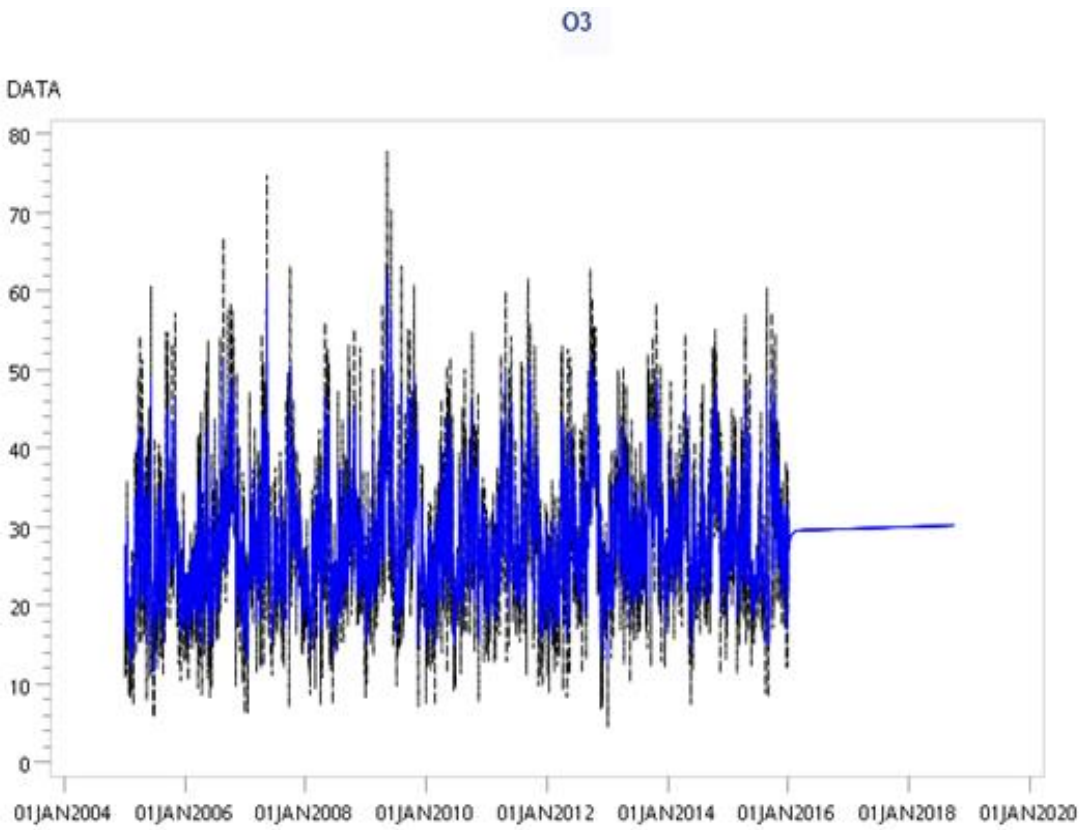


Figure 4.3: Time-Series Predictions Based on Days (O3)

Figures 4.4, 4.5 and 4.6 are the accuracy time chart of the MAPE (Mean Absolute Percentage Error) based on days, with temperature, CO and O3 as examples. Compared to Figures 4.1, 4.2 and 4.3, we can find that the temperature, CO and O3 are inaccurate for the known data in the extreme value, the temperature MAPE value of the temperature is 5.444854%, which is high accuracy. The MAPE value of CO is 20.28559%, which is reasonable. The MAPE value of O3 is 22.82577% and is reasonable.

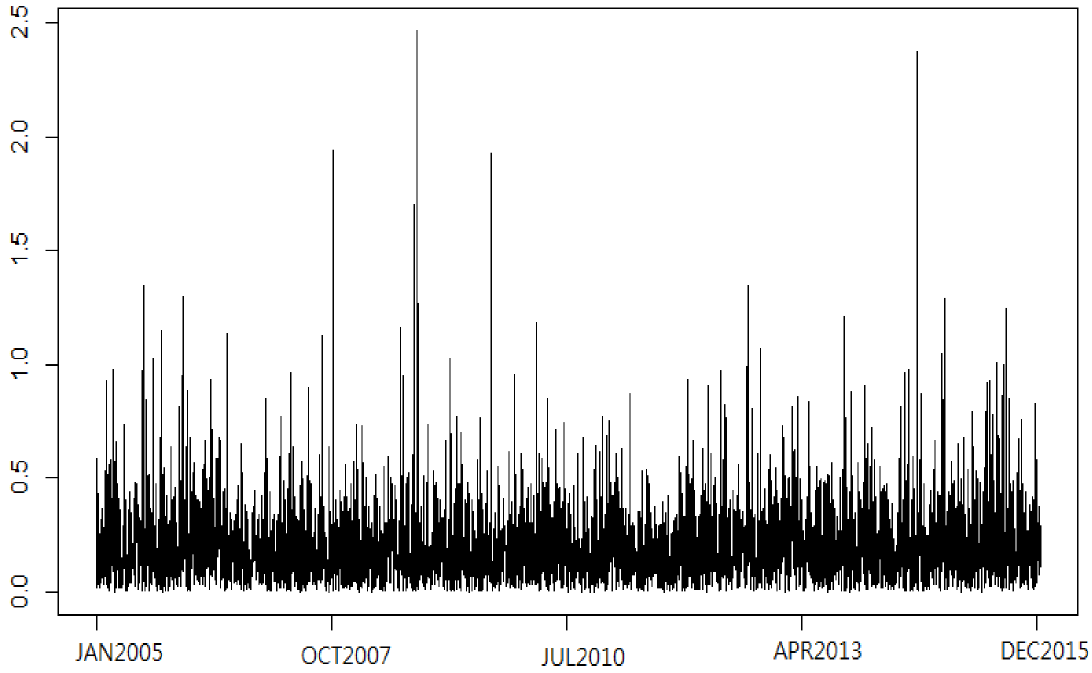


Figure 4.4: The Accuracy Time Chart of the MAPE Based On Days (Temperature)

