Doctoral Dissertation

Analyzing and Interpreting Air Quality Monitoring Data in Surabaya

ARIE DIPAREZA SYAFEI

Graduate School for International Development and Cooperation
Hiroshima University

September 2014

Analyzing and Interpreting Air Quality Monitoring Data in Surabaya

D115407

ARIE DIPAREZA SYAFEI

A Dissertation Submitted to
the Graduate School for International Development and Cooperation
of Hiroshima University in Partial Fulfillment
of the Requirement for the Degree of
Doctor of Engineering

September 2014

We hereby recommend that the dissertation by Mr. ARIE DIPAREZA SYAFEI entitled "Analyzing and Interpreting Air Quality Monitoring Data in Surabaya" be accepted in partial fulfillment of the requirements for the degree of DOCTOR OF ENGINEERING.

Committee on Final Examination:

Fujiwara Akimasa, Professor
Chairperson

Zhang Junyi, Professor

Kaneko Shinji, Professor

Kawamura Kensuke, Associate Professor

You Xueyi, Professor
School of Environmental Science and Engineering, Tianjin University, China

Date: July 23, 2014

Approved:

FUJIWARA Akimasa, Professor
Dean

広島大学大学院国際協力研究科長印

Date: September 5, 2014

Graduate School for International Development and Cooperation
Hiroshima University

（別紙様式６）
(Attached Form 6)

学位論文の要旨（論文の内容の要旨）
Summary of the Dissertation (Summary of Dissertation Contents)

論文題目
Dissertation title

Analyzing and Interpreting Air Quality Monitoring Data in Surabaya

広島大学大学院国際協力研究科
Graduate School for International Development and Cooperation,
Hiroshima University

博士課程後期 Doctoral Program	開発科学専攻 Division of Development Science	
学生番号 Student ID No.	D115407	
氏　　名 Name	ARIE DIPAREZA SYAFEI	㊞ Seal

Air pollution has become a major concern due to its significant impacts on human's health as well as human activities. Many researchers have notified the necessity of fundamental analysis to clarify the cause-effect mechanism of air quality and reflet it into countermeasures to academic researchers, environmental practitioners, stakeholders, as well as residents. In particular, policy makers at both central and local levels in developing countries are required to statistically analyze and interrete air quality monitoring data to take measures and to conduct preventive actions in order to reduce the impacts of the air pollution and ultimately control air quality. This study aimed to propose several methodologies to analyze and interprete air quality monitoring data in Surabaya City, Indonesia.

The dissertation consisted of 7 chapters with the following contents. The overview of current situation, problem statements, research objectives and scopes, and outline of the dissertation were presented in Chapter 1. Chapter 2 contained literature reviews related to air quality analysis and prediction models that have been developed in the air quality research fields. In this study, the air pollution concentrations of six indicators (NO, NO_2, O_3, SO_2, CO and PM_{10}) were continuously collected every 30 minutes for 20 months from February 1, 2001 to September 30, 2002 at five different monitoring stations that had their specific land use patterns in Surabaya. Chapter 3 described the profiles of the monitoring data, and besides, preliminary screening and complementing (missing value imputation) around 15% missing data were carried out.

The following chapters are divided into two parts, those are the intepretation part from Chapters 4 to 5 and the prediction part on Chapter 6. Firstly Chapter 4 employed multilevel model to characterize air pollution behaviors in the city allowing the identification of spatial and temporal variations over the measurements. The result showed that temporal variation highly affected air pollutant concentration. Furthermore, there was an interaction among pollutants and meteorological factors that contributed to the fluctuation of the concentrations. However, it was noted that the spatial variation was small, indicating the the fluctuation and dynamic of concentrations were insignificant because within station factors not due to between station effect.

Chapter 5 attempted to identify temporal patterns of the events in which pollutant concentrations exceeded the ambient threshold (permissible) levels. Since the observed disruption consisted of the complicated mixture of several different ones, Independent Component Analysis (ICA) was employed to decompose into unit distributions caused by emission sources such as traffic, industry and etc. The results could imply some possible policies specific to each monitoring station. By following the above outputs, this study conducted time series analysis with state-space model to find dynamically influencial factors on air pollutant concentration distributions both using weekly time scale and annual longer time scale in particular cases of NO_2 and PM_{10}. The results provided not only the information of the dynamic factors but also temporal pattern.

The short-term prediction of air quality may take another important role to persuade residents and policy makers to improve their preparedness and change their behavior timely. Therefore, Chapter 6 developed time series models: autoregression model with Bayesian Markov Switching (BMS) model and modified Bayesian Markov Switching (MBMS) model. It was found that MBMS model slightly improved the prediction power with the condition of training data set. Since this kind of model was site-specific, the performance of the model depended of data set on that particular location e.g., the fluctuation. If the training data set was less fluctuate, the Bayesian model performed better than autoregression model. In order to improve the prediction, we used AR and ARIMA models with optimum order (lag-concentrations). We also used longer time interval 4-hour instead of 1-hour. It was observed that the performance of AR and ARIMA with optimum lag information performed better than AR(1) and BMS models. Further, models using 4-hours data were better than models with 1-hour data. To further improve prediction, we tried to predict the true distribution of pollutant by extracting independent components. However, even though the results showed there was a room for improvement to predict distribution of pollutant especially using longer time interval data, the performance was not good. In order to improve the prediction by time series models, this study explored another prediction model that was Generalized Least Square (GLS) model by using original meteorological variables and variables extracted from component analysis. It turned out that the GLS models with original variables would be useful for the short-term prediction of air quality.

Finally, Chapter 7 explored and summarized some remarkable findings from this study. The author promoted some approaches to find possible policies that can be implemented in particular areas. This chapter also discussed some policies based on the assumption of feasibility of implementation. Finally the dissertation made a conclusion of the research and drawed some limitations and future directions.

備考　論文の要旨はＡ４判用紙を使用し，4,000字以内とする。ただし，英文の場合は1,500語以内とする。

Remark: The summary of the dissertation should be written on A4-size pages and should not exceed 4,000 Japanese characters. When written in English, it should not exceed 1,500 words.

Acknowledgement

First of all, I praise to the Lord, Allah for His Bless, allowing me to complete the study here, in Hiroshima University. Sholawat and salam to Prophet Muhammad SAW with his guide had brought me into a better person over time.

I would like to convey my biggest and sincere gratitude to Prof. Akimasa Fujiwara for him being my supervisor. I have extracted some philosophies of study and I learn from him best attitude and support from which he always provided to me in the middle of his busy time.

I thank to Prof. Zhang Junyi as 2nd academic supervisor, and also as a great critical reviewer of my papers before they got submitted to journals. I learnt a lot from him on logical thinking before writing any academic paper or even starting research.

I would also like to thank Associate Prof. Kawamura Kensuke for his supportive guide all this time. I still remember how he sent me spatio-temporal data from MODIS. I am still courious how I can reproduce what he had done to get such data. I look forward to having mutual collaboration in the future.

Prof. Shinji Kaneko has been my 4th academic reviewer, and I convey thank to him because of his time. I had communication with Prof. Kaneko not only during dissertation presentation but also in other chance such as GEL's jury. He indirectly taught me to concentration on fundamental first before even begin a research.

I would also like to thank to Prof. You Xueyi from Tianjin University for many useful inputs and insights for my future research direction.

I would not forget to thank all these valuable names who have tremendously been with me since the starting of my study. Their inputs have been imprinted in my mind, awaiting to be implemented to my next journey. They are: Assoc. Prof. Makoto Tsukai, Assoc. Prof. Chikaraishi, Assist. Prof. Sudarmanto Budi Nugroho, and Assist. Prof. Seya.

My labmates, they have been staying together with me. I thank you all, I wish you all the best for your future endeavour. My Indonesian friends, thanks for all warm activities in Saijo. Lastly, moslem community in Higashihiroshima. I had wonderful time and had found enlightened new spirit while I am staying here.

It is my privilege to have these financial supports during my study. Without them, I will be nowhere to step up conducting this research. I thank to General Directorate of Higher Education, Ministry of Education, Republic of Indonesia for providing full financial support during research in Hiroshima University. Further, I have published several papers and attended several conferences, these are impossible without the support from the Global Environmental Leadership (GELs) Program, Hiroshima University. Therefore I would like to express my grateful and sincere gratitude for the support.

I also wish to thank my institution Department of Environmental Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya which has been very helpful to aid any administrative process for the smooth of my study.

For the deepest of my heart, allow me to also say thank to Prof. Joni Hermana, from Department of Environmental Engineering, ITS, Surabaya. Thank you for introducing me to Prof. Akimasa Fujiwara in the Year of 2009. I still remember that day, Nopember 2009. I also would like to thank to all colleagues in my department, in particular Ir. Mas Agus Mardyanto, ME., PhD and Ir. Eddy S. Soedjono, PhD as prior and current Head of Deparment of Environmental Engineering, ITS. Prof. Lin Cheng Fang is my master's degree supervisor. Without his recommendation I would not be able to study here. I convey my special thank to him.

My family, this is for you all. My wife Nassa Arfiantinosa, my daughter Safira, my son Didi. My father Drg. Achmad Sjafei, my mother Drg. Ina Herini, my little sister Anvi and little brother Aditya. My Father and mother in-law H. Arfian Hadi and Hj. Elly Soemartini. Thank you for the support. I am ready to do more in Surabaya.

Table of Contents

Cover i

Title ii

Approval of Dissertation iii

Summary Sheet of Dissertation iv

Acknowledgement vi

Table of Contents vii

List of Tables ix

List of Figures xii

List of Abbreviations xvi

Chapter 1 Introduction 1
1.1 Background 1
1.2 Research Objectives 2
1.3 Framework, Originality and Research Scope 3
1.4 Outline of Dissertation 4

Chapter 2 Literature Reviews 7
2.1 Air Quality towards Humans Health 7
2.2 Air Quality Studies 8
2.3 Investigating the Phenomena that Affects Air Quality 9
2.4 Temporal Pattern of Pollutants 9
2.5 Prediction of air quality 10

Chapter 3 Methodology, Data, and Locations 13
3.1 Introduction 13
3.2 Monitoring Stations 13
3.3 Threshold (permissible) level of pollutant 15
3.4 Missing Data 15
3.5 Data Aggregation 20

Chapter 4 Spatial and Temporal Factors of Air Quality in Surabaya City: an Analysis based on a Multilevel Model 22
4.1 Background 22
4.2 Data 22
4.3 Multilevel model 23
4.4 Results and discussions 23
4.4.1 Diurnal variation 23
4.4.2 Multilevel model results 26
4.5 Conclusion 30

Chapter 5 Temporal Peak Time Frame and Dynamic Factors of Traffic Related Air Pollutants 31
5.1 Introduction 31
5.2 Data and monitoring stations 31
5.3 Methodology 34
5.3.1 ICA 34
5.3.2 State-space Model 35
5.4 Pollutants Distribution using ICA and Emission Source Estimation 36
5.4.1 Hypothesis of temporal patterns for five locations 36

5.4.2 Pollutions Exceeding Permissible Level (Ambient Criteria) 37
5.4.3 Diurnal variation of pollutant concentrations 45
5.4.4 Temporal Distribution based on Independent Component Analysis 48
5.4.5 Estimation of Emission Sources 73
5.4.6 Seasonal Temporal Distribution using 4 week-cycle and Annual Time Scale 78
5.5 Dynamic of Pollutants and Meteorological Factors using State-Space model 96
5.5.1 The Dynamic of Pollutants 96
5.5.2 Seasonal Change dynamically of air quality 111
5.5.3 Pollutants Concentrations Subpatterns 125
5.6 Conclusions 128

Chapter 6 Prediction of Air Quality Concentrations Based on Time Series Models and Generalized Least Square Models 130
6.1 Introduction 130
6.2 Materials and Data 130
6.2.1 Data for Time series models 130
6.2.2 Data for GLS models 132
6.3 Time series models 133
6.3.1 A first-order autoregressive (AR) model 135
6.3.2 Autoregressive Integrated Moving Average (ARIMA) model 136
6.3.3 A Bayesian markov switching model 136
6.3.4 A modified Bayesian Markov Switching Model 138
6.4 GLS Models 138
6.4.1 Generalized Least Square 138
6.4.2 Independent Component Analysis 139
6.4.3 Principal Component Analysis 140
6.4.4 Evaluation indicators 140
6.5 Results and Discussion for time series models 140
6.5.1 Internal validity: Model accuracy 140
6.5.2 External validity: Forecasting ability 147
6.6 Time Series Model with Optimum Order of Parameters 157
6.7 Results and Discussions of Time Series Models for Air Quality Data using 4 hours interval data 160
6.8 Result and Discussions for Updated Algorithms Time Series Models 163
6.9 Results and Discussion of Time Series Models to Predict True Distribution of Air Quality 164
6.9.1 Prediction of ICs using 1-hour NO_2 data 164
6.9.2 Prediction of ICs using 4-hour NO_2 data 167
6.10 Results and Discussions for GLS models 169
6.11 Conclusions 173

Chapter 7 Conclusions, Policy Suggestions, Future Direction, and Limitations 175
7.1 Summary Findings 175
7.2 Policy Suggestions 177
7.2.1 Spatial Policy for each zone 177
7.2.2 Implementing the Prediction of Air Quality 180
7.3 Limitation of Study 180
7.4 Future Directions 181

References 182

Publications 188

List of Tables

Table 1.1 Research questions and objectives of analyzing and interpreting chapters in this study 2

Table 3.1 Coordinates and elevation of each monitoring station 15

Table 3.2 Permissible (threshold) levels of five air pollutants captured by monitoring stations 15

Table 3.3 Proportion and missing values rate of parameters 16

Table 3.4 Scenario of missing values imputation 17

Table 3.5 Multiple and adjusted R^2 of linier regression comparing different dataset of imputed values based on two scenarios: partial and full dataset 20

Table 3.6 Data aggregation and pollutants used in each chapter of the study along with goals from each chapter 20

Table 4.1 Missing rates (percentage) of data availability before and after imputation 23

Table 4.2 Estimation result of multilevel model using 30-mins data 27

Table 4.3 Estimation result of multilevel model using monthly interval data 29

Table 5.1 Monitoring stations and their coordinates 31

Table 5.2 Weight loadings for independent components of NO_2 concentration 50

Table 5.3 NO_2 temporal peak time frame extracted from Independent Components 51

Table 5.4 Weight loadings for independent components of PM_{10} concentration 53

Table 5.5 PM_{10} temporal peak time frame extracted from Independent Components 55

Table 5.6 Weight loadings for independent components of O_3 concentration 55

Table 5.7 O_3 temporal peak time frame extracted from Independent Components 58

Table 5.8 Weight loadings for independent components of NO concentration 58

Table 5.9 NO temporal peak time frame extracted from Independent Components 61

Table 5.10 Weight loadings for independent components of CO concentration 63

Table 5.11 CO temporal peak time frame extracted from Independent Components 64

Table 5.12 Weight loadings for independent components of SO_2 concentration 65

Table 5.13 SO_2 temporal peak time frame extracted from Independent Components 67

Table 5.14 Temporal peak time frame from all pollutants drawn from Independent Component Analysis of each pollutant 68

Table 5.15 Seasonal pattern on weekly cycle for all pollutant profiles 70

Table 5.16 Weight loadings for independent components of 4-week cycle NO_2 concentration 80

Table 5.17 Weight loadings for independent components of 4-week cycle PM_{10} concentration 82

Table 5.18 Weight loadings for independent components of 4-week cycle O_3 concentration 84

Table 5.19 Weight loadings for independent components of 4-week cycle NO concentration 86

Table 5.20 Weight loadings for independent components of 4-week cycle CO concentration 88

Table 5.21 Weight loadings for independent components of 4-week cycle SO_2 concentration 90

Table 5.22 Weight loadings for independent components of daily average NO_2 concentration...... 93

Table 5.23 Weight loadings for independent components of daily average PM_{10} concentration 95

Table 5.24 Model performance based on Akaike's Information Criterion (AIC) *b*-value from the five hypotheses for the NO_2 concentration subpattern 126

Table 5.25 Model performance based on Akaike's Information Criterion (AIC) *b*-value from the five hypotheses for the PM_{10} concentration subpattern 126

Table 5.26 Process variation and correlation among zones 127

Table 5.27 Process variation and correlation among zones 127

Table 6.1 Data for parameter estimation, evaluation and forecasting for each station 132

Table 6.2 Descriptive statistics of levels of pollutants in Surabaya from five monitoring stations (ug/m^3) 133

Table 6.3 Statistical properties of hourly NO, NO_2, PM_{10}, and CO concentrations........ 135

Table 6.4 Estimation results of AR(1) models for all air quality parameters: NO, NO_2, PM_{10}, and CO 141

Table 6.5 Estimation results of BMS models for all parameters NO, NO_2, PM_{10}, and CO 141

Table 6.6 Min and Max Errors of BMS model for city center, trading zone, Suburban1, near highway, and Suburban2 stations for all quality parameters........ 142

Table 6.7 Model performances of internal validation based ME, MAE, and RMSE values for all site locations for all quality parameters........ 143

Table 6.8 Model performances of external validation based ME, MAE, and RMSE values for all site locations 148

Table 6.9 Internal validation of time series models with optimum order parameters on all sites........ 158

Table 6.10 External validation of time series models with optimum order parameters on all sites........ 158

Table 6.11 Internal validation of ARIMA and AR using 4 hours data for NO_2 and PM_{10}........ 160

Table 6.12 External validation of ARIMA and AR using 4 hours data for NO_2 and PM_{10}........ 161

Table 6.13 Internal validation of updated ARIMA and AR in the city center for NO_2 164

Table 6.14 External validation of updated ARIMA and AR in the city center for NO_2 164

Table 6.15 Weght loadings of ICA for NO_2 using 1-hour data........ 164

Table 6.16 Internal validation of ARIMA and AR for each ICs on 1-hour NO_2 data........ 165

Table 6.17 External validation of ARIMA and AR for each ICs on 1-hour NO_2 data........ 165

Table 6.18 Weght loadings of ICA for NO_2 using 4-hour data........ 167

Table 6.19 Internal validation of ARIMA and AR for each ICs on 4-hour NO_2 data........ 167

Table 6.20 External validation of ARIMA and AR for each ICs on 4-hour NO_2 data........ 167

Table 6.21 Pearson correlation of air pollutants and meteorological factors over five zones of monitoring stations........ 169

Table 6.22 Internal validation of each model for NO_2........ 170

Table 6.23 Performance of forecasted values in Station 1 and 5 for each model for NO_2........ 170

Table 6.24 Parameter estimates of NO_2, PM_{10}, and O_3 with lag-1 independent variables 172

Table 7.1 Possible Direct Policies based on extracted components 179

List of Figures

Figure 1.1 Research framework of the study 4

Figure 1.2 Research Scope and Framework in the present study... 5

Figure 1.3 Dissertation chapter structures... 6

Figure 3.1 Surabaya map with location of five monitoring stations 13

Figure 3.2 Five monitoring stations at five locations.. 14

Figure 3.3 Visual comparison of NO imputation in Station 2 using partial and full data set 18

Figure 3.4 Visual comparison of NO imputation in Station 5 using partial and full data set 19

Figure 3.5 Validation NO imputation based on confidence interval and distribution of imputations and observed values.. 19

Figure 3.6 Data aggregation for each pollutant.. 21

Figure 4.1 Weekly variation of average pollutants 25

Figure 4.2 Traffic distribution with interpolation using IDW in Surabaya City of: (a) motorbike, (b) private car; y ordinal indicates vehicle volumes 26

Figure 5.1 Road map of the surrounding of monitoring stations. Background image is made available from OpenStreetMap (© OpenStreetMap contributors) available under the Open Database License. Red dot is the location of the sensor of monitoring station............................. 32

Figure 5.2 Five monitoring stations in Surabaya City. The map in the trading zone is a modified image from an original image that was made available from OpenStreetMap (© OpenStreetMap contributors) available under the Open Database License. The blue dot is the location of the sensor of the monitoring station in the trading zone (Perak). 33

Figure 5.3 The mixtures of signals or raw data (a) and four independent components (b) after performing ICA, this example is taken from Malaroiu *et al.* (2000) 34

Figure 5.4 30-mins interval NO_2 concentrations from 2001-2002 at five monitoring stations........ 38

Figure 5.5 30-min interval PM_{10} concentrations from 2001-2002 at five monitoring stations 39

Figure 5.6 30-min interval of O_3 concentrations from 2001-2002 on five monitoring stations....... 40

Figure 5.7 30-min intervals of CO concentrations from 2001-2002 on five monitoring stations.. 41

Figure 5.8 30-min intervals of SO_2 concentrations from 2001-2002 on five monitoring stations.. 42

Figure 5.9 Frequency of pollutants exceeding threshold value over two years of air quality on five monitoring stations.. 43

Figure 5.10 Frequency between days of pollutants exceeding threshold value over two years of data on five monitoring stations... 44

Figure 5.11 Diurnal concentration over a week from 5th February to 11th February 2001 of NO_2.. 46

Figure 5.12 Diurnal concentration over a week from 5th February to 11th February 2001 of PM_{10} 47

Figure 5.13 Observed diurnal variations of pollutant’s concentration over five monitoring stations.. 48

Figure 5.14 Five independent component profiles from NO_2 concentration over five monitoring stations; x-axis: time, y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 50

Figure 5.15 Five independent component profiles from PM_{10} concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 53

Figure 5.16 Five independent component profiles from O_3 concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 57

Figure 5.17 Five independent component profiles from NO concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 60

Figure 5.18 Five independent component profiles from CO concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 63

Figure 5.19 Five independent component profiles from SO_2 concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday. ... 66

Figure 5.20 Historical wind direction and wind speed of captured by five monitoring stations... 75

Figure 5.21 Historical wind direction and wind speed captured by five monitoring stations when the NO_2 concentration exceeded threshold value ... 76

Figure 5.22 Historical wind direction and wind speed captured by five monitoring stations when the PM_{10} concentration exceeded threshold value. ... 77

Figure 5.23 Five independent component profiles of NO_2 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments which means 4-week cycle, y-axis: component value. ... 80

Figure 5.24 Five independent component profiles of PM_{10} concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value. ... 82

Figure 5.25 Five independent component profiles of O_3 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value. ... 84

Figure 5.26 Five independent component profiles of NO concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value. ... 86

Figure 5.27 Five independent component profiles of CO concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value. ... 88

Figure 5.28 Five independent component profiles of SO_2 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value. ... 90

Figure 5.29 Five independent component profiles of daily average NO_2 concentration over five monitoring stations; x-axis: time (days), y-axis: component value. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry ... 93

Figure 5.30 Five independent component profiles of daily average PM_{10} concentration over five monitoring stations; x-axis: time (days), y-axis: component value. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 95

Figure 5.31 Time series of estimated mean states of independent variables (thick lines). The black dot is the observation of square-root NO_2 concentration in ug/m^3. The first segment is Thursday. 98

Figure 5.32 PM_{10} Time series of estimated mean states of independent variables (thick lines). The black dot is the observation of square-root PM_{10} concentration in ug/m^3. The first segment is Thursday. 100

Figure 5.33 O_3 Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root O_3 concentration in ug/m^3. The first segment is Thursday. 103

Figure 5.34 NO Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root NO concentration in ug/m^3. The first segment is Thursday. 105

Figure 5.35 CO Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root CO concentration in ug/m^3. The first segment is Thursday. 108

Figure 5.36 SO_2 Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root SO_2 concentration in ug/m^3. The first segment is Thursday. 110

Figure 5.37 Estimated meteorological parameter coefficients for NO_2.The black dot is an observation of square-root NO_2 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 113

Figure 5.38 Estimated meteorological parameter coefficients for PM_{10}. The black dot is an observation of square-root PM_{10} concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 115

Figure 5.39 Estimated meteorological parameter coefficients for O_3. The black dot is an observation of square-root O_3 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 118

Figure 5.40 Estimated meteorological parameter coefficients for NO. The black dot is an observation of square-root NO concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 120

Figure 5.41 Estimated meteorological parameter coefficients for CO. The black dot is an observation of square-root CO concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 122

Figure 5.42 Estimated meteorological parameter coefficients for SO_2. The black dot is an observation of square-root SO_2 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry 124

Figure 5.43 Maximum likelihood (ML) estimates. Note that the estimates in (a) are equal with actual NO_2 measurement, whereas in (b) the continuous line represents the ML estimates with other lines as the actual NO_2 measurement. The time interval starts from Thursday (0-48) in the x-axis 128

Figure 6.1 Daily diurnal pollutants concentrations (square-root) on four stations 131

Figure 6.2 Hourly mean NO concentrations on four sites 134

Figure 6.3 Data for parameter estimation and evaluation in near highway station (as example) 135

Figure 6.4 Observed values, fitted values and residuals from the BMS model at City center (city center) station 145

Figure 6.5 Observed values, fitted values and residuals from the BMS model at City center station (city center) for increased sample size 147

Figure 6.6 Forecasting by AR(1), BMS, and MBMS models for NO on all sites 151

Figure 6.7 Forecasting by AR(1), BMS, and MBMS models for NO_2 on all sites 153

Figure 6.8 Forecasting by AR(1), BMS, and MBMS models for PM_{10} on all sites 155

Figure 6.9 Forecasting by AR(1), BMS, and MBMS models for CO on all sites 157

Figure 6.10 Forecasting by ARIMA and AR for NO_2 concentration on four stations 160

Figure 6.11 Forecasting by ARIMA and AR for 4-hour NO_2 concentration on four stations 163

Figure 6.12 Forecasting by updating algorithm of ARIMA and AR compared with ordinary ARIMA(2,1,1) and AR(23) for NO_2 in the city center 163

Figure 6.13 Forecasting ICs by ARIMA and AR using 1-hour NO_2 data 166

Figure 6.14 Forecasting ICs by ARIMA and AR using 4-hour NO_2 data 168

List of Abbreviations

AIC	Akaike's Information Criterion
ARIMA	Autoregression Integrated Moving Average
ANN	Artificial neural network
AQMN	Air quality monitoring network
CO	Carbon monoxide
IC	Independent component
ICA	Independent component analysis
ICi	Independent component i
ICR	Independent component regression
MAE	Mean absolute error
ME	Mean Error
MODIS	Moderate Resolution Imaging Spectroradiometer
RMSE	Root Mean Square Error
N	North
NE	Northeast
NW	Northwest
NO	Nitrogen Monoxide
NO_2	Nitrogen Dioxide
O_3	Ozone
PC	Principal component
PCA	Principal component analysis
PCi	i^{th} principal component
PCR	Principal component regression
PM_{10}	Particulate Matter with size less than 10 um
S	South
SARIMA	Seasonal Autoregression Integrated Moving Average
SE	Southeast
SO_2	Sulphur dioxide
SR	Solar radiation
SSE	Sum of the squared errors
SW	Southwest
WS	Wind speed

Chapter 1 Introduction

1.1 Background

Air quality management is a crucial issue at the national and local levels to secure residential health and living environment. Many studies have been conducted over the past decades. The field of air quality study can be divided into several specific topics: i) technology development dealing with air pollutant, e.g. Painting for photocatalytic degradation of NO_x gases from vehicles, ii) measurement and observation methods of air quality levels, iii) finding out pollutant sources and factors, and relationship among them, iv) figuring out cause-effect mechanism of air pollution, v) prediction of temporal changes in and spatial variation of air concentrations. However, all the issues still remain more challenging in developing countries, because of the lack of institutional, financial and technological limitations. In particular, the number of monitoring stations is limited, and long-term continuous monitoring is not working in many developing cities. Thus, the immature methodology of data collection, analysis and interpretation result in missing the timely and spatial countermeasures to mitigate and adapt the worsening air pollution.

Environmental Agency in Surabaya City (i.e. a target city of this study) that is the second largest city with a population of three million in Indonesia has launched an operation of five monitoring stations since 2001. The installation was highly expected to promote a cleaner air quality. Air Pollutant Index (API) comprising 5 statuses of “Good”, “Moderate”, “Unhealthy”, “Very Unhealthy”, and “Dangerous” was calculated to alert the public shortly after detections. In fact, there were 18 warnings of “Unhealthy” statuses during 2001 and 2002. However, the warning information can provide only cross-sectional (real time) states of air pollutions but its temporal changes, despite the time series observation. Also the observed distribution of air concentration is a mixture of different distributions caused by different sources. It is complicated to isolate a specific policy from mixture distribution. Hence, it is a serious and unavoidable problem that policy makers and residents cannot cope with appropriate countermeasures and prepare the expected levels of air pollution in future.

High rate of missing data is another detrimental problem in Surabaya. It makes us more difficult to ascertain causal factors and to predict the air quality at any certain time and location. Furthermore, policy makers were actually lacking in the point source information of air pollutions such as NO_x, CO_2, O_3 and PM_{10}. Actually, these incomplete monitoring data are not fully utilized by policy makers as well as researchers. There are two reasons for this fact. Firstly, the sensors of air quality often failed since 2003 two years after the operation. The extremely high missing rates made a fatal error to monitor the air quality. Secondly, because of this missing information, it was difficult for one to obtain such accurate inference about the situation of Surabaya.

Regarding the prediction of air pollution, there is a concern of displaying the future pollutant concentration. The purpose of prediction is to make residents and policy makers act promptly by displaying the expected air pollutions. For example, if residents are notified about the NO_2 concentrations at a particular location one or some hours later, users will be able to either choose another route or wear masks. In this case, a prediction model of air concentrations is required with higher accuracy. As far as author concerns, no studies have fully exploited the temporal patterns of air pollution to estimate the risk of breaching pollutant levels.

Based on the background behind the data unavailability, it is urgently required to develop an approach to interpret and predict the future air concentrations to derive inferences to make policies based on a very poor data for a limited period.

1.2 Research Objectives

This study aims at comprehensively finding factors that affect air pollutant concentration, and also estimating models that can predict the future's concentrations under the constraint of limited available data. More specifically, this study

i) examines significant temporal-spatial factors affecting air quality,
ii) decomposes the observed temporal distribution of air concentrations into some independent source distributions, and furthermore
iii) predicts the concentration of pollutants in the future using the limited data.

There are few literatures arguing time-space analysis under the above mentioned constraint of data particularly in developing cities. Table 1.1 summarizes the research questions and objectives in chapters 4 to 6.

Table 1.1 Research questions and objectives of analyzing and interpreting chapters in this study

Ch	Research questions	Objectives
4	1. General objectives: To obtain significant factors affect air pollutant levels both spatially and temporally	
	2. Which day significantly affects air pollutant (day-by-day)? Is the difference significant?	To overview the concentration difference among days
	3. What is spatial effect on the concentration?	To obtain the influence of each zone on the pollutants concentrations
		To obtain whether spatial location affect sthe fluctuation (dynamic) of the concentrations
	4. Does concentration pattern differ between seasons?	To obtain the effect of seasons on the concentrations
	Does the concentration increase significantly during the specified peak time in both morning and evening?	To obtain how air concentrations change during morning and evening peak hours
	5. Will the addition of green space reduce the pollution levels? How land use patterns affect the pollutant levels?	To obtain factors of green space in affecting the pollutant concentrations and how other land use pattern affect pollutant levels
5	1. What is temporal peak time in each zone?	To obtain temporal peak time frame where the pollutant levels reach the peak
	2. What is seasonal variation in each zone based on true distribution of pollutants?	To obtain seasonal variation effect of pollutants
	3. Can we confirm temporal peak time frame obtained from previous chapter?	To confirm temporal peak time frame
	4. What are the effects of meteorological factors toward the pollutant levels?	To obtain dynamic change of meteorological parameters to pollutant levels
6	Will time series model can predict more accurately than GLS model with exogenous variables	To obtain best prediction model between time series approach and GLS models

1.3 Framework, Originality and Research Scope

The framework of this dissertation can be shown in Figure 1.1. There are two parts of this study: one part is interpretation and another is prediction. The former interpretation part aims to understand factors affecting air quality to derive some potential policies to improve air quality. Better air quality is an ultimate goal to secure human health and quality of life. The latter part is a prediction part. Several types of models will be employed to predict the air concentrations.

While there are many efforts to analyze and interpret the air monitoring data, this study has the following originality and uniqueness of methodology:

1) Independent Component Analysis (ICA) can decompose a complicated mixture distribution of air concentration to some independent source distributions,
2) State-space model can extract the temporal pattern of pollutants as well as detect the influential meteorological factors
3) Combined time series models and ICA can predict the dynamic fluctuations of air concentrations in different land use locations.

In air quality modeling, the spatial and temporal context becomes important because of its nature that varies across different locations and time. By combining these two factors together, it forms a so-called spatio-temporal model. The problem lies in its own characteristics and criterion that must be met. For instance, in the time series model, one should be careful dealing with missing data, because not all models can handle missing data. Therefore, before running into the model, a researcher should have dealt with missing data. Another concern about time series data is the linearity. Often, time series data is approach using non-linear model. A non-linear model is a very flexible model however it sacrifices the interpreting mode from a data. In terms of spatial, we cannot avoid that if we use this model, we need more data, spatially. In other words, a researcher should have many points spread out geographically. The minimum number of locations varies among researchers, but in general, it seems that minimum 30 points is plausible. Less than that, the concern about its validation will emerge.

Due to the limitation of monitoring stations inherent in developing city Surabaya, this study can use only five stations, which are not enough to build a spatio-temporal model. Even though this study cannot fully employ a spatial analysis, but it can cover spatial factors into several chapters. In chapter 4, the study uses multilevel to find the effect of spatial location to the air pollutants concentrations. In chapter 6 we develop time series models to predict air quality and the distribution of air quality. The use of time series models because of its simplicity to forecast future values. Furthermore, we would like to predict the distribution pattern obtained from independent components with the aim to predict future distribution patterns. Furthermore, this section also tries to develop a prediction model with component analysis employing some spatial variables such as wind direction at each station.

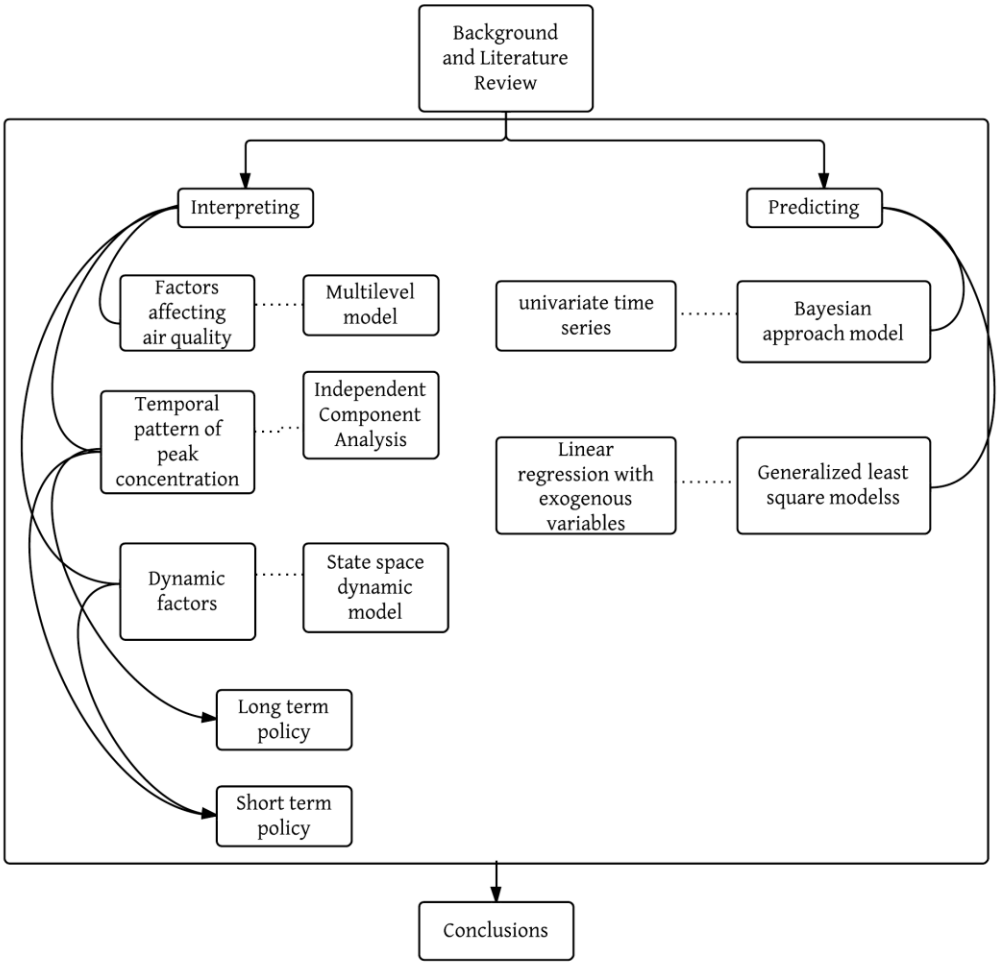

Background and Literature Review
Interpreting
Predicting
Factors affecting air quality
Multilevel model
Temporal pattern of peak concentration
Independent Component Analysis
Dynamic factors
State space dynamic model
Long term policy
Short term policy
univariate time series
Bayesian approach model
Linear regression with exogenous variables
Generalized least square modelss
Conclusions

Figure 1.1 Research framework of the study

1.4 Outline of Dissertation

There are three main types of air pollution studies: 1) generation and control of pollutants at the source, 2) transport, dispersion, and removal of gas species in the atmosphere, 3) impacts of pollutants to human beings, and 4) others likely air adsorption mechanism as a part of ecosystems. This study stands on the preliminary analysis related to the first and third scopes. The general goal of this study is thus to obtain insight how government should react to the possible event where concentrations exceeding the permissible level which will affect humans health.

Figure 1.2 explains the position of this study. In all statistical learning, the author can observe that based on its goal, statistical model comprises of interpreting and prediction part.

The first aim of this study is to characterize air pollution behaviors in the city allowing the identification of spatial and temporal variability towards the air quality measurements using multilevel model in Chapter 4. There are five monitoring stations, each represents their respective land use patterns; the multilevel model was used because its benefit of allowing estimates to vary across groups, in this case the monitoring stations. The result shows that temporal variation highly affects air pollutant concentration. Furthermore, there is interaction among pollutants and meteorological factors that contribute to the dynamic of the concentrations. However, it is noted that the spatial variation is small, indicating the fluctuation and dynamic of concentrations are because within station factors not due to between station effect.

Our second aim in the Chapter 5 is to identify temporal pattern of peak time when the concentrations exceed the permissible level. The background is that, this study finds many events when concentration of pollutants (NO_2 and PM_{10}) exceeds the permissible level (ambient threshold level). Therefore, it is mandatory to obtain 1) the main contributor of these events, and 2) when do these events likely happen. To answer those questions, this study takes these steps. It must be noted that the term "events" here means the condition where the pollutant levels exceed permissible level.

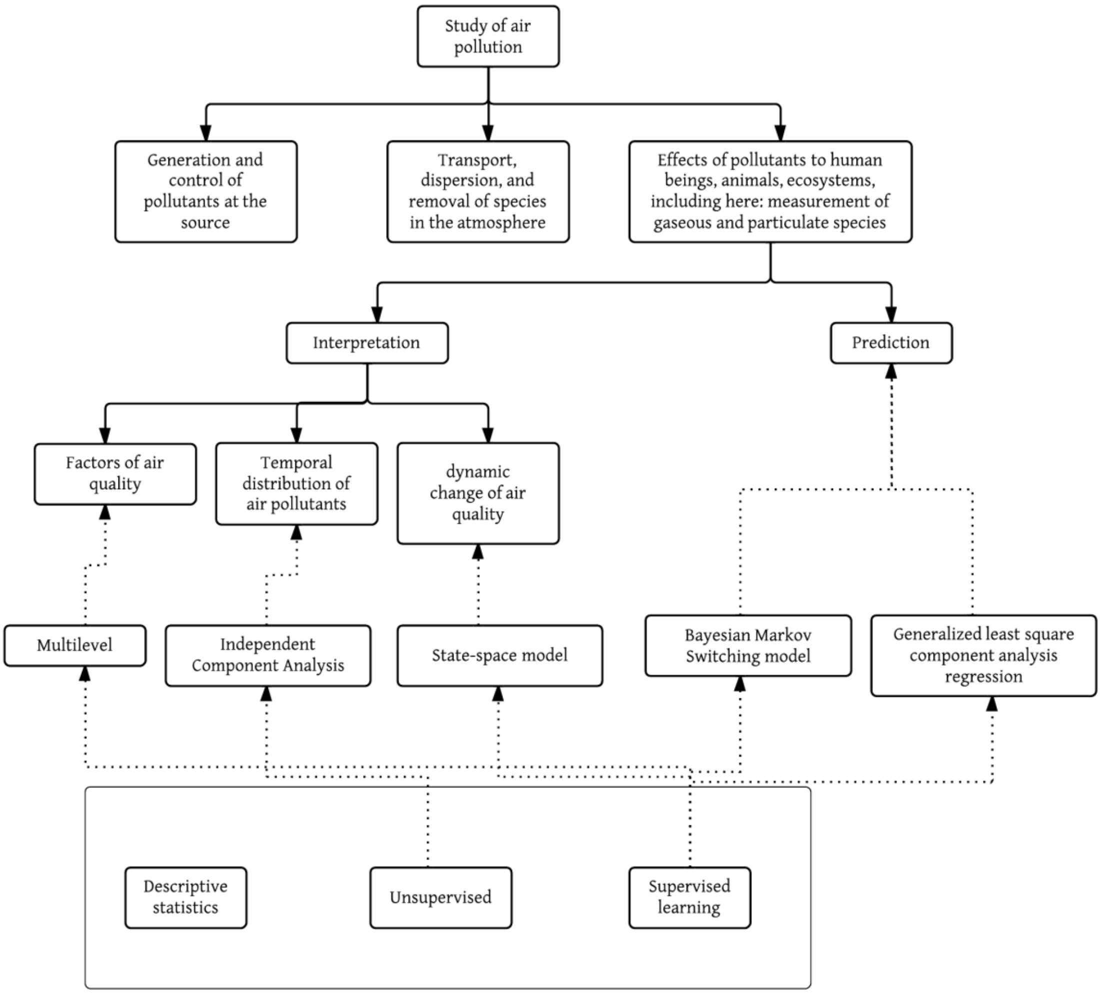

Study of air pollution
Generation and control of pollutants at the source
Transport, dispersion, and removal of species in the atmosphere
Effects of pollutants to human beings, animals, ecosystems, including here: measurement of gaseous and particulate species
Interpretation
Prediction
Factors of air quality
Temporal distribution of air pollutants
dynamic change of air quality
Multilevel
Independent Component Analysis
State-space model
Bayesian Markov Switching model
Generalized least square component analysis regression
Descriptive statistics
Unsupervised
Supervised learning

Figure 1.2 Research Scope and Framework in the present study

First, using assumption that these events are likely to happen during peak time, the author observes time frame using raw data in terms of the frequencies of events for each time interval. Conventional ordinary approach cannot obtain enough information about time frame for pollutants when the events are few. For this matter, the author employs a method called Independent Component Analysis (ICA) which is able to build components that are statistically independent. In these components, temporal peak time frame can be determined. This study combines the result of ICs (Independent Components) with wind direction and wind speed to estimate the emission sources by observing the maps. After these observations, the author discusses some policies that are specific to each site of monitoring stations which are displayed in the final chapter. To further analyze the effect of meteorological factors towards the pollutant's concentration, the author employs a time series state-space approach to obtain dynamic meteorological factors in a case of trading zone in Chapter 5. This chapter also confirms the sub-pattern of air pollutants. There are five monitoring locations each represents its own zone. The author confirms and finds out the underlying states or sub-pattern for the pollutants in the trading zone.

For air quality management, a prediction model is necessary and crucial. For this part, the author employs two different models each with different approach in Chapter 6. One is time series models and another is Generalized Least Square models. This study compares several steps of time series models.

The first one is the of lag-1 Auto Regression (AR) model with Bayesian approach model to predict several pollutants using mean hourly data. Second step is we try to improve the models by using time series models with optimum order of parameters. Third step we would like to obtain the effectiveness of using longer time interval data in this case 4-hour compared with if we use 1-hour data. We also attempt to predict longer time frame (1 week) using updating parameter algorithms.

As one of the originality of this research, we attempt to predict the true distribution of air quality by predicting independent components (ICs) that may represent certain pattern on different locations. We attempt to predict these ICs using 1-hour and 4-hour data with time series models.

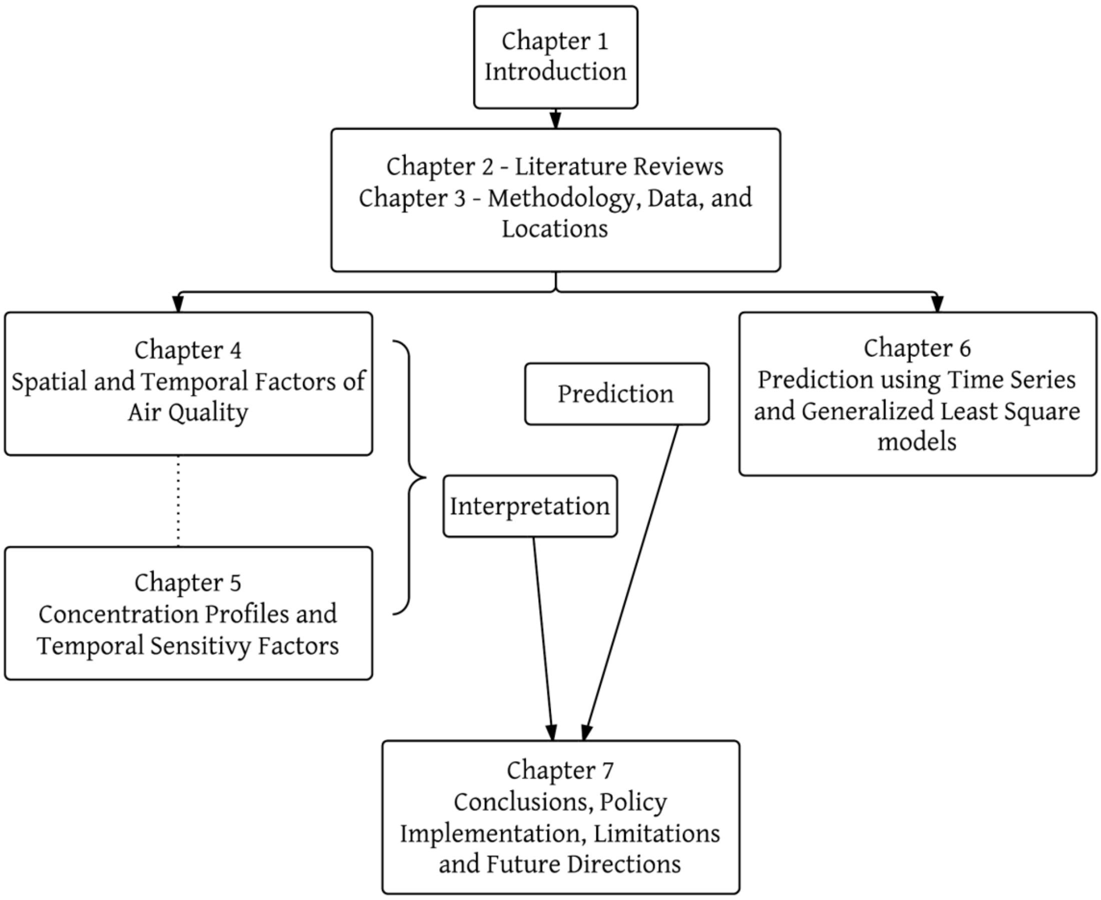

Chapter 1
Introduction
Chapter 2 - Literature Reviews
Chapter 3 - Methodology, Data, and Locations
Chapter 4
Spatial and Temporal Factors of Air Quality
Chapter 5
Concentration Profiles and Temporal Sensitivy Factors
Prediction
Interpretation
Chapter 6
Prediction using Time Series and Generalized Least Square models
Chapter 7
Conclusions, Policy Implementation, Limitations and Future Directions

Figure 1.3 Dissertation chapter structures

In the prediction model with time series data, the author must take into account residual serial autocorrelation. Unfortunately in some studies, because of the complexity of the model they used, the correlation was ignored, therefore leading to a possible bias result. To take this into account, this study uses a Generalized Least model by taking into account residual serial autocorrelation. In the regression model, multi-collinearity is a non-negligible problem that a researcher should avoid because of the correlation of independent variables may distort the standard error and lead to a bias conclusion and estimation. Because of this, this study replaces the original variables that consist of meteorological variables and air pollutants with components extracted from two different approaches: Independent Component Analysis and Principal Component Analysis. These procedures were run and explained in subsequent section in Chapter 6.