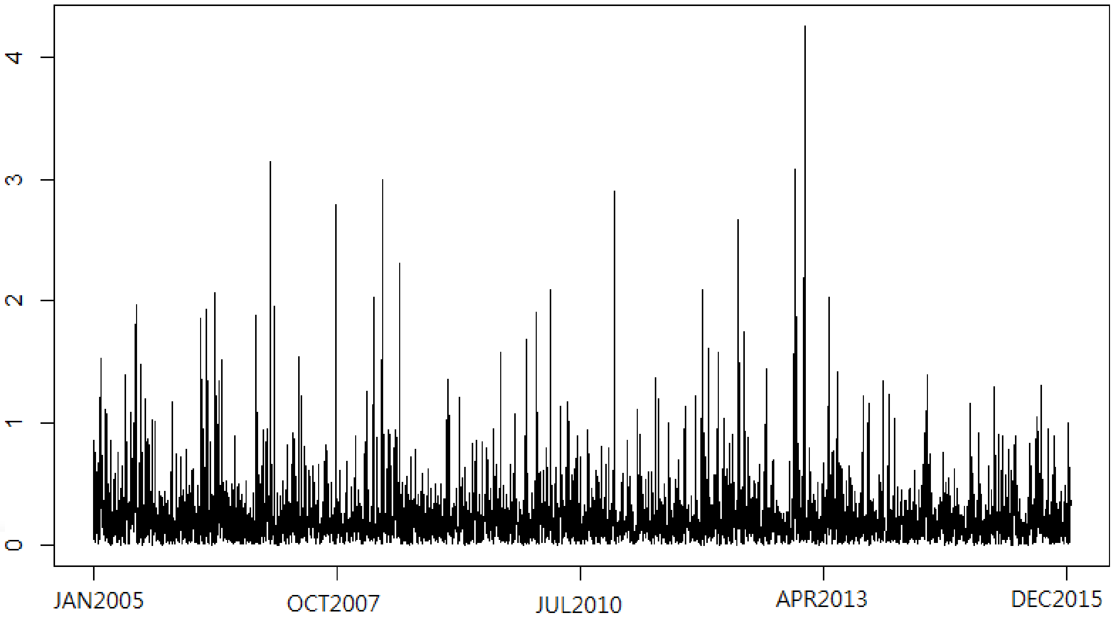


Figure 4.5: The Accuracy Time Chart of the MAPE Based on Days (CO)

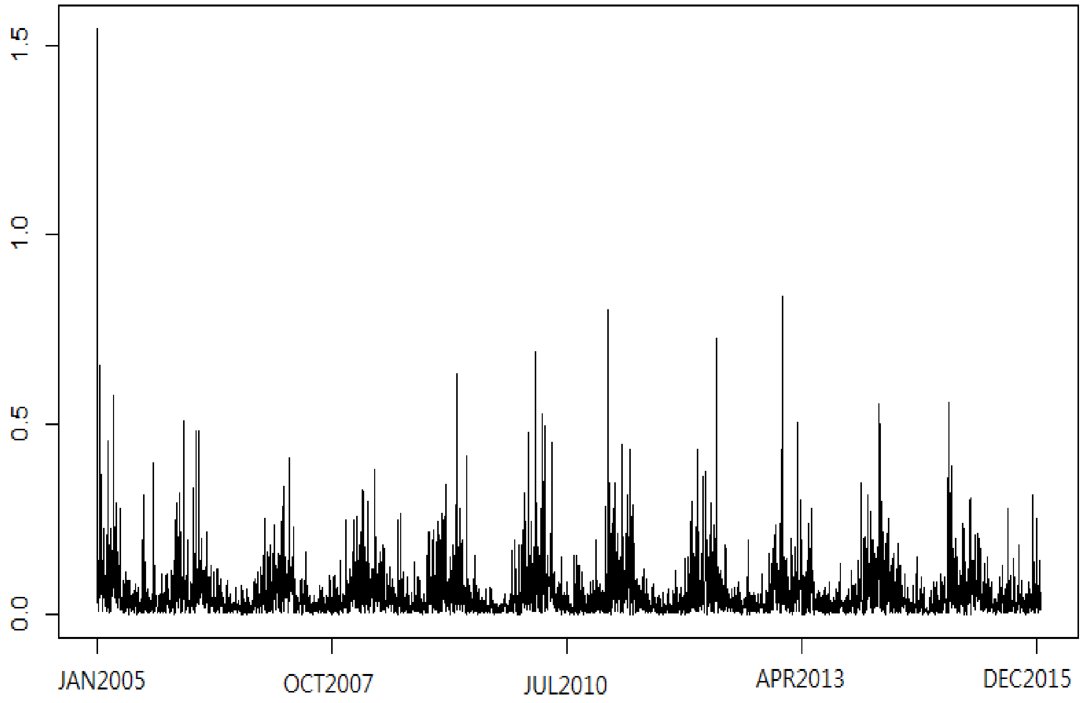


Figure 4.6: The Accuracy Time Chart of the MAPE Based on Days (O3)

Figures 4.7, 4.8 and 4.9 are time series example of temperature, CO and O3 that predict based on week, black lines are actual values, and blue lines are predictions. As shown in Figure 4.7, 4.8 and 4.9, the temperature, CO and the prediction of O3 is not accurate at the extreme value. The predicted value and the actual value of temperature data are quite close, so does the data of CO and O3. The temperature and O3 are gradually accurate at the beginning of the prediction of unknown values; and the prediction of CO at unknown values is not yet accurate.