Chapter 7 focuses on clarification and explanation about possible policy implementation in the Surabaya City and also concluding remarks with future direction of research as well as limitation of the present study. The explanation above is expressed on the flowchart above.

Chapter 2 Literature Reviews

2.1 Air Quality towards Humans Health

Our health highly depends with the quality of air we absorb. In the atmosphere, the major component of gases consist of nitrogen (N_2 ~ 78%), oxygen (O_2 ~ 20.95%), argon (Ar ~ 0.93%), carbon dioxide (CO_2 ~ 0.04%), water vapor (H20 ~ 1%) and other gas traces. These trace gases influence our health and are responsible for atmospheric chemistry. Air quality is a jargon term defined as the condition of the air we absorb which is expressed in the unit of concentrations (most of the time ppm or ug/m^3) of pollutants relative to the established threshold values. The air quality is subject to some other variables such as emissions, meteorology, and topography. The general definition of emissions is what is released into the atmosphere, whereas meteorology helps to determine what will happen to these emissions.

Important aspects that we should care about emissions are the quantity or amount of pollutants emitted, then the type, the source of emission whether it is point or non-point source. Once the pollutants are in the atmospheric layer, it is up to weather factor (meteorology) what will happen to the emission. Emission's movement is dominantly controlled by the weather (meteorological condition). There are several significant meteorological factors that affect the air quality they are solar radiation, wind (speed and direction), temperature, and humidity. However, the significance of these factors may differ across sites. Despite the complexity of atmospheric reaction to the air pollutants purification, in essence calm winds and the inversion will cause poor air quality. This condition is marked with some indicators such as the sun supplies low radiation, resulting in less warmth to the earth's surface, and then the warmer air holds cold air near the ground whereas this warmer air serves as a lid or inversion layer, and because of this layer, pollution from traffic, industries are trapped, resulting worse air quality. On the other hand, winds are responsible for transport and dispersion of pollutants. Topography has its role in the atmospheric reaction. It affects the wind direction and speed. For example, in the valleys, the temperatures are lower (cold) in the night which yields more inversion.

Air quality management is an important matter in all places, including the Surabaya City. High concentrations of air pollutants may cause health effect and interfere with human activities. For that purpose, the city government had installed five monitoring stations in the location that represent the zone and land use patterns as described in the Chapter 3. However, the use of this monitoring station has been limited. Indeed there were many air quality studies that had been investigated, however the subject of these research are being widely spread e.g., carbon stocks, dispersion modeling using deterministic model, air quality from vehicular using deterministic approach. The titles which focus on the time series information (stochastic) were very limited. One research was conducted by Djuraidah (2007). She investigated and focused on a spatio-temporal model by using an additive model with linear mixed model approach. However, she did not discuss spatial limitation on which Surabaya only has five monitoring stations which is insufficient to produce reliable spatial model. Another study was conducted by Chamida (2004), she measured PM_{10} concentrations using air dispersion model with the inputs from the use of fuel vehicles, LPG, kerosene, traffic and domestic activities. However, a thorough comprehensive study has not been done, and this present study is aimed to pioneer such study on measuring air quality based on historical measurement values which has not been done so far.

The monitoring stations capture the concentrations of several air pollutants; they are NO, NO_2, O_3, CO, PM_{10}, and SO_2. NO and NO_2 are mostly known derived from combustion sources, and in most urban areas and cities, mobile source emission sources from gasoline and diesel vehicles are the one that are responsible for emitting these pollutants (O'Donoghue, *et al.*, 2007; Gao, 2007; Wang, *et al.*, 2009). The combination between NO and NO_2 refer to NOx. These are considered as anthropogenic sources. Vehicles are considered to be the most dominant contributor of anthropogenic VOC emission and NOx emissions, roughly responsible for 47% and 55% (Gao, 2007). NO_2 is also formed through the oxidation of NO, as primary pollutant (O'Donoghue, *et al.*, 2007). Parkhurst (2004) however noted that the emission of NO_2 came most notably from bus, in the UK.

Henderson *et al.* (2007) and Wang and Oliver Gao, (2011) stated that vehicles are also sources of particulate matter. The emissions from transportation roughly account for about 30% of the PM pollution in the US. Not like particulate matters, NO, and NO_2, O_3 is formed under the condition of warm temperatures with bright sunlight. VOCs are contributor of the O_3 formation. Along with NO, VOC are precursors to secondary O_3 formation and aerosols (Wang *et al.* 2009, Parkhurst, 2004). Incomplete fuel combustion as well as stationary source of pollution may produce and increase the concentration of CO (O'Donoghue, *et al.*, 2007).

These air pollutants; if their concentrations are high, or exceeding some certain values, it will affect human's health. Not only human health that is affected but also social cost may incur (Wang *et al.*, 2009). Parkhurst (2004) further noted that NO_2 heightened mortality in the UK and O_3 is responsible of deaths and patients in hospital during the summer. Cesaroni *et al.* (2012) and O'Donoghue, *et al.* (2007) presented that these pollutants increase adult mortality due to natural diseases such as cardiovascular diseases, respiratory, and lung cancer. Cardiopulmonary is also included as the one that is caused by high concentration of NO and NO_2 (Dijkema *et al.*, 2011).

2.2 Air Quality Studies

Like other statistical studies, in broad, statistical model has been coped with two parts: interpretation and prediction. In the air quality field, there are many studies with both focusing on interpretation (Ramanathan and Feng, 2009; Gocheva-Ilieva et a l., 2013; Mamtimin and Meixner, 2011; Jorquera, 2002) and prediction time series data. Interpretation of air quality means a lot to many sides, stakeholders and even policy makers on every scale whether it is local, cities, nations, and between nations. On broader scale, we observe many studies related to the global climate change caused by anthropogenic activities. Predicting is on the other hand, is much more "exclusive" and can be said as one particular niche in air quality subject. We can say that because predicting part lets research to focus on how to forecast or predict the pollutant's concentrations ahead of time.

There are so many methods are currently being either developed or extensively applied in researcher's regional scale. The latter scale is our primary concern in the present study. In the statistical model, we can note that there two basic statistical model learning. One is supervised learning model, and the other one is unsupervised model. We do not limit ourselves to focus only in any of these methods instead we seek and use methods that may best achieve what we are targeting. Several chapters of this research has strived to obtain policy that can be implemented in order to reduce the risk of people get exposed to high concentration of pollutants.

In terms of prediction model, many researchers investigated models that best predict future concentration of pollutants. Dou *et al.* (2012) implemented a multivariate state-space model to predict next day hourly O_3 concentrations. Multivariate time series model is also used in the present study but with different goals set by authors. However, the use of multivariate time series state-space has been limited. Chapter 5 describes the use of model and shows some examples of its use in the air quality field.

Other popular model that has been extensively been developed and used is artificial neural network (ANN) (Gardner and Dorling, 1999; Chaloulakou, *et al.*, 2012; Kukkonen, 2003; Arhami, *et al.*, 2013; Cai *et al.*, 2009; Corani, 2005). Chaloulakou, *et al.* (2012) applied ANN and multiple linear regression to predict PM_{10}, the included previous information as input in the network. Kukkonen (2003) employed and tested five different neural network to predict NO_2 and PM_{10}, and he also compared the performance with other linear statistical model and a deterministic modeling system.

A multilayer perceptron neural network (MLP) was performed by Gardner and Dorling (1999) in predicting NO, and NO_2 for next hour and 24-hr ahead. Corani (2005) employed feed-forward neural networks (FFNNs), pruned neural networks (PNNs), and lazy learning (LL) to predict current day concentration of O_3 and PM_{10}.

Besides ANN based model, a time series analysis were also performed by a number of researchers. For instance, Kumar and Jain (2009) used ARMA/ARIMA to predict NO, NO_2, O_3 and CO one day ahead. Gocheva-Ilieva et a l. (2013) investigated the performance of ARIMA in predicting NO, NO_2, O_3, and PM_{10}. Moreover, Ibrahim *et al.* (2009) used and compared ARIMA and SARIMA in forecasting NO and CO.

One particular note is that in Surabaya, studies of air quality using the above mentioned above is rare. Few studies only described the difference of pattern using simple statistical methods. As explained previously, the lack of study is caused by unattractiveness of data supply from monitoring stations. There are many missing rates of data, therefore often researchers face with many limitations. Furthermore, air quality studies are often directed to use deterministic models to either explain the phenomena or predict the concentration on unmonitored location. However, still, the application of prediction model of air quality are very limited if not say almost none because no complex statistical methods had been used to explore the pollutant data from monitoring stations.

2.3 Investigating the Phenomena that Affects Air Quality

Surabaya City, as second largest city in Indonesia, is vulnerable to severe air pollution. It was reported by Chamida (2004) that there were 7, 11 and 2 days in 2001, 2002, and 2003 respectively which were categorized as unhealthy days, mainly due to transportation sector (Chamida, 2006). These facts trigger the necessity to issue various environmental policies (e.g. Blue Sky Program) and had gained interest for air pollution control strategy. There were also few studies that incorporate stochastic method to develop relationship of air quality with vehicle volume, environment, social and psychology, and health (Purwanto, 2007; Irsyada, 2011; Rachmani N., 2007; Samino, 2010; Rahmawati, 2008). However, they only account for a segment of variables and did not include a set of factors that adequately represent contribution to the concentration. To the authors' best knowledge, there is no studies have been done to comprehensively examine spatial and temporal variations and their effect to the concentrations of all air pollutants using monitoring data in this particular city.

Many models have been developed to investigate factors affecting ambient air pollutant concentrations. Models such as multiple linear and non-linear regressions and time series technique have been used (Gardner and Dorling, 2005; Geladi, 1999). Nugroho *et al.* (2013) briefly explains some drawbacks of these models, for example, a non-linear relationship among variables, interaction between variables in the model structure are not easy to be described using simple linear regression models. A study by Samino (2010) calculated level of ambient air pollution in Gayungan and Gebang Putih stations in 2008 using linear regression. Because the author just used one-year data and classical linier regression, the method was not capable to include between stations variation, therefore an alternative model is required.

At first step, we employ a multilevel model to explore factors affecting air quality. Multilevel models are a regression model with more than one level that has parameters on its own estimated from data (Gelman and Hill, 2007). We assumed if variation between-station is higher than within-station variation, then we can conclude that observed variations of air quality is mainly due to difference of locations/land-use.

2.4 Temporal Pattern of Pollutants

A part of important aspect when managing air quality is to identify points where the peak concentrations occur. When peak concentrations take place, there is a risk where the pollution levels may breach the permissible level. It is therefore crucial to obtain information about temporal peak time frame by identifying temporal pattern of pollutants on every location. Obtaining temporal pattern can be achieved via finding distribution of pollutant concentrations. There are several ways of achieving this.

A simple descriptive statistical analytics can be employed to extract temporal pattern. However this approach may not sufficiently capture the true pattern of air quality. For example, if we average all data station, there is a possibility when the unique pattern associated to each location may disappear. Therefore it is essential to use a method that is able to capture the true distribution of pollutant on each location. One method is by using component analysis and a particular emerging method is Independent Component Analysis (ICA).

When using ICA to obtain temporal time frame, we would like to identify the hidden distribution of concentration on each zone. The ordinary component analysis can be found in many literatures in terms of air quality. However to some extent, the use of ICA competes with PCA. Abdul-Wahab *et al.* (2005) used principal component analysis (PCA) to model and investigate factors of ground-level ozone, and Al-Alawi *et al.* (2008) combined principal component regression with artificial neural network for forecasting ground-level ozone. Another example of PCA is to identify air quality factors from ceramic tile clusters (Minguillón *et al.*, 2013). Pires *et al.* (2008a), Pires *et al.*, (2008b) employed principal component to optimize air quality monitoring stations in Portugal. However, there is a limit of PCA in these ordinary component analysis compared with ICA.

Back and Weigend (1997) explained that ICA can reveal better on the underlying structure and information of the stock prices than PCA. Moreover, PCA provides an orthogonal representation of the data and maximizes the variance explained by the components, however, the interpretability of the second and higher components may be limited (Westra *et al.*, 2010). Another fact, ICA is able to produce output that has very clear structure, e.g. clustered, because ICA focuses on non-Gaussian data. It is similar with the projection pursuit where it works on finding directions on which the projection of the data set show the most interesting distribution, where it is when the data has least Gaussian distribution. Hyvärinen and Oja (2000) showed that principal component analysis failed to do so, producing unclear structured output.

ICA is an alternative approach to detect some components from mixed distribution. Thus, it is often used to analyze time series data (Cheung and Xu, 2001) in many research fields including finance (Lu *et al.*, 2009), climate change (Westra *et al.*, 2010), multimedia (Long *et al.*, 2012), and air quality forecasting (Pires *et al.*, 2008c). These literatures commonly concluded that ICA is helpful in filtering out noises contained in the data. Hence this study assumes that the independent patterns of pollutant's concentrations obtained from ICA may detect noise-free temporal patterns between zones as well as maintaining independency in the case of Surabaya City.

2.5 Prediction of air quality

Air quality is one of the most complex phenomena to be forecasted because pollutant concentrations are a result of very complex interactions between meteorological factors and the influence of demographic and terrain factors as well as artificial sources of emission (e.g., vehicles and factories). Important pollutants that derive from vehicular activities are is nitrogen monoxide (NO), nitrogen dioxide (NO_2), particulate matter (PM_{10}), and carbon dioxide (CO). NO, and NO_2 (NOx) are gases that are highly correlated with transport activities because they are emitted from the majority of emissions from vehicles. NOx concentrations reflect the volume of vehicle activities and usually vary with traffic flows. Accordingly, the concentrations show temporal changes within a day. For example, the concentrations are higher during peak hours in the morning and the evening. To capture such temporal changes in a day, a time scale of hourly mean concentrations is useful from the perspective of short-term forecasting.

There are various models for air quality prediction. But these may be divided into two big groups. One group is the time series by considering the air quality data itself like e.g., autoregression model or ARIMA model. Box-Jenkins time series models involve a statistical process that incorporates the influences of past values (autoregression - AR), error values (moving average - MA), and their combinations (autoregressive integrated moving average -ARIMA). Box-Jenkins models are popular because of the ease of use and their capability in presenting both stationary and non-stationary data.

Dong *et al.* (2009) presented another prediction approach for $PM_{2.5}$ concentrations in the Chicago metropolitan area by building a hidden semi-Markov model. The trained model was shown to provide a good accuracy of prediction of $PM_{2.5}$ concentrations. An example of the use of Bayesian approaches was also available. Dou *et al.* (2012) applied a Bayesian multivariate spatial prediction model to forecast next-day hourly ground-level O_3 concentrations. They compared the Bayesian model with a model based on a state-space modeling approach and clarified that the Bayesian model is superior to the state-space modeling approach.

Stationary time series models are also widely used to predict air quality-related information. Lee *et al.* (2012) forecasted the Air Pollution Index (API) in Johor, Malaysia, using a seasonal ARIMA (SARIMA) model. The objective was to obtain best estimates for AR and MA parameters inside the SARIMA model for three monitoring stations in Johor. The model was concluded to be capable to predict and monitor air pollution index. Ibrahim *et al.* (2009) presented the use of a Box-Jenkins ARIMA model to forecast monthly maximum 1-hour CO and NO_2 concentrations in Malaysia. The purpose was to check the forecasted value with the permissible values, and analysis results showed that the predicted values will not exceed the permissible values from the regulations by either NAAQS or DOE Malaysia. Kumar and Jain (2009) also used ARMA/ARIMA models to forecast daily mean ambient air pollutants O_3, CO, NO, and NO_2 nearby roads at an urban area of Delhi, India. It was concluded that such models is suitable to the short-term forecasting for forewarning purposes. We observed that for the log-transformed NO, MAE and RMSE range from 7.3 to 8.8 and from 10.6 to 10.9, respectively.

It must be pointed out that the results in the above existing studies may be unique because different sets of data were used and therefore it is difficult to make direct comparisons in terms of model performance. Moreover, in case that a same model was employed to different data sets, worse results would be yielded if model parameters were not optimized for each data set. However, high values of prediction errors in terms of MAE and RMSE, as noted in some studies mentioned above, are also one of our motivations to developed better models that can improve prediction accuracy. Even though time series models are often applied to forecast air quality data (Sfetsos 2002), they have their own disadvantages. Bayesian Markov Switching model with state transition is thus proposed. Ji *et al.* (2013) adopted a time series state transition model to forecast short-term wind speed. Unlike other time series models, the Markov Switching model is useful because it is able to capture occasional but recurrent regime shifts in a dynamic manner. This model is also able to incorporate nonlinearities associated with regime switching (Diebold *et al.* 1994; Martínez-Beneito *et al.* 2008; Kim *et al.* 2013).

The development of air quality forecasting model has attracted many attentions especially with the involvement of exogenous variables. One type of models is built based on artificial neural network (ANN) models (Kukkonen 2003; Corani 2005; Cai *et al.* 2009; Chaloulakou *et al.* 2012; Arhami *et al.* 2013) and the other type is based on Box-Jenkins time series models (Kumar and Jain 2009; Lee *et al.* 2012). ANN models are based on connections between input layers, hidden layers, and output layers. The presence of hidden layers is useful to capture nonlinear relationships. The number of hidden layers must be predetermined to obtain optimum results. Basically, the point of applying ANN models to the forecasting is the use of training set. Training set should be large and representative otherwise it cannot be used to forecast new data. Such a feature is one of the advantages of ANN models.

Kukkonen *et al.* (2003) also proposed several models to predict hourly concentrations of NO_2 and PM_{10} in the central Helsinki. They compared five ANN models, a linear model, and a deterministic modeling system which combines a road network dispersion model and an urban dispersion model. They concluded that ANN models performed better than other models. Statistical evaluation results show that the R-squared (R^2) values of the ANN models with multilayer perceptron range between 0.36 and 0.72 from the years 1996 to 1999. On the other hand, the R^2 values for the linear models range between 0.41 and 0.5. These results suggest the necessity of further improving the prediction models. However, a significant limitation should be emphasized because the forecasting by ANN models is very sensitive to data, which are collected from different points in time and different locations. Corani (2005) presented a modified ANN model to predict ozone (O_3) and PM_{10} by integrating a feed-forward ANN model, a pruned ANN model, and a Lazy Learning model and confirmed its effectiveness. Sahu *et al.* (2011) developed a Bayesian hierarchical space-time model of ozone and revealed that it has higher model accuracy than the model without Bayesian approach. The model is able to predict ozone values a couple of hours ahead. Chaloulakou *et al.* (2012) proposed ANN models to predict daily average PM_{10} in Athens, Greece. Similar to previous studies, ANN models were confirmed to perform better than existing regression models, indicated by statistical error measure RMSE and episodic prediction ability. They further confirmed that ANN models that incorporating the influence of lagged PM_{10} has a higher accuracy than those without lagged PM_{10} information. ANN models were also employed by

Cai *et al.* (2009) in Guangzhou, China to predict hourly carbon monoxide (CO), NO_2, PM, and ozone. Cai *et al.* (2009) also demonstrated that ANN models were able to yield accurate prediction up to 10 hours, outperforming multiple linear regression models and the California line source dispersion model. Arhami *et al.* (2013) applied ANN models to predict nitrogen oxides (NOx), NO_2, NO, O_3, CO, and PM_{10} in Tehran. High correlations were obtained between the calculated and measured pollutants. They further revealed that combining ANN models and Monte Carlo simulation approaches could improve prediction accuracy. We found that mean absolute error (MAE) and the root mean square error (RMSE) of the model for NO prediction are as follows: 18.76 and 27.50, respectively, which are quite large even though the correlation coefficient shows a good agreement between predicted and observed values.

In addition from information above, linear and non-linear models have also been constantly developed, however, there was no significance difference noted between non-linear and linear models (Pires *et al.*, 2008). Liu *et al.* (2008) used a forecasting model called Bayesian hierarchical technique (Liu *et al.*, 2008) to predict CO, NOx, and dust fall . Pires *et al.*, (2008) compared five linear models to predict daily mean of PM_{10} concentrations in one site in Oporto Metropolitan Area. However, spatial variability were not concerned on that study and the regression with variables obtained independent component analysis performed the worse. Cai *et al.* (2009) employed Artificial Neural Network (ANN) to predict CO, NO_2, PM_{10} and O_3 concentrations and the performance was better compared with multiple linear regression. Wind direction was considered as independent variables but they did not separate the effect of wind direction to each prediction of pollutant, moreover, serial error correlation due to time series model was not taken into account which might cause result bias. Arhami *et al.* (2013) also used ANN to predict pollutants, but they noted less accuracy for O_3 prediction in Tehran, Iran.

Time series model is an appropriate model which avoids the problems of geographical aspects. However, the trends observed in a pollution data presents serrial error autocorrelation which generates problems in interpretation, analysis, and prediction (Touloumi *et al.*, 1994). Many researchers have performed the forecasting by regression technique but unfortunately they did not account for serial error autocorrelation.

Moreover, in a regression analysis, the correlation between independent variables (multicollinearity) may pose a serious difficulty in the interpretation of which predictors are the most influential to the response variables (Abdul-wahab *et al.*, 2005). One way to remove such multicollinearity is using component analysis method, in this case widely used a Principal Component Analysis (PCA), and the newly emerged one Independent Component Analysis (ICA). Even though these two methods have their own approach, the goal is similar is to build components that are statistically independent with each other. In regression analysis, this is particularly very useful and become good input as predictors in a regression model since they optimize spatial patterns and remove complexity due to multicollinearity (Abdul-wahab *et al.*, 2005; Shao *et al.*, 2006). ICR and PCR have been widely used in particular for plant study (Shao *et al.*, 2006), dam deformation study (Dai *et al.*, 2013), air pollutants in subway (Kim *et al.*, 2010), air quality management (Pires *et al.*, 2008a, 2008b), and O_3 prediction (Al-Alawi *et al.*, 2008).

Chapter 3 Methodology, Data, and Locations

3.1 Introduction

This chapter explains data used and obtained in the present study. Due to high missing rates of air pollutant concentrations because of sensor disorders, we focus on using data from 1st February 2001 to 30th September 2002. All information of pollutant concentrations were obtained from sensors (monitoring stations). These monitoring stations are explained in the next section below.

3.2 Monitoring Stations

In general, 80.72% of Surabaya is low land with elevation 3 – 8 m from LWS (Low Water Spring), the rest areas are hilly land (12.77%) on west and east of Surabaya (6.52%). The land slope ranges between 0-2% on lowland and 2-15% of shallow hilly lands.

To monitor air quality, Surabaya City has installed 5 monitoring stations since 2000, operating started on 2001 though. They are in (Chamida, 2004):

1. Yard of Achievement Park, Jl. Ketabang Kali (SUF 1), representation of Center of the City, housing, Office and Trading – Central Surabaya, located on Genteng District
2. Yard of Village Chief Perak Timur, Jl. Selangor (SUF 2) representation of housewares and industrial region – North Surabaya, located on Pabean Cantikan District.
3. Yard of Assistance Major Office West Surabaya, Jl. Sukomanunggal (SUF 3) representation of housing region, rural region of – West Surabaya, located on Sukomanunggal District.
4. Yard of Gayungan Subdistrict Office, Jl. Gayungan (SUF 4) represented housing region– near Surabaya Highway By Pass – Gempol – South Surabaya, located on Gayungan District.
5. Yard of Convention Hall, Jl. Arief Rahman Hakim (SUF 5) represented housing region, campus, office region– east Surabaya, located on Sukolilo District.

The locations are shown below (Figure 3.1):

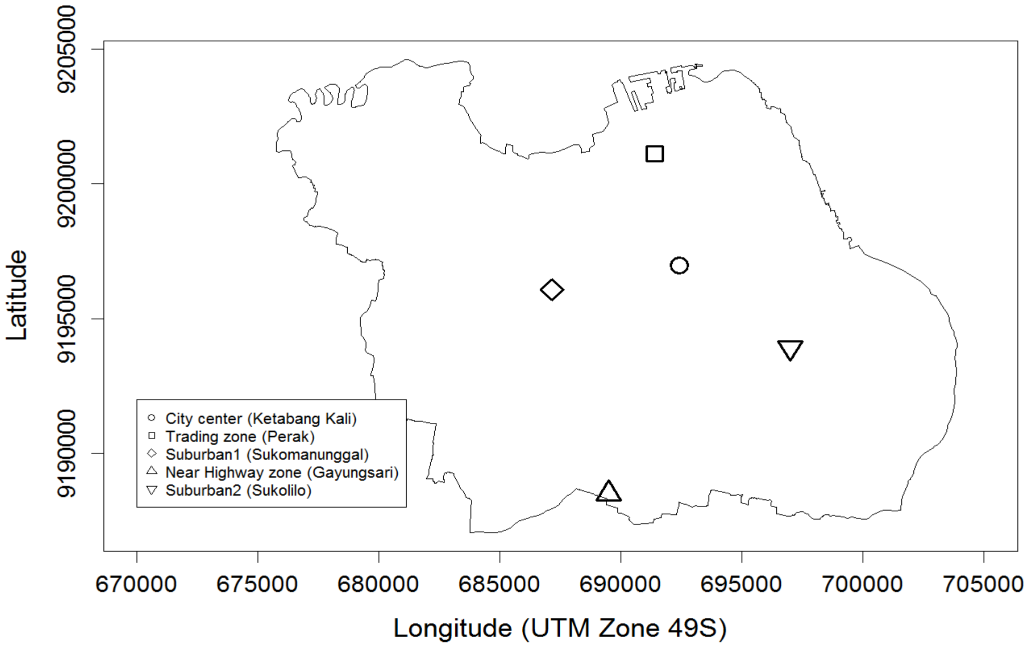
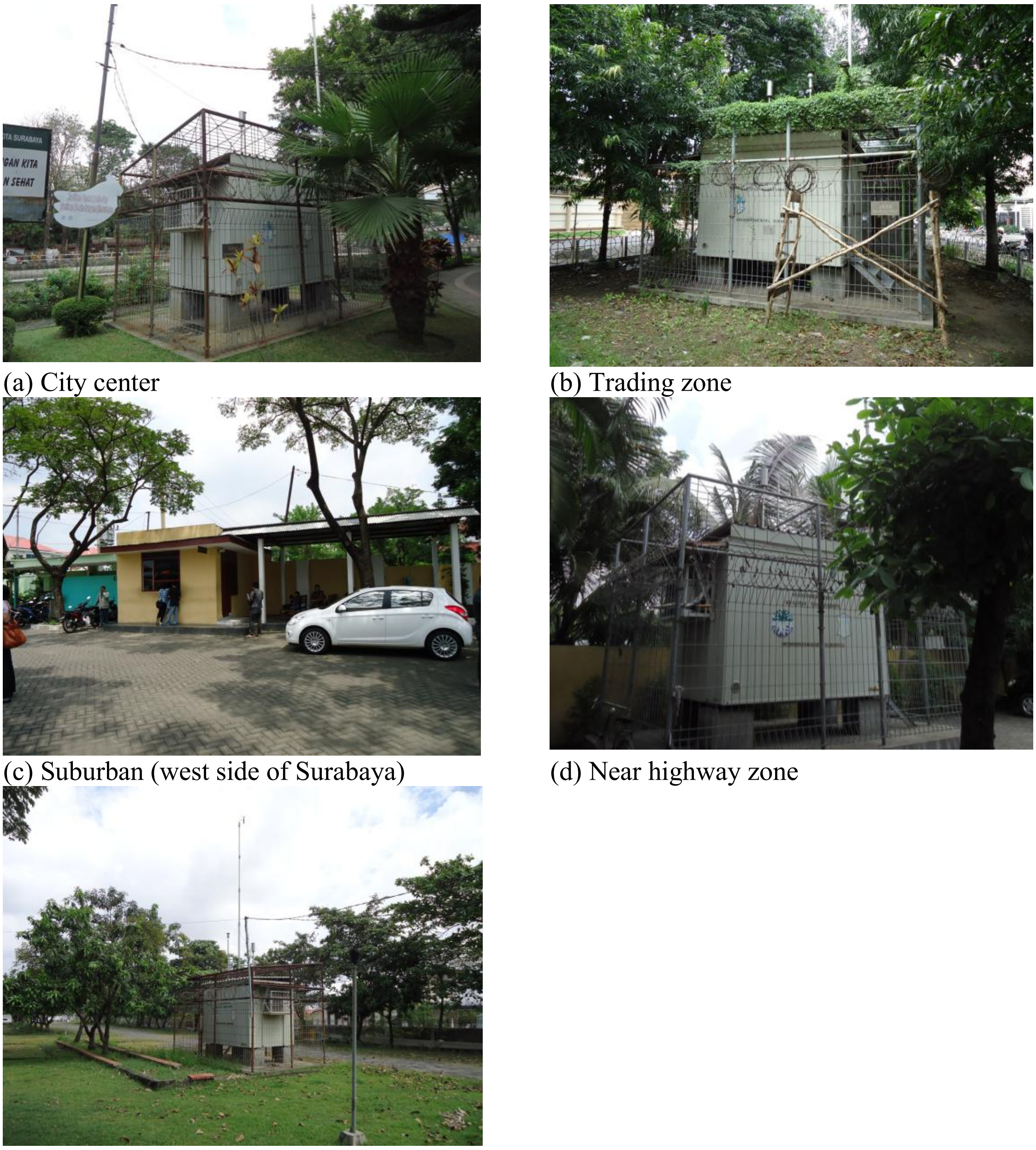


Figure 3.1 Surabaya map with location of five monitoring stations

Figure 3.2 shows the pictures of monitoring stations on five locations:



(a) City center

(b) Trading zone

(c) Suburban (west side of Surabaya)

(d) Near highway zone

(e) Suburban (east side of Surabaya)

Figure 3.2 Five monitoring stations at five locations

Sensors are located 5 m above the ground with surrounding is dominated by plants. In this study we used strong assumption that the concentrations captured may represent the ambient level. We did not take into account the effect of the surroundings towards the quality of data. There are facilities and sensors with many capabilities inside these monitoring stations, those are:

1. Sensor to capture air pollutants which consist of NO, NO_2, O_3, SO_2, CO, and PM_{10}. All concentration are in the form of ug/m^3.
2. Measurement system for meteorological information encompassing solar radiation (W/m^2), wind speed (m/sec), and wind direction (in degrees).
3. Sensors to measure temperatures (^oC) and relative humidity (in %)
4. Control system, sampling unit, reference gas supply, electricity supply system, and data transfer

The sensors work continuously, automatically, and controlled by computer system. The control system can save the data for 21 days before the old data were replaced by the current information captured by the sensor. During the 21 days period, the data are transferred through online or carried out manually via USB to the monitoring laboratory. In the laboratory, the data and the system are then be checked and be stated as a valid data. If there is malfunction of the system or if there is unclear output of concentration, then this situation must be clarified and data must be stated as valid or invalid data, and if necessary a system check and maintenance must be performed. The coordinates and elevation of each monitoring station is shown on Table 3.1. We denote the term of monitoring station in Sukomanunggal and Sukolilo as suburban1 and suburban2 to differentiate between these two zones which we will use throughout the dissertation report.

Table 3.1 Coordinates and elevation of each monitoring station

Station	Zone	Coordinates	Elevation (m)
Ketabang Kali	City center	S 07'15'41.8" E 112'44'33,9"	23
Perak	Trading	S 07'20'17.1" E 112'42'59,9"	13
Sukomanunggal	Suburban	S 07'16'41.5' E 112'44'34.0"	57
Gayungsari	Near highway	S 07'20'17.4" E 112'42'59,6"	63
Sukolilo	Suburban	S 07"17'20.3" E 112'47'48,1"	-9m

3.3 Threshold (permissible) level of pollutant

The information of the permissible level of air quality is used on Chapter 5. The permissible level or ambient threshold level for East Java Province including Surabaya that we use is based on Legislation from Governor of East Java Number 10 Year 2009. Table 3.2 shows the permissible level in the East Java Province related to the pollutants investigated in the present study. There was no threshold value for NO_2. Note for NO_2, we can use the national rule stating the permissible level for NO_2 is 400 ug/Nm^3 (1 hour measurement), 150 ug/m^3 (24-h measurement), and 100 ug/m^3 (1 year measurement). However, the rule for NO_2 is taken from National Government Rule Number 41 Year 1999 about Air Pollution Prevention (in Bahasa: Peraturan Pemerintah Republik Indonesia Nomor 41 Tahun 1999 Tentang Pengendalian Pencemaran Udara). However, in the present study, for NO_2, we use the tighter permissible level instead which is 92.5 ug/m^3.

Table 3.2 Permissible (threshold) levels of five air pollutants captured by monitoring stations

No	Parameter	Exposure Time	Threshold Ambient Level	Analysis Method	Equipments
1	Sulphur dioxide (SO_2)	24 hours	ppm (262 ug/Nm3)	Pararosanilin	Spectrophotometer SO_2 analyzer
2	Carbon monoxide (CO)	8 hours	20.00 ppm (22,600 ug/Nm3)	NDIR	CO analyzer
3	Nitrogen oxides (NOx)	24 hours	0.05 ppm (92.5 ug/Nm3)	Saltzman, NDIR	Spectrophotometer NO_2 analyzer
4	Oxidant (O_3)	1 hour	ppm (200 ug/Nm3)	Neutral Buffer potassium Iodida	Spectrophotometer
5	Dust	24 hours	0.26 mg/Nm3	Gravimetric	Hi-vol

3.4 Missing Data

The use of missing value imputation is because there are some missing values on the air quality concentration we have. Below is the descriptive statistics of the data.

Table 3.3 Proportion and missing values rate of parameters

Variables	**Proportion**	**Proportion (%)**
NO	-	15.61
NO_2	-	14.32
O_3	-	15.15
SO_2	-	15.26
CO	-	12.84
PM_{10}	-	19.29
Season		
Dry	60.3%	-
Wet	39.7%	-
Day-To-Day		
Sunday	14.33%	-
Monday	14.33%	-
Tuesday	14.17%	-
Wednesday	14.17%	-
Thursday	14.33%	-
Friday	14.33%	-
Saturday	14.33%	-
Meteorological		
Wind speed	-	5.5%
Wind direction	-	5.87%
Humidity	-	6.03%
Solar radiation	-	5.96%
temperatures	-	6%
Variables		***Mean (sd)***
NO	-	13.19 (9.55)
NO_2	-	27.62 (10.03)
O_3	-	39.48 (31.03)
SO_2	-	23.41 (26.56)
CO	-	1.045 (0.93)
PM_{10}	-	62.50 (55.14)

The missing values are around 15% for all air pollutants except for PM_{10}, and for meteorological variables the missing rates are around 6%. To input these missing values, we use Expectation-Maximization Based (EMB) Algorithm (Honaker *et al.*, 2011). All imputation processes were done on R platform using *Amelia* package (Amelia II). Amelia II assumes that the complete data are multivariate normal.

$$D \sim N_k(\mu, \Sigma) \qquad (1)$$

D represents dataset consisting of observed and unobserved part, also D has multivariate normal distribution with μ is the mean vector and covariance matrix Σ. Since the program only observes observed data, D^{obs}, so we make an assumption that the data are *missing at random* (MAR). The assumption means that the pattern of missingness is based on D^{obs}, not the D^{miss} (unobserved data). Let M to be the missingness matrix, with cells $m_{ij} = 1$ if d_{ij} Є D^{mis} and $m_{ij} = 0$ otherwise. M is a matrix that indicates missingness in cells which then we can define the MAR assumption as

$$p(M|D) = p(M|D^{obs}) \qquad (2)$$

In this algorithm, we are concerned with the complete-data parameters, θ = (μ, Σ). We denote D^{obs} as observed data and M, the missingness matrix. The likelihood of the observed data is $p(D^{obs}, M \mid \theta)$. Using the MAR assumption, we can explain:

$$p(D^{obs}, M|\theta) = p(M|D^{obs})p(D^{obs}|\theta) \quad (3)$$

As we only concern about inference on the complete data parameters, the likelihood is

$$L(\theta|D^{obs}) \sim p(D^{obs}|\theta) \quad (4)$$

Using the law of iterated expectations, we could rewrite as follows:

$$p(D^{obs}|\theta) = \int p(D|\theta)dD^{mis} \quad (5)$$

with this likelihood and a flat prior on θ, we can see that the posterior is

$$p(\theta|D^{obs}) \sim p(D^{obs}|\theta) = \int p(D|\theta)dD^{mis} \quad (6)$$

Amelia uses EMB algorithm which combines original EM method with a bootstrap to pull draws from the posterior. For each draw, the code will bootstrap the samples and simulate estimation uncertainty, and then it continue running EM to obtain the mode of the posterior for the data which is bootstrapped. Once we get draws of the posterior of the complete-data parameters, imputations are done by drawing values of D^{mis} from its distribution condition on D^{obs} and the draws of θ, which is linear regression with parameters, calculated from θ.

For imputation we use following scenarios as shown on Table 3.4 below. The use of both scenarios is to compare performance between partial dataset and full dataset in missing the value of air pollutants. Before the process we transform (square-root) skewed variables to at least approach data with normal distribution. In the process and examining the final result, we set up 5 imputation values and we allow interaction of time series and cross-section attributes, represented by Day (time attribute) and Station (cross-section), because it presented better input. We display the comparison results between if we use partial data set and if we use full data set on Figure 3.3, we can better see the difference between using partial data set and full data set on an extreme case in Station 5 (Figure 3.4).

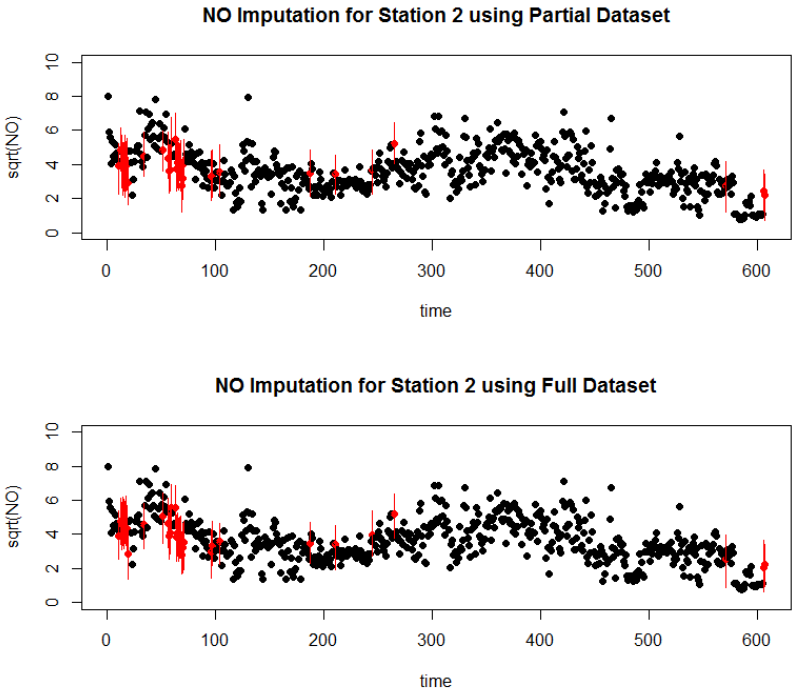
Table 3.4 Scenario of missing values imputation

Item/Variables	Scenario I	Scenario II
Day (representing temporal change with total 607 days)	V	V
Station (representing spatial pattern: categorical variable)	V	V
Meteorological		
Wind speed, wind direction, humidity, solar radiation, temperatures (5 variables)	V	V
Seasons (categorical)	-	V
Day-to-day (categorical)	-	V

For further examine, we validate the imputation and compare the result based of the difference of variables inputted to input missing data. Figure 3.5 shows validation technique by comparing the inputted values over the confidence interval and the distribution between mean imputations and observed values for NO as example, we don't show validation for other pollutants for simplicity. Relative density on full dataset data seems plausible than partial data, suggesting more variables are better for imputation process. Green confidence interval found on partial dataset is fraction missing values of covariates, suggesting that the use of full dataset helps the impute values by providing more covariates.

The output of the program is five different dataset of imputed values. We then run a linier regression and compare each set. The Table below is linier regression comparing between original dataset and dataset of NO with imputed values and we choose best R^2. Analogously, we use same procedure for other pollutants and the results are not displayed in this report.

At this step we have obtained full data consists of air pollutants (NO, NO_2, O_3, SO_2, CO, and PM_{10}) and meteorological factors (wind speed, wind direction, solar radiation, humidity, and temperatures). We then check the values and found values that are outside plausible ranges. First for air pollutants, we observe negative values for some air pollutants. For these, we omit those values and marked as NA. Second, for wind direction, some values are outside the range of 0-360 degrees, so we made appropriate conversion so that the value is within the range by subtracting to 360 if the imputed value is higher than 360 and adding with 360 if the values are negative. Third, we mark NA for imputed values of humidity if the values are above 100. After this check, we observe that the NA values are much less than before imputation. In this step, we re-run once again the EMB algorithm based on above procedure. Fortunately, after the second run, all values are filled and are within plausible ranges. From these data we generate longer time scale data: hourly, daily, and monthly, for use of analysis.



NO Imputation for Station 2 using Partial Dataset

NO Imputation for Station 2 using Full Dataset

Figure 3.3 Visual comparison of NO imputation in Station 2 using partial and full data set

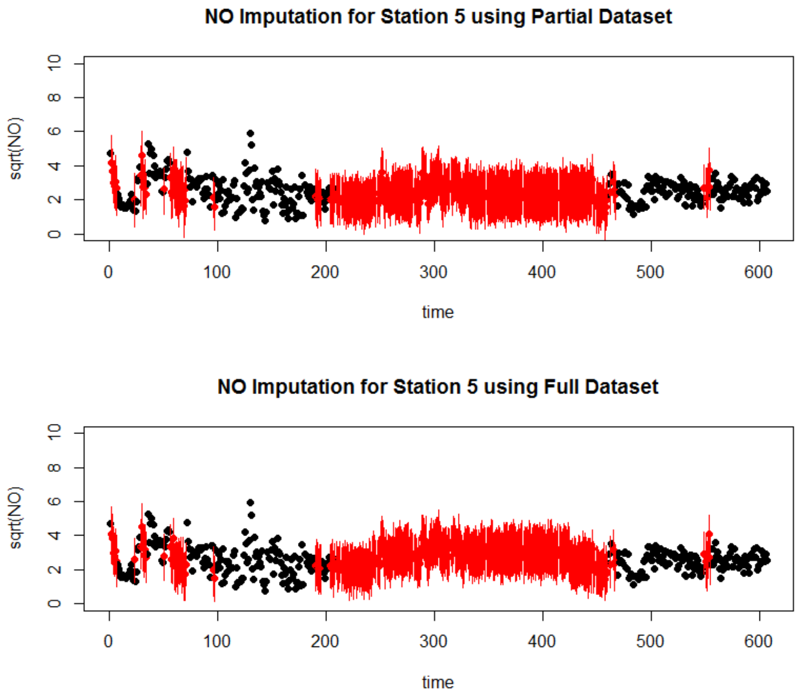


Figure β.4 Visual comparison of NO imputation in Station 5 using partial and full data set

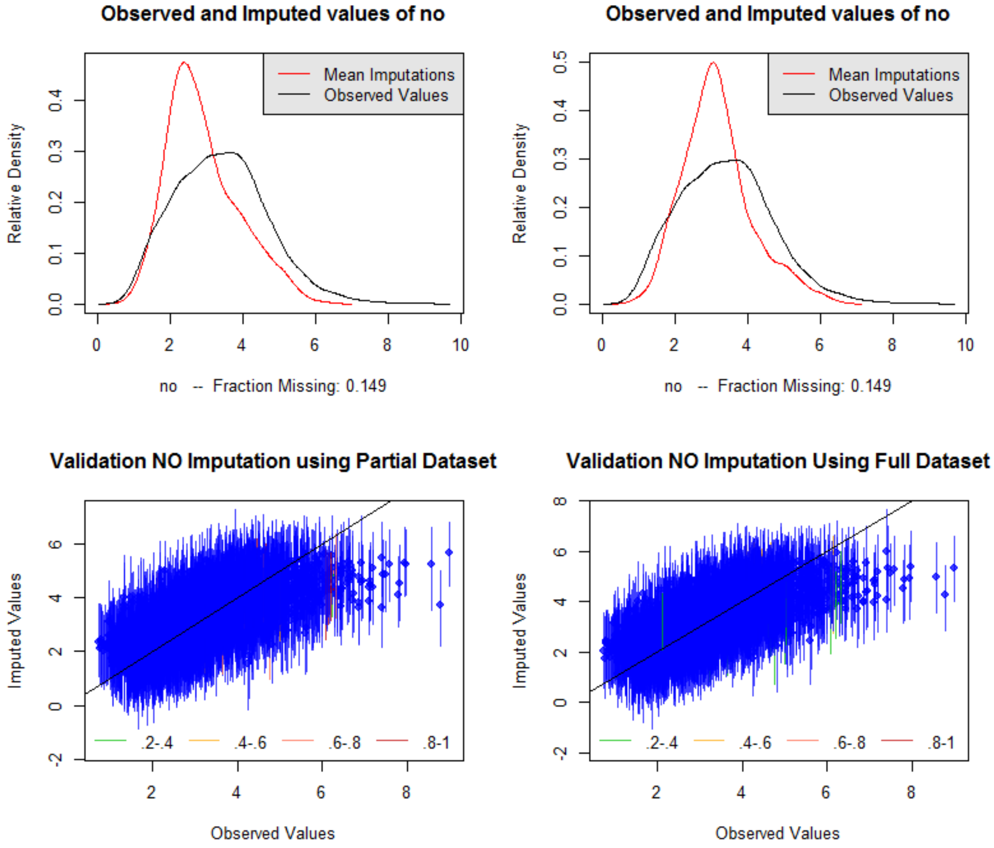


Figure β.5 Validation NO imputation based on confidence interval and distribution of imputations and observed values

Table β3.5 Multiple and adjusted R^2 of linier regression comparing different dataset of imputed values based on two scenarios: partial and full dataset

Dataset	Dataset	Multiple R^2	Adjusted R^2
Original full set		0.643	0.640
Scenario I (partial)	1	0.518	0.516
	2	0.521	0.519
	3	0.520	0.518
	4	0.503	0.502
	5	0.528	0.526
Scenario II (full)	1	0.525	0.522
	2	0.526	0.524
	3	0.521	0.519
	4	0.515	0.513
	5	0.519	0.516

3.5 Data Aggregation

We explain data aggregation in this section, but overall overview of data usage can be seen on Table 3.6 below. Data aggregation is important to determine suitable level for each pollutant. NO_2 and CO are highly affected and contributed by traffic flow, therefore short interval is better to describe this type pollutant. Because of that it is expected that using 30-mins interval, we may obtain better description and performance from NO_2 to derive any policy for this pollutant.

Table β3.6 Data aggregation and pollutants used in each chapter of the study along with goals from each chapter

Chapter	Pollutants	Aggregation	Goal
Chapter 4 Spatial and Temporal Factors of Air Quality in Surabaya City: an Analysis based on a Multilevel Model	NO, NO_2, O_3, SO_2, CO, and PM_{10}	30-mins	To obtain factors influence of spatial (zones), temporal (day-to-day, season), and meteorological factors to the concentration of air pollutants
		Monthly	To obtain effect of land use to the air pollutants
Chapter 5 Temporal Peak Time Frame and Dynamic Factors of Traffic Related Air Pollutants	NO, NO_2, O_3, SO_2, CO, and PM_{10}	30-mins, 4-week, and annual	To obtain temporal peak timeframe where the concentration may reach the peak on all location of monitoring stations To obtain the dynamic changes of air quality and the effect of meteorological factors towards the pollutant levels
		Daily and annual	To obtain seasonal variation effect to the pollutants profile
Chapter 6 Short-term Forecasting of Air Quality Concentrations Based on a Modified Bayesian Markov Switching Model	NO, NO_2, PM_{10}, and CO	Mean hourly data (1st February 2001 to 14th February 2001)	To obtain best prediction model for next 24-h in advance
Chapter 7 Prediction Model of Air Pollutant Levels using Linear Model with Component Analysis	NO_2, PM_{10}, and O_3	30-mins data from March 2002 to April 2002	To obtain best prediction model that taking into account: a) serial residual autocorrelation, b) variables extracted from component analysis (ICA and PCA)

On the other hand, PM_{10} come from more complex activities, not only from transportation but also from other activities such as business related activities, government related activities, and school related activities. These activities are high on a certain time frame in the morning and high on a certain time frame in the evening or night session. This pattern occurs every day, therefore, daily data aggregation may be best option for us to be able to better derive policies. Figure 3.6 shows approximate proper data aggregation for each pollutant and Table 3.6 shows data aggregation used in the present study.

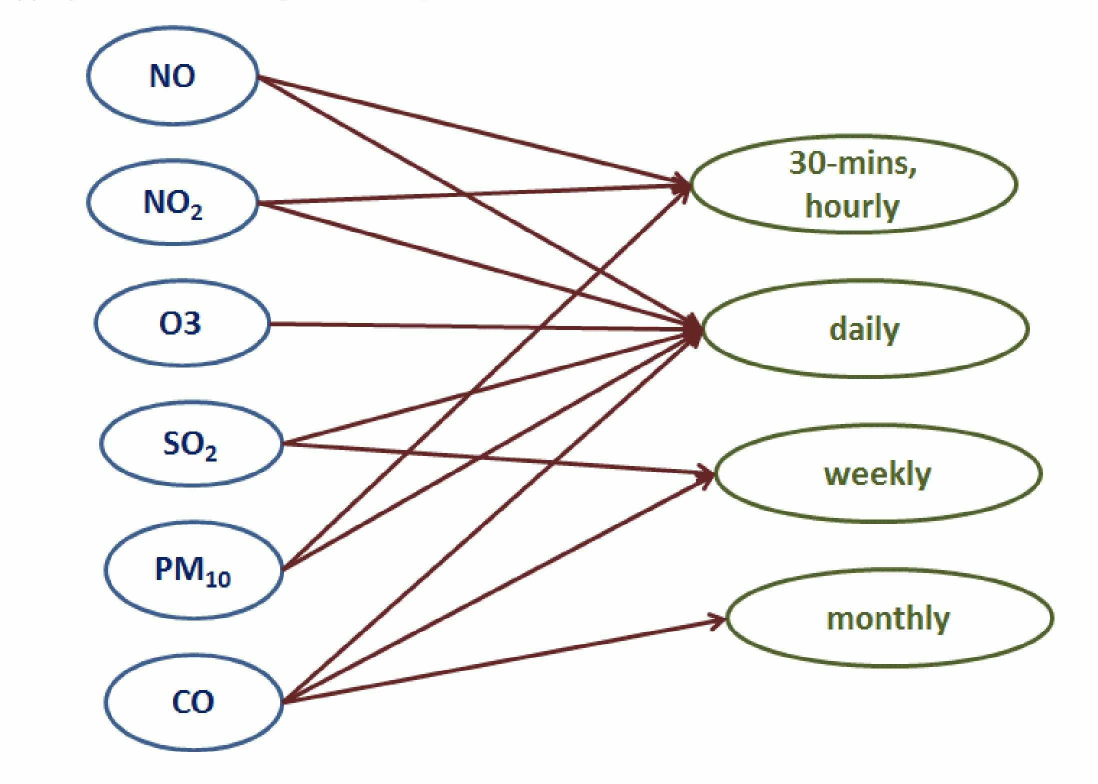


Figure β.6 Data aggregation for each pollutant

Chapter 4 Spatial and Temporal Factors of Air Quality in Surabaya City: an Analysis based on a Multilevel Model

4.1 Background

There are many factors affect air quality, and if these factors are not well known, it is difficult for policy makers to formulate environmental policies which aim to reduce the risk of people get exposed to the pollution that may harm to human health.

There are factors that known to affect pollutants concentrations actively such as meteorological factors, other pollutants. The characteristic of pollutants may differ across different sites or locations, depends on surrounding, and may also differ between days and seasons. The difference between spatial and temporal may affect what countermeasures are going to be taken from policy makers. For example, the concentrations in the city center will be different with in the trading zone due to different land use. In the trading zone, there are many business activities, and also there is a port harbor which leads to a very crowd traffic volume near this monitoring station. Since the characteristics between locations may differ, it is therefore the temporal pattern may also differ between days. Therefore, it is required to obtain how these differences affect the pollutant levels.

Furthermore, the interaction of pollutants had been noted to relate each other. Each pollutant may increase or decrease the concentration of other pollutants. This kind of relationship is lack in particular in Surabaya. Therefore because of this background we conduct this study to investigate various factors that affect pollutant levels in Surabaya. To run and obtain the result, we will use data from monitoring station. Since there are data from five different stations, it is necessary to use models which allow variation between groups, in this case station. Therefore we chose multilevel model to explain the phenomena of air quality.

The main objectives of this chapter are: 1) to overview the concentration difference between days, 2) to obtain the influence of each zone towards the pollutants concentrations, 3) to obtain whether spatial location affect the fluctuation (dynamic) of the concentrations, 4) to obtain the effect of seasons towards the concentrations, 5) to obtain how concentrations change during peak time both morning and evening session, 6) to obtain the role of green space in affecting the pollutant concentrations and how other land use pattern affect pollutant levels

4.2 Data

We investigate all pollutants that are monitored by the network taken from all five monitoring stations. The air pollutants are particulate matter 10 um (PM_{10}), Sulphur dioxide (SO_2), ozone (O_3), nitrous oxide (NO), carbon monoxide (CO), and nitrogen dioxide (NO_2). The supportive parameters are wind direction (in degrees, starting from north as 0 degrees), wind speed (m/s), humidity (%), temperature (degree Celsius), and global radiation (W/m^2).

We employ 30-mins interval data from 1 February 2001 to 30 September 2002. There are 145,681 data obtained from each station. However, there are missing values as observed on Table 4.1 below and missing value imputation using Expectation Maximization Based algorithm (EMB-algorithm) available on a *Amelia* package on R open source software was run as explained in the previous chapter. To form a multivariate normal distribution of variables, NO data was square-root-transformed. Other variables were also square root-transformed: O_3, SO_2, PM_{10}, CO, and wind speed. After such an imputation, missing rates of data were reduced as shown in Table 4.1 below.

Table 4.1 Missing rates (percentage) of data availability before and after imputation

Rate of missing values (%)	NO	NO_2	O_3	SO_2	CO	PM_{10}
Before imputation	15.61	14.32	15.15	15.26	12.84	19.3
After imputation	1.75	2.14	0.0535	0.345	0.041	0.0316

4.3 Multilevel model

We develop a multilevel model with explanatory variables which is expressed as follows:

$$y_{ij} = \alpha_{j[i]} + \beta xij + \mu_{oj} + \epsilon_i \qquad (1)$$

where y_{ij} is measurable data of air quality in ug/m^3 (either NO, NO_2, O_3, SO_2, CO, and PM_{10}) *i* at monitoring station *j,* $\alpha_{j[i]}$ and β are unknown parameters that is to be estimated, x_{ij} indicates explanatory variables:

1. dummy variables of spatial locations (city center as base, trading zone, suburban1, near highway zone, suburban2)
2. Seasons (dry and wet seasons), where dry season is set as base (dry = 0)
3. Day-to-day variation when Monday is set as base (Monday = 0)
4. interaction between all pollutants and
5. Meteorological factors that consist of wind direction (in degrees, 0 degree is north), wind speed (m/s), humidity (%), solar radiation (W/m^2), and temperatures (in ^{o}C)
6. Long holiday, defined as more than 3 days off consecutively.
7. Peak morning time defined from 7am to 9am and peak evening time defined from 5pm to 7pm.

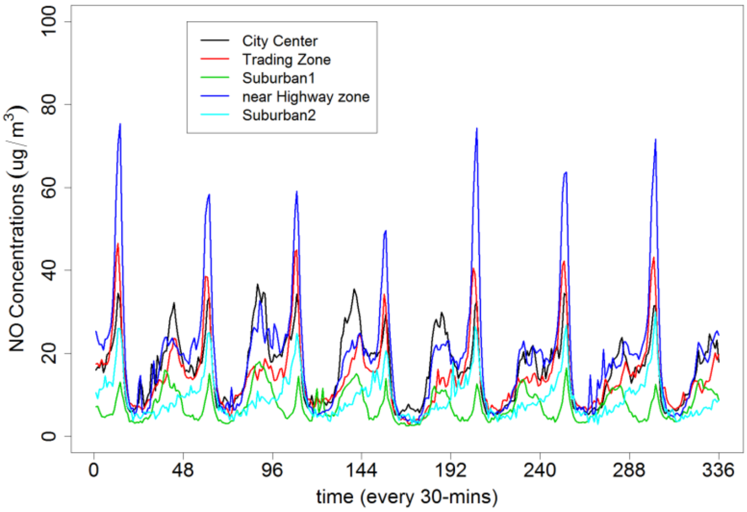
μ_{oj} and ϵ_i represent random components indicating between-station variation and within-station variation respectively. We excluded PM_{10} from interaction among pollutants as PM_{10} is inert gaseous particles where its deposition mainly due to physical properties. All estimation in this study were conducted using the *R* statistical software with `lmer()` function (a multilevel code in a package).

4.4 Results and discussions

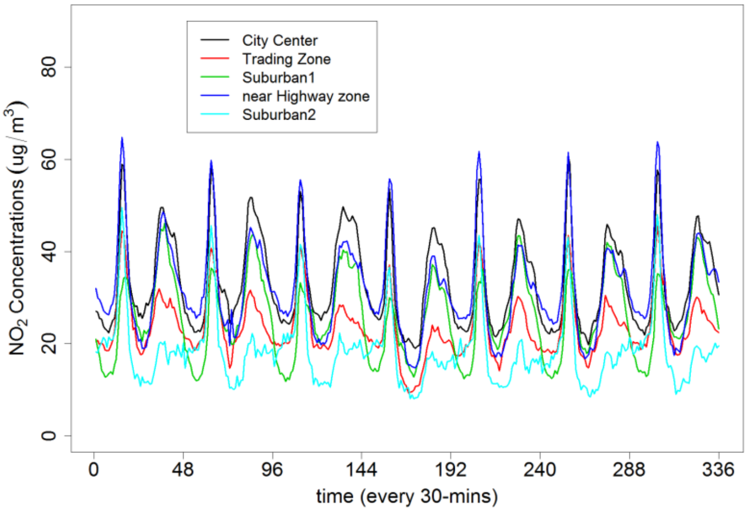
4.4.1 Diurnal variation

Figure 4.1 below shows weekly diurnal variation of each pollutant from Thursday to Wednesday, averagely taken from February 2001 to September 2002 using 30-mins interval data. As previously mentioned, 30-min interval shows most accurate description, and business and industry activities are diversified by week. The diurnal cycle of NO and NO_2 is shaped like double waves, with morning peak is higher in magnitude then evening peak. The decrease of NO and NO_2 correlates with an increase in O_3 (Han, 2011). The concentrations of NO increased from 5am to 8am in all stations, however we also observe another increase, although not as high as in the morning on city center, trading, highway, and suburban1 (Sukomanunggal area) zones. These results are similar with diurnal concentration on Tianjin, China, as reported by Han (2011). Average weekly diurnal cycle for O_3 were generally lower in trading zone, while higher in suburban area and city center. The concentration increased and reached its peak between 9.30-1.30am in all locations, took place on mid-day, in agreement with Han (2011). The average concentration appeared to be higher on suburban1 (Sukomanunggal), similar with in Jakarta and Tianjin, China (Nugroho *et al.*, 2013, and Han, 2011).

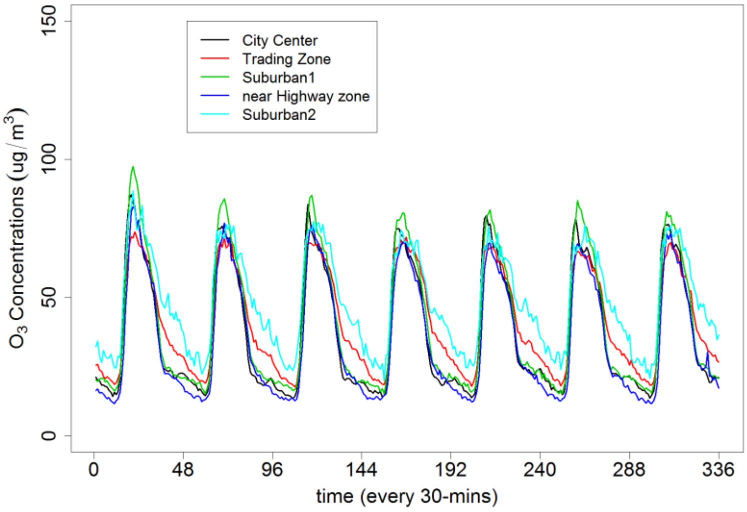
Furthermore, we found that SO_2 concentrations were generally lower in suburban area, while higher in city center and trading zone. The concentrations of CO increased and reached its peak between 6.30-8am in the morning and between 6-10pm in all stations. The average concentration of CO also showed seasonal variation. The concentrations of PM_{10} increased and reached its peak in around 5.30-8.30am, indicating peak flow of activities. The highest average of PM_{10} concentration was measured on highway zone.


City Center
Trading Zone
Suburban1
near Highway zone
Suburban2
NO Concentrations (ug/m^3)
0
20
40
60
80
100
0
48
96
144
192
240
288
336
time (every 30-mins)

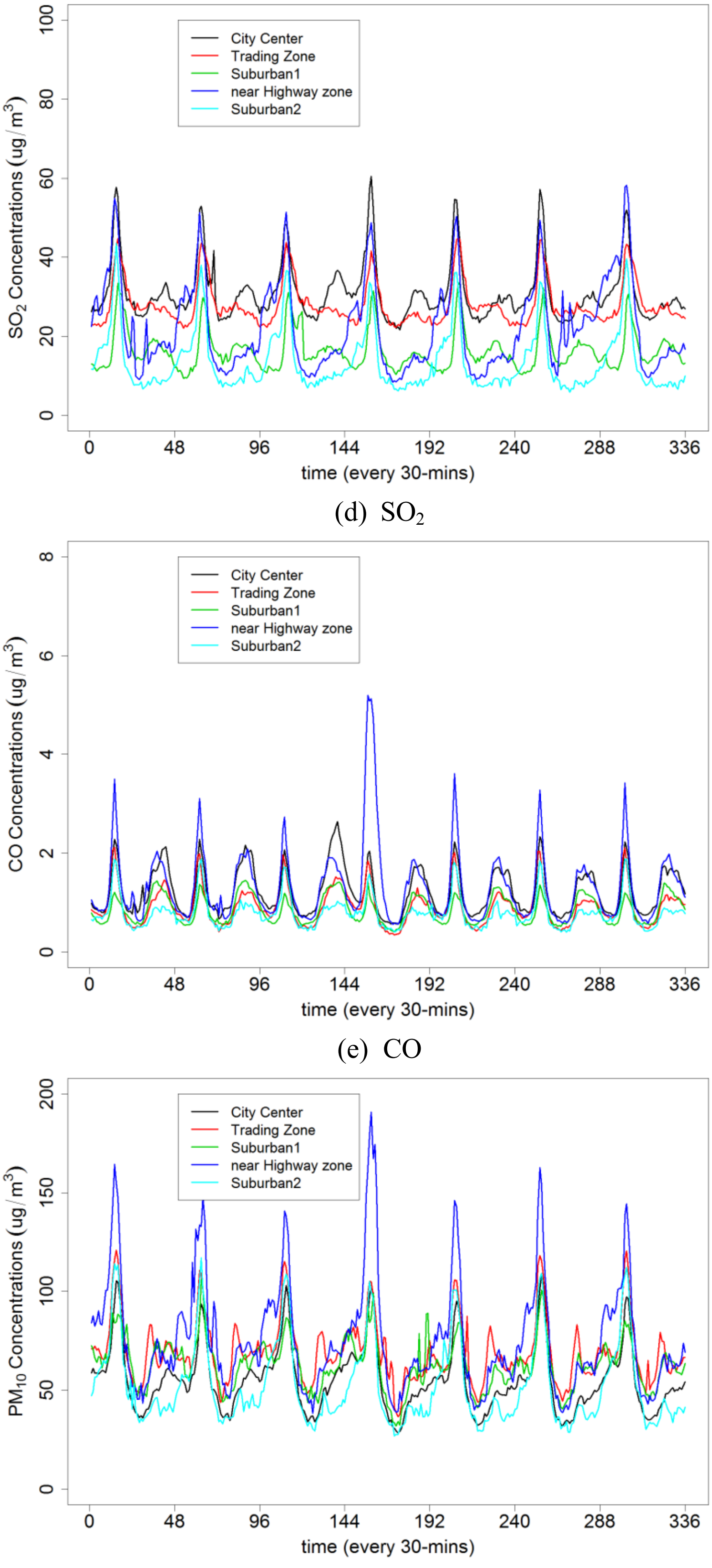
(a) NO


City Center
Trading Zone
Suburban1
near Highway zone
Suburban2
NO_2 Concentrations (ug/m^3)
0
20
40
60
80
0
48
96
144
192
240
288
336
time (every 30-mins)

(b) NO_2


City Center
Trading Zone
Suburban1
near Highway zone
Suburban2
O_3 Concentrations (ug/m^3)
0
50
100
150
0
48
96
144
192
240
288
336
time (every 30-mins)

(c) O_3



(d) SO_2

(e) CO

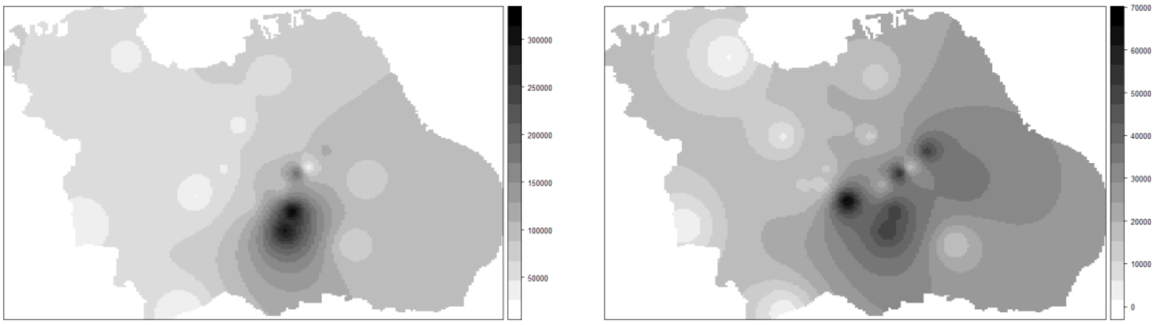
(f) PM_{10}

Figure 4.1 Weekly variation of average pollutants

Average weekly diurnal cycle for PM_{10} was generally lower in suburban, while higher in highway. One note that must be addressed, even though there is temporal pattern we can capture, however, these patterns over regions are alike and similar, leading to a suspect. The difference is only in terms of magnitude and considering the pollutants sources are unique for each location, which we expect different peak time interval, therefore we should identify and confirm further the temporal pattern using another method which will be described in the next chapter.

4.4.2 Multilevel model results

Table 4.2 shows the model result explaining all variables. In terms of spatial distribution, the estimate for O_3 on highway zone is noticed to be the largest, being the city center with lowest estimate. The result is in agreement with Zhang and Oanh (2002) where O_3 levels were observed to be high on areas except locations where the traffic is high and when there is a curbside on that particular location. Figure 4.2 shows traffic distribution over Surabaya City from motorbike, private car on year 2009, taken on June, from 5am to 9pm using Inverse Distance Weight (IDW) interpolation. It can be seen that high volume of vehicles concentrated on city center. This distribution is in agreement with O_3 level along with the proportion of NO_2 and NO where the estimate on suburban area is lower than city center and highway. Also, it shows tight correlation between NO_2 and traffic volume, in particular motorbike and private car. Higher traffic volume in highway, city center, and suburban2 area resulted higher concentration of NO_2.



(a) (b)

Figure 4.2 Traffic distribution with interpolation using IDW in Surabaya City of: (a) motorbike, (b) private car; y ordinal indicates vehicle volumes

Regarding SO_2, which is mainly from fuel-related combustion activities, tthe estimate is lowest in suburban1 with highest parameter on the trading zone which makes sense since the zone consists of dense industrial activities. On the other hand, the estimate of PM_{10} is found to be highest on trading zone, clearly indicates high volume of activities e.g., transport, industries.

Based on the observation on day-to-day variation result, Sunday is found to have lowest estimate for NO and NO_2 (estimate -2.006 for NO_2, statistically significant, Table 4.2), indicating lower activities of transportation. Moreover, we did not see clear pattern with regard to SO_2, however, we found that O_3 on Sunday, Monday and Tuesday were lower compared other days, indicating a complex reaction in ozone formation that takes time. The estimate for PM_{10} on Monday is observed to be lowest, with no significant difference to other days. The concentration tended to increase over time (daily) for O_3 and SO_2, indicating the increase of atmospheric problems in the future, unless sustainable environmental policy is put into practice. The increase of O_3 was in agreement with an increase of O_3 in Jakarta (Nugroho *et al.*, 2013). The negative estimate for NO and NO_2 are relatively small and not significant. We note that there is an increase of concentration in wet season for NO, NO_2, and O_3, suggesting there was an increase of chance of smog occurs. The emission may come from a traffic behavior that is affected by local weather condition, e.g., rain. As expected for PM_{10}, since they are inert particle, droplets in wet season helped in the deposition of PM_{10}.

Additional note is for negative AIC value of CO model result. AIC is calculated by using the following formula:

AIC = -2*Ln(L) + 2k

where L is the maximized value of any Likelihood function defined for that particular model and k is the number of parameters in the model. If -2*Ln(L) + 2k is less than 0, it means that the log-likelihood at the maximum point is > 0 indicating that the likelihood at the maximum is > 1. Positive log-likelihood is not a problem and it is misconception that the log-likelihood has to be negative. This occurrence may happen if the likelihood is derived from a probability density that can reasonably exceed 1 which means that log-likelihood is positive, therefore the AIC and the deviance and are negative.

Table 4.2 Estimation result of multilevel model using 30-mins data

Parameter	NO	NO_2	O_3	SO_2	CO	PM_{10}
Fixed Part						
Intercept	2.9e-01	18.5	2.43	4.20	4.91e-02	1.47e+01
	(2.53)	(17.25)	(18.45)	(27.38)	(2.82)	(69.63)
Spatial (D) with city center:0						
Trading	2.79e-01	-4.30	3.44e-01	1.1e-01	-4.12e-02	9.55e-01
	(13.66)	(-22.44)	(14.65)	(1.04)	(-13.25)	(31.60)
suburban1	-2.52e-01	0.55	3.62e-01	-7.83e-01	-1.36e-02	5.9e-01
(Sukomanunggal)	(-12.03)	(2.81)	(15.02)	(-7.42)	(-4.28)	(18.93)
Highway	2.14e-01	0.13	4.66e-02	-1.02	3.20e-02	1.10
	(10.40)	(0.67)	(1.97)	(-9.70)	(10.24)	(36.10)
Suburban2 (Sukolilo)	3.18e-01	-3.59	1.06	-1.13	5.04e-03	1.15e-01
	(14.94)	(-17.97)	(43.50)	(-10.73)	(1.55)	(3.70)
Days (Dummy) with Monday: 0						
Sunday	-2.02e-01	-2.41	-3.92e-02	9.92e-02	4.2e-02	8.75e-02
	(-15.90)	(-20.20)	(-2.68)	(6.56)	(21.71)	(3.68)
Tuesday	6.54e-02	0.18	-8.85e-03	-1.26e-02	-1.14e-02	2.63e-02
	(5.17)	(1.54)	(-0.61)	(-0.84)	(-5.92)	(1.11)
Wednesday	6.74e-02	1.05	6.50e-02	2.39e-03	-1.71e-02	1.1e-01
	(5.30)	(8.83)	(4.45)	(0.16)	(-8.86)	(4.61)
Thursday	1.12e-01	1.26	1.73e-01	-4.59e-02	-1.39e-02	2.92e-01
	(8.84)	(10.53)	(11.88)	(-3.05)	(-7.18)	(12.28)
Friday	9.65e-02	1.40	6.48e-02	-7.43e-02	-1.48e-02	1.65e-01
	(7.63)	(11.77)	(4.46)	(-4.95)	(-7.71)	(6.95)
Saturday	8.81e-02	0.52	1.22e-01	-1.81e-02	-6.75e-03	2.25e-01
	(6.95)	(4.39)	(8.35)	(-1.20)	(-3.50)	(9.47)
Day (temporal)	-4.77e-04	-0.004	3.26e-04	3.31e-03	-1.06e-04	-3.80e-04
	(-21.64)	(-17.84)	(12.86)	(135.89)	(-31.64)	(-10.00)
Seasonal (Dummy):	2.79e-01	0.25	4.28e-01	-6.56e-02	-1.12e-02	-5.70e-01
Dry: 0; Wet: 1	(33.40)	(3.17)	(44.78)	(-6.61)	(-8.79)	(-36.88)
Interaction:						
i) NO	-	0.23	-4.85e-01	2.85e-01	7.74e-02	-
		(8.39)	(-159.90)	(85.18)	(203.32)	
ii) NO2	2.60e-03	-	5.22e-03	4.08e-02	7.92e-03	-
	(8.39)		(14.63)	(117.00)	(192.33)	
iii) O_3	-3.67e-01	0.35	-	1.53e-01	1.77e-02	-
	(-159.90)	(14.63)		(51.54)	(46.38)	
iv) SO_2	2.03e-01	2.55	1.44e-01	-	-8.72e-03	-
	(85.18)	(117.00)	(51.54)		(-23.42)	
v) CO	3.35	30.16	1.01	-5.31e-01	-	-
	(203.32)	(192.33)	(46.38)	(-23.42)		
Meteorological:						
Wind Direction	2.6e-03	0.013	-8.56e-04	1.11e-03	-3.12e-04	4.99e-03
	(54.06)	(29.73)	(-15.33)	(19.25)	(-42.61)	(56.53)
Wind Speed	1.21e-01	-0.76	-2.08e-01	-3.27e-01	-4.04e-02	-8.26e-01
	(20.84)	(-13.96)	(-31.17)	(-47.84)	(-46.09)	(-80.20)

Parameter	NO	NO_2	O_3	SO_2	PM_{10}	CO
Global radiation	8.27e-04 (43.51)	-0.0045 (-25.22)	1.21e-03 (55.78)	1.96e-04 (8.63)	-1.30e-04 (-45.22)	-1.64e-04 (-4.74)
Humidity	-5.41e-03 (-11.42)	-0.2 (-46.71)	-4.27e-02 (-80.51)	-6.80e-03 (-12.12)	3.13e-03 (43.83)	-2.70e-02 (-32.00)
Temperature	1.05e-02 (3.50)	-0.64 (-22.92)	2.12e-01 (62.40)	-6.87e-02 (-19.33)	1.19e-02 (26.07)	-1.93e-01 (-35.14)
Effect of long holiday (dummy: non-long holiday: 0)						
Long holiday	4.38e-02 (2.57)	-1.22 (-7.64)	2.02e-01 (10.32)	-2.79e-01 (-13.87)	-3.62e-03 (-1.40)	-9.54e-01 (-30.31)
Peak event (D) non-peak time:0						
Peak morning time	-4.47e-01 (-41.31)	1.91 (18.72)	3.12e-01 (24.96)	6.05e-01 (47.31)	1.44e-01 (89.88)	2.11 (113.89)
Peak evening time	-0.30 (-24.15)	3.64 (31.36)	-2.33e-01 (-16.29)	-2.25e-01 (-15.24)	4.82e-02 (25.54)	1.86e-01 (8.04)
Random Part						
Between Stations	0.000	0.013	0.000	0.006	3.54e-06	0.000
Within Stations	1.356	119.261	1.791	1.906	0.031	4.941
Model Performance						
AIC	370921	898812	403718	411042	-73444	542992
BIC	371182	899073	403979	411304	-73183	543215
-2 * Log likelihood	-185433	-449379	-201832	-205494	36749	-271473

(t-test in parentheses)

Not only concentrations are affected by spatial and temporal, the interaction between pollutants has significant roles. Based on the interaction estimates, we summarize that using monthly timescale (Table 4.3), the increase of NO and NO_2 concentration might lead to the decrease of O_3 formation, although O_3 is affected by ratio of NO and NO_2. Interactions with NO for pollutants NO_2, O_3, SO_2, and CO are statistically significant using 30-mins data. The positive dependency of O_3 production on the solar radiation and the negative relationships with NO_2 are due to the NO_2 photolysis process (Monoura, 1999). O_3 was also weakly affected by CO, which was in agreement with Monoura (1999). Judging interaction of SO_2 with other pollutants, we can conclude that only CO has negative estimate to SO_2 with estimate -5.31e-01 (statistically significant). It indicates that NO and O_3 had significant interaction with SO_2. Madhavi Latha and Highwood (2006) found that there was positive correlation between PM_{10} and SO_2 concentrations over UK study area. SO_2 had negative impact to the concentration of CO.

Wind speed, global radiation and temperature had positive estimates for NO, indicating that the increase of those variables will increase the concentration. For NO_2, except wind direction, wind speed, global radiation, humidity, and temperatures have negative impact. The increase of those four variables will decrease the concentration of NO_2. O_3 is negatively affected by wind speed and humidity with estimates, -2.077e-01 and -4.267e-02, respectively, all statistically significant. However, O_3 is positively relevant with global radiation, which in agreement with Monoura (1999), and also positively correlated with temperature (Han, 2011).

The increase of NO, NO_2, SO_2, and PM_{10} is consistent with the increased frequency of stronger easterly winds. The decrease of O_3 and CO is consistent with the increase frequency of stronger westerly winds. Estimates of daily average concentrations of all pollutants are significantly affected by temperature. We found out an increase of mean temperature over 20 months of observations of all locations (not shown). It is therefore necessary to add green spaces to balance and attenuate temperature as related to urban heat island subject. Furthermore, the addition of green spaces will surely assist the reduction of emission (Table 4.3). To confirm the necessity of central government policy regarding long holiday, the estimation result of green spaces suggested a decrease of concentration during holiday (NO_2), which fits our initial expectation. Usually in long holiday, people tend to have trip to outside the city, therefore causing an increase of traffic volume before the holiday and at the last day of holiday. The estimates for SO_2 and PM_{10} are negative (-0.278 and -0.887, statistically significant). The decrease of transport and industrial activities had positive contribution to the decrease of PM_{10} concentration.

Table 4.3 Estimation result of multilevel model using monthly interval data

Parameter	NO	NO_2	O3	SO_2	CO	PM_{10}
Fixed Part						
Intercept	-1.18e-01 (-0.061)	2.52e+01 (1.20)	9.79 (3.26)	-4.47e-01 (-0.087)	-8.21e-02 (-0.29)	1.76e+01 (5.61)
Land use						
Trade and services	1.42e-05 (2.51)	-1.20e-04 (-1.93)	7.92e-07 (0.083)	4.10e-05 (2.76)	3.74e-07 (0.45)	-2.81e-05 (-3.02)
Public facilities	-4.01e-07 (-0.24)	6.76e-07 (0.04)	-6.76e-07 (-0.25)	6.3e-07 (0.14)	6.24e-07 (2.71)	3.80e-07 (0.14)
Green spaces	-3.21e-06 (-2.2)	3.63e-05 (2.29)	3.61e-06 (1.49)	-1.39e-05 (-3.77)	-4.69e-07 (-2.24)	-5.88e-06 (-2.85)
Residences	-3.17e-06 (-3.35)	2.90e-05 (2.77)	-1.00e-06 (-0.61)	-5.47e-06 (-2.10)	2.3e-07 (1.61)	-9.37e-08 (-0.062)
Seasonal (Dummy): Dry: 0; Wet: 1	2.08e-01 (1.605)	3.84 (2.81)	6.1e-01 (2.98)	-5.43e-01 (-1.58)	-2.88e-02 (-1.55)	2.14e-01 (1.03)
Interaction:						
i) NO	-	2.02 (1.75)	-6.02e-01 (-3.64)	3.52e-01 (1.23)	7.38e-02 (5.54)	-
ii) NO_2	1.71e-02 (1.75)	-	-1.91e-02 (-1.18)	9.14e-02 (3.72)	6.57e-03 (5.3)	-
iii) O_3	-2.24e-01 (-3.64)	-8.42e-01 (-1.18)	-	7.75e-01 (5.02)	1.67e-02 (1.8)	-
iv) SO_2	4.99e-02 (1.23)	1.53 (3.72)	2.95e-01 (5.02)	-	-1.76e-02 (-3.19)	-
v) CO	3.60 (5.54)	3.78e+01 (5.3)	2.18 (1.8)	-6.05 (-3.19)	-	-
Meteorological:						
Wind Direction	9.57e-04 (0.53)	-3.97e-03 (-0.20)	-5.89e-05 (-0.02)	3.04e-03 (0.63)	-4.32e-04 (-1.70)	-5.69e-03 (-1.86)
Wind Speed	-1.06e-02 (-0.14)	2.12e-01 (0.26)	-2.86e-01 (-2.39)	-2.67e-01 (-1.34)	-9.66e-03 (-0.9)	1.53e-01 (1.25)
Global radiation	2.31e-03 (2.33)	2.54e-03 (0.23)	4.07e-03 (2.53)	-1.11e-02 (-4.57)	-1.72e-04 (-1.19)	1.60e-03 (1.02)
Humidity	3.3e-02 (2.46)	-7.54e-01 (-5.96)	-5.95e-02 (-2.73)	4.1e-02 (1.12)	5.97e-03 (3.18)	-5.5e-02 (-3.46)
Temperature	-4.39e-02 (-0.78)	-1.63e-01 (-0.27)	-3.92e-02 (-0.42)	2.16e-01 (1.46)	6.42e-03 (0.8)	-1.75e-01 (-1.78)
Random Part						
Between Stations	0.009	1.038	0.023	0.062	0.000	0.027
Within Stations	0.066	7.765	0.176	0.464	0.001	0.207
Performance Criterion						
AIC	201.5	602.3	284.4	365.5	-124.9	283.6
BIC	245.8	646.5	328.6	409.8	-80.66	317.5
-2 * Log likelihood	-83.74	-284.1	-125.2	-165.8	79.47	-128.8

(t-test in parentheses)

In this model, we also investigate the distribution of concentration during peak time. The result is in agreement with our general assumption that the concentration during peak time are higher than non-peak time, even though we only observe the tendency of O_3 increase only during morning peak time. The estimate of NO_2 in the morning is positive (statistically significant). In the evening, the estimate is also positive. Another observation regarding the effect of land use towards the concentration is we note that the concentration of ozone was negatively relevant with public facilities and residential (Table 4.3), vice versa for trade and services and green spaces. Zhang and Oanh (2002) mentioned that O_3 concentrations tend to be lower on locations that are marked by high traffic volume and curbside, where O_3 destruction is significant. The result indicates quite markable traffic on public facilities and residential land use.

4.5 Conclusion

This study analyzes factors that affecting concentrations of six air pollutants in Surabaya City: NO, NO_2, O_3, SO_2, CO, and PM_{10} measured over 607 complete days from February 2001 to September 2002. Diurnal analysis figures show seasonal pattern of all air pollutants, we can observe a high concentration located on highway zone. The result from multilevel model shows that the air quality related to traffic volume (NO_2) was generally lower in suburban area (Sukomanunggal) and the concentrations had positive correlation with traffic volume. We observe the traffic volume was relatively high on city center, highway, and suburban2 zones and from the model the estimates over these zones are high. Important note taken is traffic volume has positive relationship with traffic-related emission.

We note that the difference between days had been significance for all pollutants, suggesting unique pattern and emission each day. Moreover, as expected, the concentration the traffic-related pollutant, NO_2, was lower in Sunday but not in Saturday. This result suggests that Saturday may be regarded as another working day, particularly it is difficult to consider Saturday as part of weekends. Moreover, we observe high concentration of PM_{10} on Sunday, suggesting the source of PM_{10} on this day was dominated by other sources other than transport.

The concentration is also noted to be different among zones, meaning that each zone has different concentration patterns and it is significant. The dynamic of pollutants was affected by the factors within station, it had nothing to do with other stations, as indicated by spatial variation in the model result.

Interesting fact is noted from NO_2 that during wet season, the concentrations tended to be higher than concentration during dry season. This may reflect changes of travel behavior, for example during wet season traffic is easily to be congested, therefore leading to more emission. As part of traffic management, it is important to identify the different concentrations occur during peak time hour. In this case, we found that the concentrations of pollutants are significantly higher during both peak morning time and peak evening time. Based on this result, we obtain information that during peak morning and evening time, the concentrations were significantly high, however we didn’t know the exact time frame as the peak time was predetermined before in the model. Secondly, the dynamic of meteorological variables cannot be fully explained using multilevel as it require further group definition. Therefore we explore these problems and investigate them in the next chapter.

In one of our variables based on monthly data, we explored the effect of land use towards the concentrations. It was found that green spaces effectively reduce emission of traffic emission air quality. The other land uses did not significantly affect the concentration of pollutants.

We observe high interactions among NO, NO_2, and O_3. Since there is a tendency for temporal increase of O_3, not only policy related to vehicles must be issued due to the increase of traffic-related emission, but land use management is crucial. Adding green spaces is important to help reducing temperature in many spots. Additionally, NO_2 concentrations tended to be higher if there are more spaces for trade and services.

Chapter 5 Temporal Peak Time Frame and Dynamic Factors of Traffic Related Air Pollutants

5.1 Introduction

One important aspect of air quality management is to be able identifying crucial time point when there is a risk of environmental pollution occurring. The results from previous chapter have shown that during peak time, pollutant concentrations are higher than non-peak time. During this time interval, people are exposed to pollution on which there is a possibility that the pollution breach the permissible level and therefore pose dangerous effect to the human health. Therefore it is crucial as part of air quality management to identify temporal pattern on certain locations. Not only daily interval cycle must be obtained, the effect of seasons towards the concentration levels is also necessary to grasp insight of air quality phenomena from different seasons.

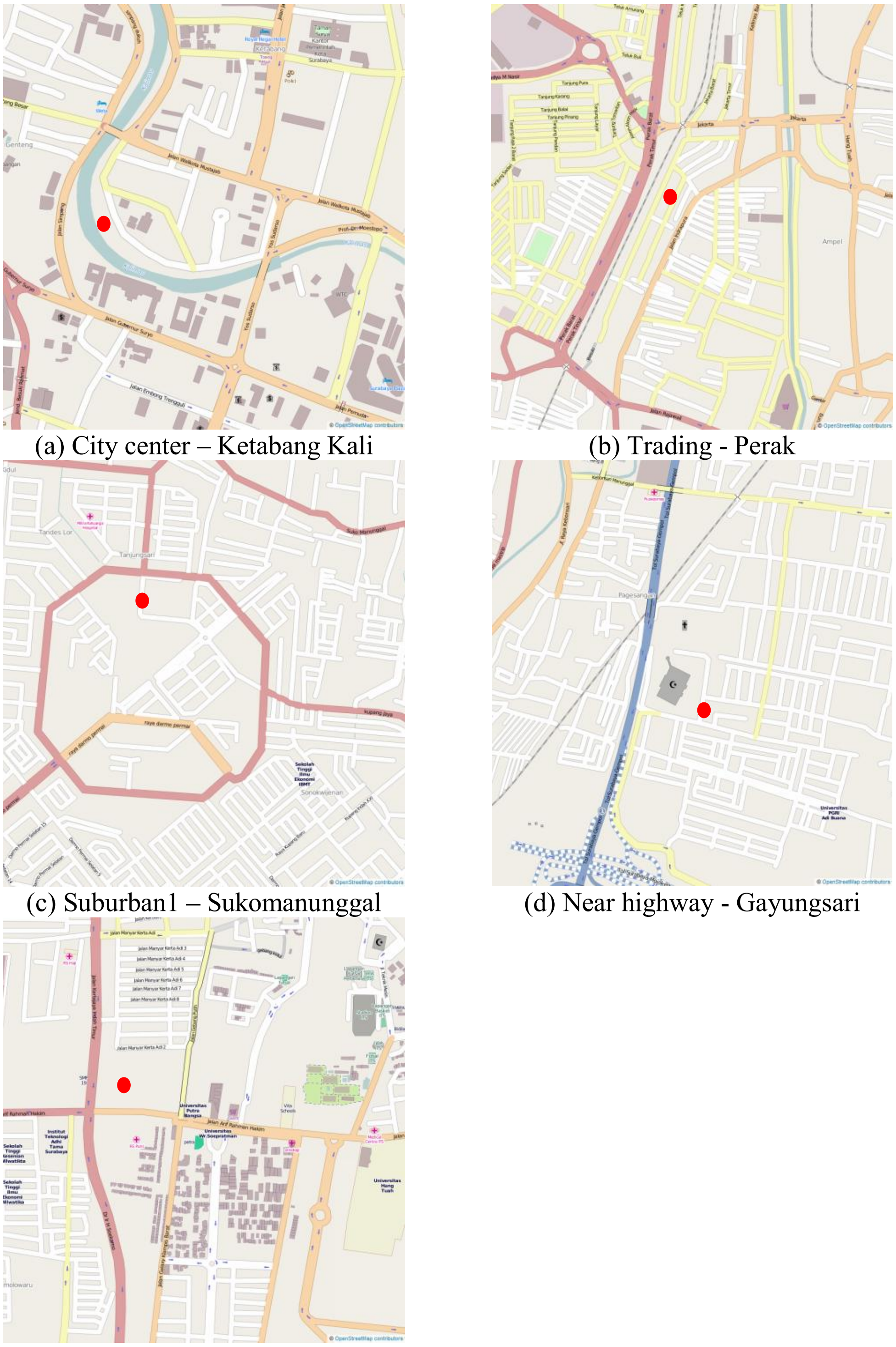
It is therefore the goal of the present chapter consist of: 1) identifying the temporal peak time frame pattern of pollutants; 2) comparing and judging the effectiveness of the use of different time-scale and data aggregation to obtain seasonal variation and temporal patterns; 3) confirmation of temporal time frame during which the pollutant's concentrations are high using a dynamic linear model; 4) determining the dynamics of pollutants and the effects of meteorological factors; 5) determining subpatterns of NO_2 as a way to assess monitoring efficiency. Number 1 and 2 are carried out using ICA, whereas number 3 to 5 are accomplished using a state-space model. Moreover, the discussion of number 3 to 4 is limited to trading zone only, as we observe the highest possibility of policy implementation on this zone within short time.

5.2 Data and monitoring stations

In addition to information provided in Chapter 3, Table 5.1 shows the characteristics of zones where each monitoring stations locate. Figure 5.1 shows the situation around monitoring stations.