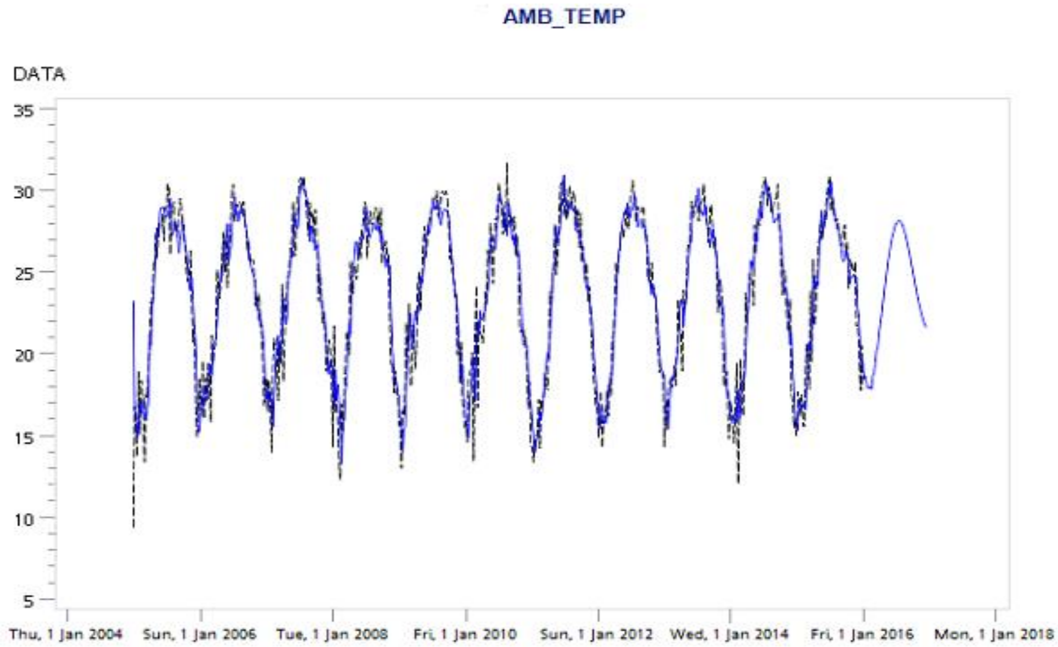


Figure 4.7: Time-Series Predictions Based on Weeks (Temperature)

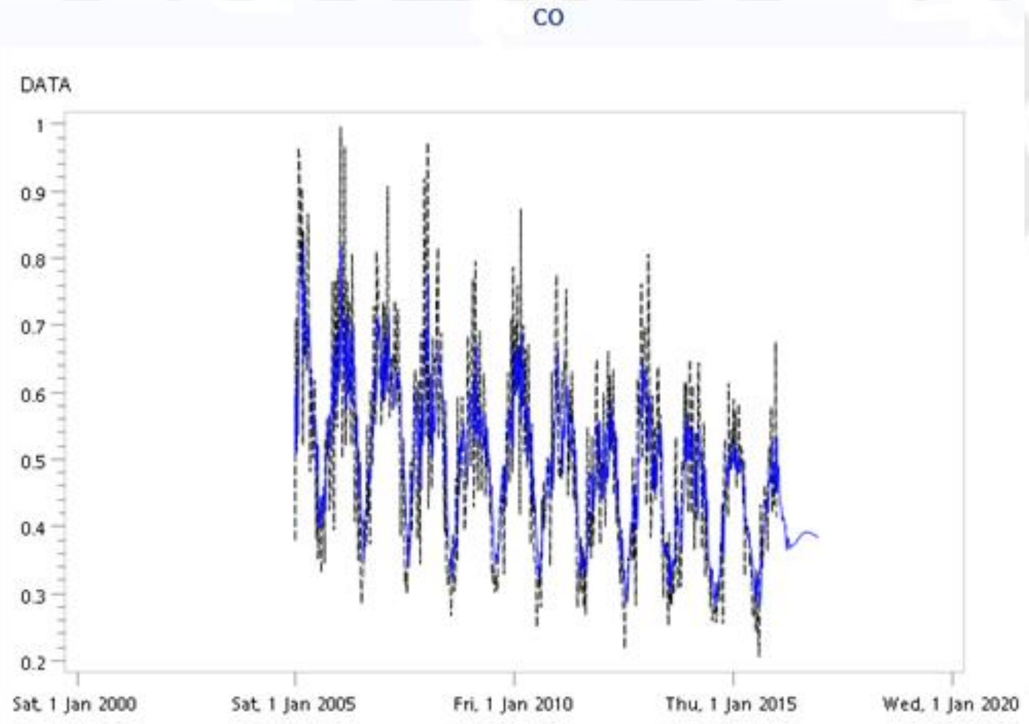


Figure 4.8: Time-Series Predictions Based on Weeks (CO)

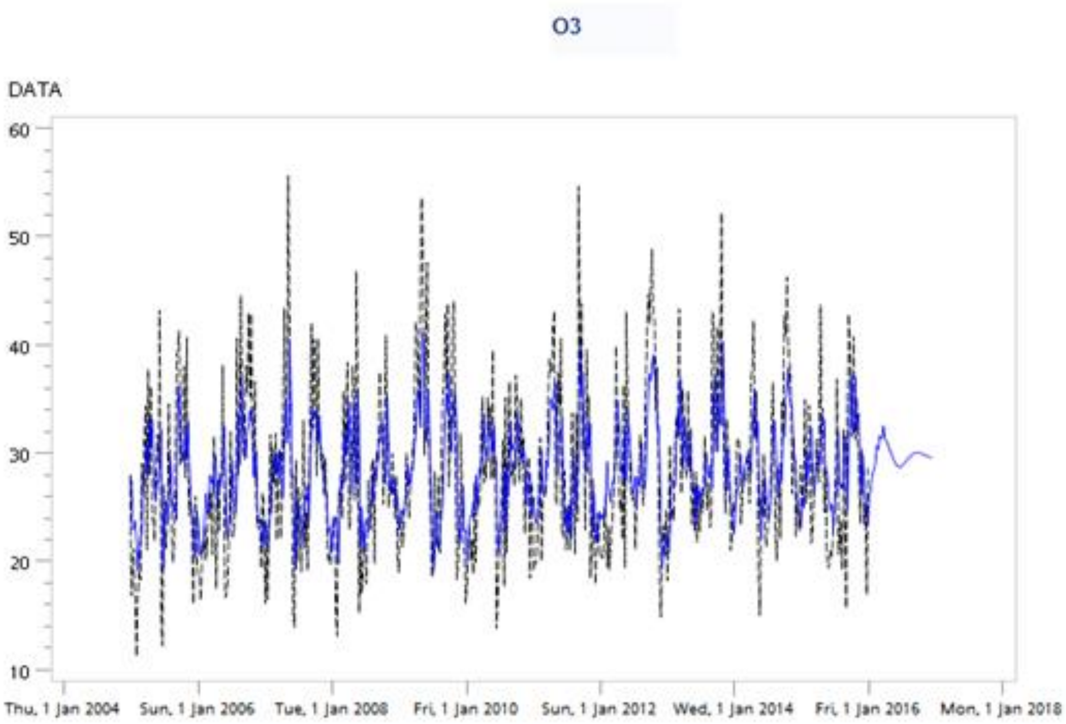


Figure 4.9: Time-Series Predictions Based on Weeks (O3)