Table 5.1 Monitoring stations and their coordinates

Zone (Station)	Land use	Population size (ca)	Topography	Average wind speed (m/s)	Dominant Wind direction
City center (Ketabang Kali)	Green space, parks, hotels, offices, houses, businesses	64,400-65,296	Low land	1.850	East
Trading (Perak)	Green spaces, houses, offices, industries, gas station	89,348-90,418	Low land, near coastal area	2.161	Southeast to east
Suburban (Sukomanunggal)	Offices, few factories, houses, shopping malls	85,012-86,421	Hilly land	3.373	East
Near highway (Gayungsari)	Religious building, businesses, houses	83,644-85,012	Low land	2.181	Between east and southeast
Suburban (Sukolilo)	Convention halls, offices, universities, offices, houses	75,021-77,362	Low land	4.144	East and southeast



(a) City center – Ketabang Kali

(b) Trading - Perak

(c) Suburban1 – Sukomanunggal

(d) Near highway - Gayungsari

(e) Suburban2 – Sukolilo

Figure 5.1 Road map of the surrounding of monitoring stations. Background image is made available from OpenStreetMap (© OpenStreetMap contributors) available under the Open Database License. Red dot is the location of the sensor of monitoring station

We employ 30-minute (min) intervals of NO_2, PM_{10}, O_3, NO, CO and SO_2 concentrations data from 1 February 2001 to 30 September 2002. These data are in the form of 30-min interval data. All missing data were imputed using EM algorithm (Honaker *et al.*, 2011) run through an R program using the *Amelia* package. We averaged the data to form weekly diurnal information, which is 336 time steps (48 time steps each day multiplied by 7 days = 336 times). Therefore, time *t* herein refers to (t_1, t_2…, t_{336}). The 336 values are derived by averaging a pollutant's concentrations over 607 days in the same time position, e.g., we averaged all values for the first 30 mins followed by the next time plot. To obtain more results and compare the effectiveness of the time scale, we also tried using 30-min data but with a different time scale, that is, a 4-week-cycle data. In this way, instead of 336 time steps, we extended it to 1344 time steps. We also took a look towards a longer time scale using longer aggregation. We aggregated 30-mins of data into daily data for 2 years and ran ICA. We would like to see seasonal pattern and seasonal effects on the concentrations. However, we also wanted to see the effectiveness of using these different time scales and aggregation levels. The definition of effectiveness is when a particular result (from different time scales and aggregations) has given us the necessary answer about the temporal time frame pattern, whether it is daily or not.

Data for ICA is original, whereas, for state-space models, all concentrations were square-root transformed to reduce data long tails. We then focused on the use of the result using NO_2 and PM_{10} data for sources estimation. Besides air quality, we used meteorological information, that is, wind direction and wind speed (m/s) to obtain from which direction the wind blew, when the concentrations exceeded the threshold value. We imputed any missing data using Expectation-Maximization algorithm, as described in Chapter 3.

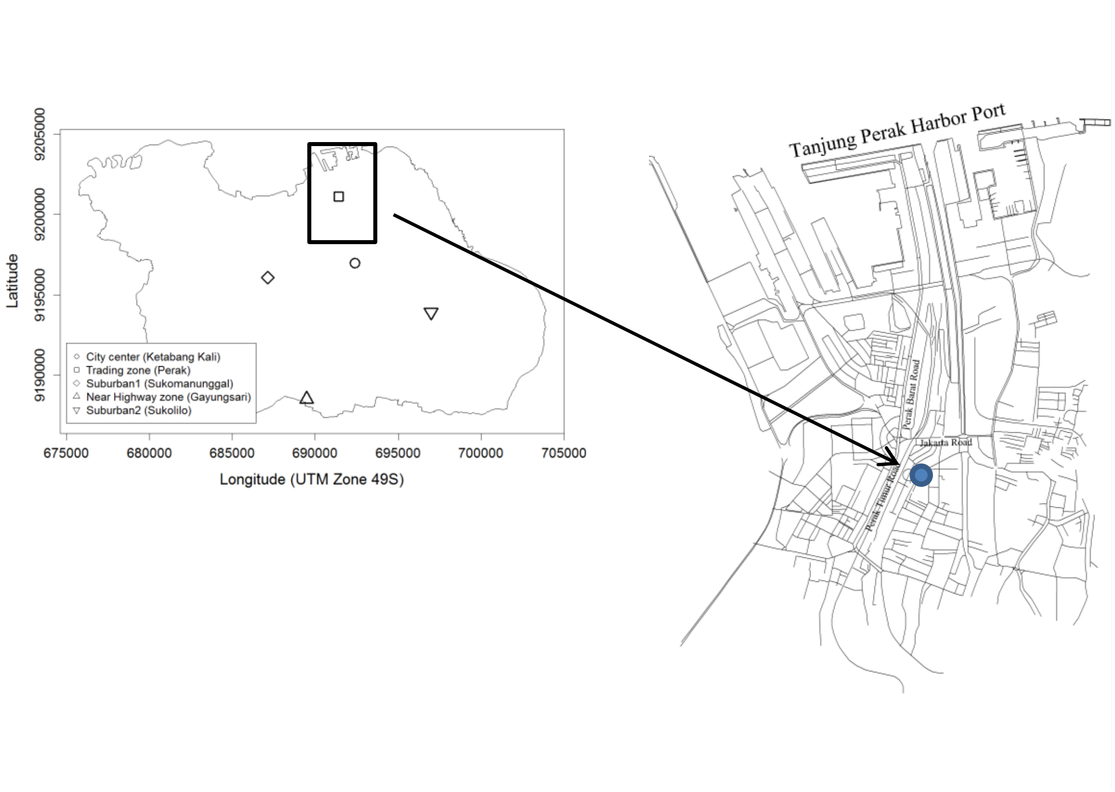


Figure 5.2 Five monitoring stations in Surabaya City. The map in the trading zone is a modified image from an original image that was made available from OpenStreetMap (© OpenStreetMap contributors) available under the Open Database License. The blue dot is the location of the sensor of the monitoring station in the trading zone (Perak).

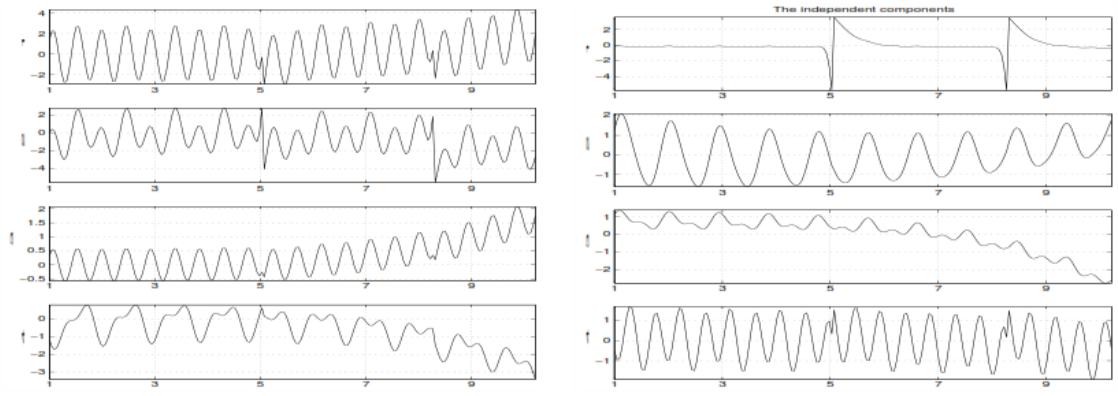
To confirm the temporal time frame, we applied a state-space model, specifically to the trading zone data. The selection of this zone is related to its surroundings. Near the zone is a large road, suspected of being responsible for air pollution. This road connects the city to the harbor port, and, therefore, the majority of vehicles are from companies that utilize large vehicles. Since this road is dominated by trucking company vehicles, we can apply traffic-adjustment by limiting vehicle distribution specifically from companies. This approach seems to be more plausible than other zones that require a more complex approach and longer steps. Therefore, an additional goal of the use state-space model was to decide day variation by observing the dynamics of a temporal time frame. We would like to determine if policy is going to be implemented if it will be for every day or for certain days, e.g., working days only. Figure 5.2 shows the monitoring station on the trading zone.

In the second model, we square-root transformed all pollutants, which allowed us to meet our assumption of normally distributed observation errors. We also standardized the meteorological variables, such that the independent variables or covariates have an empirical mean of 0 and a variance of 1.

5.3 Methodology

5.3.1 ICA

ICA can be considered an advanced method of PCA. The latter focuses on identifying components based on covariance and second-order statistics, while ICA uses higher-order statistics, which allow the algorithm to find components that are statistically independent (Westra *et al.*, 2009). The basic algorithm of ICA is explained by Figure 5.3. The mixtures (Figure 5.3a) are the original signals that wish to be decomposed. Figure 5.3b shows after decomposition by ICA, there four components resulted.



(a) (b)

Figure 5.3 The mixtures of signals or raw data (a) and four independent components (b) after performing ICA, this example is taken from Malaroiu *et al.* (2000)

The form of ICA consists of observation matrix, X, which is derived through the mixing of an *n*-dimensional source matrix, S = $(s_1, \ldots, s_n)^T$, with a temporal dimension of *t* for every component $s_1, \ldots, s_n$ referred to ICs, with *n* independent components extracted. Assuming that the mixing is both linear and stationary, a typical ICA model is expressed as:

$$X = SA \tag{1}$$

where A is the mixing matrix of dimension *n* x *n* or *n* x *m*, where $m \leq n$. The objective of ICA is to estimate A and S, knowing only the observations matrix X. This is achieved up to some scalar multiple of S, since any constant multiplying an independent component in equation can be cancelled by dividing the corresponding column of the mixing matrix A by the same constant.

The independent components S in the model are found by searching matrix W, such that $S = WX$ up to some indeterminacies. The *FastICA* algorithm is used for independent component analysis, using the iterative fixed-point algorithm for finding one unit, which is:

$$\widetilde{w}_{n+1} = E\{x(w_n x) * g(|w_n x|^2)\} - E\{g(|w_n x|^2) + |w_n x|^2 g'(|w_n x|^2)\} w_n \tag{2}$$

where $w_{n+1} = \frac{\widetilde{w}_{n+1}}{\|\widetilde{w}_{n+1}\|}$

Getting the estimate of w, we can obtain an IC by $S = WX$. As part of the process is decorrelation of outputs w_1x, … w_nx after every iteration. Using *FastICA* algorithm, we can estimate A and S from observations X, where $A = W^1$.

One note about ICA is that it is sensitive to the choice of n. It is a key parameter that must be determined prior to using ICA for interpretive applications. The process of finding components must undergo preprocessing stages, which are centering and whitening. Details are explained elsewhere (Hyvärinen and Oja, 2000). The process of parameter estimation is using maximum likelihood estimation.

5.3.2 State-space Model

A state-space model is used to model and estimate different temporal patterns and to estimate parameters, as well as to describe the process and observation variability. The temporal variability of a pollutant's concentrations are represented by the process variability in the model. This variability is due to environmental stochasticity. Observational variation encompasses sampling error, measurement error, and other variability caused by meteorological factors that are forming complex atmospheric reactions. Sampling error occurs from only a portion of a population being sampled, while measurement error of concentrations may be due to inaccurate measurement (e.g., sensor sensitivity, sensor disorder, misspecification of instruments).

State-space models have been used for many time series applications because of its ability to separate out two sources of variation, without the need for prior estimates of observation variance (Ward *et al.*, 2010). In the multivariate time-series studies, the use of a state-space model for air quality studies has seen limited application. One example we found was the use of dynamic factor analysis to identify factors influencing $PM_{2.5}$ in Southern Taiwan by Kuo *et al.* (2011). Some applications of a state-space model were on the subject of ecological studies (Ward *et al.*, 2010). The majority of air quality studies had focused on non-state-space time series analysis, e.g., time series with and without exogenous variables (Aneiros-Pérez *et al.*, 2004; Pollice and Lasinio, 2010), and univariate time series. The Multivariate Autoregressive State-Space (MARSS) here is used to establish a multi-dimensional setting (Holmes, Ward, and Wills, 2012).

In the MARSS framework, we follow the following formula for population or state process:

$$X_t = BX_{t-1} + u + \eta_t \qquad (3)$$

Whereas the observation process in the model is given below:

$$Y_t = a + ZX_t + \varepsilon_t \qquad (4)$$

In the model, we use n to represent the number of monitoring sites, and m to represent latent, unknown states or substructure of the process. Xt denotes the vector of length m representing the fluctuation of NO_2 concentrations, and η_t is also a vector with length m representing the process error in time t. We assume the process errors are uncorrelated in time, with a multivariate normal distribution with mean zero and variance-covariance matrix **Q**. When **Q** is a diagonal matrix, the trajectories of substructures are independent. Several concentration subpatterns may be correlated. Therefore, we set off-diagonal elements of **Q** to be non-zero in the models.

y_t is the n-element vector of NO_2 concentrations at the n monitoring sites at the time t. We also used the n-element vector $\boldsymbol{a}$ to display mean bias between monitoring sites. ε_t is a n-element vector that represents observation errors, which we also assumed to be serially uncorrelated in time, with distribution multivariate normal distribution, a mean of 0, and variance-covariance matrix as **R**.

To obtain information about temporal variation, we employed a dynamic linear model (DLM), where the parameters in the regression model are treated as time-varying. In this model, the regression parameters are dynamic in that they "change" over time. For a single observation, the formula is:

$$y_t = F_t^T \theta_t + v_t \qquad (5)$$

where $\mathbf{F}_t$ is a column vector of regression variables at time *t*. The regression variables in the present study are wind direction, wind speed, solar radiation, humidity, and temperature. θ_t is a column vector of regression parameters at time *t* and $v_t \sim N(0,r)$. The equation expresses the uniqueness of the predictor variables at every time *t* (i.e., $\theta = \{\theta_1, \theta_2, \theta_3, \dots, \theta_T\}$). To address the shortcoming of the above model as presented by Holmes *et al.* (2014), the following model was formulated:

$$\theta_t = G_t \theta_{t-1} + W_t \qquad (6)$$

$\mathbf{G}_t$ is the parameter matrix, and $\mathbf{w}_t$ is a vector of process errors, in which $\mathbf{w}_t \sim \text{MVN}(\mathbf{0},\mathbf{Q})$. The elements of $\mathbf{G}_t$ may be known and fixed *a priori*, or estimated from the data, or unknown. Although $\mathbf{G}_t$ is time-varying, we typically assume that it is time invariant. In the present study, for the square-root pollutant's concentrations in time *t*:

$$y_t = \alpha_t + \beta_{1t}F_{1t} + \cdots + \beta_{5t}F_{5t} + v_t {\sim} N(0,r) \qquad (7)$$

with F_t consisting of $(F_{1t,} \dots, F_{5t})$, representing wind direction, wind speed, solar radiation, humidity, and temperatures, respectively, for one full week in time *t*. Both the intercept and slope are time-varying:

$$\begin{aligned}
&\alpha_t = \alpha_{t-1} + w_t^{(1)}, w_t^{(1)} {\sim} N(0,q_1) \\
&\beta_{1t} = \beta_{1(t-1)} + w_t^{(2)}, w_t^{(2)} {\sim} N(0,q_2) \qquad (8) \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \\
&\beta_{5t} = \beta_{5(t-1)} + w_t^{(6)}, w_t^{(6)} {\sim} N(0,q_6)
\end{aligned}$$

If we define $\theta_t = (\alpha_t, \beta_{1t}, \dots, \beta_{5t})^T$, $G_t = I \forall t$, $\boldsymbol{w}_t = \left(w_t^{(1)}, \dots w_t^{(6)}\right)^T$, and $\mathbf{Q}$ = diag(q1,…q6), we will get equation (6). By defining y_t = pollutant's concentrations, and $\mathbf{F}_t = (1, F_{1t}, \dots, F_{5t})^T$, we can write out the state-space model as follows:

$$\begin{aligned}
&\theta_t = \boldsymbol{G}_t \theta_{t-1} + w_t \text{with } w_t \sim \text{MVN}(\mathbf{0},\mathbf{Q}); \\
&y_t = \boldsymbol{F}_t^T \theta_t + v_t \text{ with } v_t \sim \text{N}(0,r); \qquad (9) \\
&\theta_0 {\sim} \text{ MVN}(\pi_0, \Lambda_0).
\end{aligned}$$

The above equation is equivalent to the above formula (5) and (6), and, therefore, we can fit with the MARSS modeling framework set up.

5.4 Pollutants Distribution using ICA and Emission Source Estimation

During the period specified in the monitoring data, we identified events when the concentrations exceeded the environmental criteria. We first developed a hypothesis to help us make decisions or for making any data-driven verdicts.

5.4.1 Hypothesis of temporal patterns for five locations

To determine temporal pattern of pollutants, we can observe each weekly diurnal pattern and compare these between locations. Therefore, we establish hypotheses of the pollutant pattern for each zone because each zone may represent a unique profile. In the city center, since it is located in the midpoint of the city, surrounded not only by houses but also with government office activities, we can hypothesize that the pattern in this zone is related to working and school activities, indicated by peak hours in both morning and evening periods. It is logical that in this zone we mark two daily cycle peaks, during the week, with the possibility of lower value activity during weekends or on Sundays.

In the trading zone, we assume that the pattern is different from the city center because the majority of land use there involves fewer urban industries, business offices, and it is near the harbor, which is the second largest port harbor in Indonesia. Therefore, our hypotheses would be that there are several short intervals during the day that show peak concentration, indicating the movement of large vehicle volume that may be concentrated during those time intervals. The dominant emission source is estimated from a road that is heading toward the port harbor. In Suburban1, west of Surabaya, we assume that the pattern will be similar with Suburban2 due to similar locations. However, the monitoring station in Suburban1 is surrounded by highly dense residential dwellings. The location in Suburban2 is surrounded not only by houses but also by universities, business-related buildings, and offices. In these two regions, we hypothesize there is wider interval of high concentration of pollutants, especially NO_2 which is related to traffic volume.

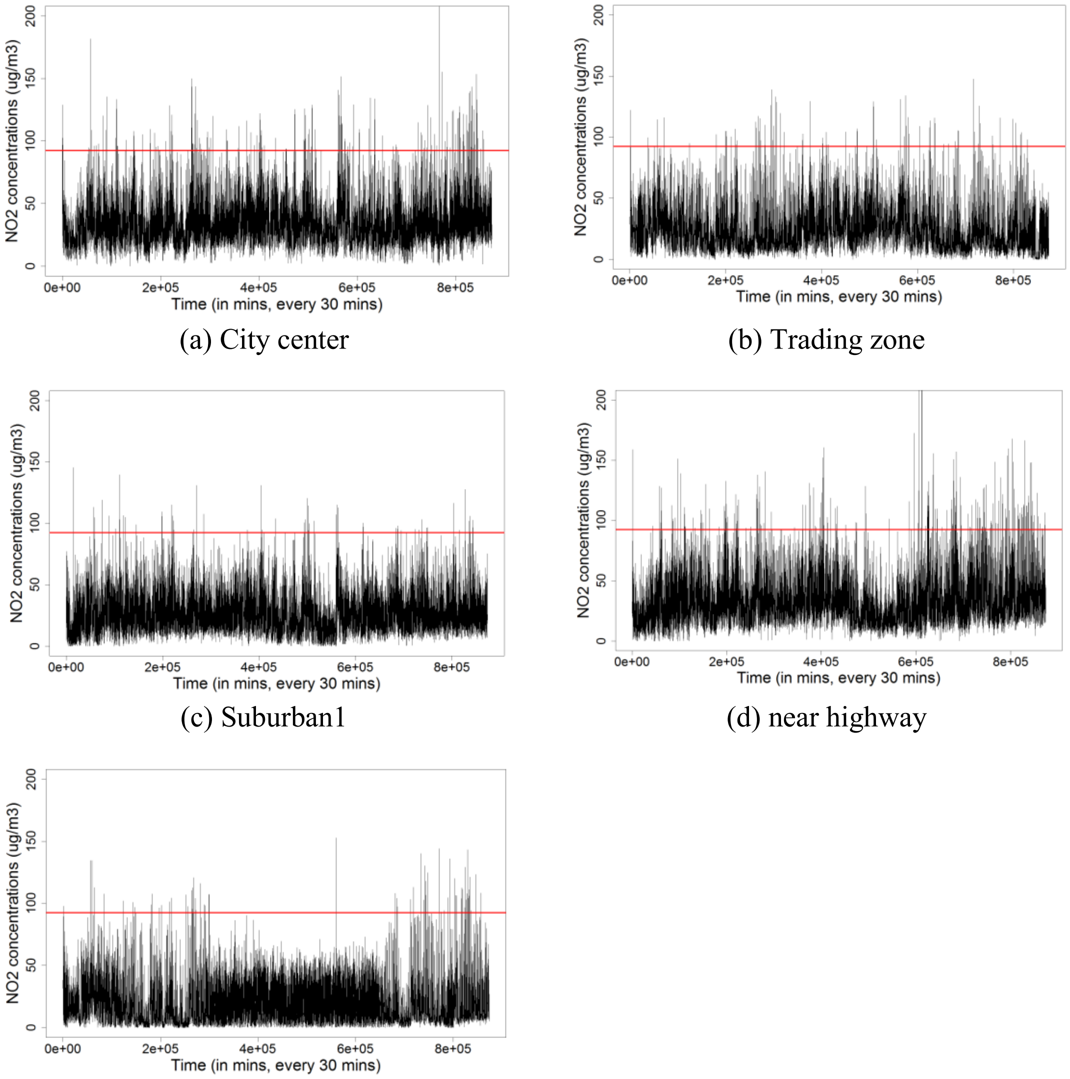
Lastly, in the zone near the highway, we assume an earlier temporal peak time frame in the morning and late temporal peak time during the evening. Our hypothesis is based on a situation where the highway connects Surabaya to other neighboring cities. Therefore, when people go for work, the tendency would be for them to depart earlier in the morning and return later in the evening. Based on our description above, our preliminary conclusion is that these zones may have unique pollutant characteristics based on the land use that each monitoring stations represents. In the next section, we will analyze the course of pollutants when they breach the permissible level.

5.4.2 Pollutions Exceeding Permissible Level (Ambient Criteria)

In this section, we will focus on displaying events where a pollutant's concentration has exceeded the permissible level during the prescribed duration (2001 to 2002). As shown in Figure 5.4, there were many times when NO_2 concentrations exceeded the ambient criterion (permissible level), which is 92.5 ug/m^3 based on air pollution regulations in East Java. These concentrations were drawn from five monitoring stations in Surabaya City. Furthermore, we observed the times when PM_{10} concentration has exceeded the threshold ambient value of 260 ug/m^3 as shown in Figure 5.5. The exceeding values of PM_{10} may came from a mixture from various sources. Figure 5.6, Figure 5.7, and Figure 5.8 show the exceeding values of O_3, SO_2, and CO concentrations, respectively. We observed that the over-limit events were rare for SO_2 and CO. Moreover, O_3 is not a result of direct emissions. Instead, it is formed through a series of reactions between NO and NO_2. The concentration of NO is represented by NO_2. Therefore, we did not focus on NO. Figure 5.4 to Figure 5.8 show the evidence of events where pollutants exceeded the regulatory threshold levels.

Since our objective is to identify the temporal pattern of these events, we begin by utilizing the descriptive statistics from data by averaging two years' of value forming one week of data. We can observe some patterns for all pollutants. For example, with NO_2, we observe the frequency of events was high between 8am to 10am for all zones except the trading zone (Figure 5.9a). On the other hand, the time frame of events during the evening session was observed between 5pm to 9pm at the highway and the city center, whereas, an unclear pattern was found for the remaining zones. The patterns of other pollutants are displayed on Figure 5.9b-e.

By observing these figures, the information that can be derived is quite limited. First of all, if we take $NO_{2,}$ for example, the temporal pattern is similar for four zones, except for the trading zone. This result violates our hypothesis and subject to further investigation. Furthermore, we also found a similar PM_{10} pattern for all zones. It is therefore difficult to obtain a distinct temporal pattern for each location. We also experienced difficulty in obtaining temporal patterns for NO, CO, and SO_2 because the data is limited. The events for those parameters exceeding the threshold limit were few. Therefore, little information can be extracted from Figures 5.9c, d and e. In the next figure, we attempt to answer the following question, that is, what was the dominant day when these events occured?



(a) City center

(b) Trading zone

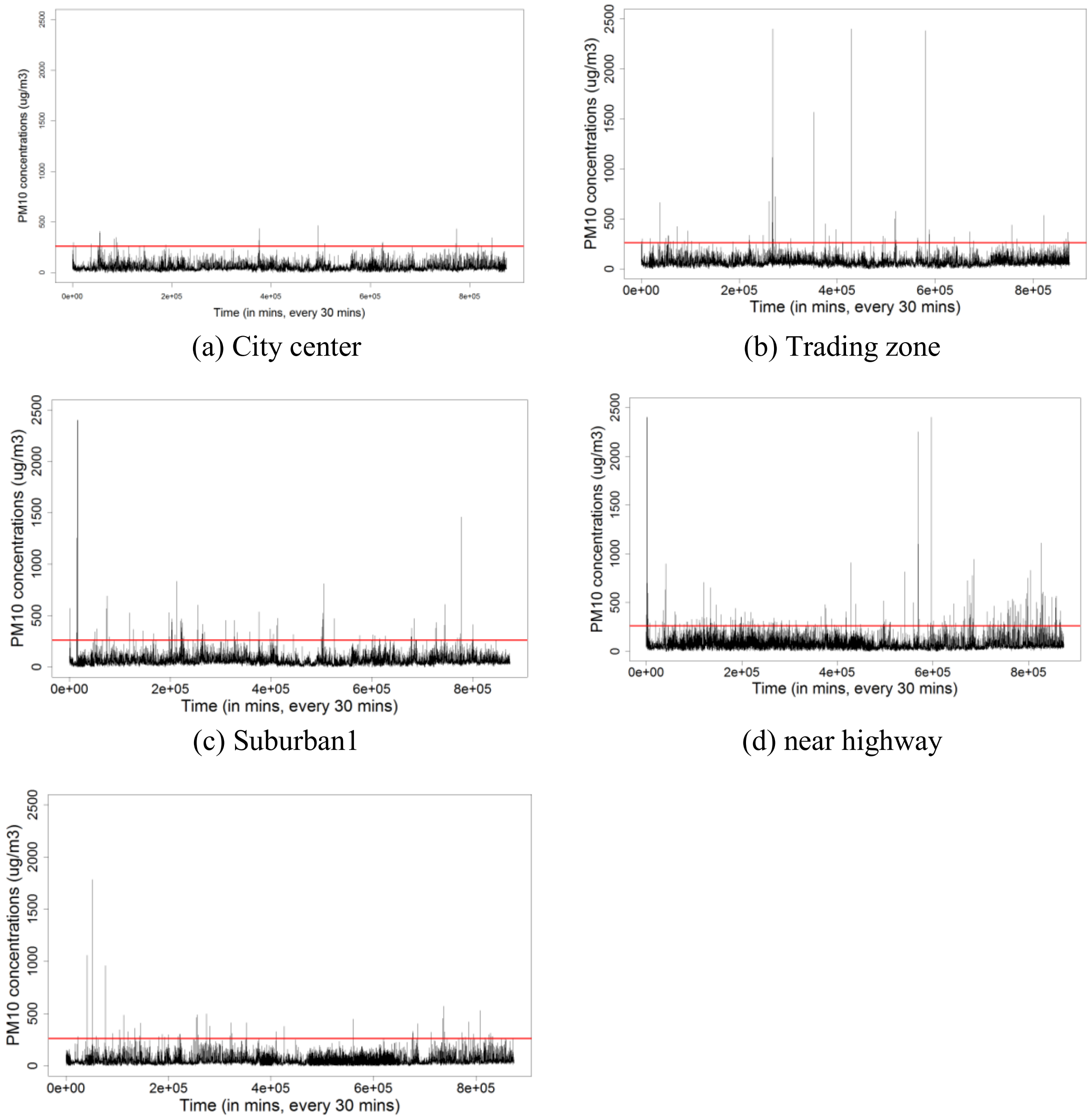
(c) Suburban1

(d) near highway

(e) Suburban2

Figure 5.4 30-mins interval NO_2 concentrations from 2001-2002 at five monitoring stations

In this section, we also want to determine on which day events largely occured. This is because if we want to propose regulation or policy, we need to know whether we should implement the policy for every day or just for certain days. Figure 5.10 shows the frequency of the events on particular days of a week. We can observe that for NO_2, the most frequent event took place on Wednesday, followed by Thursday. It also shows lower risks over the weekends. On the other hand, the events took place more frequently on Sunday for PM_{10}, suggesting activities other than transportation contributing to emissions. The other figures are self-explanatory. We should note that these figures are only able to explain a small part of temporal activities, as they can only show the days when the events happened over a two-year period.



(a) City center

(b) Trading zone

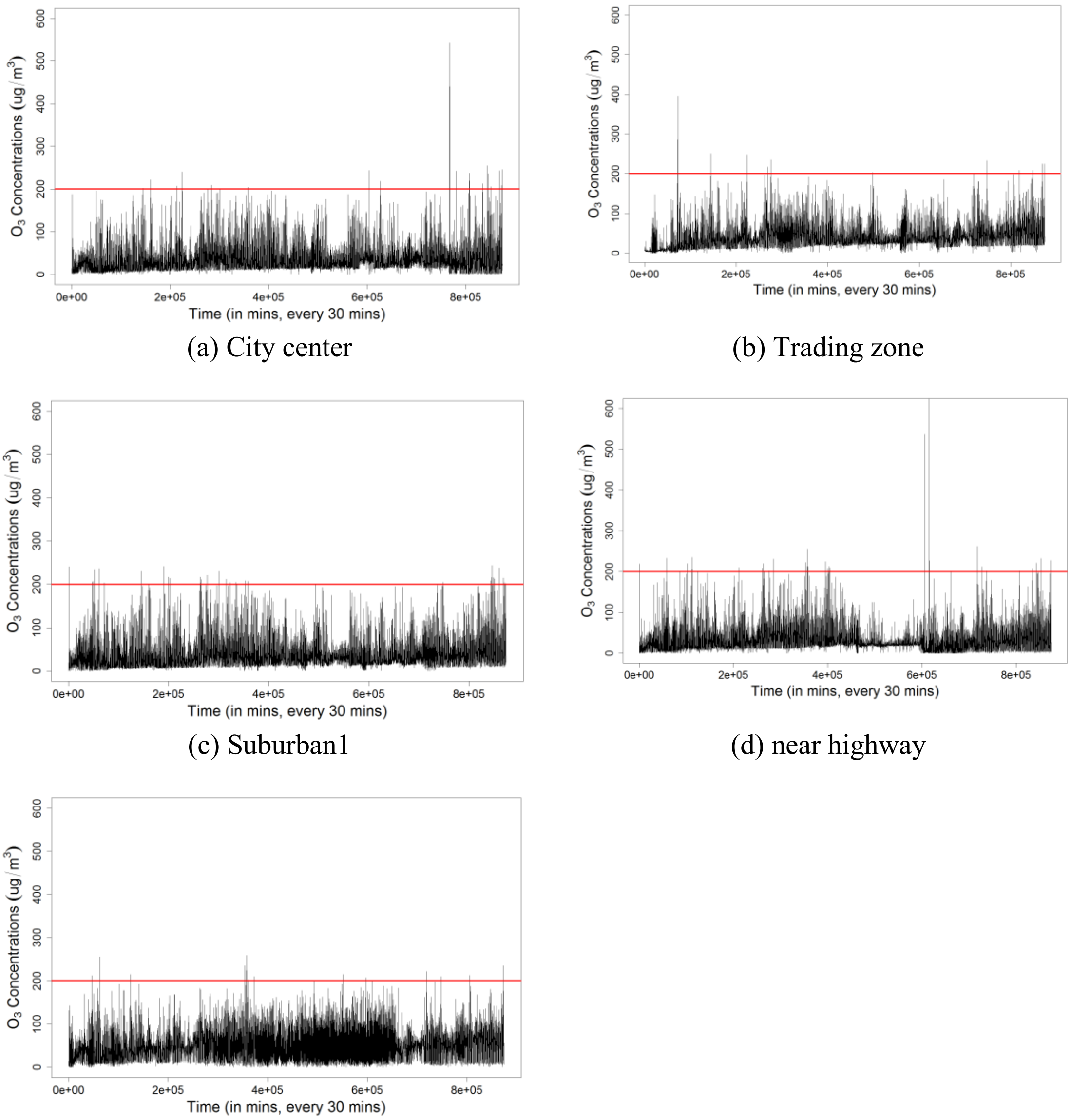
(c) Suburban1

(d) near highway

(e) Suburban2

Figure 5.5 30-min interval PM_{10} concentrations from 2001-2002 at five monitoring stations

Figure 5.10 has been of value in determining specific days when events occured. We can see that events happened every day for both NO_2 and PM_{10}. In some zones, we observed low frequency, e.g., in the trading zone on Sunday (NO_2), and in the city center on Monday for PM_{10}. We observed similar cases with other parameters and zones. However, we noted that in the working days data, the events were lower than on weekends based on PM_{10} for all locations. Therefore, this may violate commonly held beliefs that weekends usually generate lower concentrations. Even though the activities contributing to PM_{10} may be higher during weekends, the difference between weekends and working days looked quite clear. Therefore, further investigation is required.



(a) City center

(b) Trading zone

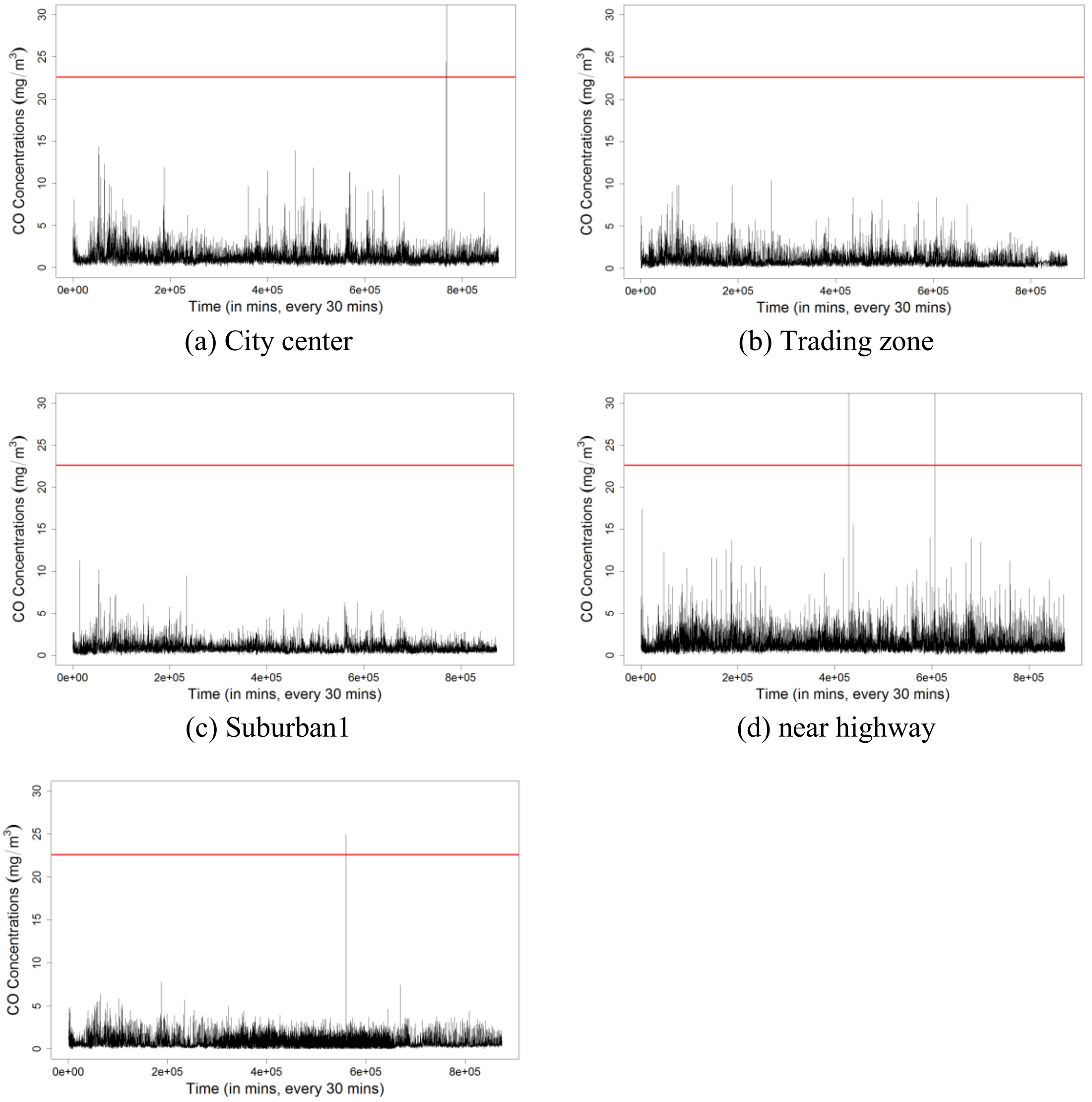
(c) Suburban1

(d) near highway

(e) Suburban2

Figure 5.6 30-min interval of O_3 concentrations from 2001-2002 on five monitoring stations

For other parameters, such as O_3 and CO, we noted that there were several days where the frequency was zero. For example, for O_3, we observed that in the trading zone on Thursday, Wednesday, and Sunday, the frequency was very low or almost zero. This fact is interesting because during working days, the contribution of NO_2 and NO is higher. Therefore, it seems logical that after a photochemical reaction, this should lead to increased O_3 formation. This perplexing result also indicates the need for further investigation.



(a) City center

(b) Trading zone

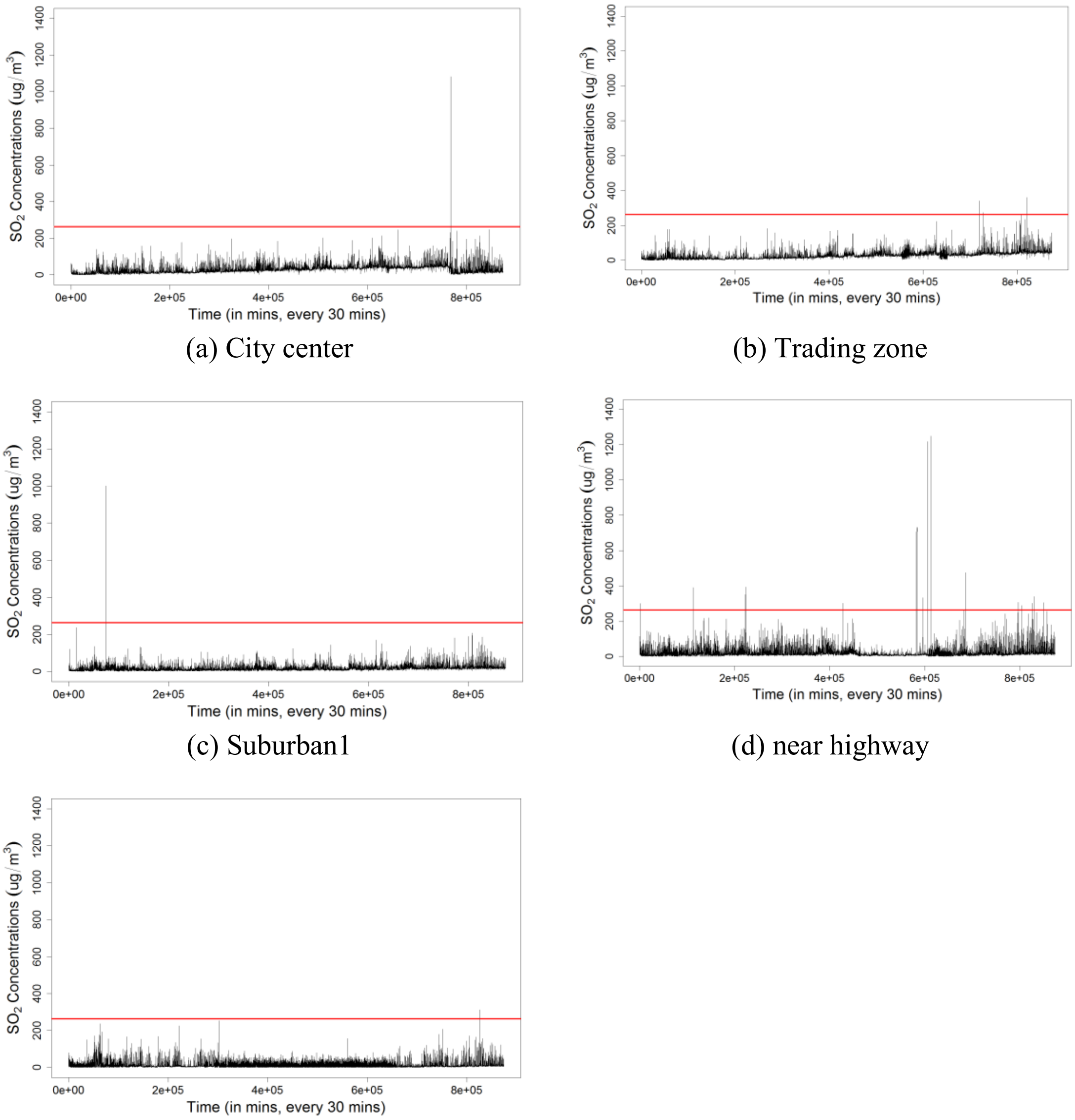
(c) Suburban1

(d) near highway

(e) Suburban2

Figure 5.7 30-min intervals of CO concentrations from 2001-2002 on five monitoring stations

If these results are going to be used to formulate any form of policy, they will be inadequate for the following reason. We know that for O_3, NO, and SO_2, there are several variables where the frequency was almost zero. If we should put forth a specific policy, we need to know whether the policy should be implemented every day or not, or whether the policy should only be implemented during working days only or not. Then the levels of O_3, NO_2 and SO_2 will not be helpful because zero frequency may bring about a decision that a policy be implemented for a particular day, when it may not be necessary. For example, NO, where we noted in Suburban2, its frequency was almost zero for most days (Figure 5.10d). If a policy for Suburban2 will be issued, a decision would be that the policy will only be implemented on Monday and Sunday because the frequency in these days are higher than other days, whereas, in fact, they may say different. This may not represent the true data because although the frequency was lower on these days, people will actually be exposed to a degree of risk on other days.



(a) City center

(b) Trading zone

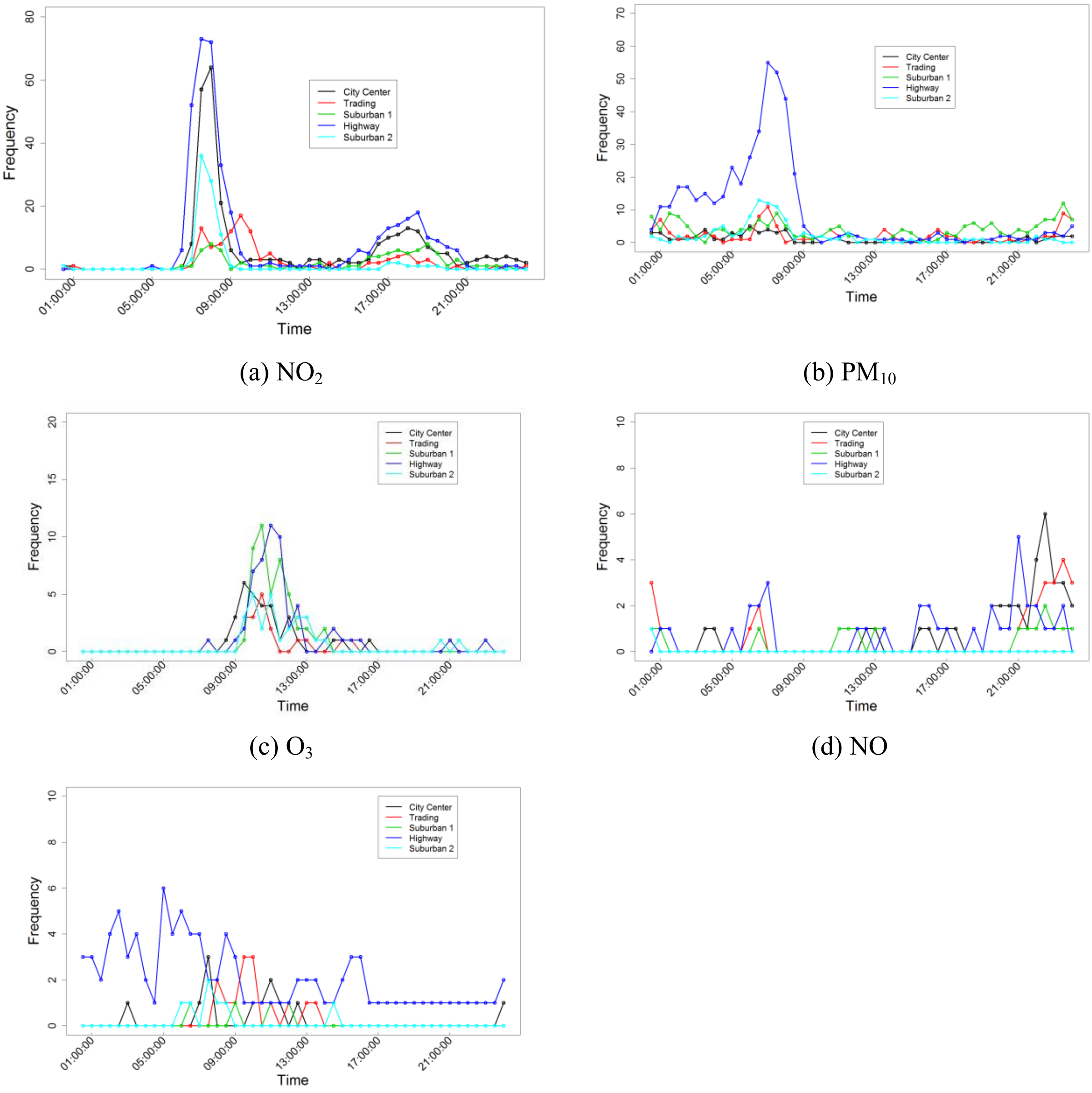
(c) Suburban1

(d) near highway

(e) Suburban2

Figure 5.8 30-min intervals of SO$_2$ concentrations from 2001-2002 on five monitoring stations

In terms of NO$_2$ and PM$_{10}$, since there are many events on all days, we have no problem in judging that a policy can be implemented for all days because the frequency appears on all days, but for parameters where we don't have the events, we may conclude the policy may not be implemented on that day, and this may be misleading the true fact. Therefore this kind of analysis is limited and we shall require another method to investigate seasonal variation daily into more details.



(a) NO_2

(b) PM_{10}

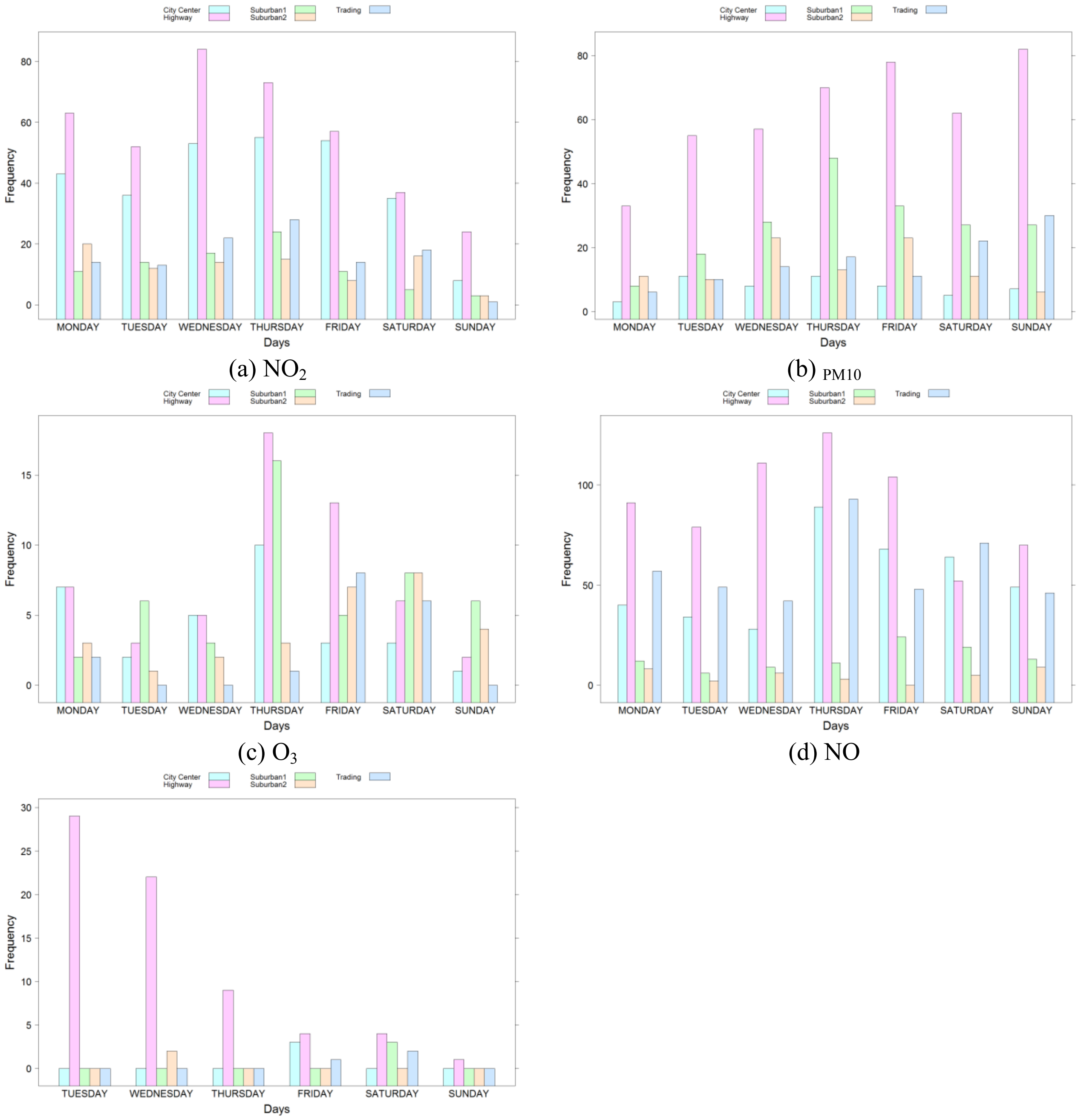
(c) O_3

(d) NO

(f) SO_2

Figure 5.9 Frequency of pollutants exceeding threshold value over two years of air quality on five monitoring stations

Both Figure 5.9 and Figure 5.10 help us understand the phenomena of pollutants. However, we have noted several things that need further investigation. First, Figure 5.9 violates our assumption of the pollutants' patterns on all locations. The pattern on the figure clearly shows similarity, whereas, based on our hypothesis, each location should have its own unique pattern. Secondly, although Figure 5.10 shows that almost every day people are exposed to the risk of health problems due to the frequency of events for both NO_2 and PM_{10}, this is because there are facts to show when the events occur every day. Thus, we may conclude that there is a need to be mindful every day of these pollutants. Combining this with Figure 5.9, we can conclude that within a certain time frame, and on a daily basis, people were at risk. Therefore, if a policy is going to be implemented, it should be a daily policy. On the other hand, for O_3, NO, and SO_2, we found that there were days where the frequency is almost zero. If one concludes that on these days there was no risk, that assumption would be misleading and may prevent any action that will minimize health problem risk to people on the days where there are no frequencies observed. Based on these analyses, it is important and crucial to obtain true distributions for all locations which we are able to observe not only temporal pattern (time frame) but also days where people may have an increased risk of getting exposed to high concentration of pollutants.



(a) NO_2

(b) PM10

(c) O_3

(d) NO

(e) SO_2

Figure 5.10 Frequency between days of pollutants exceeding threshold value over two years of data on five monitoring stations

For this reason, we need to explore further and obtain a true distribution of pollutants so that we are able to extract an accurate temporal time frame, with days on which there is a significant probability the events are likely to occur. First, we will try using an ordinary approach by using weekly raw data (Section 5.4.3) and by averaging the data over two years on a daily scale (Section 5.4.3) and weekly scale (Section 5.4.4). Another approach is that we use the approach of the Independent Component Analysis to form statistically independent components, which will help us explore patterns more thoroughly (Section 5.4.5). Temporal variation of the events can be explored based on ICA. Furthermore, it is important to identify why, from what direction these events (i.e., high values of pollutants) occur in order to discuss possible policy that can aid in an improved air quality management system. However, based on the Figure 5.9, it is observed that only NO_2 and PM_{10} that have most frequent of pollution breaching the permissible levels. Therefore, in the policy suggestion steps, we would focus and highlight to policies related to these two pollutants. In this study, we will also try to explain and investigate the source of pollution and the timeframe in which the events take place using ICA. The rest of paper will further explore this question.

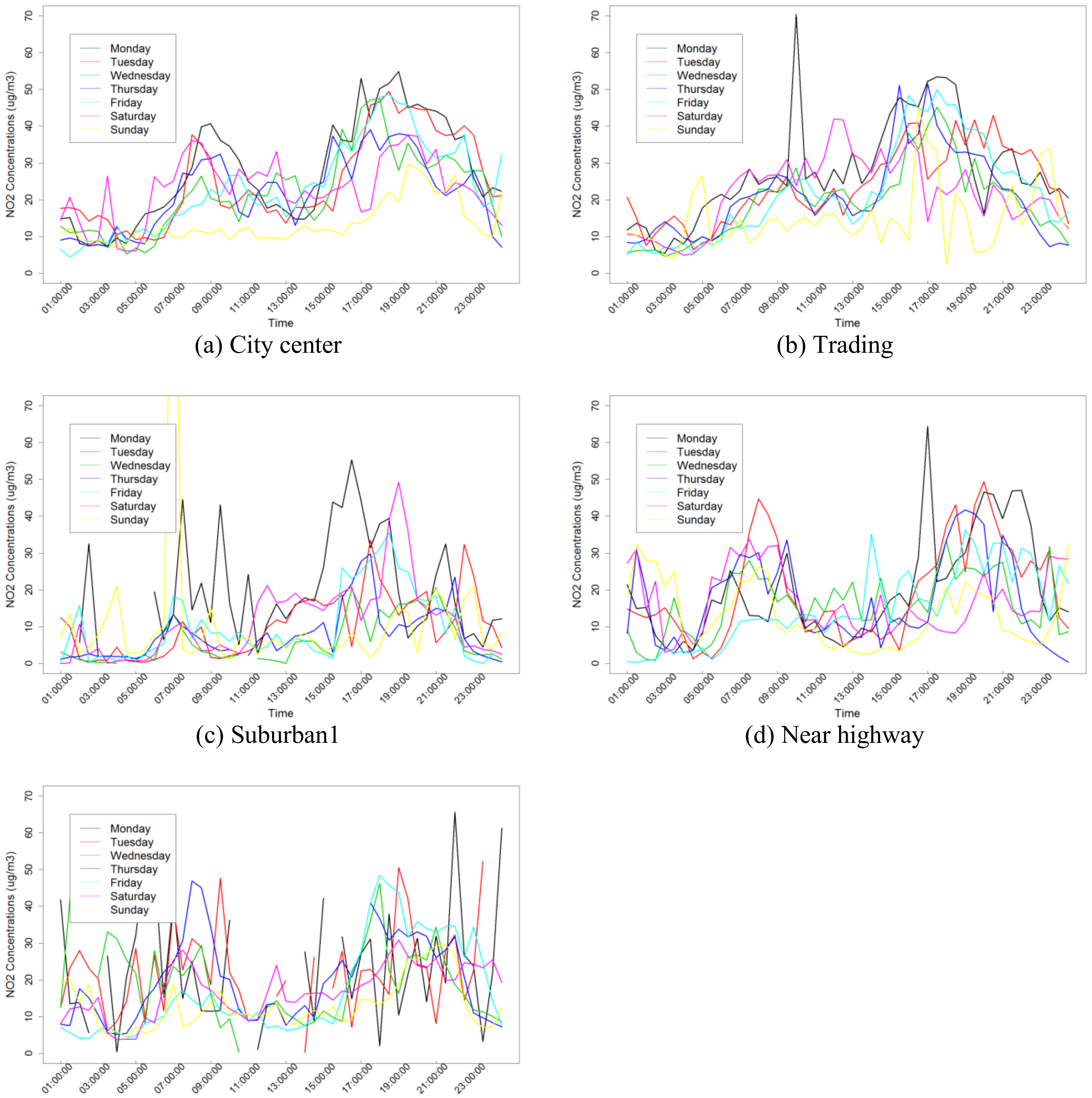
5.4.3 Diurnal variation of pollutant concentrations

In order to identify and determine the general diurnal variations of pollutants, we first tried to capture the diurnal variation using raw data. In this attempt, we used NO_2 and PM_{10} as an example. For NO_2 and PM_{10} concentrations, we will first observe temporal patterns taken on a certain week as shown in Figure 5.11 (NO_2) and Figure 5.12 (PM_{10}). We also analyzed the variation using weekly-scale aggregated data taken from Figure 4.1 in the Section 4.4.1, and, finally, we compared the analysis with daily aggregation data (Figure 5.13).

Based on Figure 5.11, we found it difficult to extract temporal peak time frame pattern clearly. However, to the best of our observation, we may see that in the morning, NO_2 concentrations rose between 6am to 9am in Ketabang Kali, except on Sunday, which is a holiday. During the evening session, we can observe that the concentration rose at around 3pm until about 10pm. This result fits with travel behavior of people within the city. In Perak, we didn't see an exceptional peak during the morning session, although we could see an increase after 3pm. However, these patterns over days were not clearly observed. In the Sukomanunggal region, the concentrations in the evening were relatively higher than in the morning, when the concentrations rose around 3pm until 10pm. In the Gayungsari site, near the highway, we can observe the pattern better than in the Sukolilo region. The timeframe during morning peak concentrations was between 6am to 10am while during evening session, the peak was between 5pm to 11pm. We noted that the NO_2 temporal patterns can be observed better than the profile of PM_{10} as shown on Figure 5.12. The temporal patterns highly fluctuated as a result of mixed sources of emission. Overall, the use of raw data to determine temporal peak time frame is not recommended.

We tried to capture the temporal time frame and the uniqueness of each pattern using a weekly-scale view aggregated from two years of data to establish weekly cycle distribution. Figure 4.1 in Chapter 4 shows weekly variation of six pollutants, and Section 4.4.1 describes the weekly diurnal cycle of all pollutants. However, we cannot distinguish temporal patterns between locations because the peak interval time, even though different in terms of magnitude (concentrations), is alike.

To assist in deciding general time frames, where both concentrations may reach their peaks exceeding the ambient criteria, we shortened the time scale. Figure 5.13 shows the daily diurnal variations of NO_2 and PM_{10} concentrations at five monitoring stations. The two-peaked distribution obtained is similar to those observed by Bigi and Harrison (2012) at a site in London, and Han (2011) in Tian Jin, China. In more detail, the NO_2 concentrations in our case started to rise around 6am and reached their peak at 7:30am, and then they declined. The time interval of peak concentrations for NO_2 was between 6am to 9am. These patterns may be explained by travel behavior because the time intervals fitted closely with the morning peak school traffic. Moreover, in the evening, the concentrations started to rise again around 4:30pm, which was in parallel with working hours. In terms of PM_{10} diurnal variations, as in Figure 5.13b, the morning concentrations of PM_{10} increased from 6am, similar to those of NO_2. They then reached their peaks at 7:30am and declined afterwards. The interval time of PM_{10} can be safely assumed to be between 6am to 9am, comparable with NO_2, suggesting similar patterns. For the evening session, the concentrations started to increase at 4pm, similar to NO_2 patterns, since 4pm was the end of working hours. Interestingly, unlike NO_2, the concentrations reached their peaks at around 5:30pm, and they became relatively stable, although there was a decrease observed at the Perak site (trading zone), when after the concentration reached their peak, slightly declined. In other zones, the concentrations after the peak were relatively stable.



(a) City center

(b) Trading

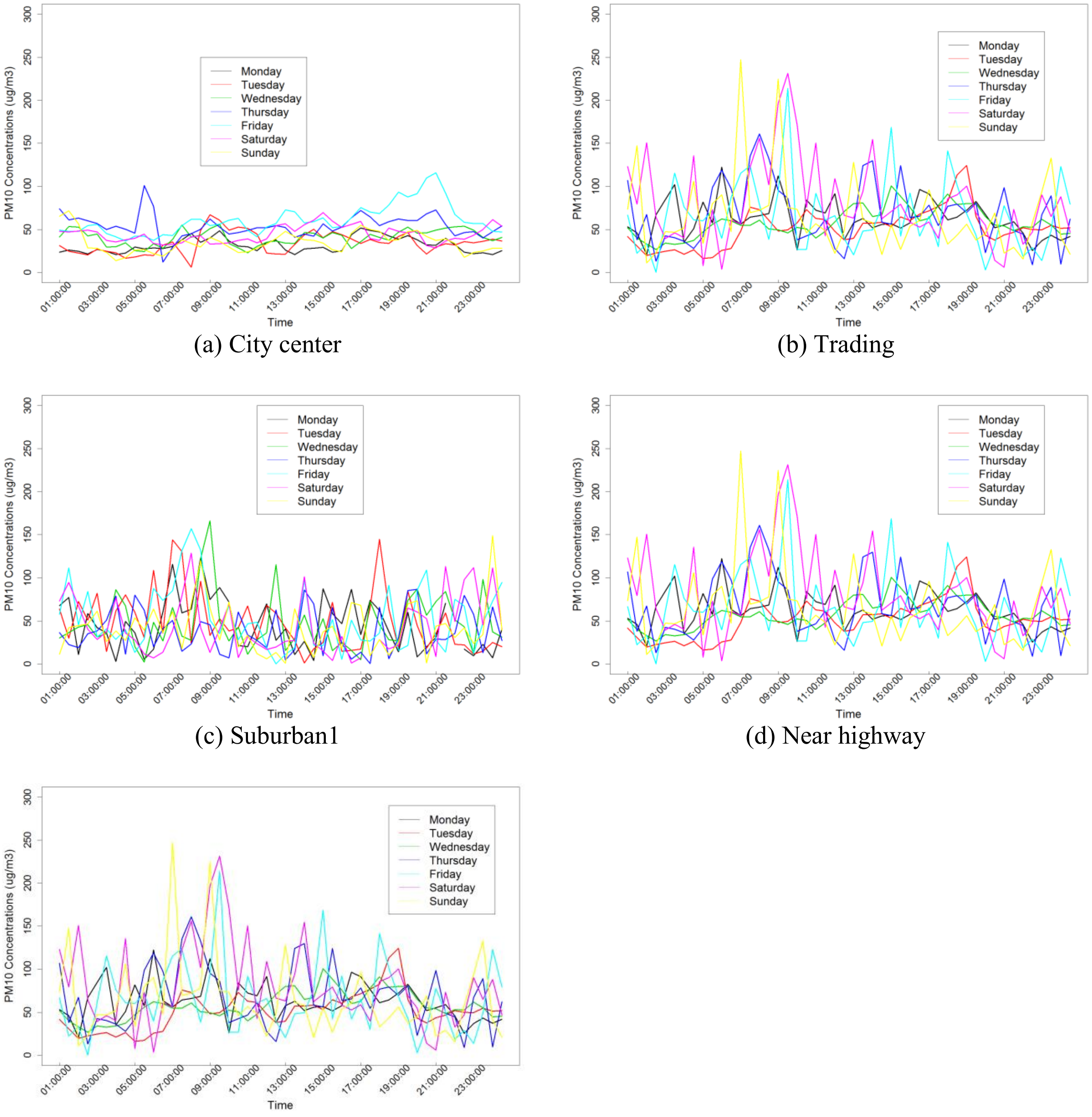
(c) Suburban1

(d) Near highway

(e) Suburban2

Figure 5.11 Diurnal concentration over a week from 5th February to 11th February 2001 of NO_2

There are several conclusions that can be drawn from the results above. First, information using raw data as shown in Figure 5.11 and 5.12 are complicated. It is crucial to investigate and explore the temporal patterns of the concentrations more clearly. We observed data noises, and that the temporal patterns cannot be observed clearly, in particular for PM_{10}, because of its fluctuation. The average diurnal concentration displays limited information (Figure 4.2 and Figure 5.13). The averaging process includes factors that may not represent the concentration clearly. As such, we can only draw conclusions of the different magnitudes between stations and timeframes of peak concentrations. The averaged value may distort the different patterns between stations. Therefore, we need a method that can capture the uniqueness of the temporal pattern of pollutants at each location, which is why ICA is useful to display such the desired patterns.



(a) City center

(b) Trading

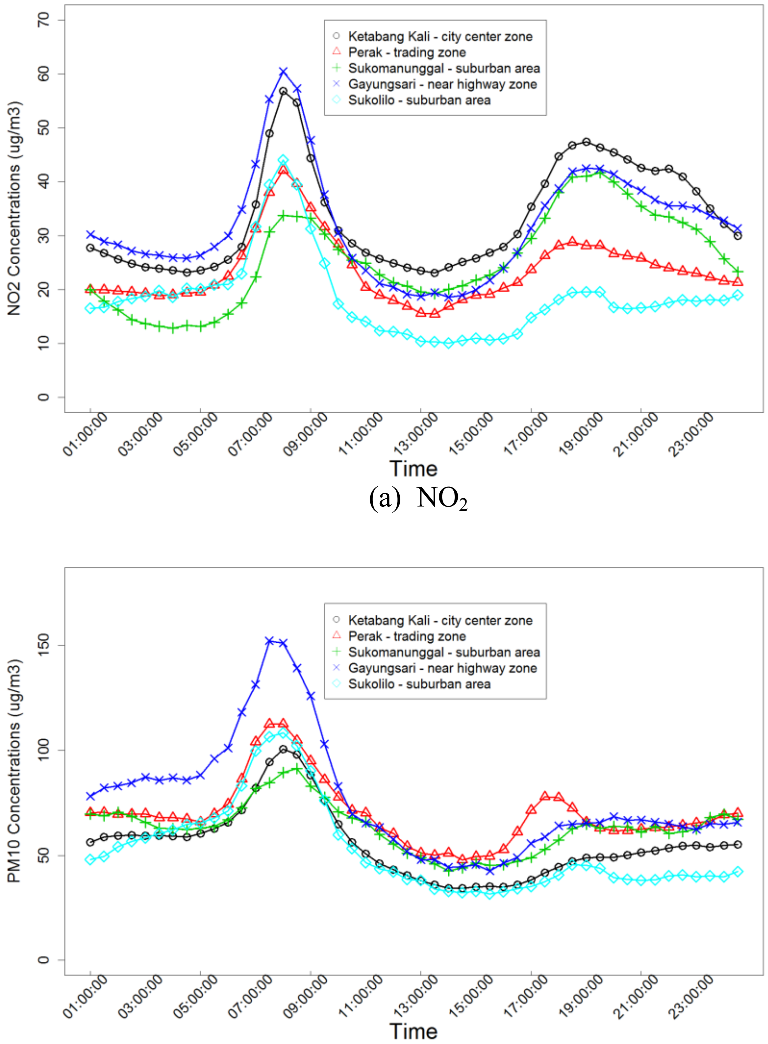
(c) Suburban1

(d) Near highway

e) Suburban2

Figure 5.12 Diurnal concentration over a week from 5th February to 11th February 2001 of PM_{10}

For this reason, we shall employ an ICA method for weekly average distributions in order to identify: a) the different temporal patterns between days (working days and weekends), and b) the profile of interval times between peak concentrations at each station. These are done for all pollutants. ICA is better than PCA in terms of the output produced. Hyvärinen and Oja (2000) showed that principal component analysis failed to produce clear structured output. as produced by ICA, because ICA performs with non-Gaussian distribution data. The non-Gaussian data have better characteristics for data projection. When using ICA, it must be noted that we assume that there is no spatial correlation of concentrations between stations. Thus, the output solely identifies the similarity in the temporal patterns. Combined with wind direction and speed profile, we observe the possible cause of high concentrations.



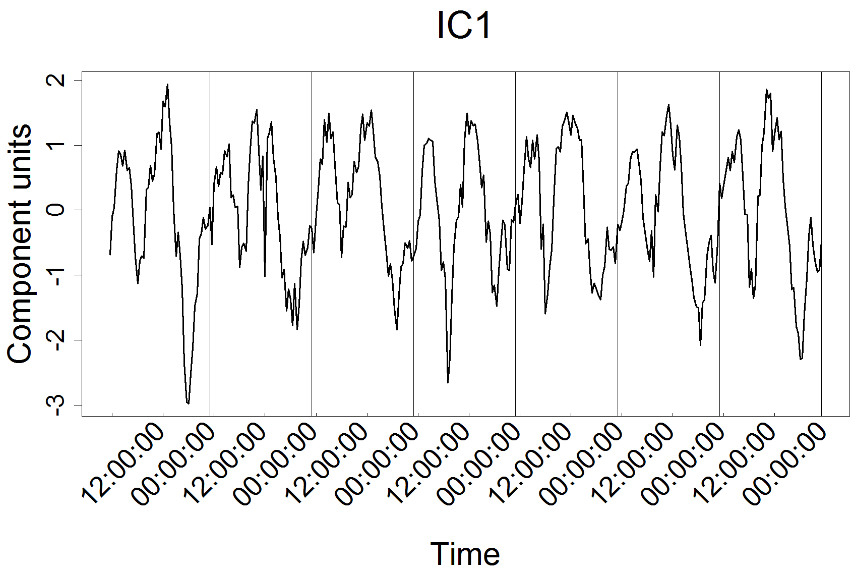
(a) NO_2

(b) PM_{10}

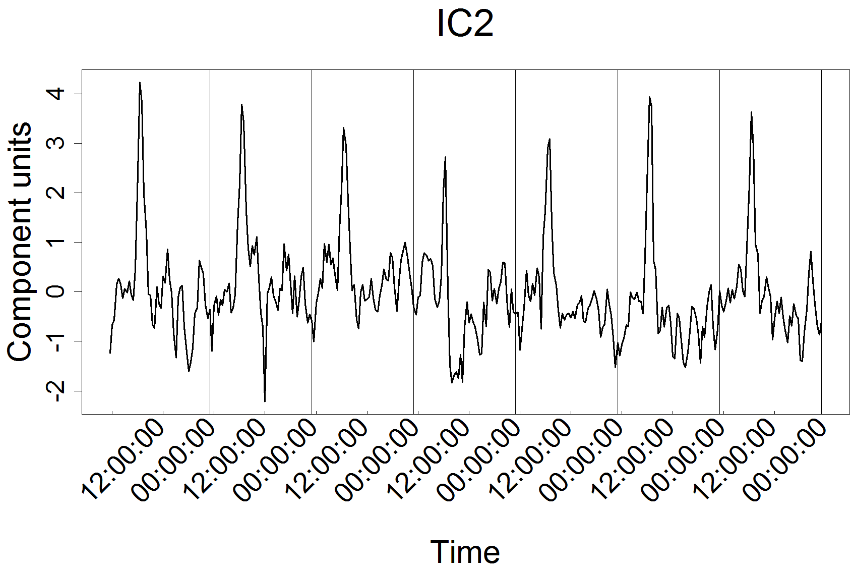
Figure 5.13 Observed diurnal variations of pollutant’s concentration over five monitoring stations

5.4.4 Temporal Distribution based on Independent Component Analysis

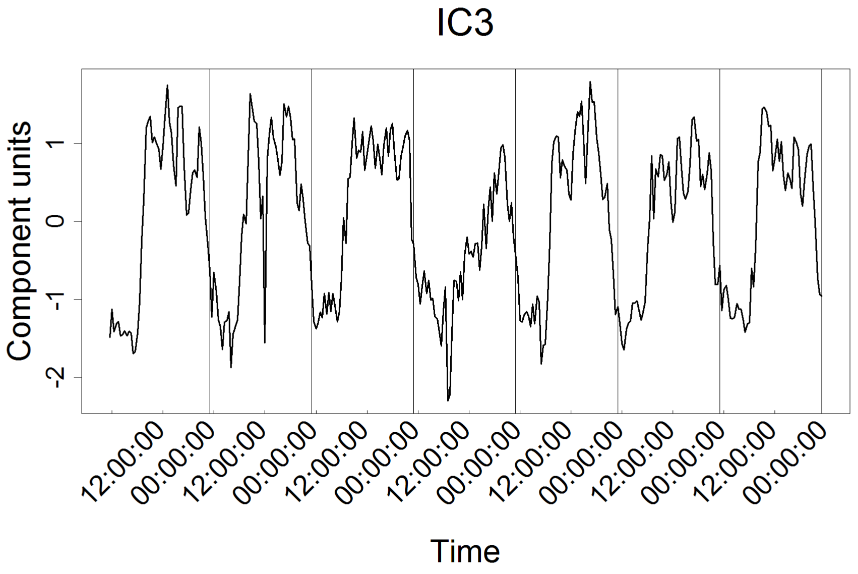
In this section, we display the ICA results consisting of weight loadings of each component, and figures showing component units. From these figures, we identified temporal patterns, especially. We also determined the time frame when the component units are high. Note that if the weight loading has a negative value, it means that the negative component units in the figure will reflect to positive (higher) air pollutant concentrations.



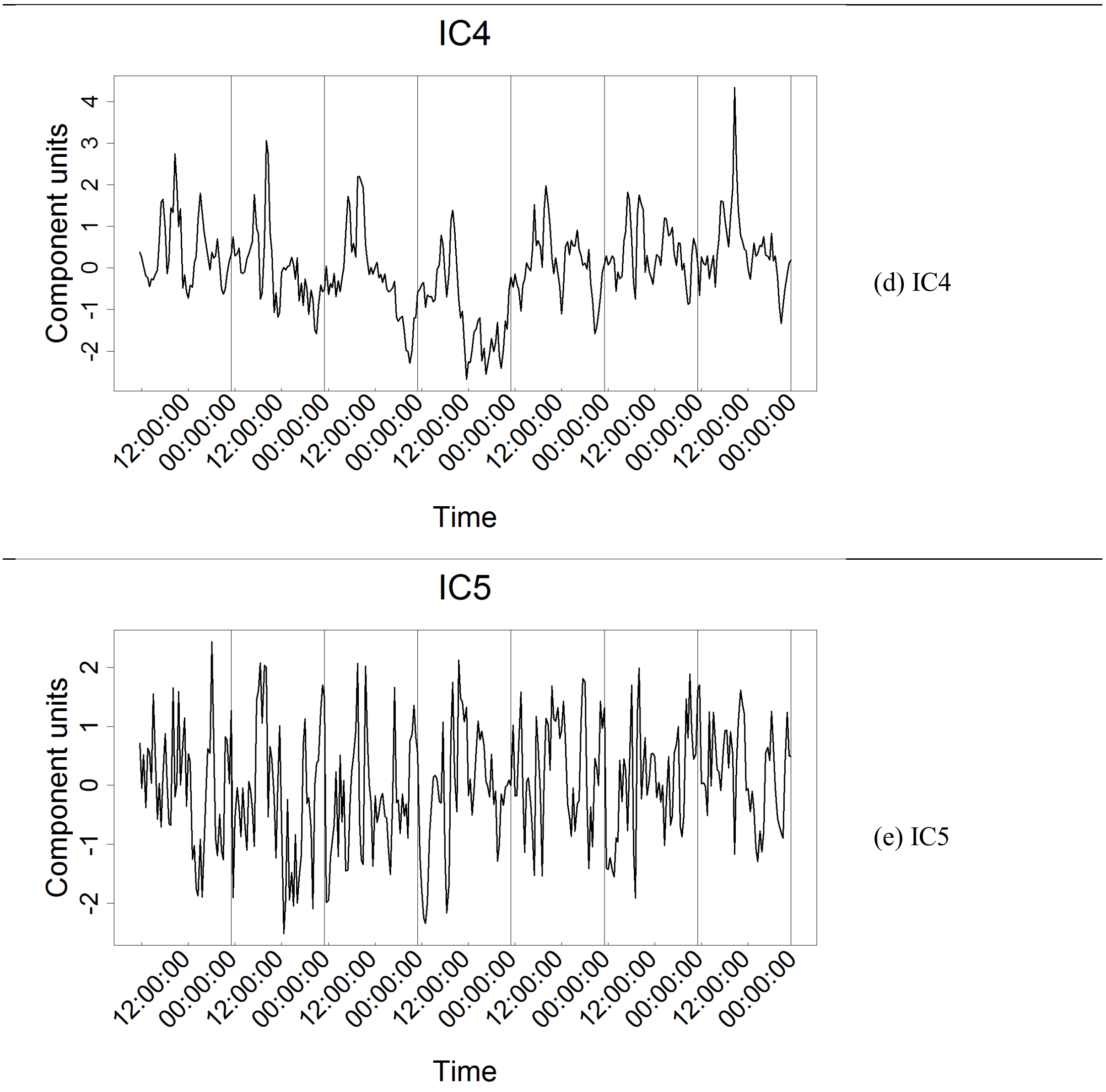
(a) IC1



(b) IC2



(c) IC3



(d) IC4

(e) IC5

Figure 5.14 Five independent component profiles from NO_2 concentration over five monitoring stations; x-axis: time, y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

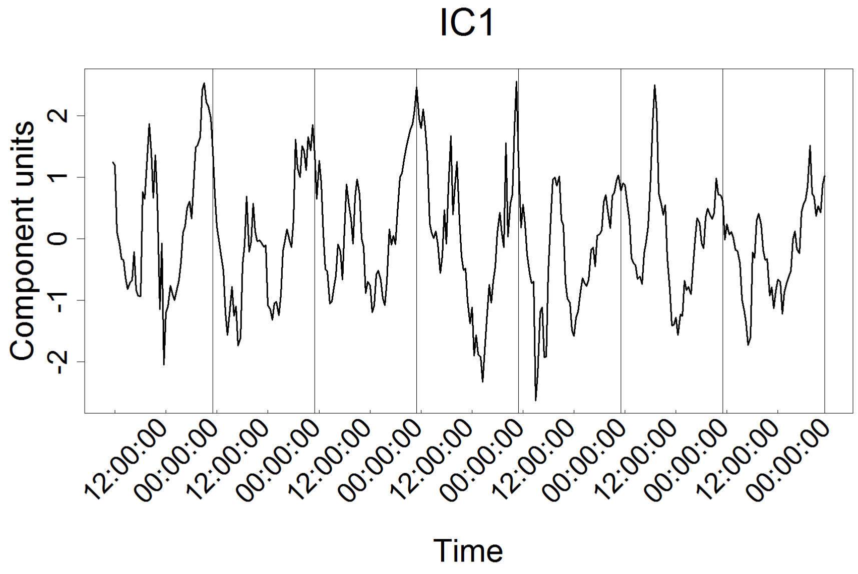
Table 5.2 Weight loadings for independent components of NO_2 concentration

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**-7.639**	**5.404**	3.164	-0.357	0.041
Trading (Perak)	-3.741	4.546	0.949	**2.912**	0.044
Suburban1 (Sukomanunggal)	**-6.693**	1.561	**5.428**	0.096	**1.050**
Near Highway (Gayungsari)	**-8.067**	**6.134**	-0.685	1.584	0.439
Suburban2 (Sukolilo)	-2.817	**6.237**	-2.527	1.844	**1.293**

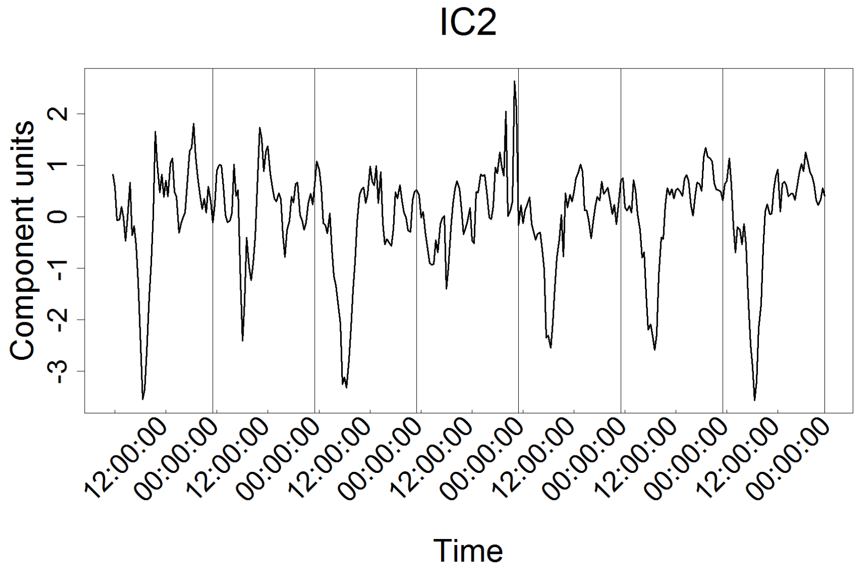
Table 5.3 NO_2 temporal peak time frame extracted from Independent Components

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	Concentration is up between 530am to 10am	Concentration is up between 6~7am to 10am	-	-	-
	Concentration is up between 330pm to 12pm				
Trading (Perak)	-	-	-	Concentration is up between 9am-10am	-
	-	-	-	Peak concentration around 5pm except weekends	-
Suburban1 (Sukomanunggal)	Concentration is up between 530am to 10am	-	Concentration is up steadily between 8am to 11pm except for Sunday	-	-
	Concentration is up between 330pm to 12pm	-	-	-	-
Near Highway (Gayungsari)	Concentration is up between 530am to 10am	Concentration is up between 6am to 10am	-	-	-
	Concentration is up again between 4pm to 12pm		-	-	-
Suburban2 (Sukolilo)	-	Concentration is up between 6am to10am	-	-	-

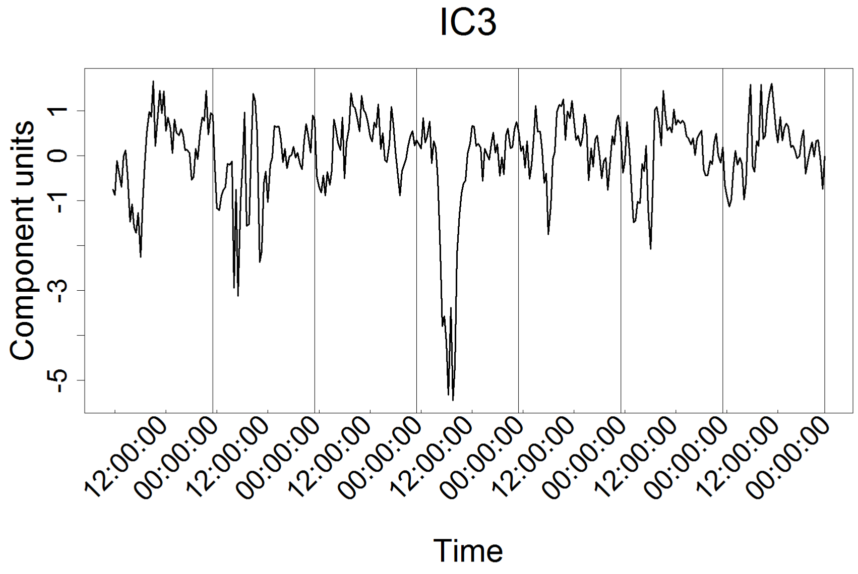
NO_2 emission comes from traffic-related sources. Therefore, the pattern formed by ICA (Figure 5.14) may partially explain the pattern of the traffic. Based on the weight loadings (Table 5.2) of NO_2, the city center pattern can be characterized by IC1 and IC2. The ICs that explain the remaining zones can be observed in Table 5.2. To assist in understanding the usefulness of ICA, and how we interpret the result from Figure 5.14, we extracted some important information in particular regarding the temporal pattern of air pollutant c oncentration. One specific feature that we can derive from the figure is that we can spot the increase in concentration temporally. So we can identify the time frame when the concentrations are high. Therefore, we focus on these particular time frames and propose what policy would be suitable for a particular zone based on the timeframe extracted in the last chapter (Chapter 7). In the IC1 and IC2, there are several patterns found. We found that the concentration was high between 5:30am to 10am (morning session), whereas, in the evening, the high concentrations occured between 3:30pm to 12pm. This information was found from IC1 (Figure 5.14a). However, based on Table 5.2, the city center pattern can also be found on IC2. From the IC2 figure (Figure 5.14b), we observed high concentrations between 6-7am to 10am. Since this time interval was between 5:30am to 10am, as noted from IC1, then we concluded the following. High concentrations of NO_2 in the morning session were between 5:30am to 10am, whereas, during evening sessions, the high concentrations were found between 3:30pm to 12pm. The analysis for the remaining zones can be observed in Table 5.3.



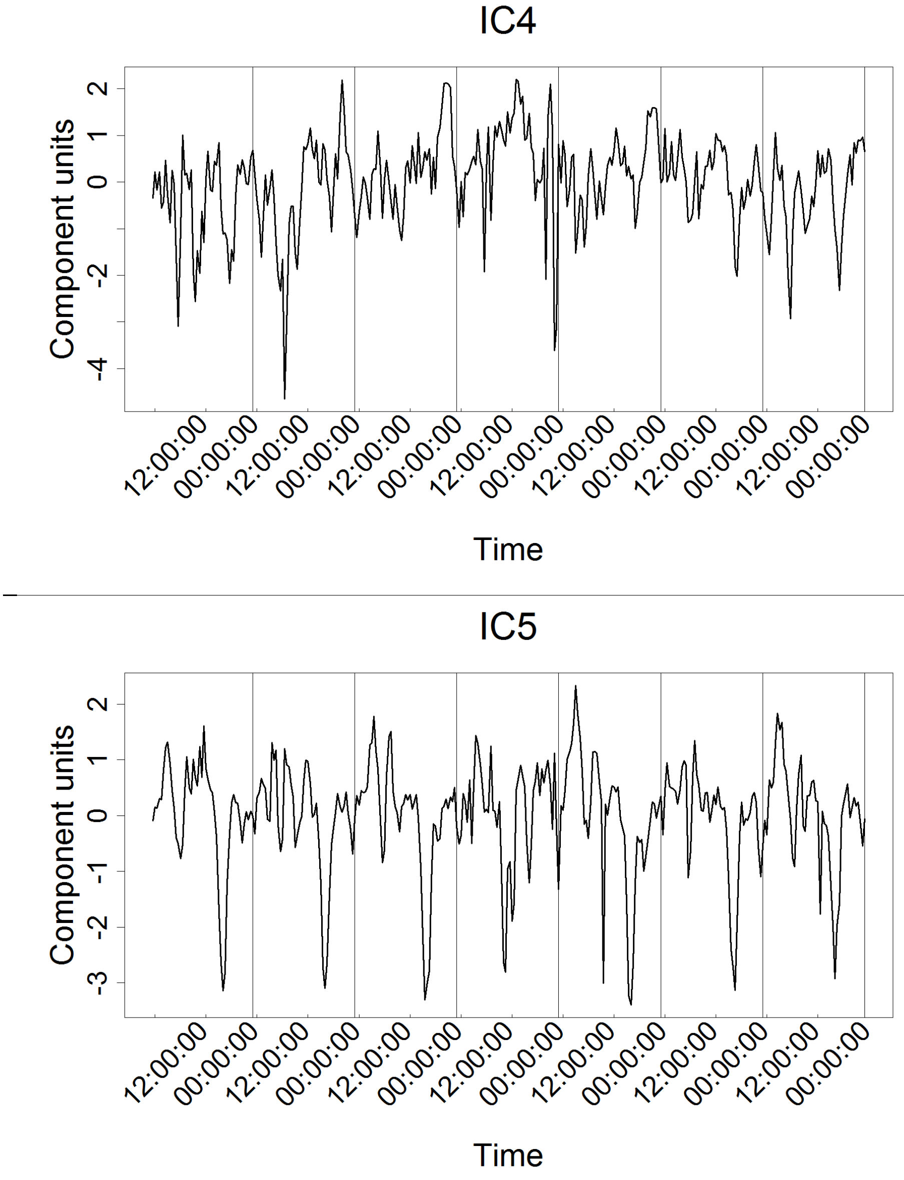
(a) IC1



(b) IC2



(c) IC3



(d) IC4

(e) IC5

Figure 5.15 Five independent component profiles from PM_{10} concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

Table 5.4 Weight loadings for independent components of PM_{10} concentration

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**6.710**	-12.707	-6.150	-3.957	**4.746**
Trading (Perak)	**5.172**	-13.136	-4.607	-5.537	-2.758
Suburban1 (Sukomanunggal)	**7.393**	-5.861	-3.636	**-7.238**	3.165
Near Highway (Gayungsari)	**5.339**	**-18.657**	**-18.882**	**-8.043**	**6.011**
Suburban2 (Sukolilo)	1.027	**-17.148**	-7.234	**-7.789**	**6.063**

Judging from all the derived information, there was the same daily pattern between the city center and the highway zone for both the morning session and evening session. The interval between the city center and the highway was similar with Suburban1, except there was an additional pattern observed, which occured from Monday to Saturday. The additional pattern was steady high concentrations between 8am to 11pm, suggesting there was an additional traffic flow captured in Suburban1. In the trading zone, the peak concentrations occured daily between 9am-10am in the morning session. However, for the evening session, the pattern was only for workdays, which was around 5pm when the concentration reached its peak. The results for the evening session indicated that traffic activities were less during weekends than on working days in the trading zone. In Suburban2, the daily pattern observed was only for the morning session between 6am to 10am.

If we observe spatial differences, we may observe that in the city center the profile or concentration patterns mimic closely routine patterns, e.g., departures and returns, daily, that can be due to working or school activities. In the meantime, we observed tight intervals of high concentrations in the trading zone, suggesting there was a certain time frame where the traffic flow reached its peak, e.g., based on the Table 5.3, between 9-10am (morning) and around 5pm (evening). The profiles of the near highway zone and Suburban2 (Sukolilo) were similar to the pattern in the city center. However, we observed relatively stable high concentrations between 8am to 12pm in the Sukomanunggal (Suburban1) region, suggesting a long-term policy approach may be appropriate for this zone because the interval was very wide.

Table 5.4 shows components and composition of each component in terms of PM_{10} concentration whereas, Figure 5.15 displays the profile of each component. We also draw some important temporal value based on peak concentration for PM_{10} as shown in Table 5.5. Based on that information, we observed that there was a delay peak hour for PM_{10} than city center, as indicated by the IC1. Based from the IC1, the concentrations in the morning were high between 7am to 11am, and, in the evening, the concentrations were high between 7pm to 1am. This pattern was also found in the trading zone, Suburban1, and near highway zone.

In the Suburban1 zone, based on IC1 and IC4, the concentrations were high between 7am to 11am in the morning session. Whereas, for the evening zone, the concentrations were high between 7pm to 1am. However, we observed a peak component unit at 5:30pm. The situation in the morning between Suburban1 and near highway was similar because the interval time was also between 7am to 11am because these two zones shared the same IC, which was IC1. However, during the night, we observed the interval time was between 7pm until dawn, suggesting emission still occured until dawn. These emissions may come from activities that are not related to traffic but related to other activities such as business and entertainment. These suggested complex sources of PM_{10}. As one of our concerns is policy formation from this information, we can combine the time interval extracted from NO_2 and PM_{10} because PM_{10} emission is also from traffic.

O_3 formation highly depends on photochemical reactions in the atmospheric layer. Therefore, we may expect wide intervals of temporal peak time of concentrations. In the city center, IC1, IC3 and IC4 made up the composition. The remaining composition can be observed in Table 5.6. The component units for each IC are displayed in Figure 5.16.

Based on information from those results, we note the following. In the city center, there was an increase of concentrations between 7:30 to 9:30am (IC1) before the concentrations go to a steady high (IC4). This pattern occured during workdays because during weekends we observe a lower pattern. A similar pattern was found in the trading zone, Suburban1, and the near highway zone, where the concentrations were relatively high between 9am to 5pm, and these occured during workdays only. However, in the Suburban2, the O_3 concentrations tended to be relatively high between 1pm to 11pm except on the weekends (Table 5.7).

Along with NO_2, the reaction between NO and NO_2 has an important role for O_3 formation. High concentrations of O_3 may interfere with visibility. For the NO pollutant, in the city center, where IC1, IC2 and IC4 represented the pattern in this zone (Table 5.8), we observed high concentrations in the morning between 1:30am to 7:30am whereas, in the evening, the concentrations were high between 5pm to 11pm (Figure 5.16). The pattern was highly different with NO_2, suggesting possible different emission source types. In the trading zone, we gathered limited information because the patterns were not clear based on IC3 and IC4. For example, we only found a temporal time frame in the morning between 4:30am to 7am. Moreover, the concentrations were lower on Sunday.

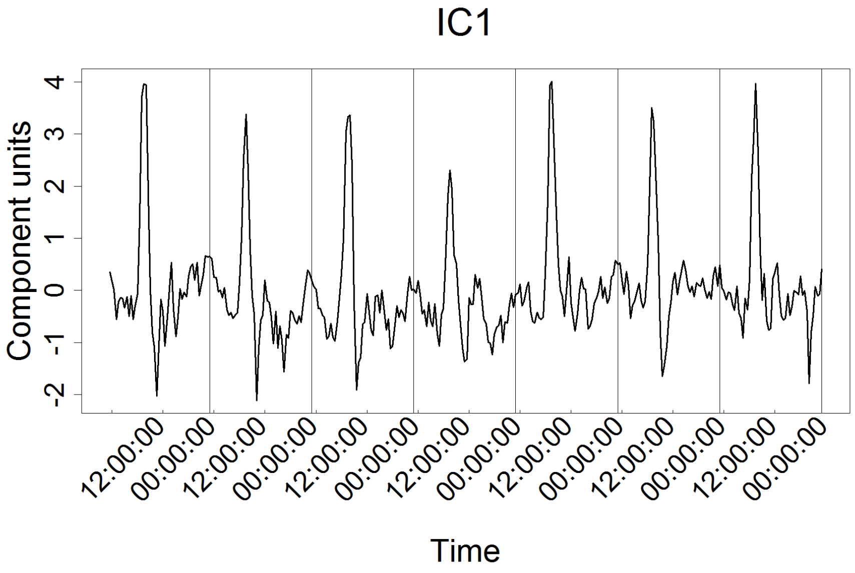
In the Suburban1 zone, the temporal peak time was not found, based on ICA, because there were no weight loadings observed for this particular zone (Table 5.8). In the highway zone, concentrations were high in the morning between 1:30am to 7:30am. However, based on IC2, the temporal time frame in the morning was from 5:30am to 7:30am. There was no information found for Suburban2 that can be extracted from ICA.