Figures 4.10, 4.11 and 4.12 are the accuracy time chart of the MAPE (Mean Absolute Percentage Error) based on weeks, with temperature, CO and O3 as examples. Compared to Figures 4.7, 4.8 and 4.9, we also can find that the temperature, CO and O3 are inaccurate for the known data in the extreme value, the temperature MAPE value of the temperature is 6.498594 %, which is high accuracy. The MAPE value of CO is 16.08986%, which is good. The MAPE value of O3 is 17.95183%, which is good.

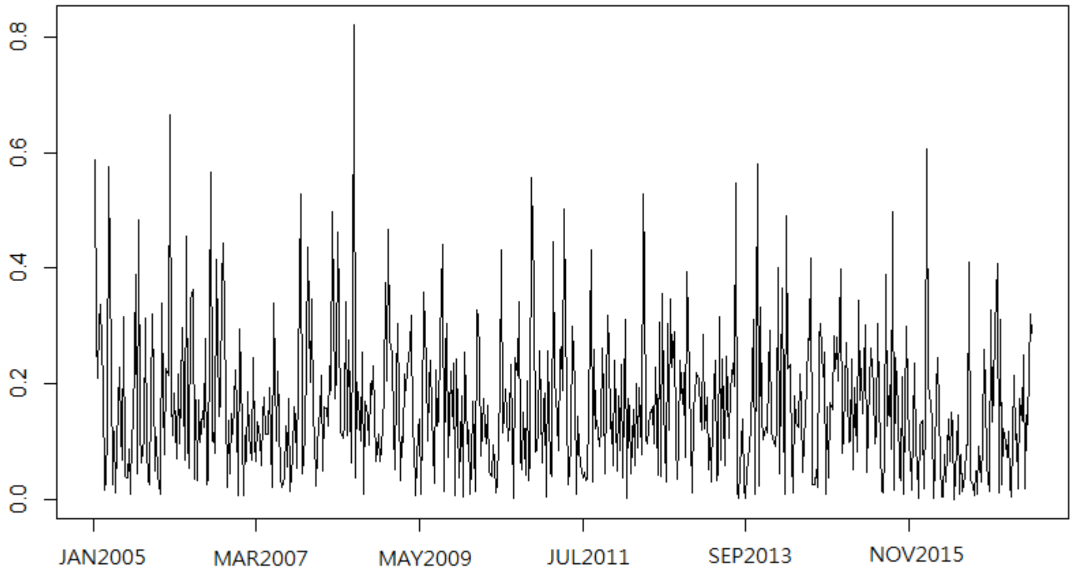


Figure 4.10: The Accuracy Time Chart of the MAPE Based on Weeks (Temperature)

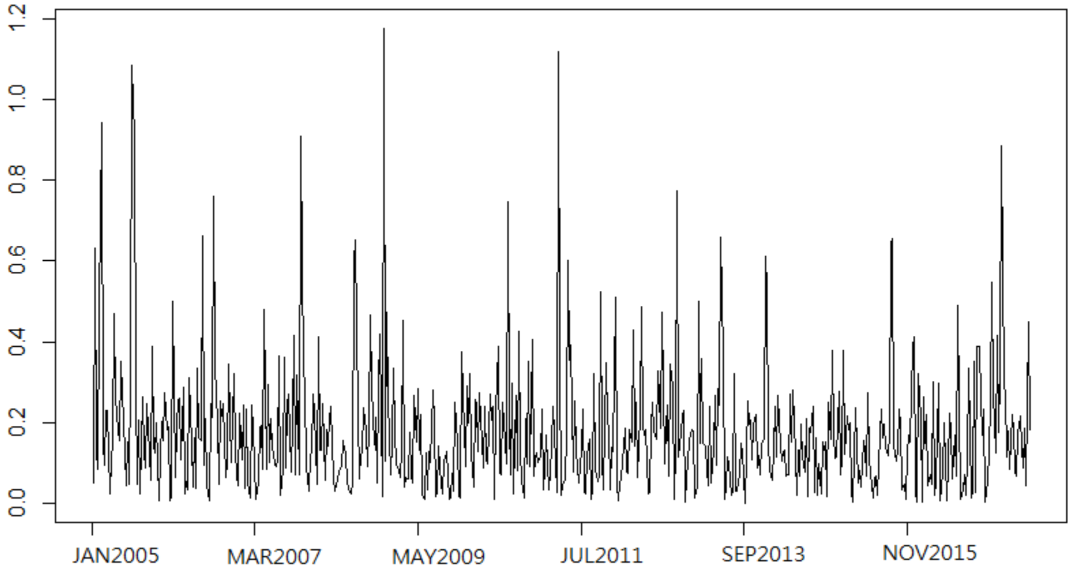


Figure 4.11: The Accuracy Time Chart of the MAPE Based on Weeks (CO)