Table 5.5 PM_{10} temporal peak time frame extracted from Independent Components

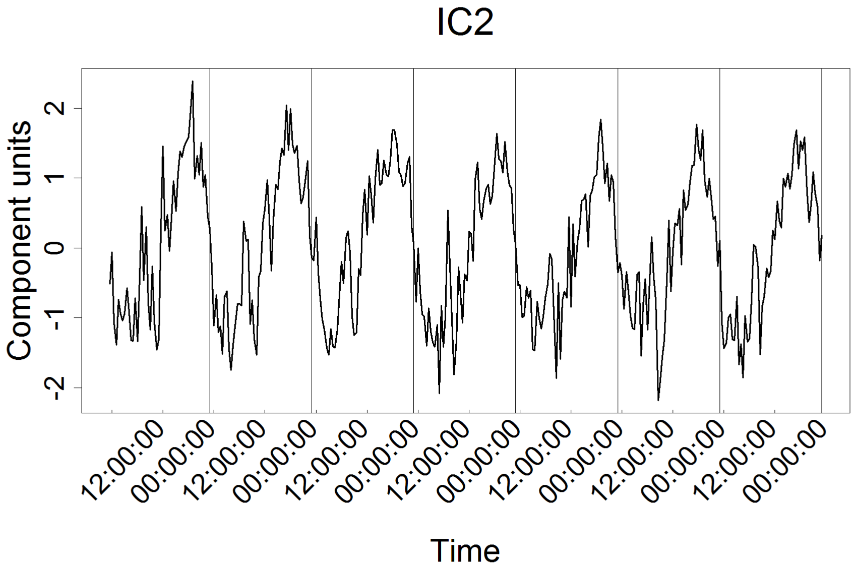
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	Concentration is up between 7am to 11am	-	-	-	Concentrations high between 8am to 3pm
	Concentration is up between 7pm to 1am	-	-	-	-
Trading (Perak)	Concentration is up between 7am to 11am up	-	-	-	-
	Concentration is up between 7pm to 1am	-	-	-	-
Suburban1 (Sukomanunggal)	Concentration is up between 7am to 11am up	-	-	Concentration peak is around 6:30-7am in the morning except for weekends	-
	Concentration is up between 7pm to 1am	-	-	Another concentration peak is around 11am in the morning, except for weekends	-
	-	-	-	Concentration peak in around 5:30 pm on working days	-
Near Highway (Gayungsari)	Concentration is up between 7am to 11am	Concentration is high between 6am to 9:30am except on Sunday	Peak concentration around 6-7am	The peak concentration is on 6:30am and 10:30am	Concentration is down around 5pm
	Concentration is up between 7pm to 1am	-	-	-	Concentrations high between 8am to 3pm
Suburban2 (Sukolilo)	-	Concentration is high between 6am to 9:30am, except on Sunday	-	The peak concentration is on 6:30am and 10:30-11 am	Concentration is down around 5pm
					Concentrations high between 8am to 3pm

Table 5.6 Weight loadings for independent components of O_3 concentration

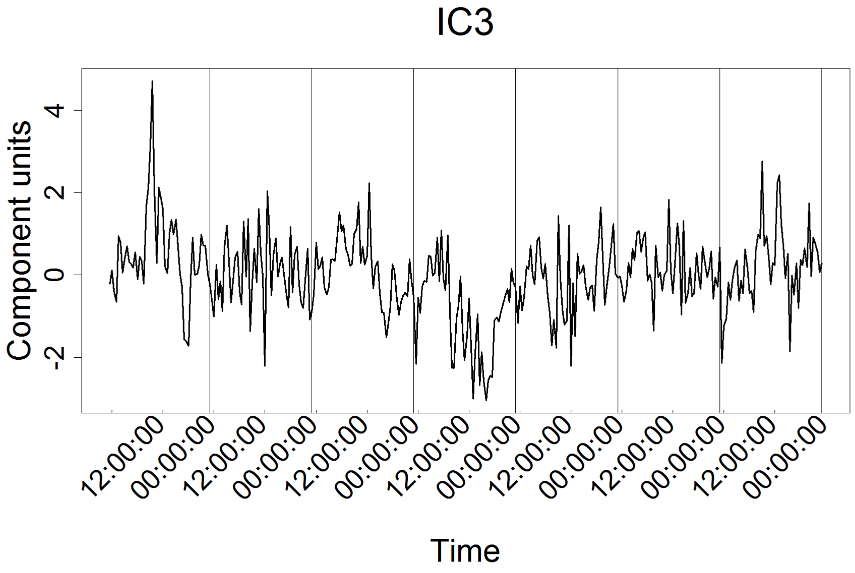
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**5.254**	-1.392	**2.688**	**-21.327**	-3.464
Trading (Perak)	1.705	2.325	0.332	**-17.590**	-2.701
Suburban1 (Sukomanunggal)	2.029	-1.313	**2.997**	**-22.930**	**-4.922**
Near Highway (Gayungsari)	0.586	0.566	**2.877**	**-20.940**	-2.788
Suburban2 (Sukolilo)	2.251	**4.687**	**2.651**	-16.470	-3.031



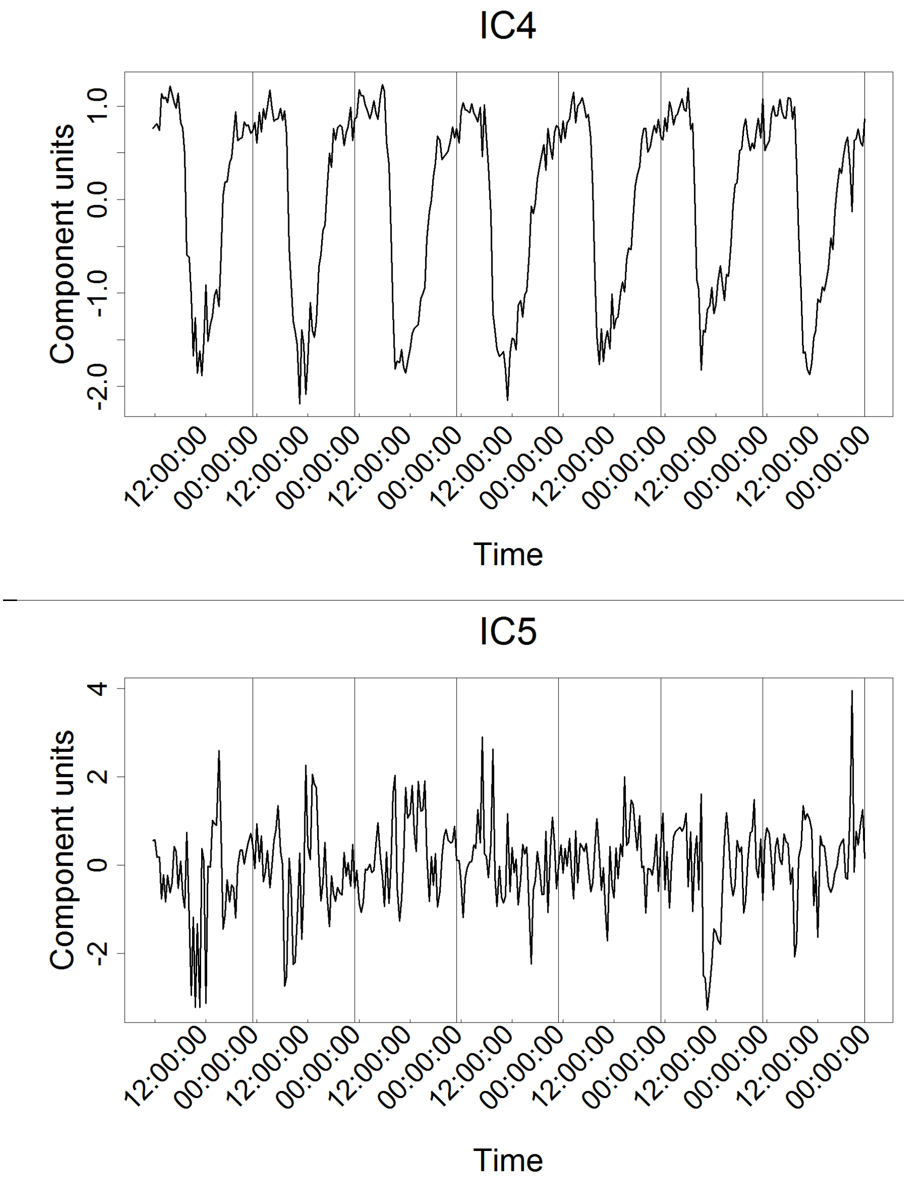
(a) IC1



(b) IC2



(c) IC3



(d) IC4

(e) IC5

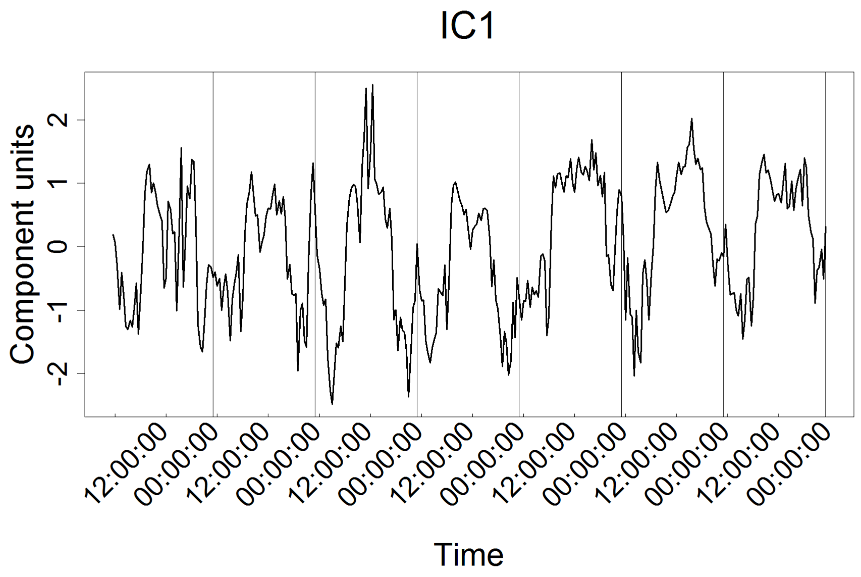
Figure 5.16 Five independent component profiles from O_3 concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

Table 5.7 O_3 temporal peak time frame extracted from Independent Components

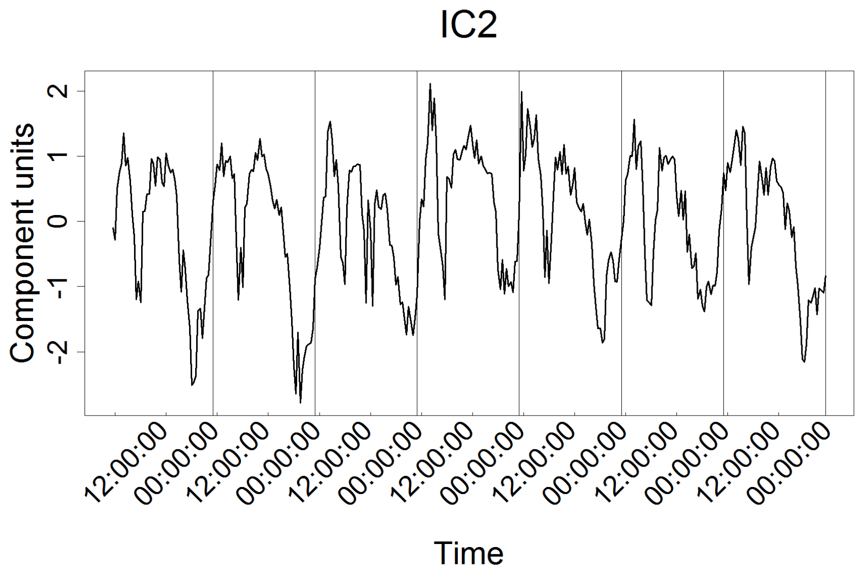
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	Sharp increase between 7:30am to 9:30am	-	No apparent pattern but lower on Saturday and Sunday (lowest)	High Concentrations between 9am to 5pm	-
Trading (Perak)	-	-	-	High concentrations between 9am to 5pm	-
Suburban1 (Sukomanunggal)	-	-	High concentrations between 9am to 5pm No apparent pattern but lower on Saturday and Sunday (lowest)	-	No clear pattern is found
Near Highway (Gayungsari)	-	-	No apparent pattern but lower on Saturday and Sunday (lowest)	High concentrations between 9am to 5pm	-
Suburban2 (Sukolilo)	Concentrations are high between 1pm to 11pm (except on Sunday)	Concentrations are high between 3pm to 11pm	No apparent pattern but lower on Saturday and Sunday (lowest)	-	-

Table 5.8 Weight loadings for independent components of NO concentration

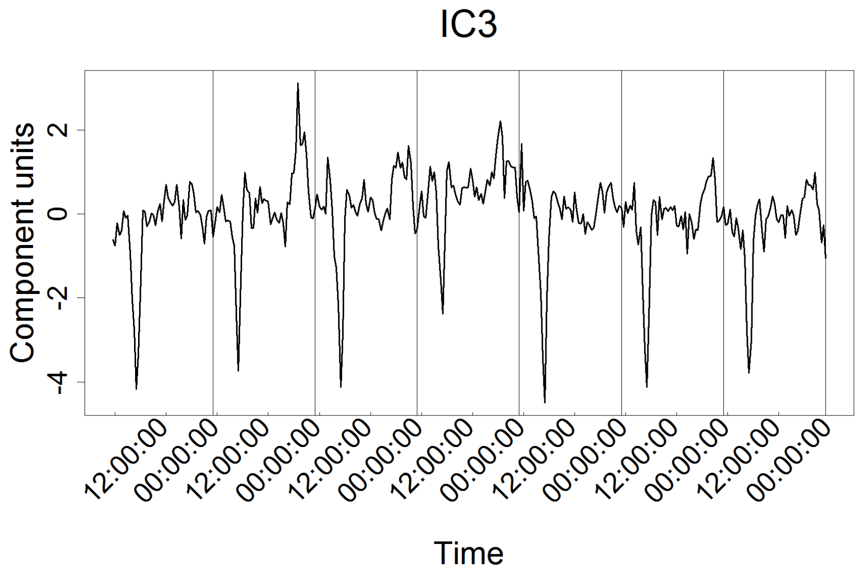
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**-6.220**	**-4.773**	-0.705	**-1.017**	-1.828
Trading (Perak)	-4.293	-2.611	**-5.702**	**-1.776**	-2.812
Suburban1 (Sukomanunggal)	-0.625	-3.332	0.599	-0.629	-0.914
Near Highway (Gayungsari)	**-6.731**	**-5.095**	**-7.153**	0.004	**-8.205**
Suburban2 (Sukolilo)	-2.508	-0.043	-2.284	**-1.220**	-3.582



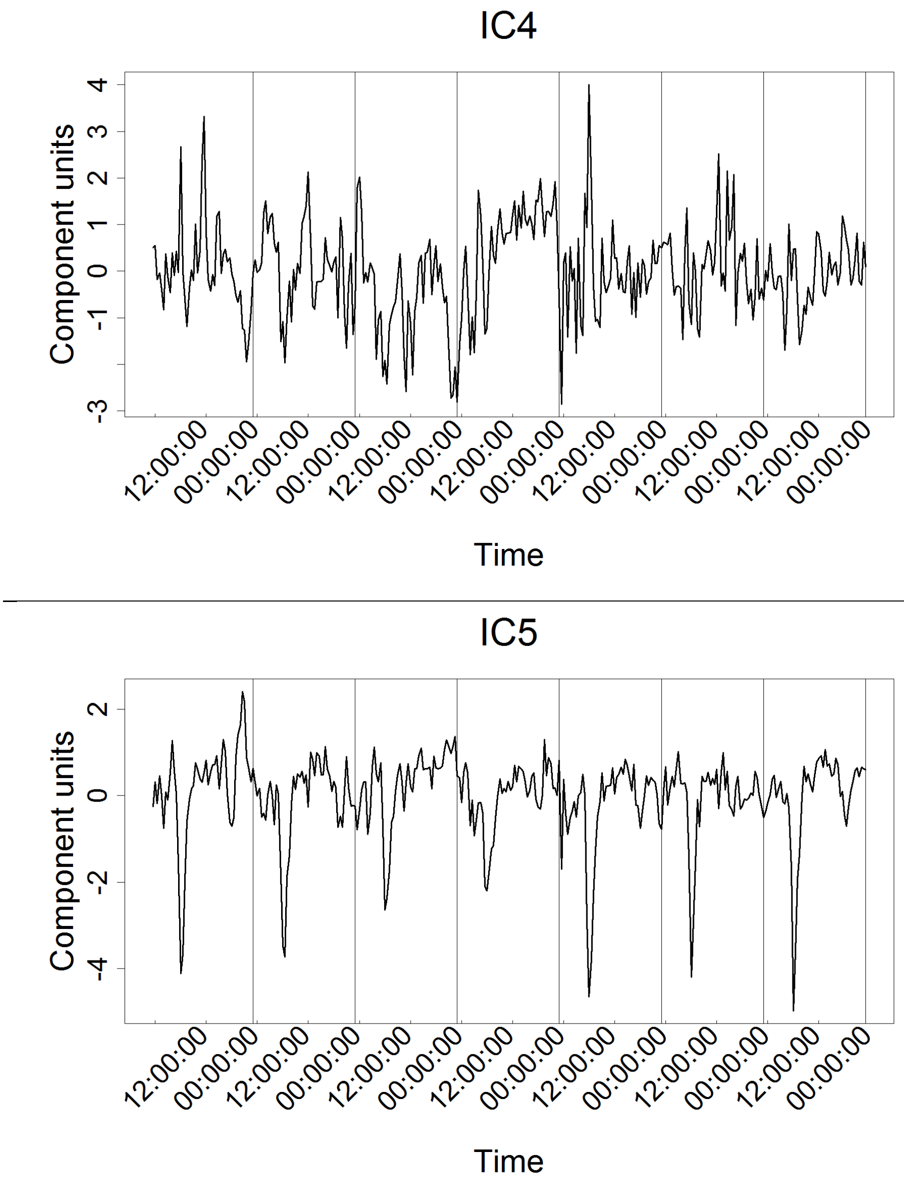
(a) IC1



(b) IC2



(c) IC3



(d) IC4

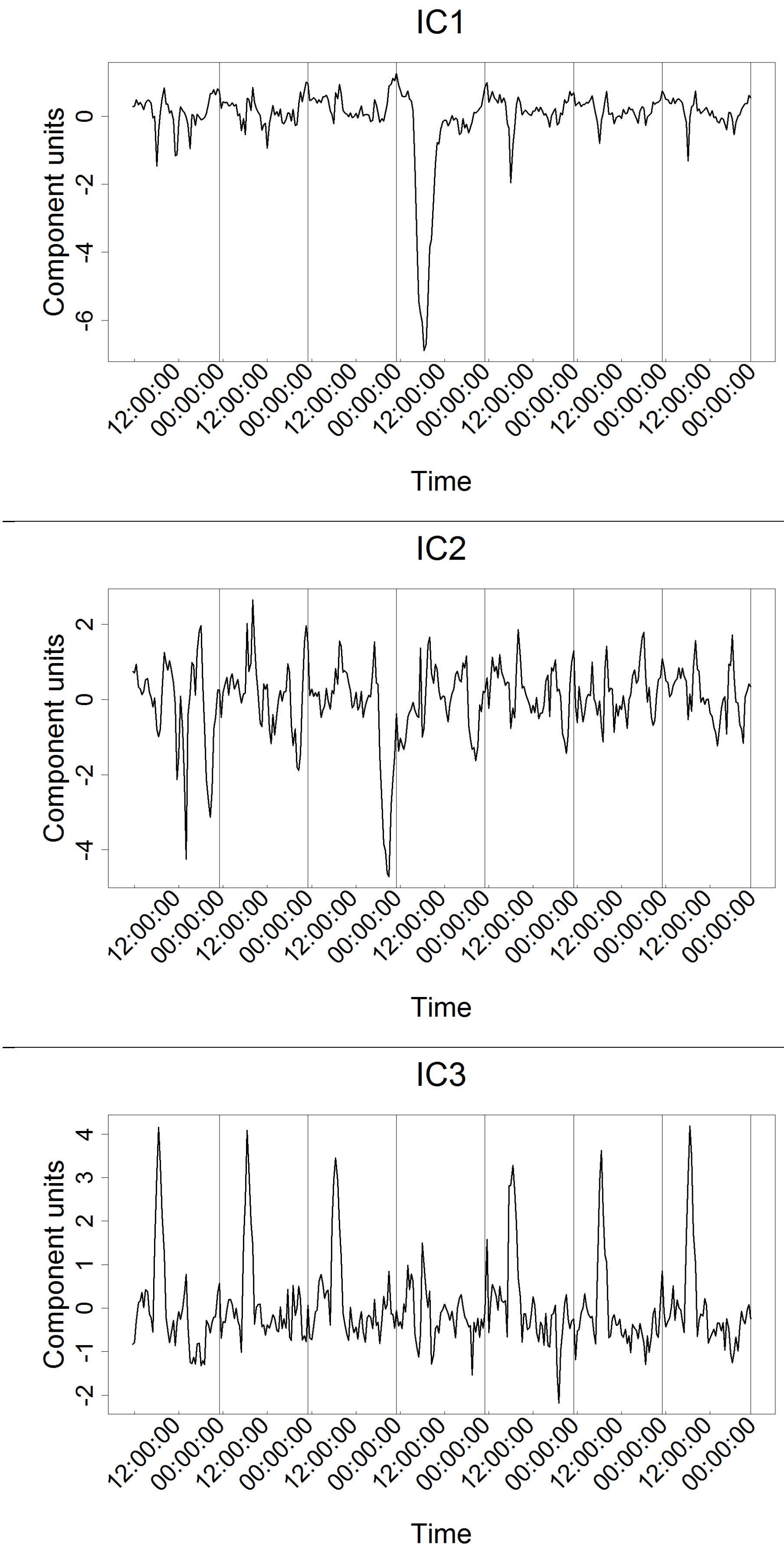
(e) IC5

Figure 5.17 Five independent component profiles from NO concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

Table 5.9 NO temporal peak time frame extracted from Independent Components

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	Concentrations are high between 1:30am to 7:30am	Concentrations are high between 5:30am to 7:30am		Concentrations are relatively higher in morning and night, lower on Sunday. No apparent consistent pattern found	
	Concentrations are high between 8pm to 10pm (wider interval on Sunday)	Concentrations are high between 5pm to 11pm (lower unit on Sunday)			
Trading (Perak)	-	-	Concentrations are up from 4:30am to 7am (Sunday is the lowest)	Concentrations are higher in morning and night, lower on Sunday. No apparent consistent pattern found	
Suburban1 (Sukomanunggal)	-	-	-	-	-
Near Highway (Gayungsari)	Concentrations are high between 1:30am to 7:30am	Concentrations are high between 5:30am to 7:30am	Concentrations are up from 4:30am to 7am (Sunday is the lowest)	Concentrations are higher in morning and night, lower on Sunday. No apparent consistent pattern found	Concentrations up between 6am to 9am
	Concentrations are high between 8pm to 10pm (wider interval on Sunday)	Concentrations are high between 5pm to 11pm (lower unit on Sunday)			
Suburban2 (Sukolilo)	-				

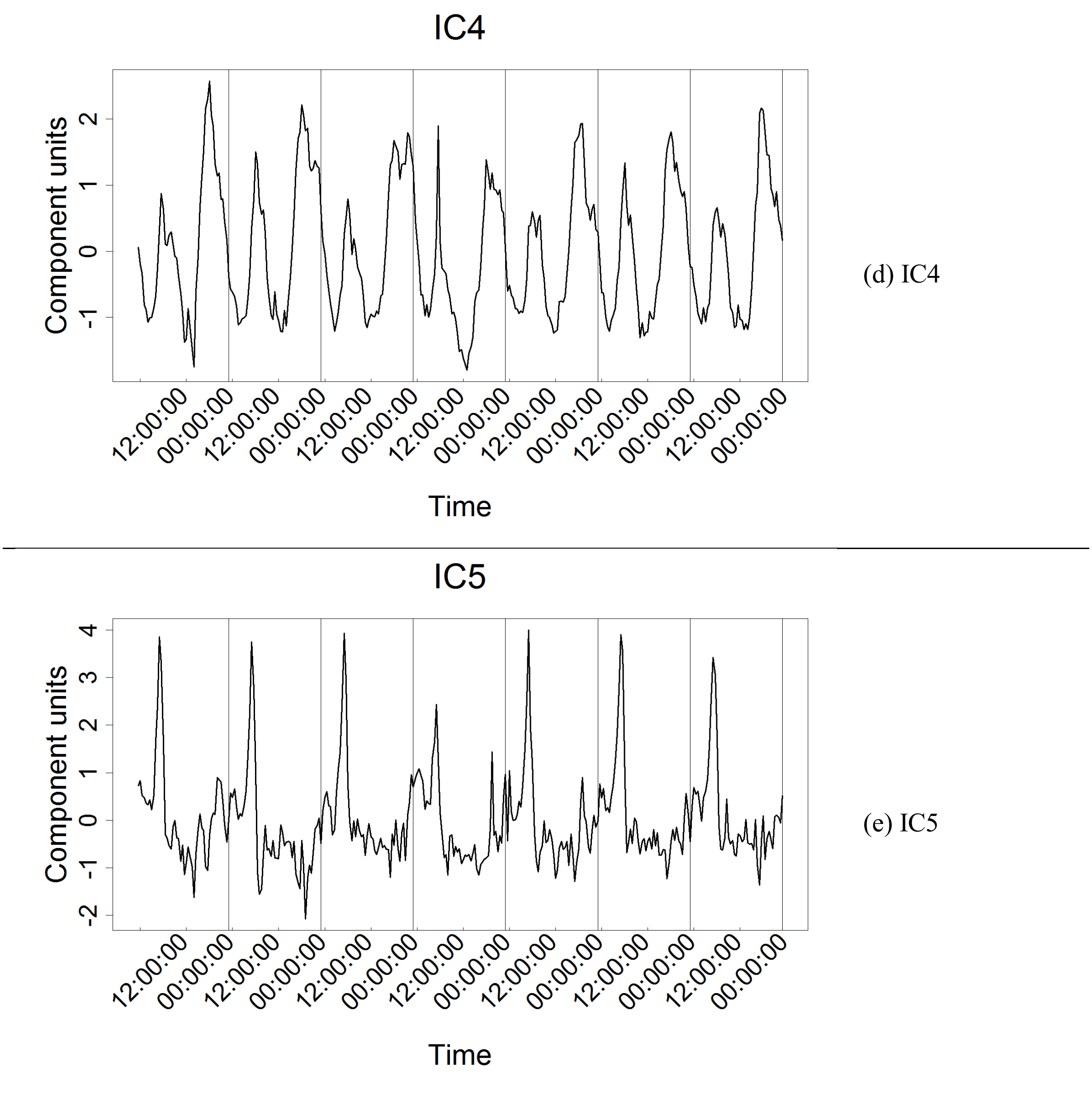
Traffic also contributes towards the concentration of carbon monoxide (CO). In this section, we shall discuss the temporal pattern of CO. However, we observed that the magnitude of concentration for CO was small, and the fluctuation was relatively small as well when compared with NO_2 and PM_{10}. The concentration of CO in the city center may be represented by IC2 and IC4 as indicated by the weight loadings in Table 5.10. In this zone, we only observed a distinct pattern on Thursday and Sunday. The concentrations on Thursday were high between 7am to 9am (morning) and between 12:30pm to 3:30pm and between 8:30pm to 10:30pm in the night (Figure 5.18 and Table 5.12). On Saturday, the concentrations were high between 8pm to 11:30pm.



(a) IC1

(b) IC2

(c) IC3



(d) IC4

(e) IC5

Figure 5.18 Five independent component profiles from CO concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

Table 5.10 Weight loadings for independent components of CO concentration

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	-0.052	**-0.174**	0.175	**0.394**	0.034
Trading (Perak)	-0.065	-0.059	0.189	0.256	**0.192**
Suburban1 (Sukomanunggal)	-0.025	-0.034	0.045	0.271	-0.017
Near Highway (Gayungsari)	**-0.528**	-0.018	**0.287**	**0.432**	**0.134**
Suburban2 (Sukolilo)	-0.065	0.010	**0.243**	0.164	0.091

SO_2 is released primarily from combustion of fossil fuels, that is, from both coal and oil. It is then oxidized during the transport to sulfur trioxide (SO_3) in the troposphere. The SO_3 will react with water vapor to form sulfuric acid. The sources of SO_2 are different from each zone. Therefore, we would like to obtain a temporal pattern for SO_2 from each zone. Our hypothesis was that there was a low contribution of SO_2 in the city center and suburban area that may be indicated by short intervals of SO_2 concentration. In the meantime, we expected higher contributions of SO_2 in the trading zone and highway zones due to industrial activities nearby.

Table 5.11 CO temporal peak time frame extracted from Independent Components

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	-	Thursday – high between 12:30pm to 3:30pm	-	High between 7am to 9am	-
	-	Thursday – high between 8:30pm to 10:30pm	-	High between 5pm to 11pm	-
	-	Saturday – high between 8pm to 11:30pm	-	-	-
Trading (Perak)	-	-	-	-	High between 5am to 7am
Suburban1 (Sukomanunggal)	-	-	-	-	-
Near Highway (Gayungsari)	High concentrations on Sunday between 5am to 10am	-	Sharp increase between 7am to 9am (except Sunday)	High between 7am to 9am	High between 5am to 7am
Suburban2 (Sukolilo)	-	-	Sharp increase between 7am to 9am (except Sunday)	High between 5pm to 11pm	-

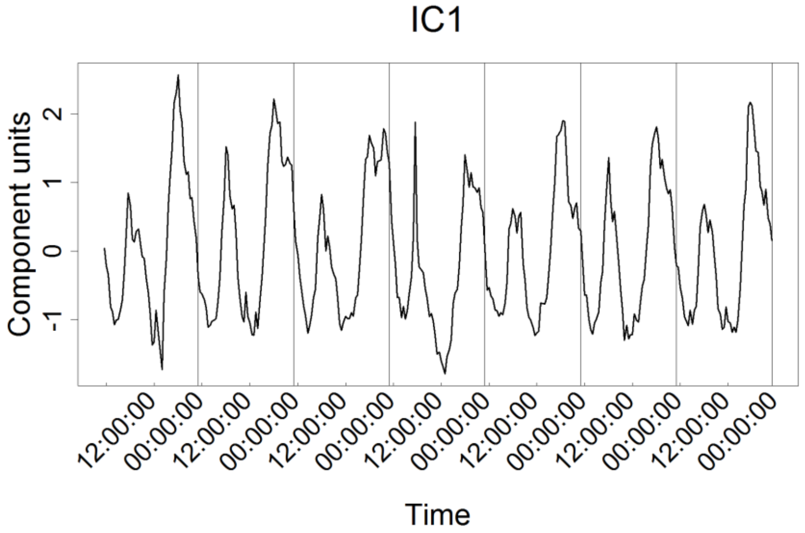
Based on the ICA, the city center was represented by IC1 and IC5. The remaining ICs that explain each zone are shown in Table 5.12. These ICs may explain the different sources of SO_2 for each zone. We observed a temporal pattern as follows, based on Figure 5.19. The temporal pattern results are found in Table 5.13.

The SO_2 pattern in the city center was characterized by high concentrations between 6am to 9am and 5pm to 10pm, and there was one unique characteristic captured, that was, we observed high concentrations on Saturday between 8pm to 12pm, indicating other industrial activities within that zone. Interestingly, there was short interval of high concentrations between 5am to 7am. The temporal difference is obvious between the trading zone and city center. A similarity of SO_2 occurs between the city center and Suburban1, which indicated a similar source pattern emission. Even though we can observe a similarity in the highway zone, we observed, based on other ICs, that the interval was extended between 5am to 9am, except on Sunday, when no pattern ws found, suggesting no industrial activities were taking place. In the evening, the temporal pattern was similar to the city center and Suburban1. The temporal pattern of Suburban2 was between 6am to 9am, similar to the highway, city center, and Suburban1, except that in the city center and Suburban1, the concentrations were stable for all days, whereas for the highway zone and Suburban2, we noted the concentrations were low on Sunday.

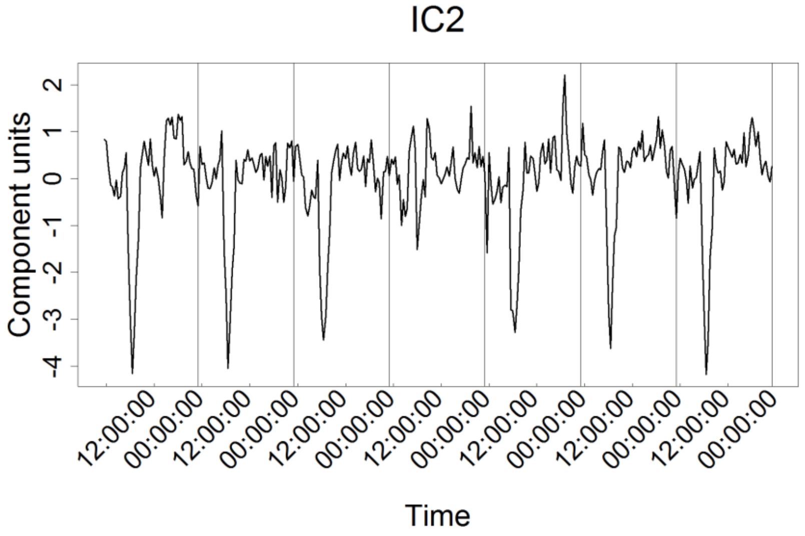
Therefore we concluded the following: The SO_2 temporal pattern can be divided into three groups: city center and suburban1, near highway and suburban2, and trading zone. The shortest interval is in the trading zone.

Table 5.12 Weight loadings for independent components of SO_2 concentration

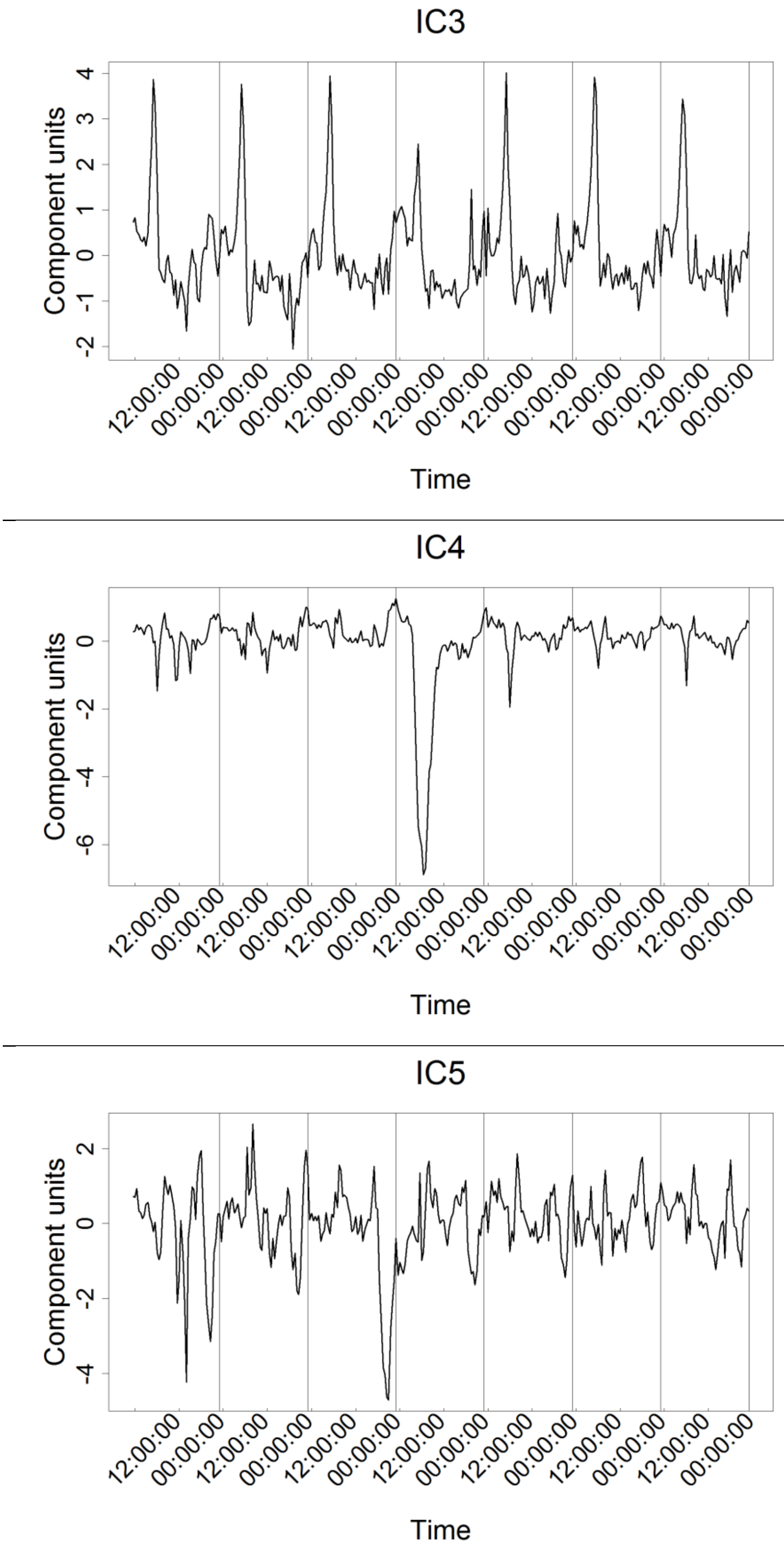
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**0.395**	-0.171	0.040	-0.051	**-0.174**
Trading (Perak)	0.256	-0.186	**0.195**	-0.065	-0.059
Suburban1 (Sukomanunggal)	**0.272**	-0.041	-0.013	-0.025	-0.035
Near Highway (Gayungsari)	**0.434**	**-0.282**	**0.141**	**-0.527**	-0.018
Suburban2 (Sukolilo)	0.166	**-0.241**	0.094	-0.065	0.011



(a) IC1



(b) IC2



(c) IC3

(d) IC4

(e) IC5

Figure 5.19 Five independent component profiles from SO_2 concentration over five monitoring stations; x-axis: time (in 30 mins), y-axis: component value. The vertical line is the division indicating each day, starting from Thursday.

Table 5.13 SO_2 temporal peak time frame extracted from Independent Components

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	High concentrations between 6am to 9am	-	-	-	High concentration between 8pm to 12pm in Saturday
	High concentrations between 5pm to 10pm	-	-	-	-
Trading (Perak)	-	-	Short interval of high concentrations 5am to 7am (except for Sunday being lower)	-	-
Suburban1 (Sukomanunggal)	High concentrations between 6am to 9am	-	-	-	-
	High concentrations between 5pm to 10pm	-	-	-	-
Near Highway (Gayungsari)	High concentrations between 6am to 9am	High concentrations between 6am to 9am (except Sunday)	Short interval of high concentrations 5am to 7am (except Sunday being lower)	High concentrations between 5am to 11am only on Sunday	-
	High concentrations between 5pm to 10pm	-	-	-	-
Suburban2 (Sukolilo)	-	High concentrations between 6am to 9am (except Sunday)	-	-	-

For easier navigation and explanation we combine extracted components from Table 5.3 to 5.13 into a single table as shown on Table 5.14 below. It is helpful to identify seasonal or daily cycle pattern of pollutants on each zone. For example NO_2 in the suburban1, we found two patterns. First pattern was similar with pattern in the city center and near highway zone (morning and evening session), however the second pattern was unique. The second pattern shows additional source of emission which relatively high within wide interval between 8am to 11pm and this pattern occured from Monday to Saturday, indicating that the source of emission did not occur on Sunday. This information is weekly cycle and unique for this particular zone only. Comparing with NO_2, the profile of PM_{10} is different. In the city center, there were two daily patterns observed. First pattern encompased interval on both morning (7am to 11am) and evening session (7pm to 3pm), whereas the other pattern included interval time between 8am to 3pm. In the trading zone, the daily cycle pattern was also found, similar with city center. All in all, Table 5.15 shows how ICA is helpful to obtain weekly-cycle pattern every day which cannot be captured by other method.

Table 5.14 Temporal peak time frame from all pollutants drawn from Independent Component Analysis of each pollutant

Parameters	**City Center (Ketabang Kali)**	**Trading (Perak)**	**Suburban1 (Sukomanunggal)**	**Near Highway (Gayungsari)**	**Suburban2 (Sukolilo)**
NO_2	• 5am to 10am • 330pm to 12pm	• 9am~10am • around 5pm except weekends	• 530am to 10am • 330pm to 12pm • up steadily between 8am to 11pm except for Sunday	• 530am to 10am • 330pm to 12pm	• 6am to10am
PM_{10}	• 7am to 11am • 8am to 3pm • 7pm to 1am	• 7am to 11am • 7pm to 1am	• 7am to 11am • 7pm to 1am • 630am ~ 7am except weekend • 11am except weekends • around 530 pm in working days	• 7am to 11am • 7pm to 1am • 6am to 930am except on Sunday • peak at 630am and 1030am • 8am to 3pm	• 6am to 930am except on Sunday • Peak on 630am and 1030~11 am • down around 5pm • 8am to 3pm
O_3	• 730am to 930am • lower in the weekends • 9am to 5pm	• 9am to 5pm	• 9am to 5pm (working days)	• No apparent trend but lower during weekends • 9am to 5pm	• 1pm to 11pm (except in Sunday) • 3pm to 11pm • Weekends is low

Table 5.14 Temporal peak time frame from all pollutants drawn from Independent Component Analysis of each pollutant (Cont'd)

Parameters	**City Center (Ketabang Kali)**	**Trading (Perak)**	**Suburban1 (Sukomanunggal)**	**Near Highway (Gayungsari)**	**Suburban2 (Sukolilo)**
NO	• 130am to 730am • lower in the Sunday • 5pm to 11pm (lower unit in Sunday)	• 430am to 7am • Sunday is the lowest	-	• 130am to 9am • lower in the Sunday • 5pm to 11pm (lower unit in Sunday)	-
CO	• 7am to 9am • 5pm to 11pm • Thursday –1230pm to 330pm • Thursday –830pm to 1030pm • Saturday –8pm to 1130pm	• 5am to 7am	-	• Sunday - 5am to 10am • 7am to 9am (except Sunday) • 7am to 9am • 5am to 7am	• 7am to 9am (except Sunday) • 5pm to 11pm
SO_2	• 6am to 9am • 5pm to 10pm • 8pm to 12pm in Saturday	• 5am to 7am (except Sunday)	• 6am to 9am • 5pm to 10pm	• 6am to 9am • 5pm to 10pm • 6am to 9am (except Sunday) • 5am to 7am (except Sunday) • 5am to 11am only in Sunday	• 6am to 9am (except Sunday)

Table 5.15 Seasonal pattern on weekly cycle for all pollutant profiles

	Zones	Every day	Only workdays	Only weekends	Monday to Saturday	Certain days
NO_2	City center	• 530am to 10am • 330pm to 12pm				
	Trading zone	9am~10am	• around 5pm except weekends			
	Suburban1	• 530am to 10am • 330pm to 12pm			8am to 11pm	
	Near highway	• 530am to 10am • 330pm to 12pm				
	Suburban2	6am to10am				
PM_{10}	City center	• 7am to 11am • 8am to 3pm • 7pm to 1am				
	Trading zone	• 7am to 11am • 7pm to 1am				
	Suburban1	• 7am to 11am • 7pm to 1am	• 630am ~ 7am • peak at 11am • peak at 530 pm			
	Near highway	• 7am to 11am • 7pm to 1am • peak at 630am and 1030am • 8am to 3pm			6am to 930am	
	Suburban2	• peak at 630am and 1030am • 8am to 3pm			6am to 930am	

Table 5.15 Seasonal pattern on weekly cycle for all pollutant profiles (Cont'd)

	Zones	Every day	Only workdays	Only weekends	Monday to Saturday	Certain days
O_3	City center	• 730am to 930am • 9am to 5pm		Lower concentrations		
	Trading zone	9am to 5pm				
	Suburban1	9am to 5pm		Lower concentrations		
	Near highway	9am to 5pm		Lower concentrations		
	Suburban2	3pm to 11pm		Lower concentrations	1pm to 11pm	
NO	City center	• 130am to 730am • 530am to 730am			• 8pm to 10pm (wider interval in Sunday) • 5pm to 11pm (lower in Sunday)	
	Trading zone				430am to 7am (Sunday is the lowest)	
	Suburban1					
	Near highway	• 130am to 730am • 530am to 730am • 6am to 9am			• 8pm to 10pm (wider interval in Sunday) • 5pm to 11pm (lower in Sunday) • 430am to 7am (lower in Sunday)	
	Suburban2					

Table 5.15 Seasonal pattern on weekly cycle for all pollutant profiles (Cont'd)

	Zones	Every day	Only workdays	Only weekends	Monday to Saturday	Certain days
CO	City center	• 7am to 9am • 5pm to 11pm				• Thursday –1230pm to 330pm • Thursday – high between 830pm to 1030pm • Saturday – high between 8pm to 1130pm
	Trading zone	5am to 7am				
	Suburban1					
	Near highway	• 5am to 7am • 7am to 9am			7am to 9am (except Sunday)	Sunday between 5am to 10am
	Suburban2	5pm to 11pm			7am to 9am (except Sunday)	
SO_2	City center	• 6am to 9am • 5pm to 10pm				8pm to 12pm in Saturday
	Trading zone				5am to 7am (except Sunday being lower)	
	Suburban1	• 6am to 9am • 5pm to 10pm				
	Near highway	• 6am to 9am • 5pm to 10pm			• 5am to 7am (except Sunday) • 6am to 9am (except Sunday)	5am to 11am only in Sunday
	Suburban2				6am to 9am (except Sunday)	

To fully explore the usefulness of the results obtained in this chapter, one can use the information to extract direct policies from each component. We display those results in Chapter 7. We composed policies by observing the temporal air pollutant concentration patterns and propose suitable transport-related policy.

5.4.5 Estimation of Emission Sources

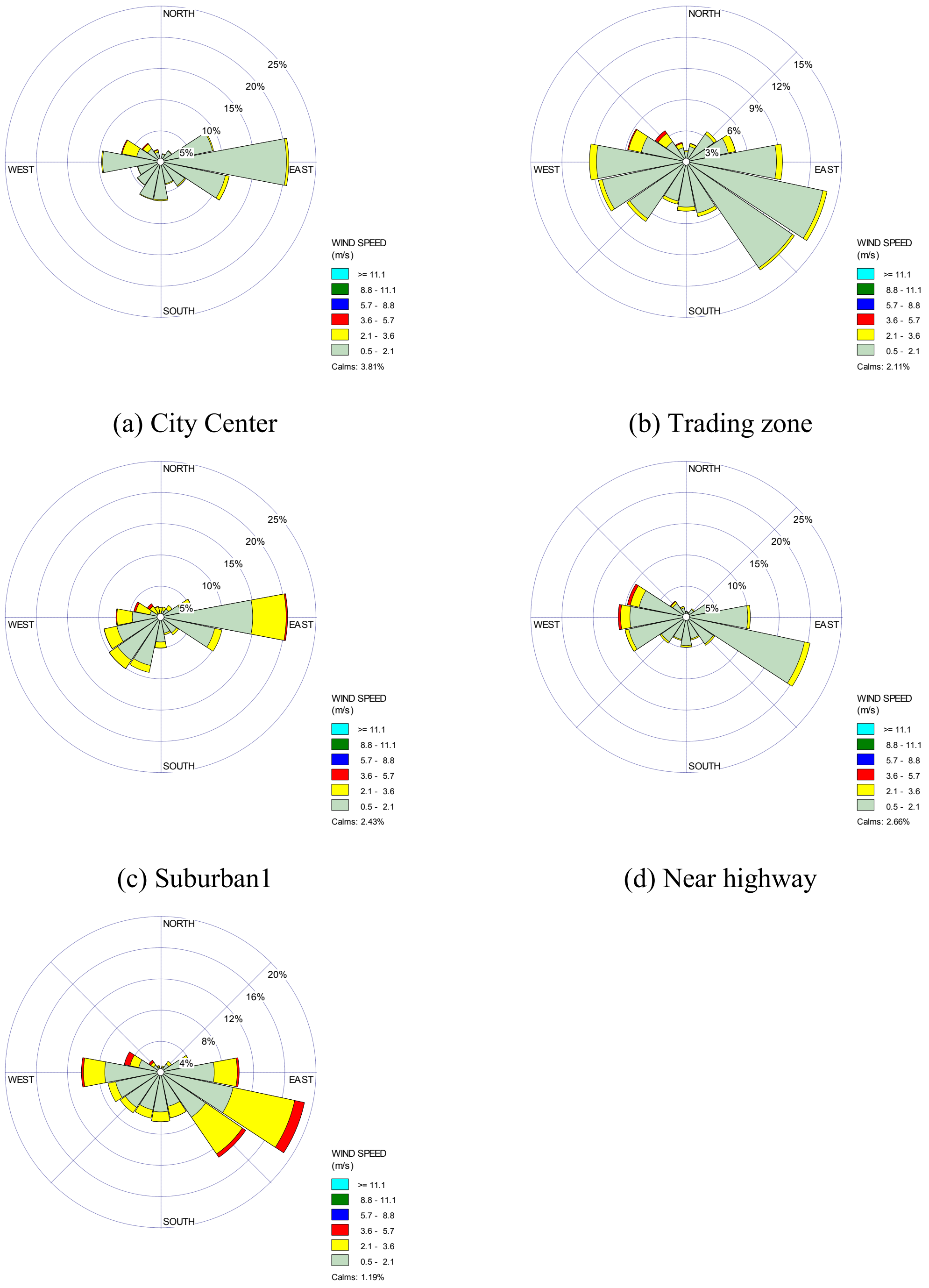
In this section, we focus on NO_2 and $PM_{10,}$ as it was shown that there are many events for both these pollutants when the concentrations exceed the permissible threshold level. City center (Ketabang Kali) station is located at the center of the city through which many vehicles pass. Figure 5.1a shows the map displaying the surroundings of the monitoring sensor in Ketabang Kali. The nearest protocol road, with this station, is Gubernur Suryo Street (St.) on the south (S) side, while the nearest normal road is Simpang Dukuh Road (Rd.) on the southwest (SW) side, A small road, Ketabang Kali Rd. is located on the north (N) side. Considering wind direction and speed, dominant pollutants came from the east (E) side (Figure 5.20a). When the NO_2 concentration exceeded the ambient level, we noted that the wind largely came from the southeast-south (SE-S) side, as shown in Figure 5.21a. In this instance, the wind was dominantly coming from Gubernur Suryo St. It is a protocol road where many vehicles pass through, and where the building of the Governor is located. Therefore, the concentration of pollution was mainly due to traffic flow. On the other hand, when the concentrations of PM_{10} were higher than the threshold value (260 ug/m^3), dominant wind direction was from the SW-S side (Figure 5.22a). The area on the SW-S side comprised of office buildings, a big shopping mall, and highly-dense dwellings. These entities contributed greatly to PM_{10} contamination, along with emissions from traffic passing through on Gubernur Suryo St. and Simpang Dukuh Rd. Based on the facts above, the events where NO_2 exceeded ambient level were mainly due to traffic flow. On the other hand, for PM_{10}, the main causes were not only from traffic but also from business-oriented and residential emissions.

The monitoring sensor, located in the trading zone, is in the Perak region on Johor Rd., which is on the north of Surabaya (Figure 5.1b). The wind mostly came from the direction between east (E) and southeast (SE) (Figure 5.20b). The dominant wind speed was observed to be between 0.5-2.1 meters per second (m/s), a smaller portion with a speed between 2.1-3.6 m/s. When the concentrations of NO_2 exceeded the threshold value (92.5 ug/m^3), the wind was coming mostly from a SE direction, whereas a smaller portion came from the northwest (NW) and northwest-west (NW-W) (Figure 5.21b). The wind speed was between 2.1-3.6 m/s for winds coming from the SE. The SE side of the site monitoring station has several protocol roads, residential areas, and large scale industries. Since NO_2 is decidedly attributed to traffic emission, we observed that emissions came from roads on the SE side. Roads in this direction were Johor Rd., Indrapura St., and Kalimas Barat St. The largest street among those mentioned is Indrapura St. because vehicles heading to the north of Surabaya, including the neighboring city, Gresik, and the Tanjung Perak Port, must use this street, making it the main street with major traffic flow. When PM_{10} concentrations exceeded the ambient level (260 ug/m^3), the dominant wind direction was from the S-SW side and the SW-W side (Figure 5.22b). The wind speed from the S-SW side was quite high, with the highest interval between 3.6-5.7 m/s, indicating that potential sources were quite far from the monitoring station. The area was comprised of houses and many business-related activities (restaurants, medical clinics, pharmacies, etc.), and a few small-scale industries. For the SW-W side, the wind speed was below 3.6 m/s, suggesting that potential sources were not far from the sensor. This area was dominated by a highly-dense residential area. However, from this direction as well, there is one large protocol road (Tanjung Perak Barat St.) with the final destination the Tanjung Perak Port. Therefore, in this zone, the sources of PM_{10} came mainly from traffic, residential, and business-related activities.

The next station is the Sukomanunggal site, located in a suburban area on the west side of Surabaya (Figure 5.1c). Based on the period of this study, the wind flew from the E side, which was dominated by residential areas (Figure 5.20c). When the NO_2 concentrations exceeded the threshold ambient value, the wind mostly came from the E side, as well (Figure 5.21c). In this direction, in front of the monitoring site, a busy street (Sukomanunggal Jaya Rd.) exists, with vehicles using that road to go through the west area of Surabaya. When concentrations were high, the wind speed was relatively slow. The highest winds were between 2.1-3.6 m/s. Therefore, we can conclude that the majority of the cause of high levels of NO_2 was from traffic flow. However, we observed somewhat different causes for PM_{10} concentrations. When the PM_{10} concentrations exceeded the threshold ambient level, the direction of the wind came largely from SW-W sites (Figure 5.22c), where on that particular location, there are offices, green fields, modern shopping malls, apartments, towers, and also housing areas. However, no factories exist in this zone. We observed very high wind speed intervals (5.7-8.8 m/s) that carried PM_{10} pollutants to the sensor.

The Gayungsari monitoring site is located in the south of Surabaya. The surroundings of this site are shown in Figure 5.1d. This monitoring site is located ~350m from the highway. Thus, it refers to a zone near the highway. During the observation period, wind direction was mostly from the SE-E, where the location is dominated by a housing area. When the NO_2 (Figure 5.21d) and PM_{10} (Figure 5.22d) concentration exceeded the threshold level, the dominant wind direction came from the SW-W, where it lays near a highway connecting neighboring cities from and to Surabaya. This result is as expected, as NO_2 and PM_{10} come mainly from traffic in this region. For PM_{10}, there are times when the concentration exceeded the threshold ambient level when the wind speed was between 3.6-5.7 m/s, which was high.

The monitoring site in Sukolilo is located east of Surabaya, a suburban area, and surrounded by a big convention hall, small hospital, houses, universities, and offices. Even though the location is in a suburban area, the surroundings of the sensor were not dominated by residential (houses) but instead by businesses and offices. If we look at the wind direction over the data-collection period, we observed that the wind came largely from the SE-E, where the wind speed was between 0.5-5.7 m/s. When the NO_2 concentration exceeded the threshold level, the most dominant wind direction was from the SE-E, with the highest wind speed was observed when the exceeding concentrations were between 3.6-5.7 m/s (Figure 5.21e). In this direction are offices as well as roads connecting to some universities, and where the Arief Rahman Hakim Rd. is the main road for these universities (Figure 5.1e). There are also some residential areas on the east side of Surabaya. Opposite this, when PM_{10} concentrations exceeded the threshold ambient level, wind direction came mostly from the S-W and partially from the SW direction (Figure 5.22e). There were 42 events recorded where PM_{10} concentrations were higher than the threshold ambient level. The area from which the wind came consisted of houses, a small clinic, and universities. The wind speed was also up to the interval of 3.6-5.7 m/s (S-W direction) and >11.1 m/s (SE direction), as shown by Figure 5.22e.



(a) City Center

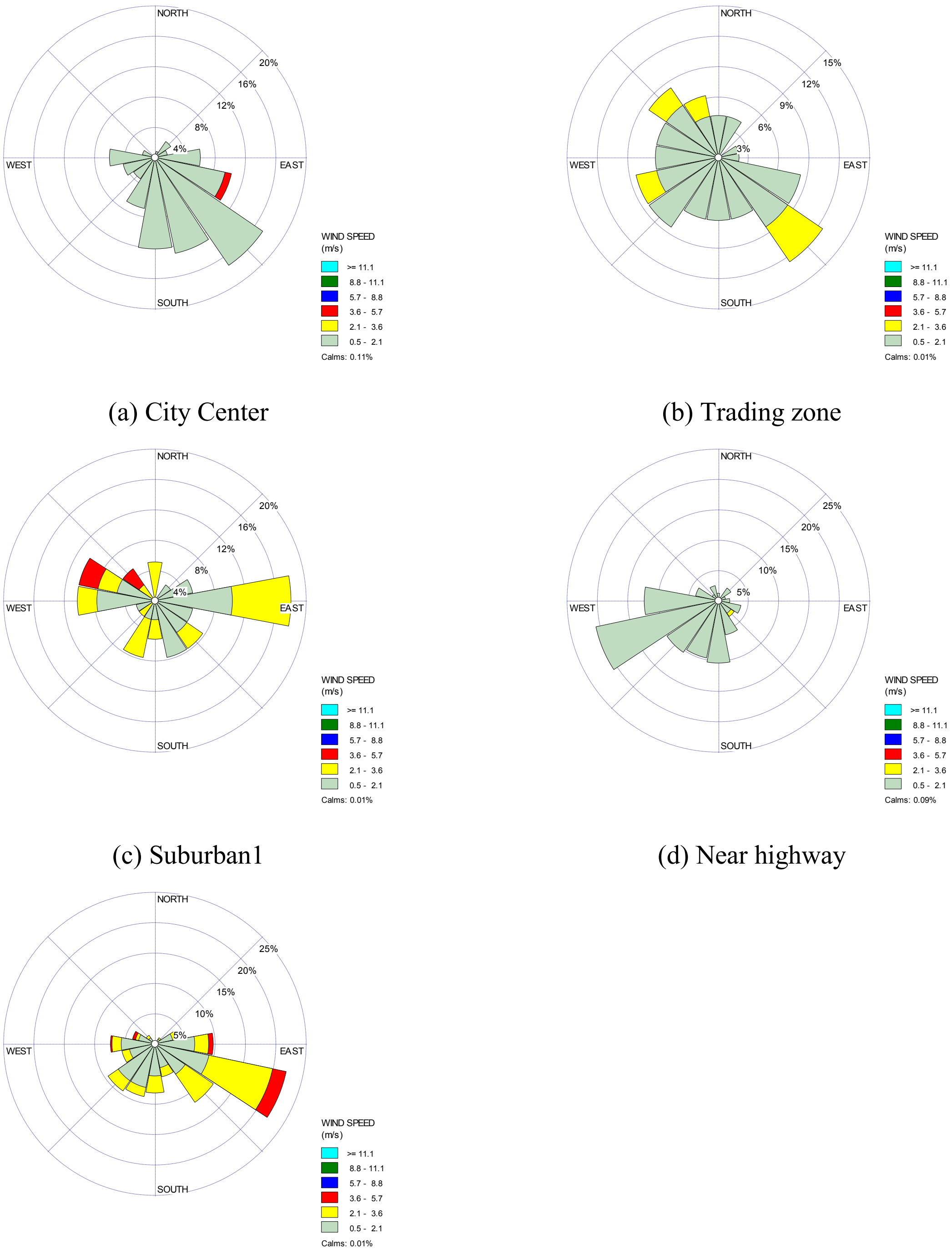
(b) Trading zone

(c) Suburban1

(d) Near highway

(e) Suburban2

Figure 5.20 Historical wind direction and wind speed of captured by five monitoring stations



(a) City Center

(b) Trading zone

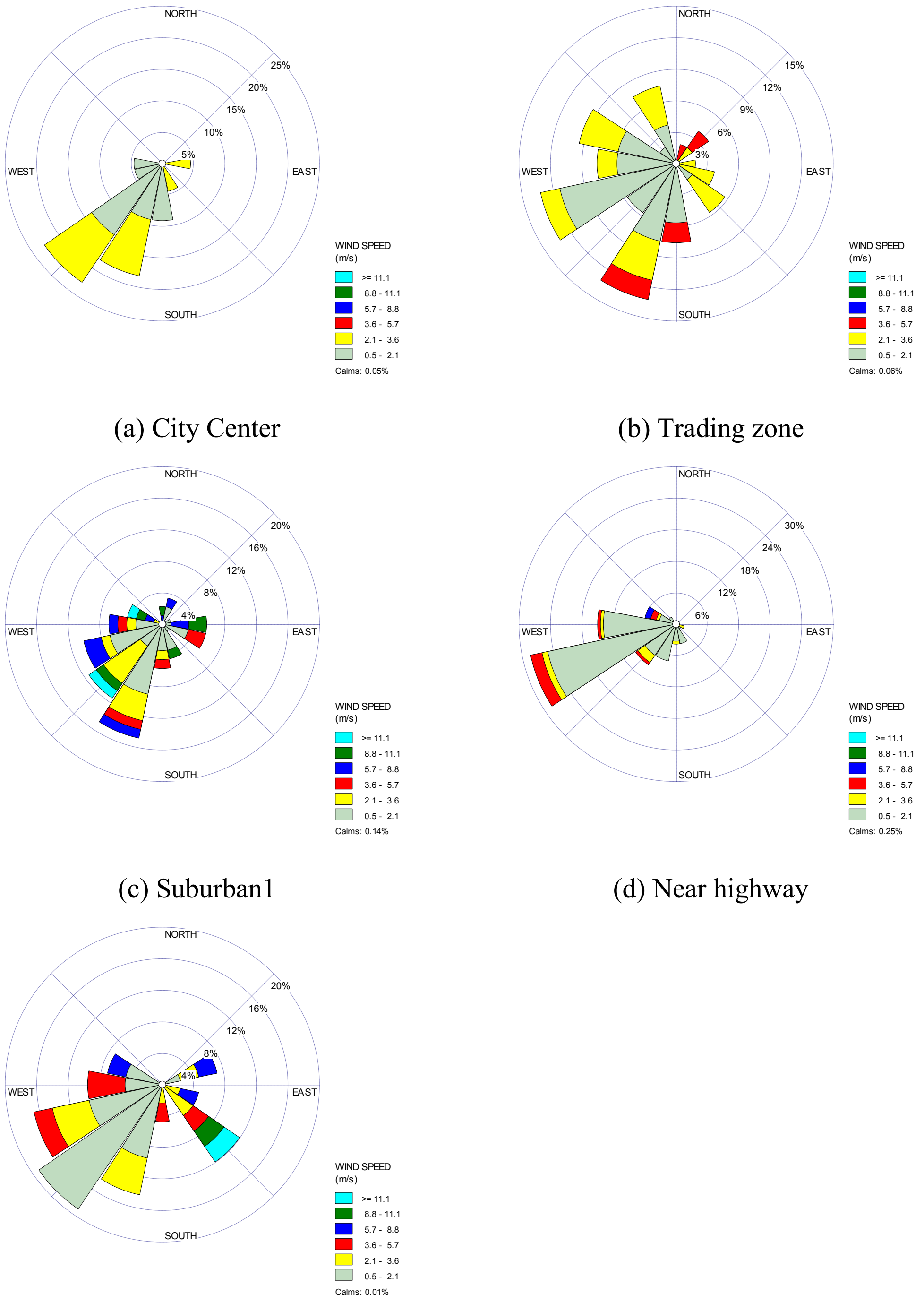
(c) Suburban1

(d) Near highway

(e) Suburban2

Figure 5.21 Historical wind direction and wind speed captured by five monitoring stations when the NO_2 concentration exceeded threshold value

The figure below is PM_{10}, only for all stations exceeding values.



(a) City Center

(b) Trading zone

(c) Suburban1

(d) Near highway

(e) Suburban2

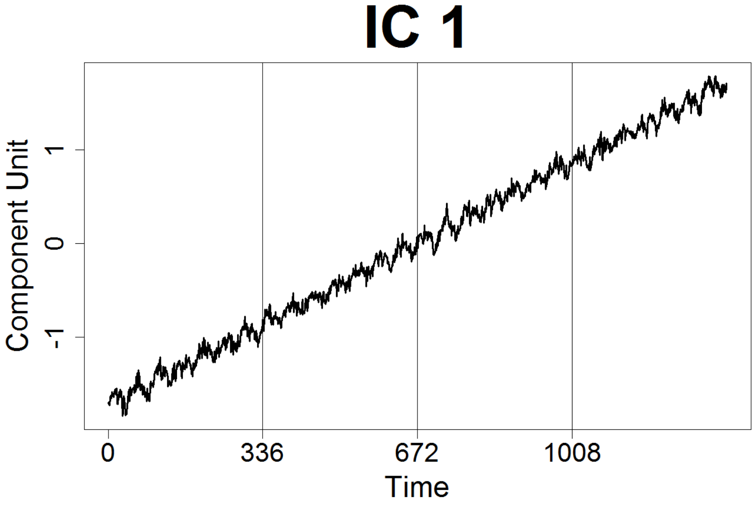
Figure 5.22 Historical wind direction and wind speed captured by five monitoring stations when the PM_{10} concentration exceeded threshold value.

5.4.6 Seasonal Temporal Distribution using 4 week-cycle and Annual Time Scale

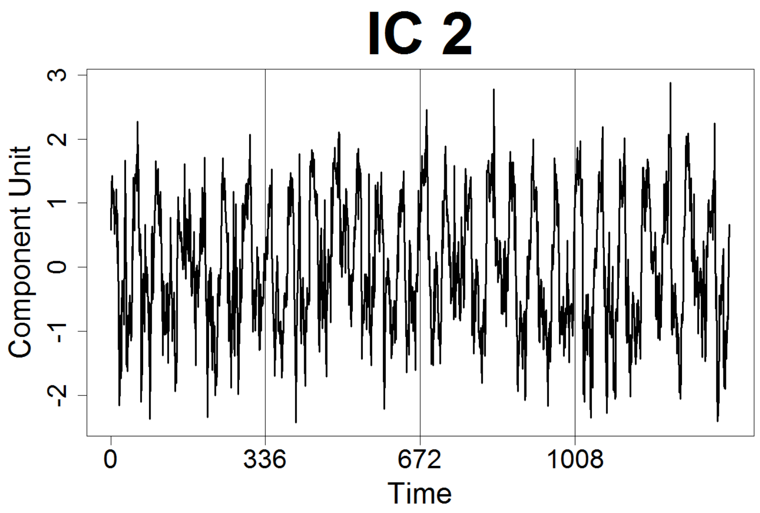
Among ways to interpret and study the phenomena of air quality by observing air quality data from monitoring stations, one crucial concern is how we select the best time-scale to obtain results that are meaningful. In the last section, we showed the results of ICA, using weekly-scale figures. In this section, we are showing results from 4-week-cycle data (Figure 5.23 to Figure 5.28) and annual information using daily-average concentrations (Figure 5.29 to Figure 5.30). The goal is to compare the results between different time scales, and to decide which one is the best to extract the needed information that is a temporal pattern.

Firstly, we discuss the distribution found for NO_2. One interesting pattern we found was the temporal pattern, which was increasing in the city center, not only for NO_2 but also for all pollutants. This pattern was observed on Figure 5.23a (NO_2), Figure 5.24a (PM_{10}), Figure 5.25a (O_3), Figure 5.26b (NO), Figure 5.27b (CO), and Figure 5.28e (SO_2). This pattern may explain why the pollutant levels were increasing following a 4-week cycle. A daily cycle was also found on IC2 (Figure 5.23b), which was similar in results. This is shown on IC1 NO_2 in Figure 5.14a. The discussion will be similar to that previously shown, that is, in general, we observe two-peak waves within a day, and the pattern that concentrations were lower on Sunday. The main disadvantage of using monthly-scale air quality is that we cannot obtain distribution in the trading zone, near the highway zone, and for Suburban2. Furthermore, ICs are dominated by the city center pattern as shown by Table 5.16, leading to the conclusion that by using this time scale, we were not able to obtain important information for all sites and for all pollutants.

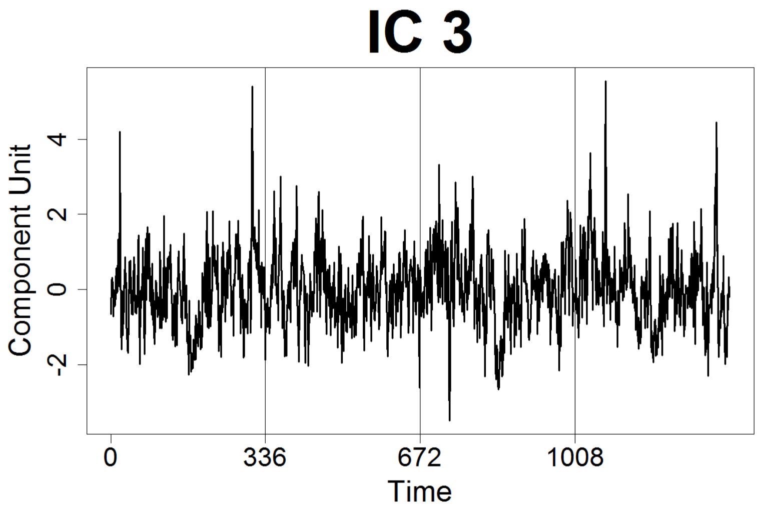
Figure 5.24 shows the ICA results for PM_{10}. We found similar trends like NO_2 (IC1), and we noted that there were spikes of PM_{10} concentration on week 2 and week 3, as shown by IC2 (Figure 5.24b). We also noted several spikes for the city center and near highway zone on week 1 and week 4 (IC5). The rest of the pollutants have similar patterns as shown by Figure 5.25 (O_3), Figure 5.26 (NO), Figure 5.27 (CO), and Figure 5.28 (SO_2).



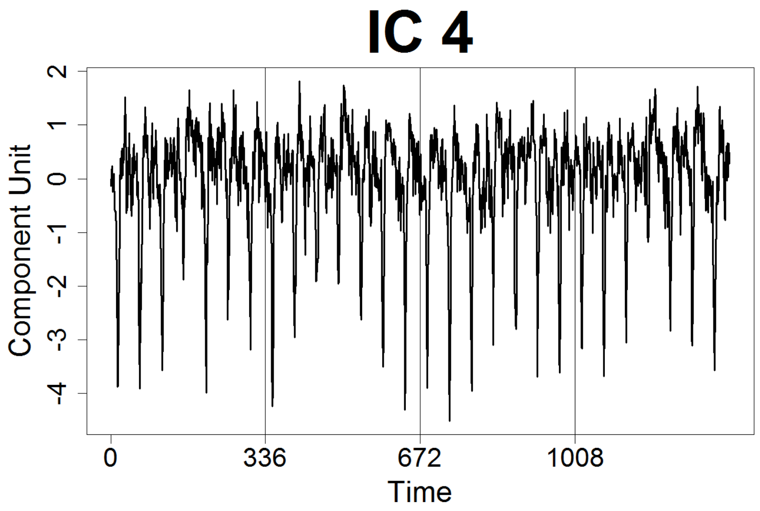
(a) IC1



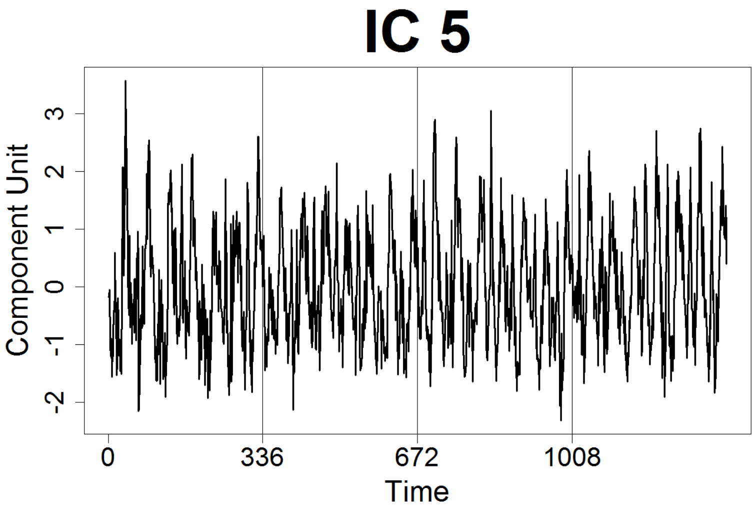
(b) IC2



(c) IC3



(d) IC4

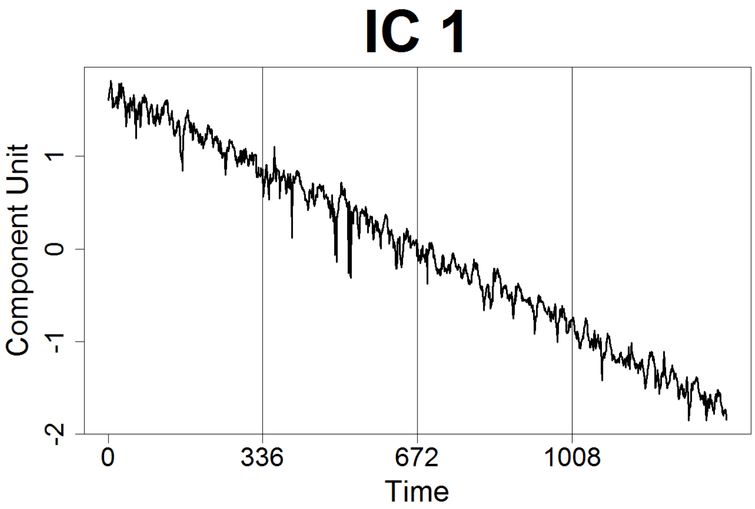


(e) IC5

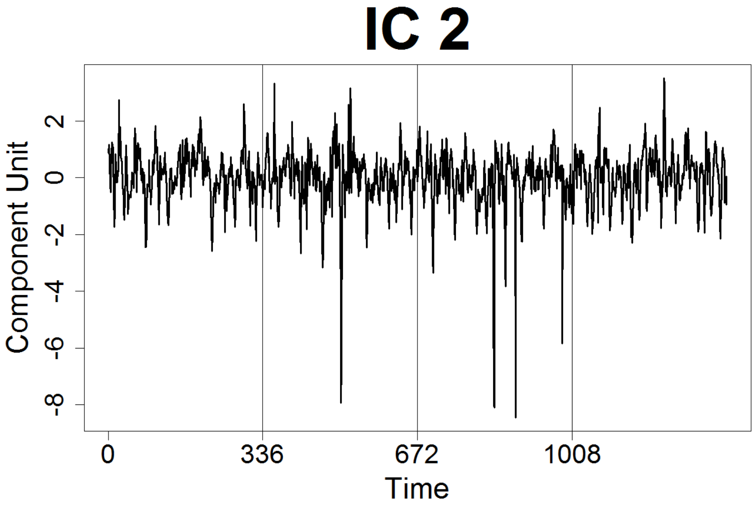
Figure 5.23 Five independent component profiles of NO_2 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments which means 4-week cycle, y-axis: component value.

Table 5.16 Weight loadings for independent components of 4-week cycle NO_2 concentration

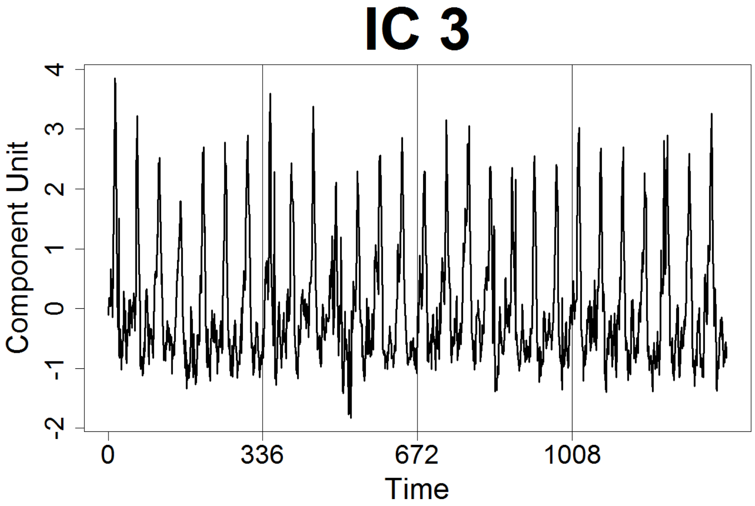
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**386.824**	-4.146	**9.589**	**15.055**	**23.649**
Trading (Perak)	0.550	-1.206	3.259	-5.529	2.684
Suburban1 (Sukomanunggal)	0.076	**-6.185**	0.474	-2.439	6.275
Near Highway (Gayungsari)	0.201	0.426	1.219	-7.587	7.322
Suburban2 (Sukolilo)	0.499	1.371	-0.168	-7.592	1.249



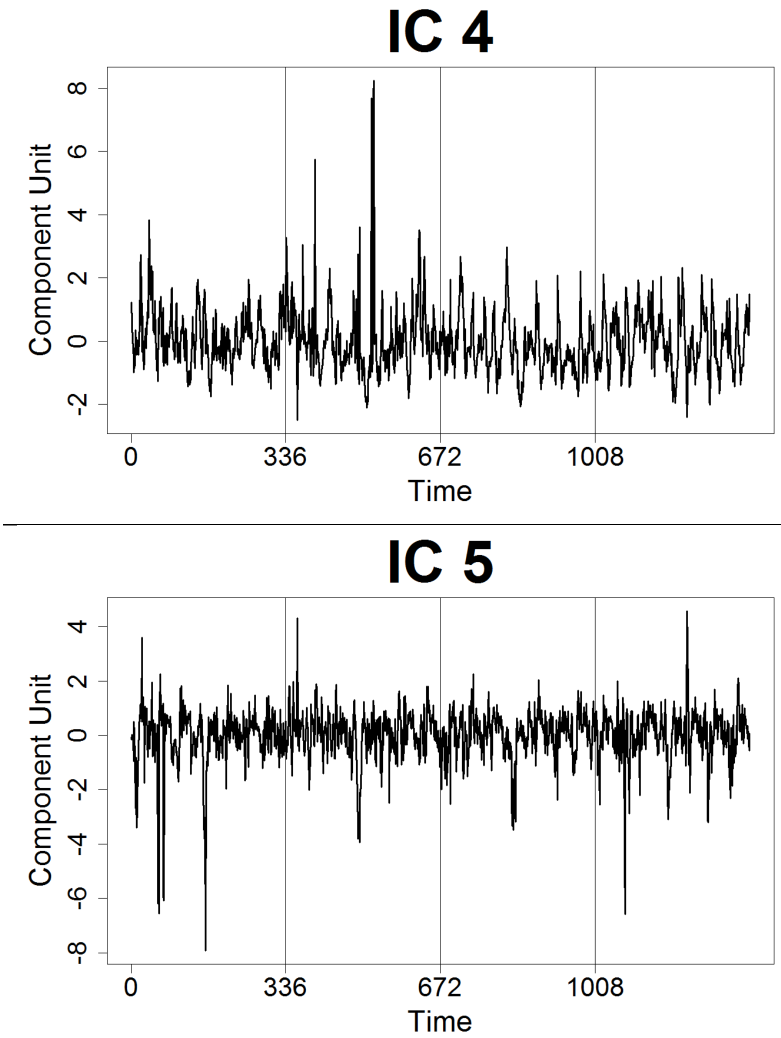
(a) IC1



(b) IC2



(c) IC3



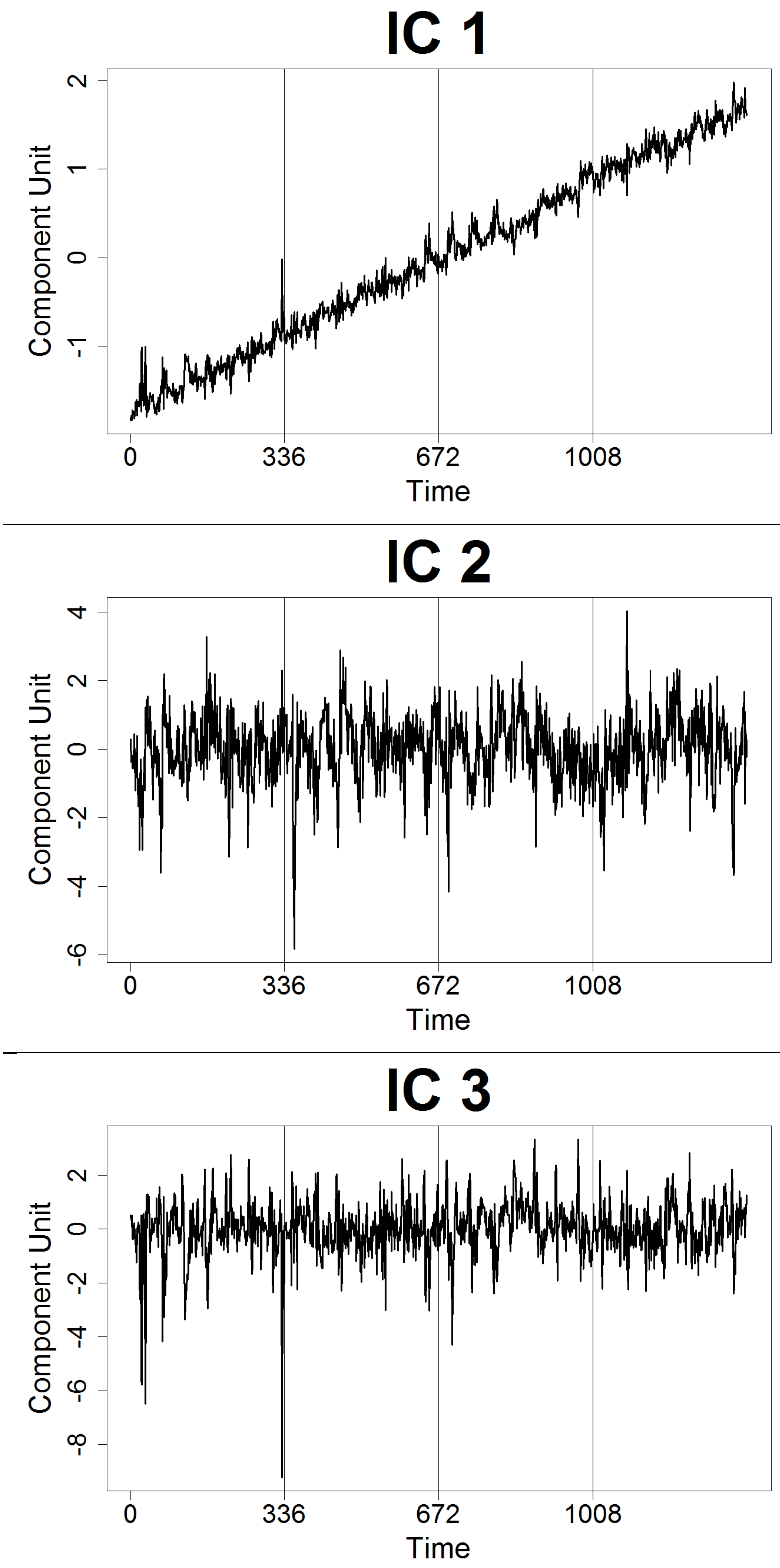
(d) IC4

(e) IC5

Figure 5.24 Five independent component profiles of PM_{10} concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value.

Table 5.17 Weight loadings for independent components of 4-week cycle PM_{10} concentration

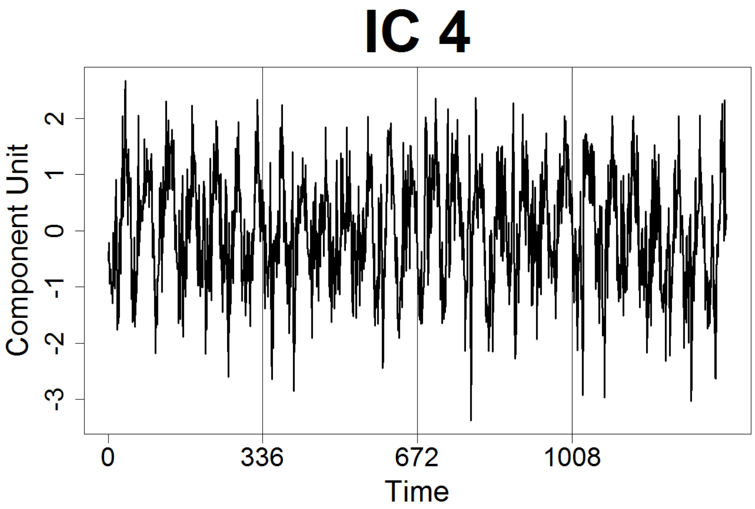
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**-385.594**	**-9.892**	-12.107	**-34.523**	**20.220**
Trading (Perak)	-1.771	**-8.049**	**15.470**	5.113	-2.317
Suburban1 (Sukomanunggal)	-1.579	2.611	8.388	12.548	-1.989
Near Highway (Gayungsari)	-2.025	4.330	**24.050**	4.194	**-18.207**
Suburban2 (Sukolilo)	-1.368	5.450	**21.203**	1.082	-1.798



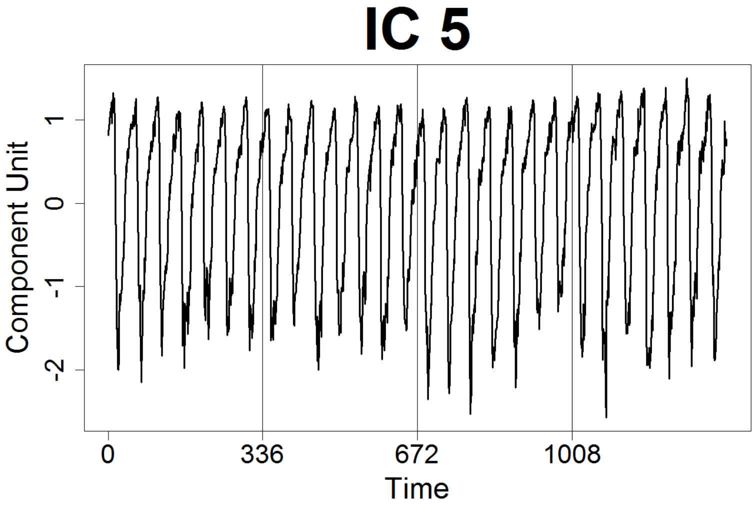
(a) IC1

(b) IC2

(c) IC3



(d) IC4

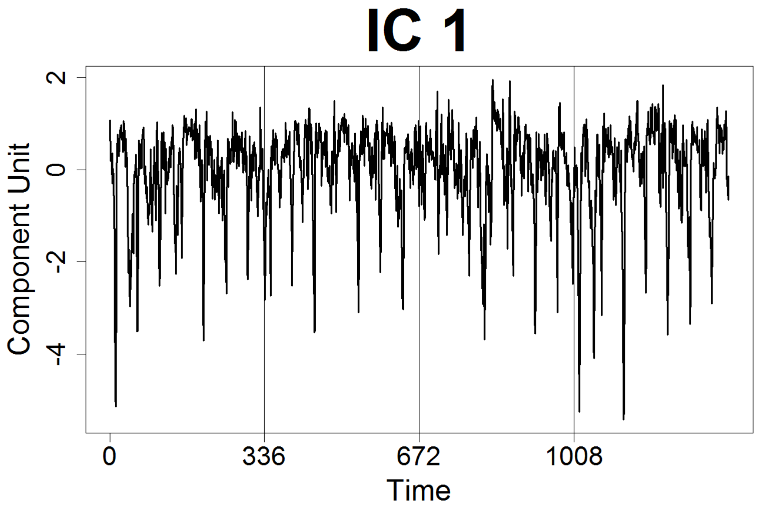


(e) IC5

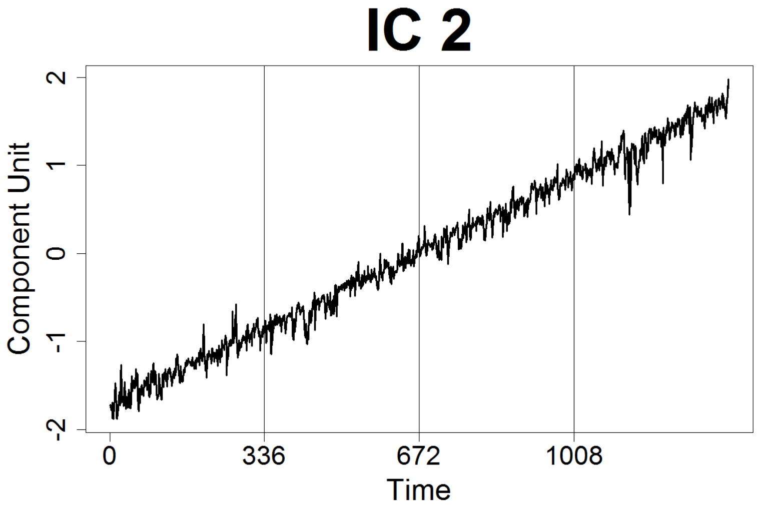
Figure 5.25 Five independent component profiles of O_3 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value.

Table 5.18 Weight loadings for independent components of 4-week cycle O_3 concentration

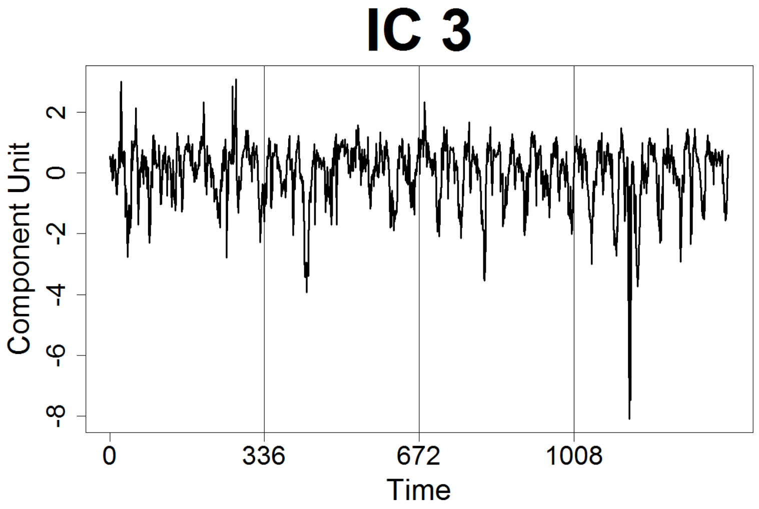
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**385.219**	**11.991**	**40.248**	**-18.765**	4.257
Trading (Perak)	0.498	-0.371	-0.616	0.885	-18.291
Suburban1 (Sukomanunggal)	0.528	-5.209	-1.515	-2.974	**-23.142**
Near Highway (Gayungsari)	0.905	-2.704	-4.557	-1.021	**-20.893**
Suburban2 (Sukolilo)	0.495	-3.827	-1.656	4.979	-16.956



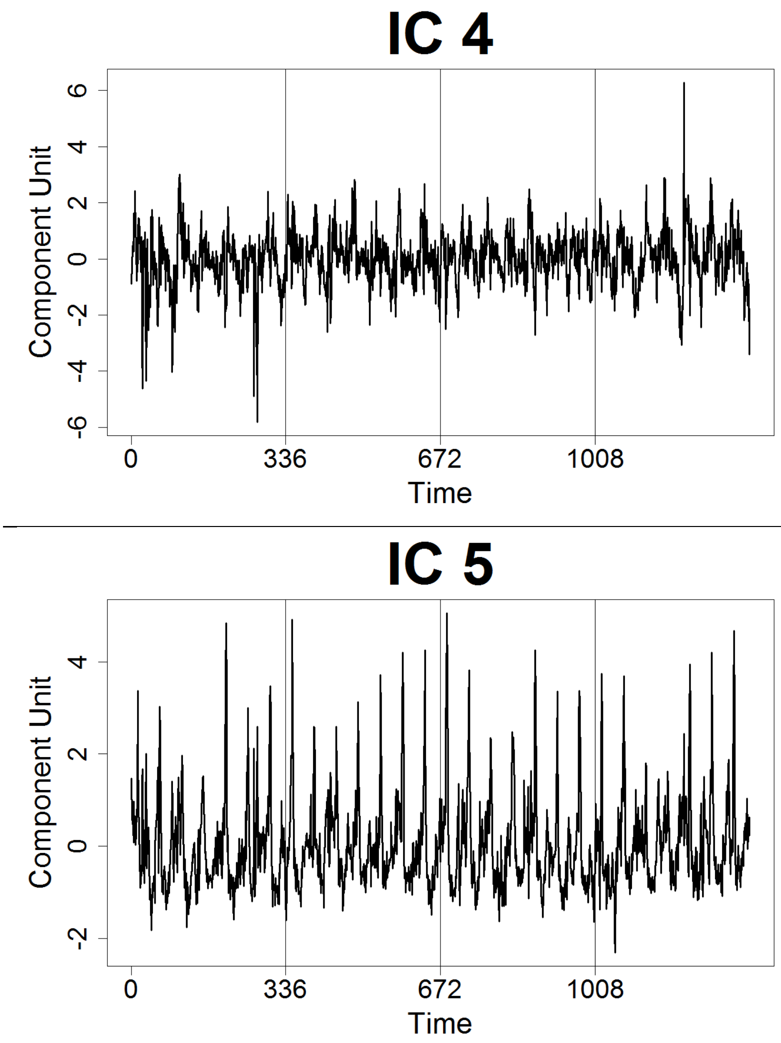
(a) IC1



(b) IC2



(c) IC3



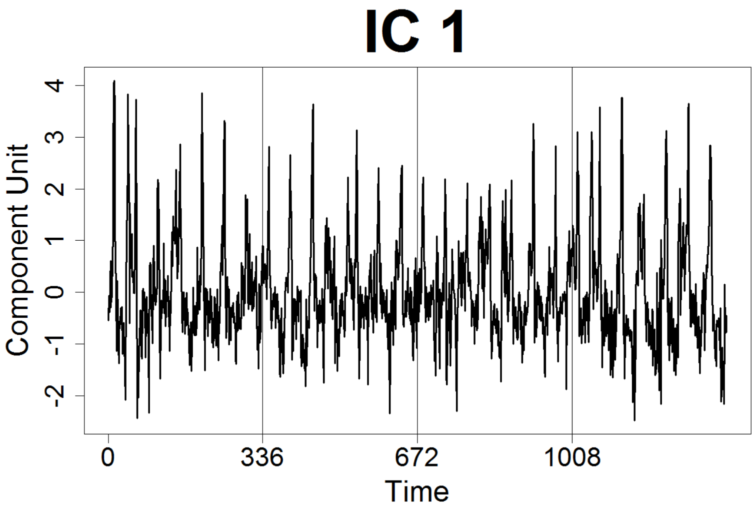
(d) IC4

(e) IC5

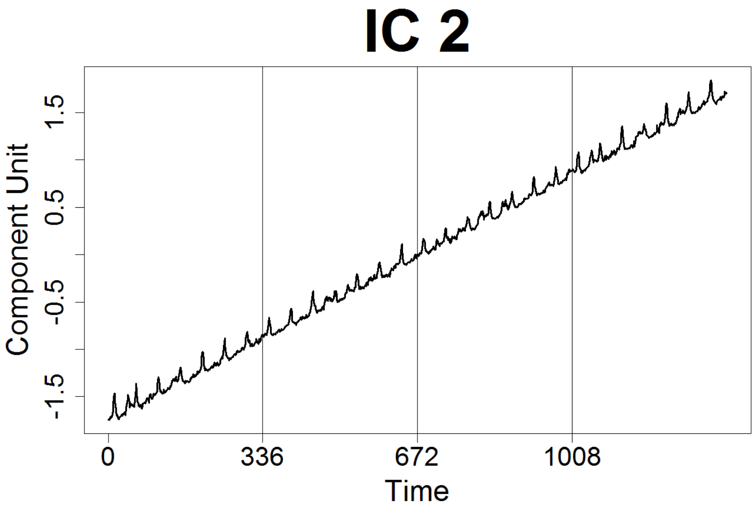
Figure 5.26 Five independent component profiles of NO concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value.