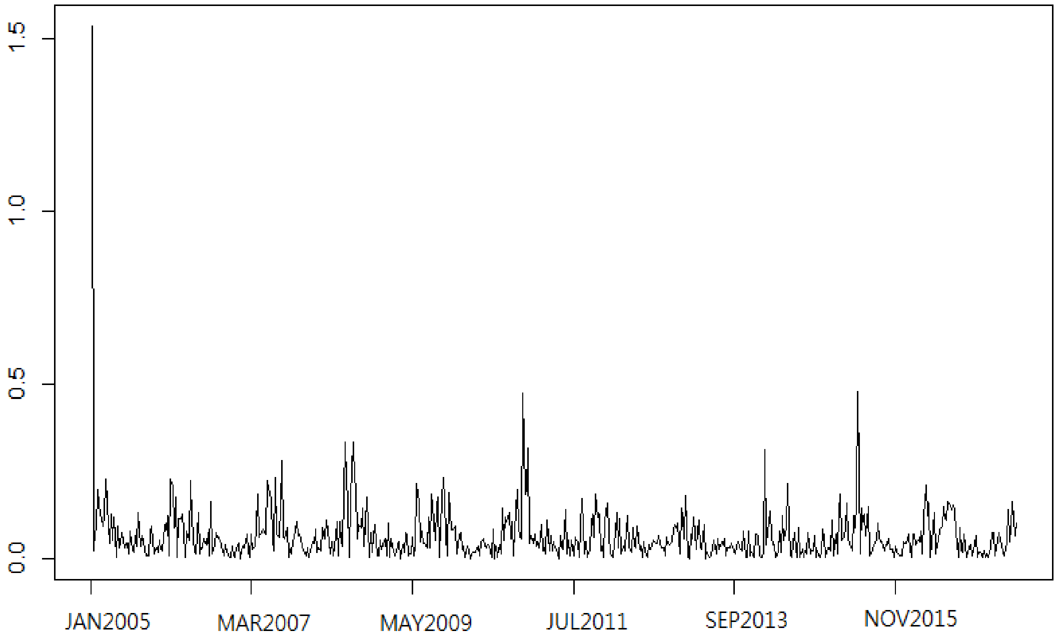


Figure 4.12: The Accuracy Time Chart of the MAPE Based on Weeks (O3)

Figures 4.13, 4.14 and 4.15 are time series example of temperature, CO and O3 that predict based on month, black lines are actual values, and blue lines are predictions. As shown in Figures 4.13, 4.14 and 4.15 which is based on monthly research, the prediction of the temperature in the extreme value is gradually accurate and the prediction of CO and O3 are not accurate at the extreme value. The predicted value and the actual value of the temperature are quite close, so do the data of CO and O3. The temperature at the unknown value is predicted to be accurate; and the prediction of the CO and O3 at the unknown value is accurate.

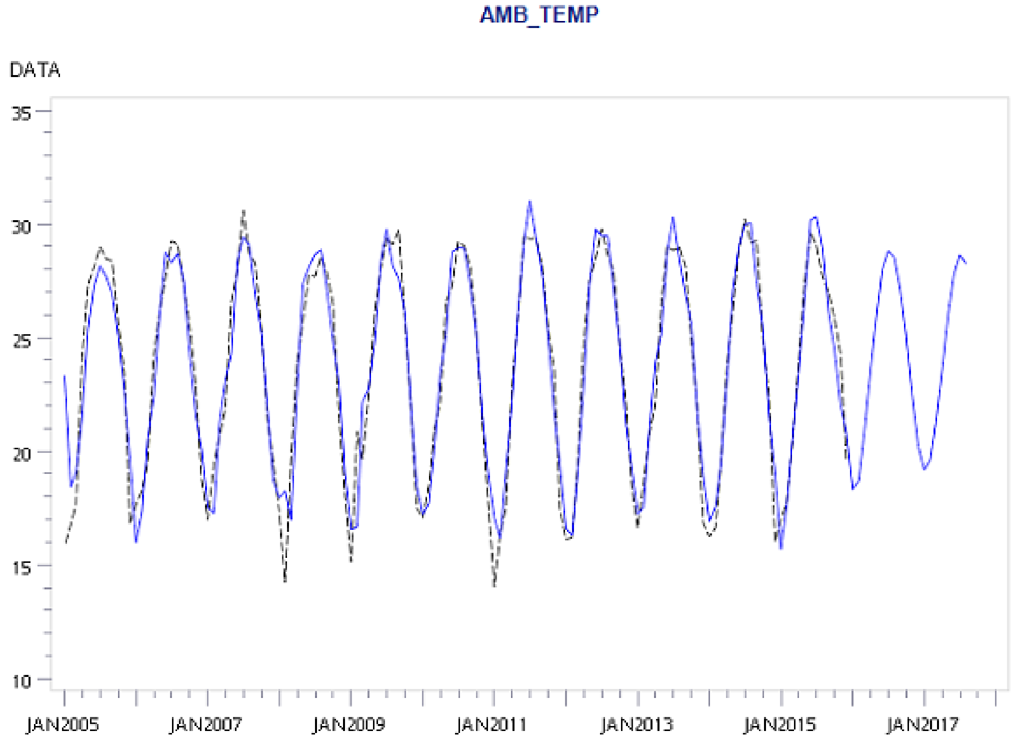


Figure 4.13: Time-Series Predictions Based On Month (Temperature)

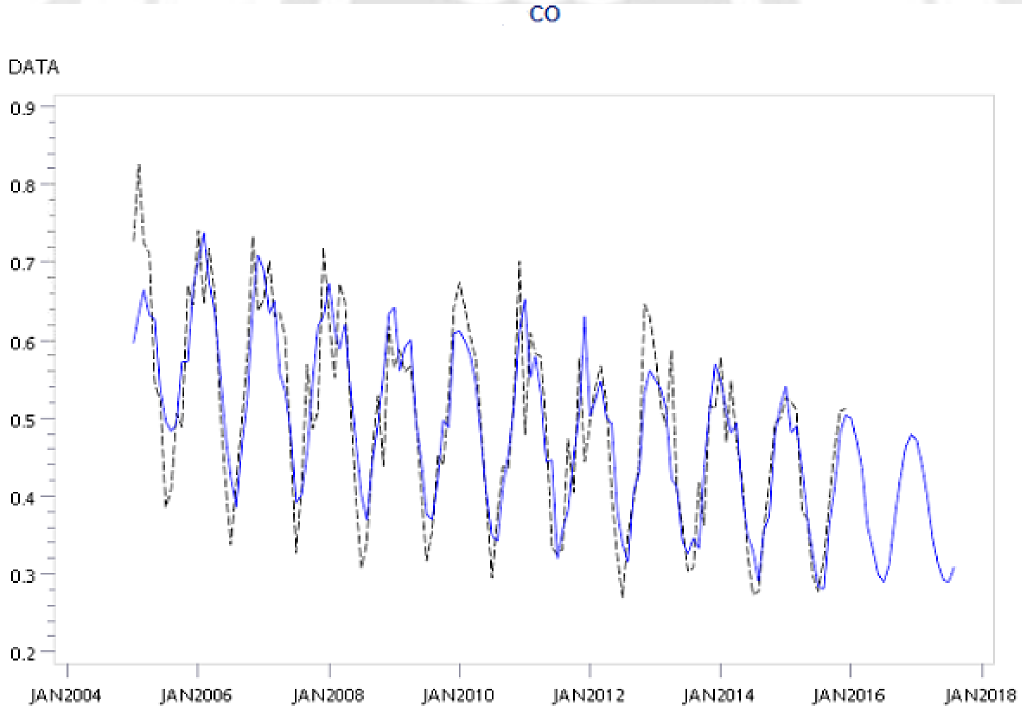


Figure 4.14: Time-Series Predictions Based On Month (CO)

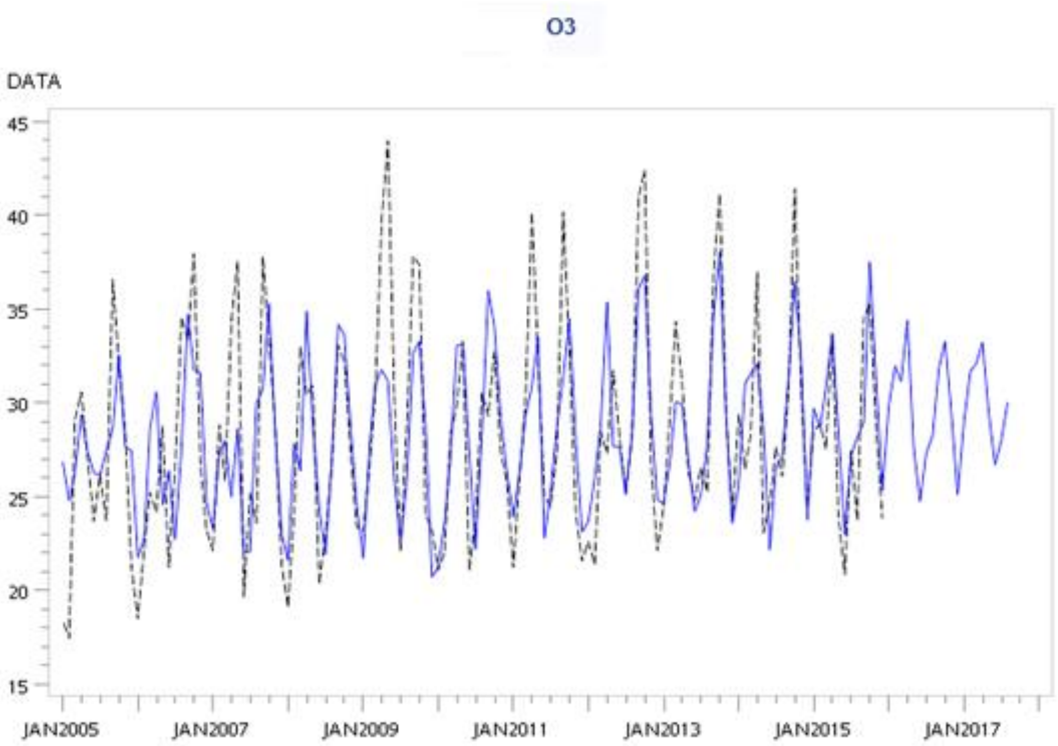


Figure 4.15: Time-Series Predictions Based On Month (O3)

Figures 4.16, 4.17 and 4.18 are the accuracy time chart of the MAPE (Mean Absolute Percentage Error) based on weeks, with temperature, CO and O3 as examples. Compared to Figures 4.16, 4.17 and 4.18, we also can find that the temperature, CO and O3 are inaccurate for the known data in the extreme value, the temperature MAPE value of the temperature is 5.073653%, which is high accuracy. The MAPE value of CO is 10.14667%, which is good. The MAPE value of O3 is 10.58128%, which is good.

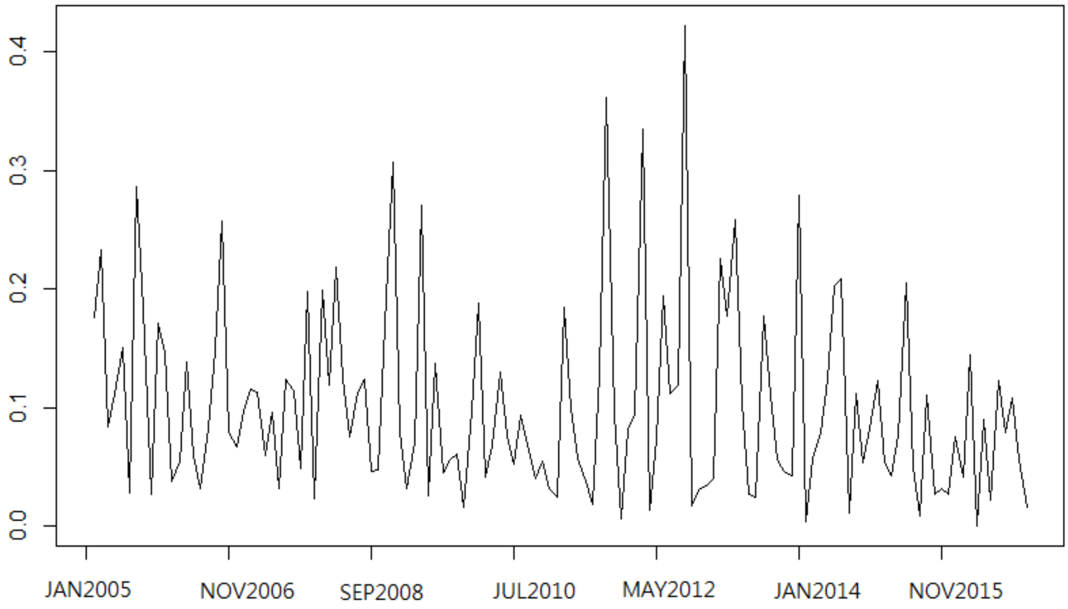


Figure 4.16: The Accuracy Time Chart of the MAPE Based on Months (Temperature)

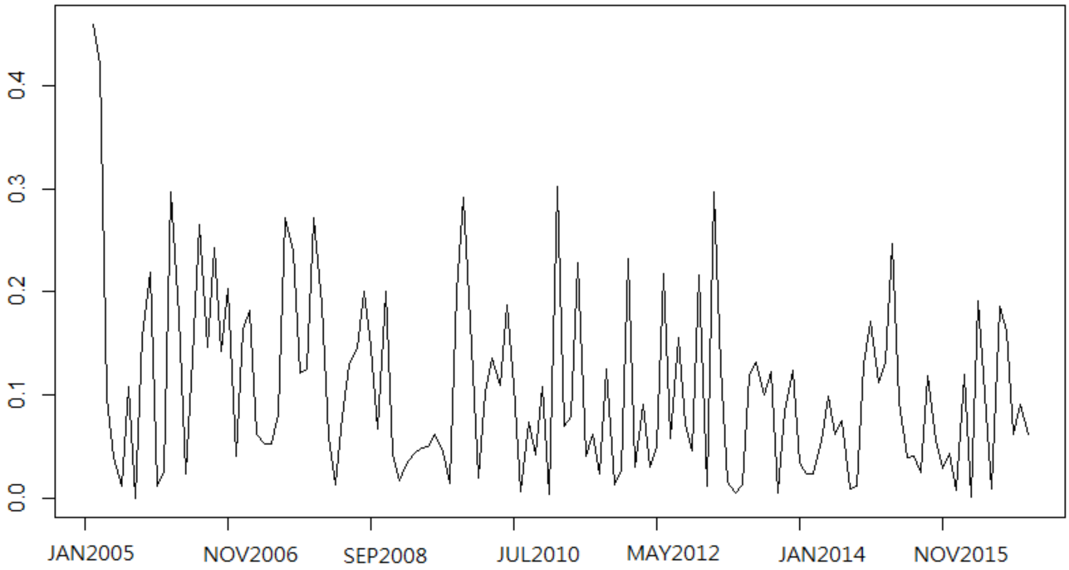


Figure 4.16: The Accuracy Time Chart of the MAPE Based on Months (CO)

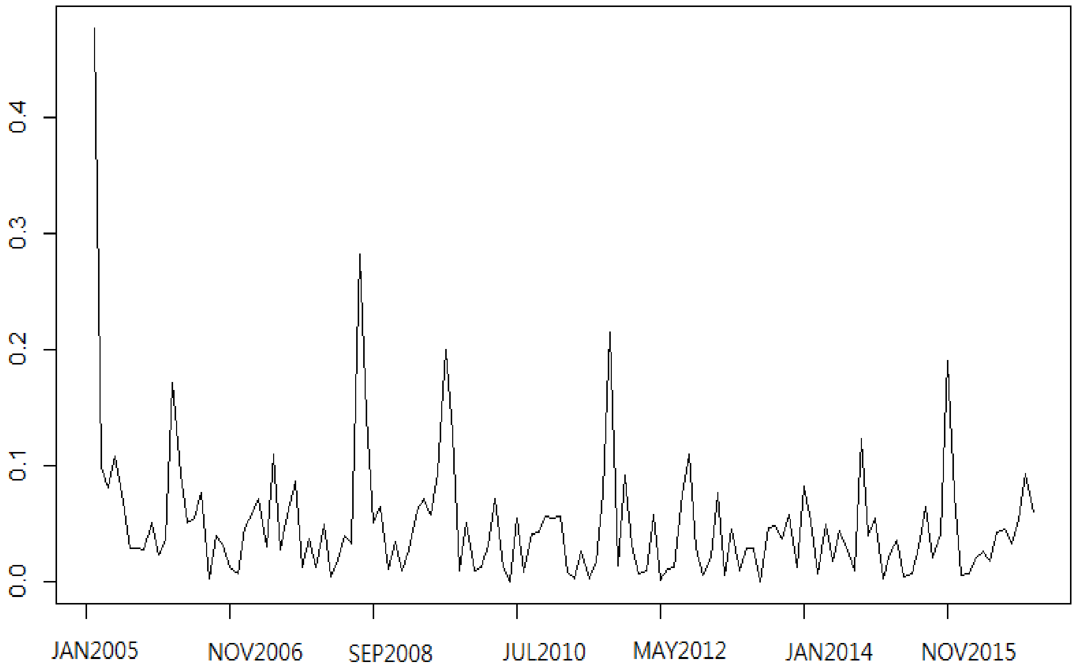


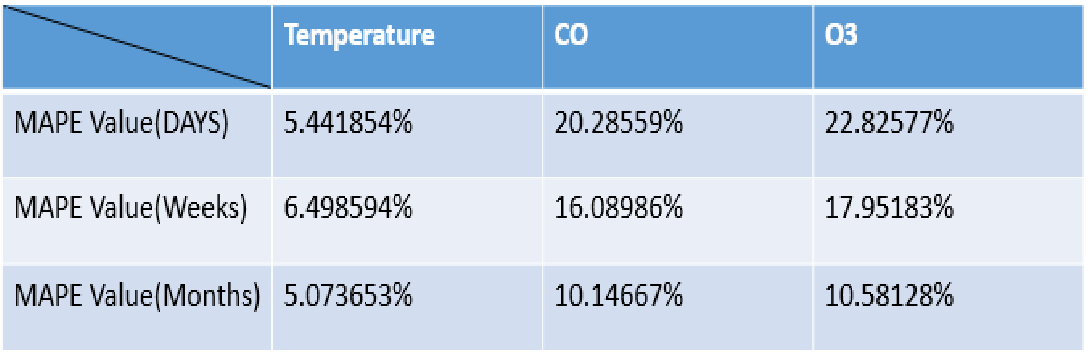
Figure 4.16: The Accuracy Time Chart of the MAPE Based on Months (O3)

Table 4 shows the MAPE values calculated based on days, weeks and months of temperature, CO, and O3. It can be found from the table that the MAPE value based on weeks is more accurate than based on days, and the MAPE value based on months is more accurate than based on weeks.

Although the forecast based on days is relatively inaccurate, but the forecast based on weeks and months are acceptable, people or the government can know that the changing of the previous week’s air pollution or the previous month's air pollution to do some precautionary measures in advance to reduce health hazards.

However, we found that the temperature prediction in time series is accurate, because the temperature changes with the seasons.

Table 4: MAPE Value



	Temperature	CO	O3
MAPE Value(DAYS)	5.441854%	20.28559%	22.82577%
MAPE Value(Weeks)	6.498594%	16.08986%	17.95183%
MAPE Value(Months)	5.073653%	10.14667%	10.58128%

Chapter 5
Conclusions and Future Work

It can be found that the time series prediction is not accurate based on day, because the stepwise autoregressive analysis is the historical time series of the predicted target in different periods. Moreover, the existence relation between the values is to establish the regression equation to predict. However, use the day for the research unit were changes too large that it cannot accurately predict in this data.

The future experiments will use mean absolute percentage error test method to test predictive accuracy. For different time series patterns will with different analytical methods to do the analysis and prediction.

References

[1] Paul Bourke, “autoregressive analysis (AR),” November 1998. http://paulbourke.net/miscellaneous/ar/

[2] Seng Hansun, “A New Approach of Moving Average Method in Time Series Analysis,” http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6708545

[3] P. Stoica, T. Soderstrom, and B. Friedlander, “Optimal instrumental variable estimates of the AR parameters of an ARMA process,” IEEE Trans. Automat. Control, vol. 30, no. 11, pp. 1066–1074, Nov. 1985

[4] Regression Analysis, 2017 https://en.wikipedia.org/wiki/Regression_analysis

[5] L. A. Zadeh, “Fuzzy sets,” Inf. Control, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[6] I. H. Kuo, S. J. Horng, T. W. Kao, T. L. Lin, C. L. Lee, and Y. Pan, “An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization,” Expert Syst. Appl., vol. 36, no. 3, pp. 6108–6117, Apr. 2008. DOI: 10.1016/j.eswa.2008.07.043.

[7] Q. Song and B. S. Chissom, “Forecasting enrollments with fuzzy time series-Part 1,” Fuzzy Sets Syst., vol. 54, no. 1, pp. 1–9, Feb. 1993.

[8] S. M. Chen, “Forecasting enrollments based on fuzzy time series,” Fuzzy Sets Syst., vol. 81, no. 3, pp. 311–319, Aug. 1996.

[9] W. Qiu, X. Liu, and L. Wang, "Forecasting shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees," *Expert Systems with*

Applications, vol. 39, pp. 7680-7689, 2012.

[10] C. Li and J. W. Hu, "A new ARIMA-based neuro-fuzzy approach and swarm intelligence for time series forecasting," *Engineering Applications of Artificial Intelligence*, vol. 25, pp. 295-308, 2012.

[11] E. Bai, W. K.Wong, W. C. Chu, M. Xia, and F. Pan, "A heuristic time-invariant model for fuzzy time series forecasting," *Expert Systems with Applications*, vol. 38, pp. 2701-2707, 2011.

[12] Data mining, 2017 https://en.wikipedia.org/wiki/Data_mining

[13] Big data, 2017. http://en.wikipedia.org/wiki/Big_data.

[14] SAS, 2015 https://www.sas.com/zh_tw/home.html

[15] H. Esmaeil, G. Arash, S. Kamran, and A. N. Salman, "Tourist arrival forecasting by evolutionary fuzzy systems," *Tourism Management*, vol. 32, pp. 1196-1203, 2011.

[16] W. K. Wong, E. Bai, and A. W. Chu, "Adaptive Time-Variant Models for Fuzzy-Time-Series Forecasting,"*IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, pp. 1531-1542, 2010.

[17] C. H. Wang, "Predicting tourism demand using fuzzy time series and hybrid grey theory," Tourism Management, vol. 25, pp. 367-374, 2004.

[18] Chen, S. (1996). Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems, 81(3), pp.311-319.

[19] Doi.org. (2017). A new ARIMA-based neuro-fuzzy approach and swarm intelligence for time series forecasting. [online] Available at: http://doi.org/10.1016/j.engappai.2011.10.005 [Accessed 25 Apr. 2017].

[20] Doi.org. (2017). An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. [online] Available at: http://doi.org/10.1016/j.eswa.2008.07.043 [Accessed 25 Apr. 2017].

[21] Doi.org. (2017). Forecasting shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees. [online] Available at: http://doi.org/10.1016/j.eswa.2012.01.051 [Accessed 25 Apr. 2017].

[22] Doi.org. (2017). Predicting tourism demand using fuzzy time series and hybrid grey theory. [online] Available at: http://doi.org/10.1016/S0261-5177 (03)00132-8 [Accessed 25 Apr. 2017].

[23] Ieeexplore.ieee.org. (2017). A new approach of moving average method in time series analysis - IEEE Xplore Document. [online] Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6708545 [Accessed 25 Apr. 2017].

[24] Paulbourke.net. (2017). Auto-regression (AR). [online] Available at: http://paulbourke.net/miscellaneous/ar/ [Accessed 25 Apr. 2017].

[25] Sas.com. (2017). Analytics, Business Intelligence and Data Management. [online] Available at: https://www.sas.com/zh_tw/home.html [Accessed 25 Apr. 2017].

[26] Song, Q. and Chissom, B. (1993). Forecasting enrollments with fuzzy time series — Part I. Fuzzy Sets and Systems, 54(1), pp.1-9.

[27] Bolger, F. and D. Onkal-Atay (2004) "The Effects of Feedback on Judgmental Interval Predictions", International Journal of Forecasting, Vol.20, pp.29~39.