Table 5.19 Weight loadings for independent components of 4-week cycle NO concentration

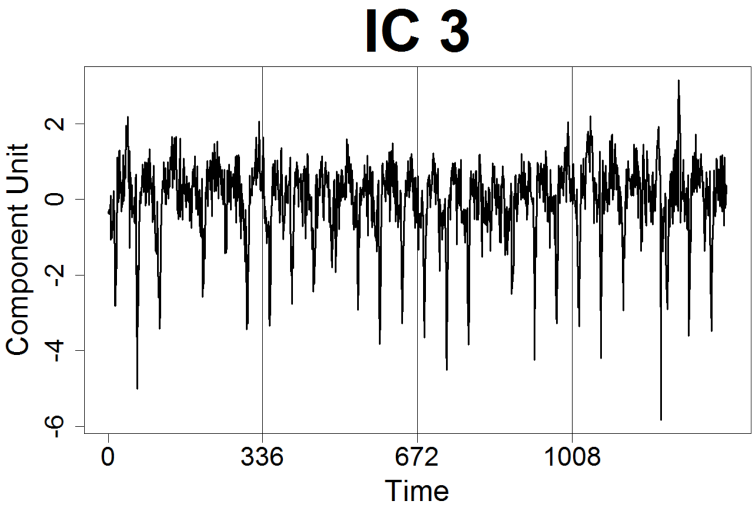
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**13.914**	**385.695**	**-31.884**	**20.743**	**11.258**
Trading (Perak)	-7.831	0.359	-0.689	1.171	4.215
Suburban1 (Sukomanunggal)	-0.910	-0.042	-3.907	-1.242	1.242
Near Highway (Gayungsari)	-8.082	-0.155	-0.470	-2.208	**11.619**
Suburban2 (Sukolilo)	-2.014	0.018	0.008	2.309	4.315



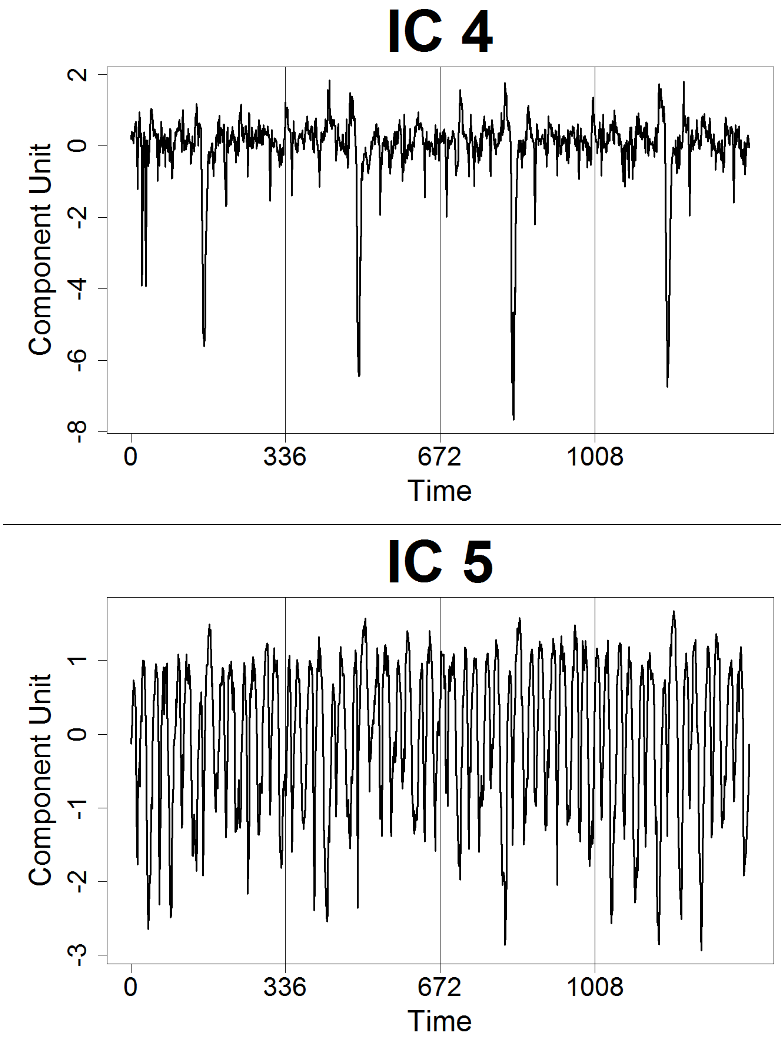
(a) IC1



(b) IC2



(c) IC3



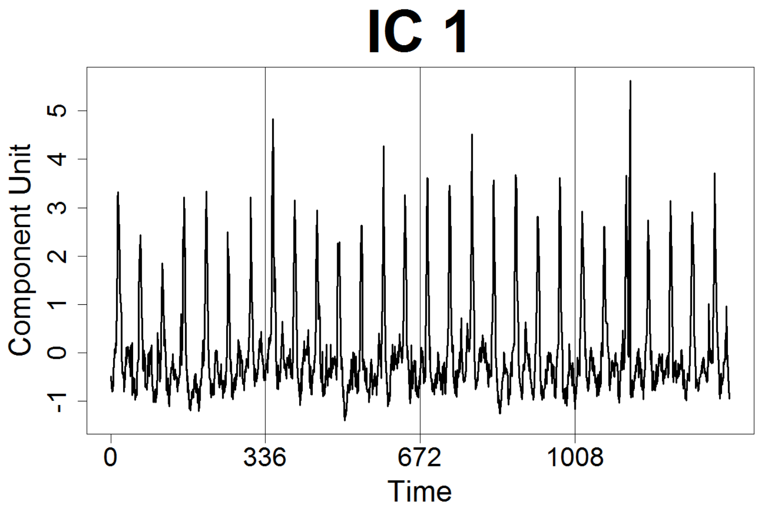
(d) IC4

(e) IC5

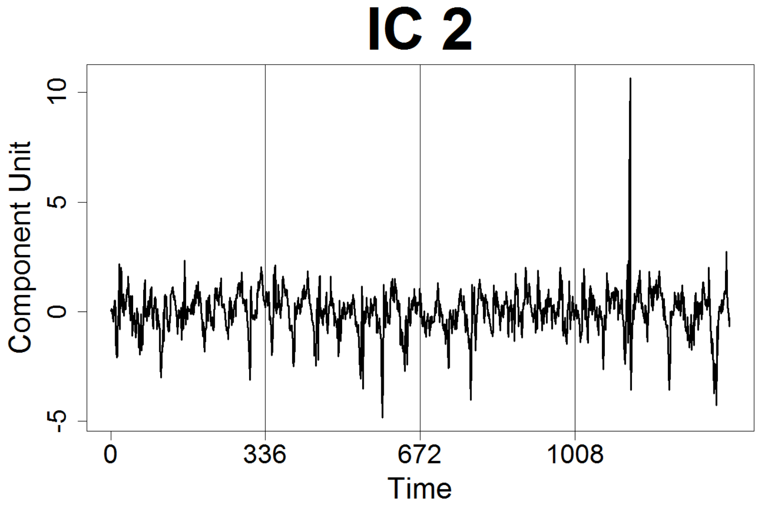
Figure 5.27 Five independent component profiles of CO concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value.

Table 5.20 Weight loadings for independent components of 4-week cycle CO concentration

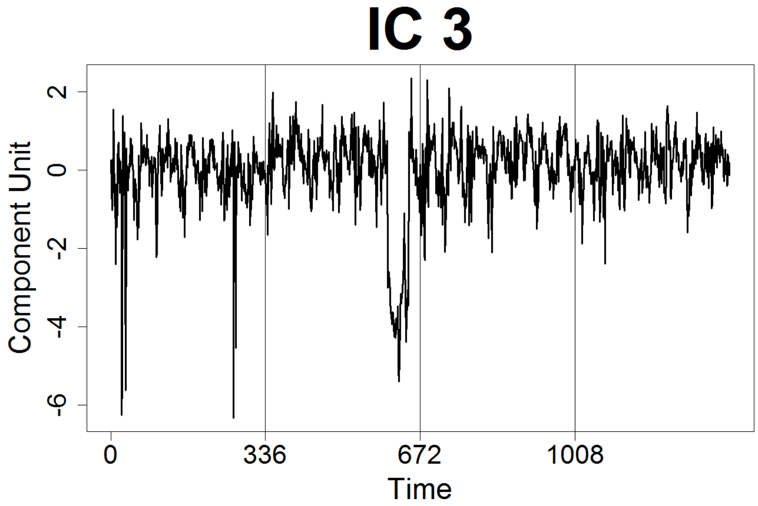
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**-17.541**	**387.413**	**8.486**	**2.736**	**7.228**
Trading (Perak)	0.215	0.039	-0.148	-0.048	-0.294
Suburban1 (Sukomanunggal)	-0.026	0.022	-0.004	-0.009	-0.296
Near Highway (Gayungsari)	0.130	0.035	-0.215	-0.530	-0.495
Suburban2 (Sukolilo)	0.074	0.032	-0.239	-0.058	-0.190



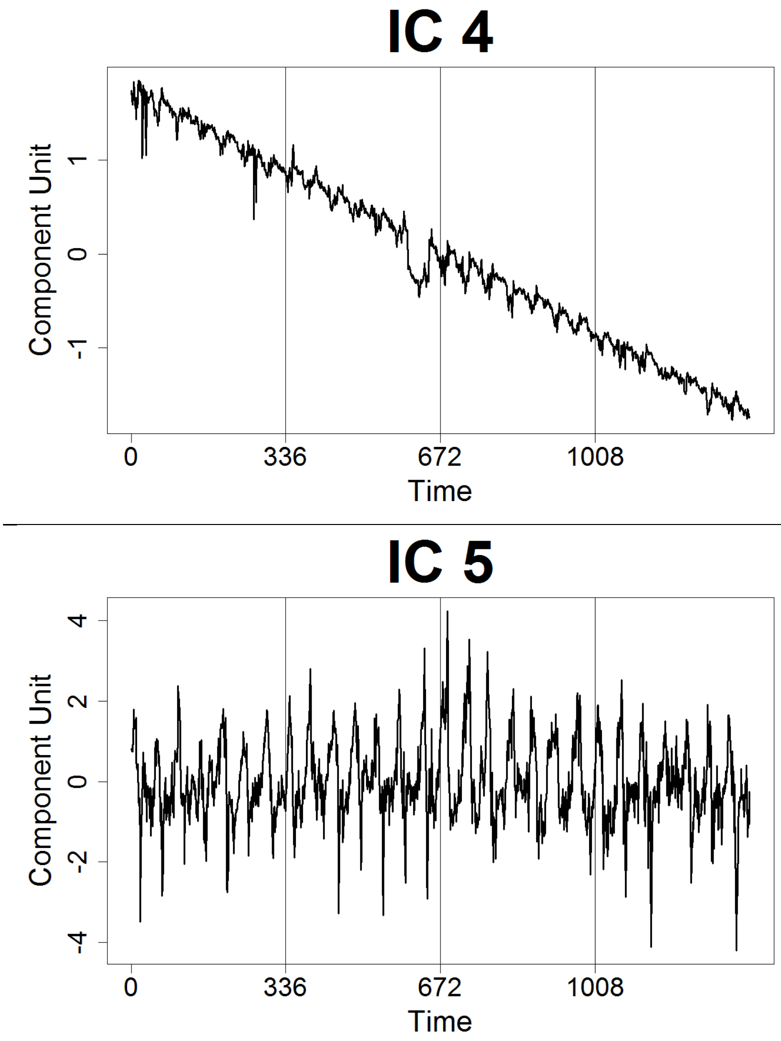
(a) IC1



(b) IC2



(c) IC3



(d) IC4

(e) IC5

Figure 5.28 Five independent component profiles of SO_2 concentration over five monitoring stations; x-axis: time (in 30-mins), divided into 4 segments, which means a 4-week cycle, y-axis: component value.

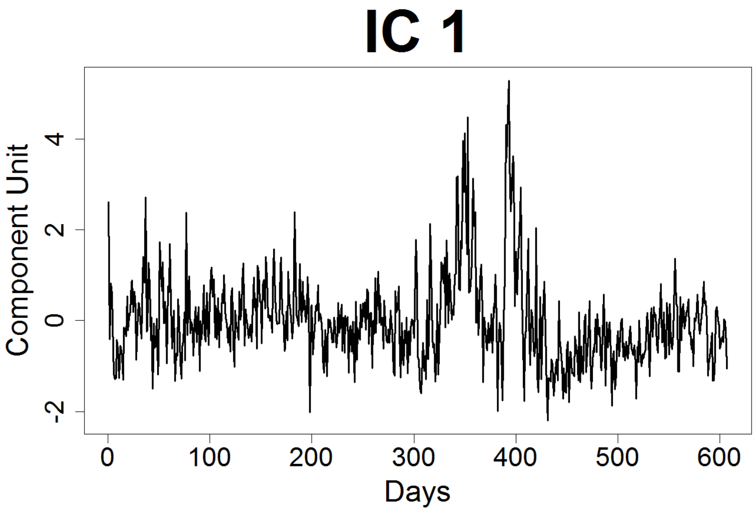
Table 5.21 Weight loadings for independent components of 4-week cycle SO_2 concentration

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**7.095**	1.602	**41.971**	**-385.441**	**-12.203**
Trading (Perak)	4.730	-2.267	0.095	0.080	-2.687
Suburban1 (Sukomanunggal)	4.952	1.584	-0.070	0.094	-2.151
Near Highway (Gayungsari)	**7.689**	**-5.161**	-8.043	-0.416	4.739
Suburban2 (Sukolilo)	6.535	-3.717	0.171	-0.019	3.186

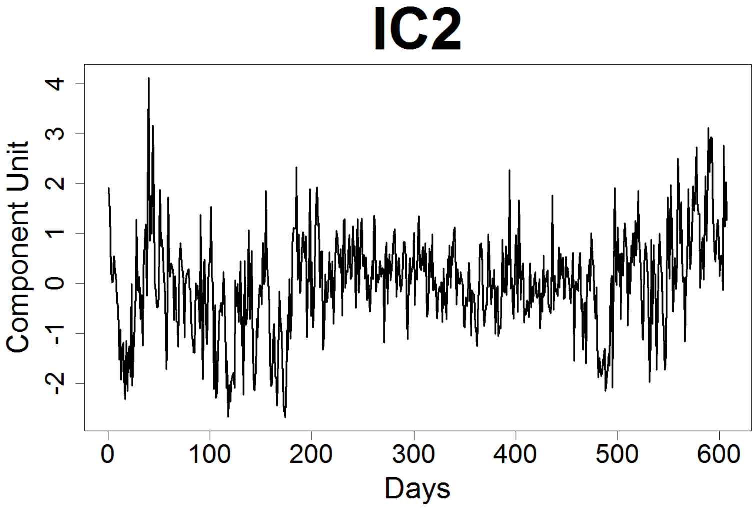
Overall, it is difficult to obtain significant information from 4-week-cycle data because: 1) the weight loadings were not distributed equally, and most ICs only explain the city center pattern (Table 5.16-Table 5.21), and 2) the distribution pictures were difficult to interpret clearly. It is, therefore, clear that using this level of data may not be useful compared to using weekly-scale information. This is because the weekly-scale information is richer, and we can draw similar information if we use monthly-scale data. However, one particularly significant result, if we use monthly-scale data, is that we found a pattern where the concentrations of pollutants are, in fact, increasing following a 4-week cycle. Moreover, NO_2 concentration tends to be higher during the wet season than during the dry season.

As mentioned above, we also explored the use of daily aggregated data to investigate the effect of season and how these differ among locations for 2 years of data. By using a 2-year period, we can also observe annual patterns. Based on the result of ICs components (Figures 5.29a-5.29b), we can summarize as follows. Based on IC1 (Figure 5.29a), we noted that there was sharp increase in the city center, trading zone, and Suburban1, west of Surabaya between January 2002 and March 2002. However, this was not seasonal. This event duration was taking place during the rainy season. From Figure 5.29b, in the city center, near the highway zone, and Suburban2 (east of Surabaya) based on weight loadings of IC2 (Table 5.22), there was a sharp increase observed in March 2001, then decreasing between April-July 2001, and increasing again starting in August 2002 through September 2002. Furthermore, based on IC3 (Figure 5.29c), the tendency of concentration was high during August 2001 to December 2001 in these regions: the trading zone and the near highway zone. We observed quite lower trends between January 2002 and March 2002 on the Suburban1 and near the highway zone (Table 5.22), based on Figure 5.29d, representing the IC4 component. The last, seasonal declining trend in the trading zone and Suburban2 started between March and September. All these results indicated that during the dry season, NO_2 concentrations decreased. This result is in agreement with results discussed in Chapter 4. The NO_2 concentrations during the dry season tended to be lower than during the wet (rainy) season.

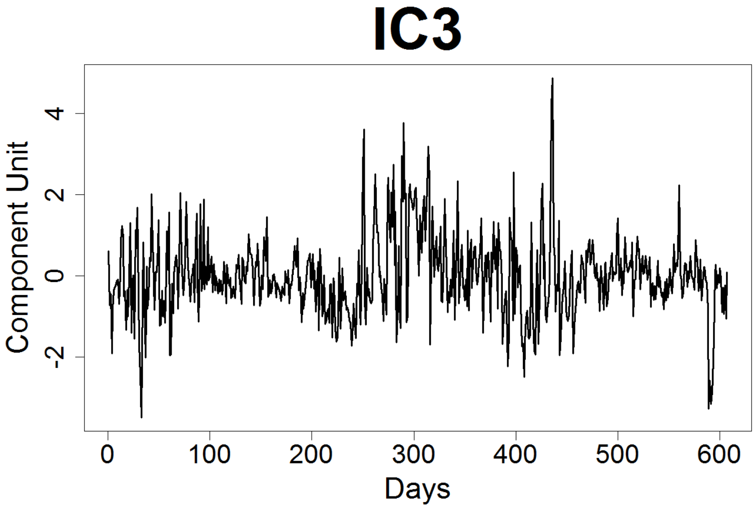
Fluctuation with no apparent pattern for IC1 (Figure 5.30a) showed that this component represented the city center, trading zone, Suburban1, and near highway area (Table 5.23), suggesting a similar concentration pattern in terms of PM_{10} observed. Furthermore, we noted that there was a decrease during wet season (IC2, Figure 5.30), which is in agreement with the results of the multilevel model from Chapter 4. This is due to the washout process by rain (Vallero, 2008), as part of agglomeration and sedimentation of PM_{10}. When the two particulates collide in the air, because of attractive surface forces, the two particles will adhere and attach to each other, therefore forming larger particles by agglomeration. The larger particles become, the greater their weight, and, therefore, are likely to fall to the ground. This pattern was apparent in the Suburban1 Sukomanunggal area (Table 5.23).



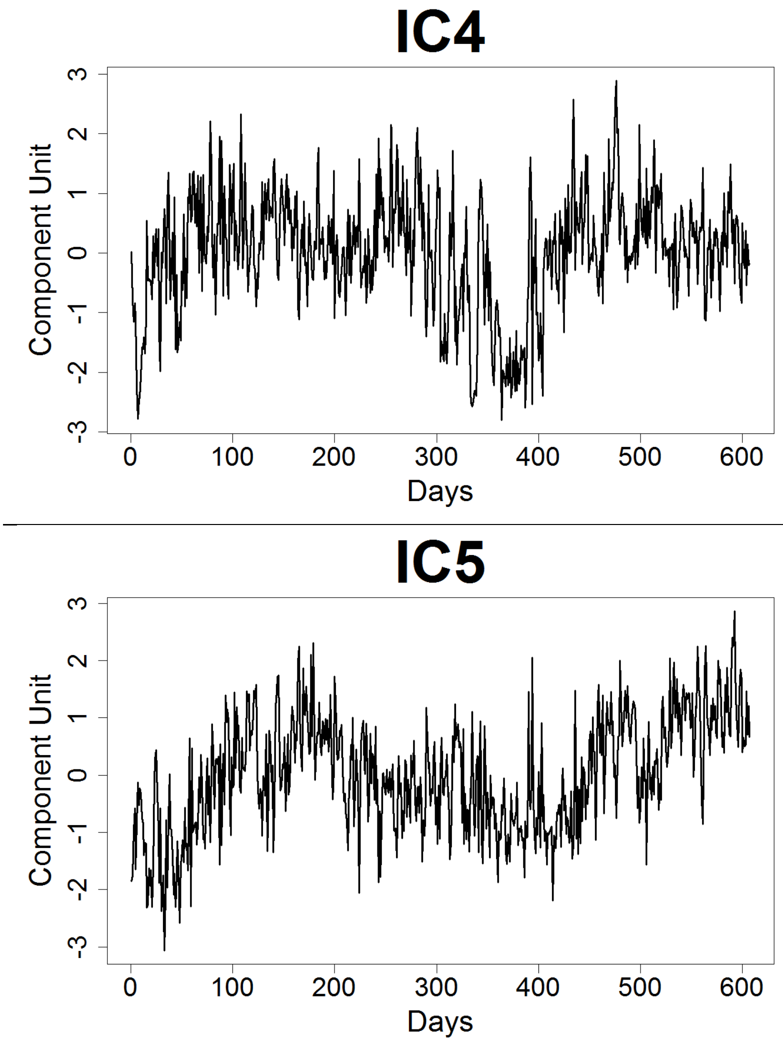
(a) IC1



(b) IC2



(c) IC3



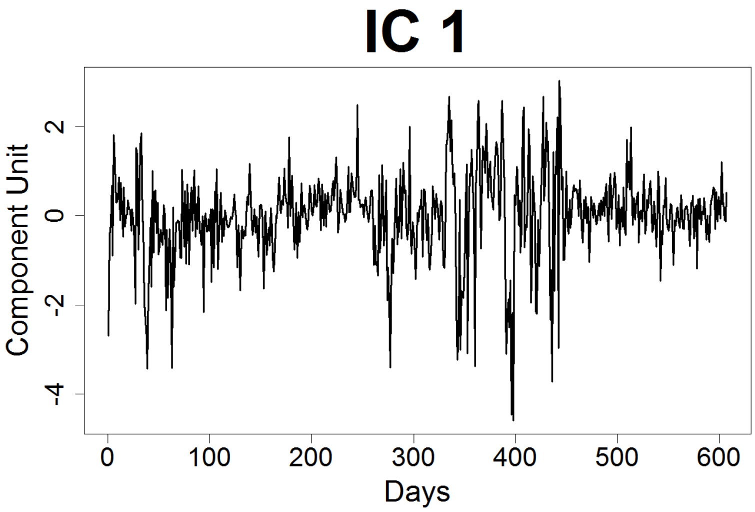
(d) IC4

(e) IC5

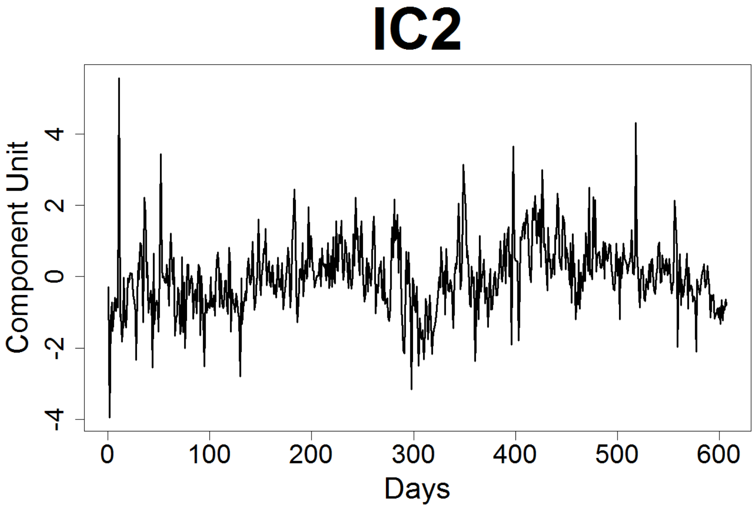
Figure 5.29 Five independent component profiles of daily average NO_2 concentration over five monitoring stations; x-axis: time (days), y-axis: component value. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry

Table 5.22 Weight loadings for independent components of daily average NO_2 concentration

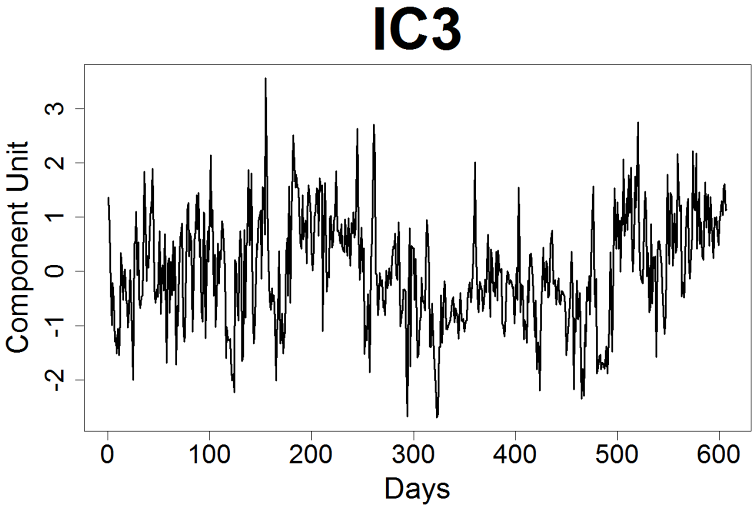
Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**4.089**	**5.632**	3.083	2.266	0.290
Trading (Perak)	**3.881**	2.828	**5.271**	-0.276	**-5.081**
Suburban1 (Sukomanunggal)	**3.018**	1.336	0.358	**5.753**	-0.506
Near Highway (Gayungsari)	-1.178	**4.462**	**4.219**	**6.207**	1.307
Suburban2 (Sukolilo)	0.265	**4.696**	0.577	0.199	**-3.668**



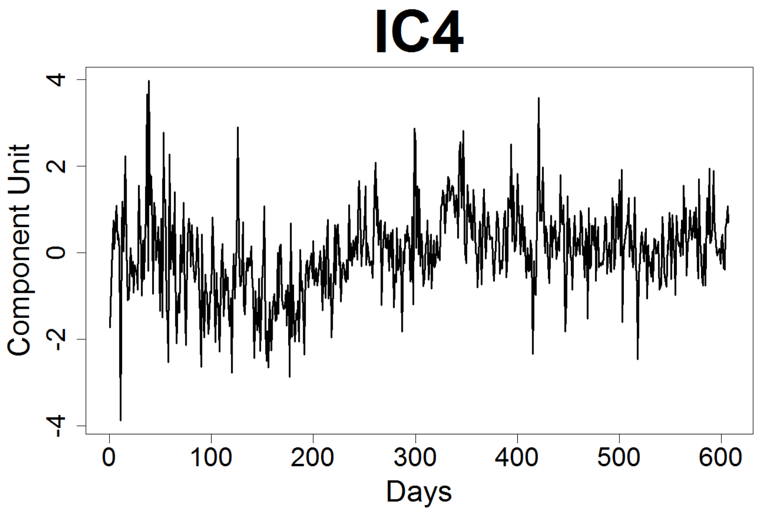
(a) IC1



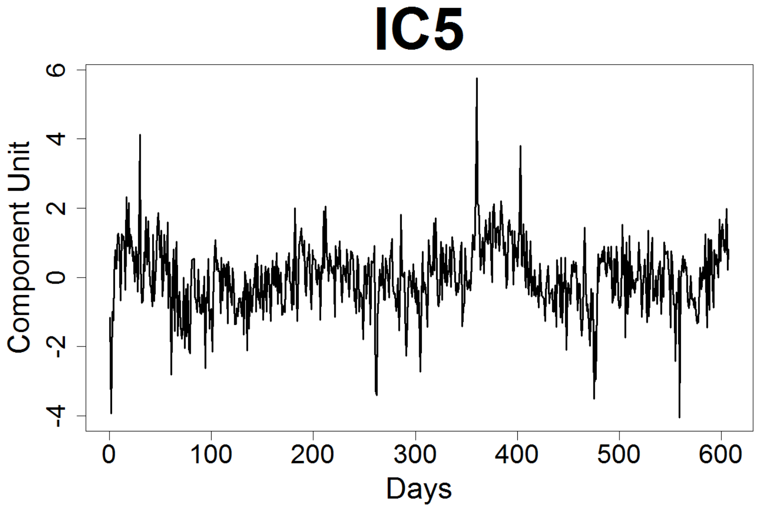
(b) IC2



(c) IC3



(d) IC4



(e) IC5

Figure 5.30 Five independent component profiles of daily average PM_{10} concentration over five monitoring stations; x-axis: time (days), y-axis: component value. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry

Table 5.23 Weight loadings for independent components of daily average PM_{10} concentration

Sites	IC1	IC2	IC3	IC4	IC5
City Center (Ketabang Kali)	**-8.338**	2.654	9.061	**5.878**	-3.130
Trading (Perak)	**-8.055**	0.473	**14.554**	0.122	2.359
Suburban1 (Sukomanunggal)	**-7.324**	**11.704**	10.884	-3.388	-4.329
Near Highway (Gayungsari)	**-7.328**	-4.421	**13.617**	-3.979	**-12.201**
Suburban2 (Sukolilo)	1.252	1.657	**12.694**	**4.220**	-2.858

Daily data collected over a two-year period is useful if we want to find seasonal patterns and how seasons affect air quality. Here we have shown how NO_2 and PM_{10} change during dry and wet seasons. However, if we focus on determining the temporal pattern on a daily basis, using 30-mins data on a weekly-cycle as shown above, we find this better and adequate to capture all of the necessary information.

5.5 Dynamic of Pollutants and Meteorological Factors using State-Space model

5.5.1 The Dynamic of Pollutants

To observe and investigate the dynamic of pollutant's concentration over five different zones, we employed dynamic linear models (DLMs) by treating the parameters in the regression model as time-varying. The function of the dynamic linear model is to obtain the dynamic of parameters, especially meteorological parameters, to the concentration of pollutants. In our first discussion, we discussed data taken from 30-minute aggregate data collected over two years. We define the dynamic by observing the fluctuation of regression parameters of each parameter. Furthermore, even though the unit scale may be different for each figure, we define fluctuation to be high if there is change in a scale of 1-10 in a figure. The changing below 1-point unit is not considered as highly dynamic even though there is the dynamic pattern. In this case, we only view and focus on the temporal time of the change.

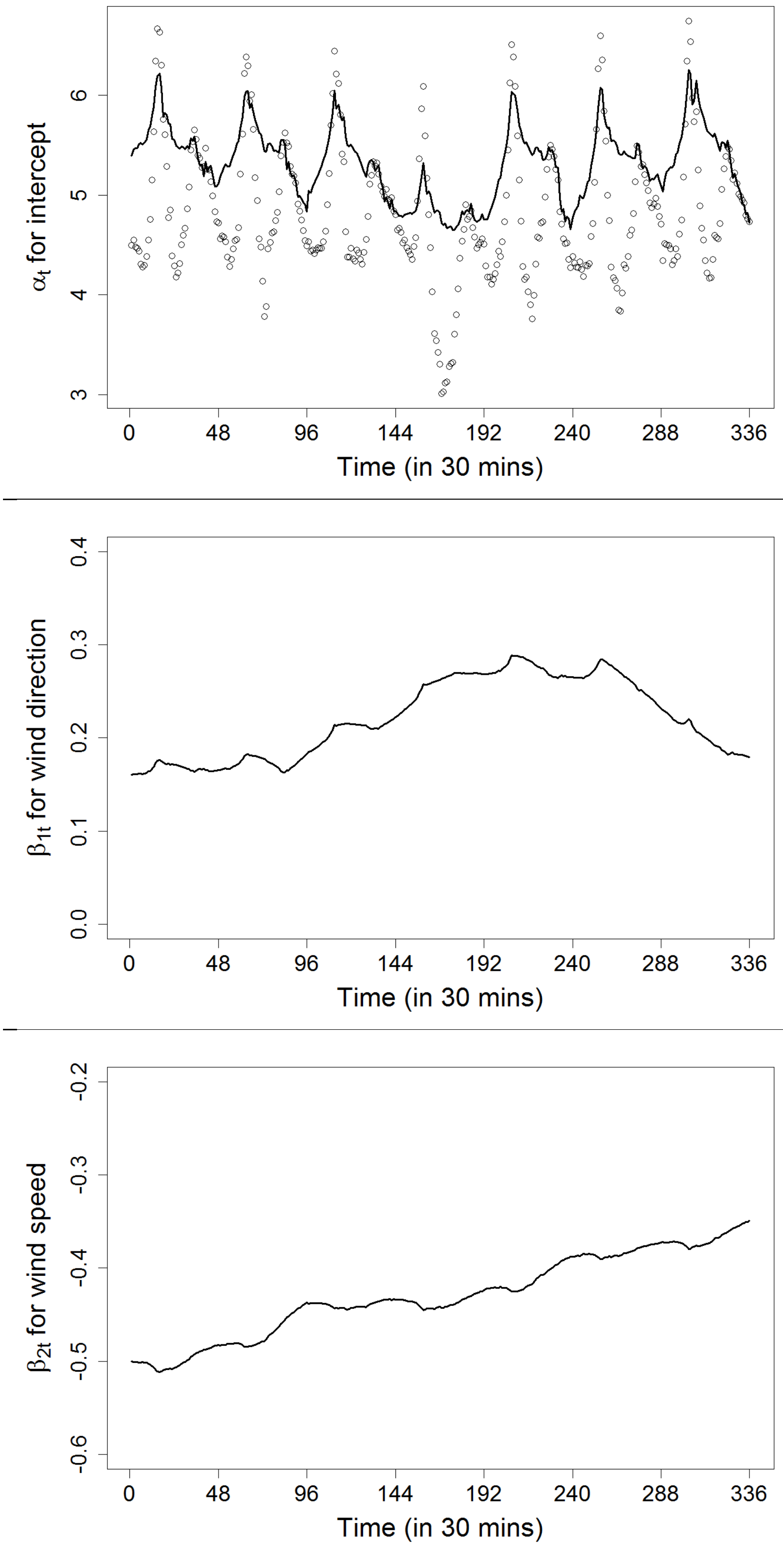
The estimated parameters for intercept and independent variables are shown in Figure 5.31 for NO_2. We noted that the intercept was much more dynamic over time. Besides the intercept, the effect of temperatures was also very much dynamic, whereas the dynamic of other variables were very small. We may observe that the dynamic of NO_2 concentrations were relatively low on Sunday. The patterns were similar with other days including Saturday. Figure 5.31 suggests that the magnitude of NO_2 concentrations was higher during the morning session than in the evening. If this phenomenon is going to be translated into policy, this indicates that the policy on the trading zone must take this fact into account. For example, the policy is time-sensitive, which means that its implementation in the morning session must be emphasized.

In the trading zone, the activities include business-related activities and traffic. This zone is near the second biggest harbor port in Surabaya, called Tanjung Perak. Heavy logistics, as well as many passengers, occupy this port, thus making it the busiest port in Indonesia. Because of its status, transportation to that area becomes important, as the effects of traffic flow towards the port becomes extremely crowded, especially with large vehicles. One of the concerns is in minimizing the adverse effect of high pollution of NO_2 emitted from these vehicles. By judging the figure, one can manage the flow of vehicles. We observed that during the time frame in the morning between 7am to 10pm, the intercept parameters were big during Monday until Saturday. If we are going to use this fact for policy implementation, we could propose limiting the entrance of big vehicles between those time frames.

Another fact we can observe is that we noted the solar radiation seems to have smallest effect towards to NO_2 concentration because of flat line observed. This is in agreement with the result of previous chapter for NO_2 on Table 4.2 page 29. It is seen for NO_2, the effect of solar radiation is smaller than other meteorological variables.

The estimated parameters for intercept and independent variables for PM_{10} are shown in Figure 5.32. We also noted that the intercept was much more dynamic over time than the slopes of independent variables. The effect of temperatures was also very much dynamic. One difference in the result was that the dynamic changing was relatively stable for all days. This result suggested that the proposal for a policy as described above should be recommended to be applied for all days, not just working days. The implementation is expected to reduce the risk of being exposed to high concentrations of both NO_2 and PM_{10}.

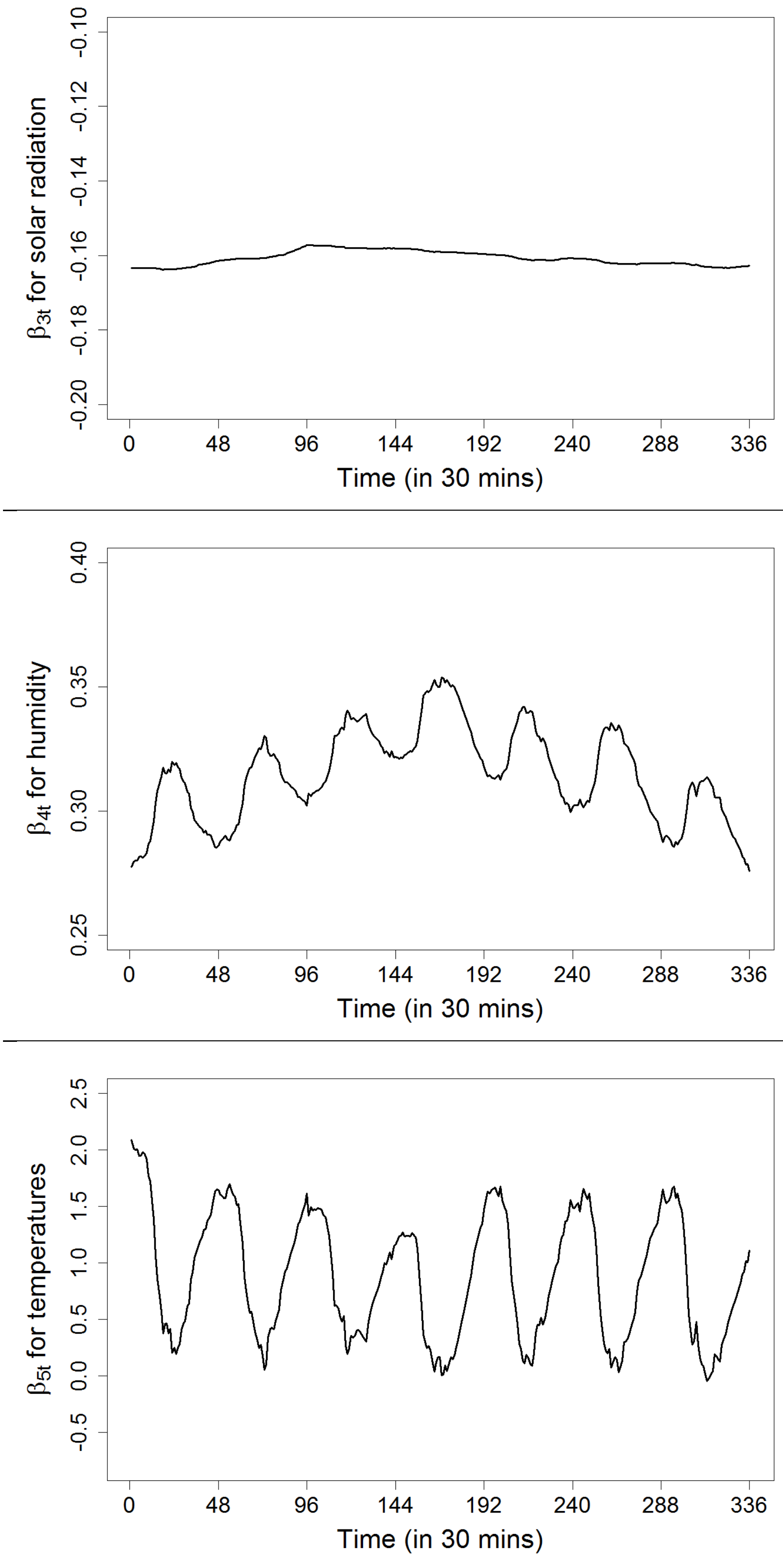
Furthermore, we noted that all five meteorological variables seem to have much more dynamic effect than NO_2. The direction of wind had big role to decide the PM_{10} concentrations. Moreover, the tendency of wind speed was, during working days, higher wind speed reflected in significant reduction of PM_{10}. This tendency also applied for other meteorological parameters such as solar radiation, and humidity. This result was also in agreement with the result obtained from the multilevel model. An interesting finding about the use of state-space dynamic model was confirmed by the followings. If we observe the effect of the temperatures towards the PM_{10} concentration (Table 4.2 page 29), we see that the temperature has negative effect, meaning that higher temperature will likely reduce the concentrations. However, Figure 5.32f shows that only during daytime the increase of temperature will reduce the concentrations. During nighttime, the regression parameter become positive, therefore the increase of temperature will lead to the increase of PM_{10} concentrations. In this case, we have provided the usefulness of state-space dynamic model to capture the limitation of linear model on the previous chapter. All these results are in agreement with the finding obtained by multilevel models with some additional discussion like shown above.



(a) intercept

(b) wind direction

(c) wind speed

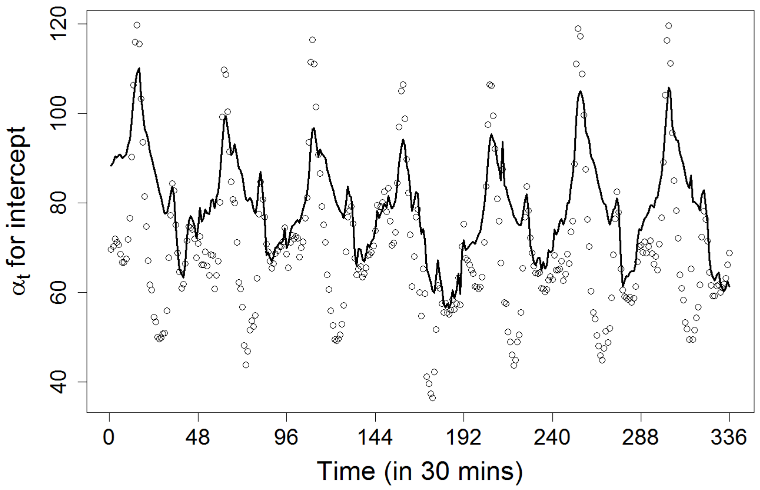


(d) solar radiation

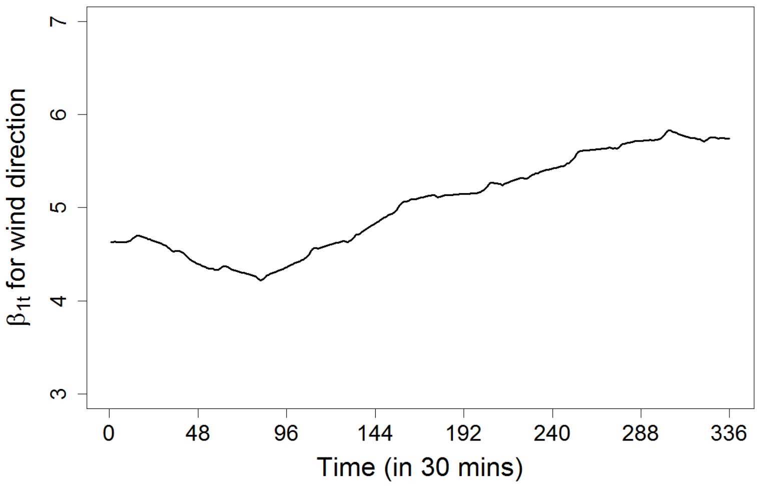
(e) humidity

(f) temperatures

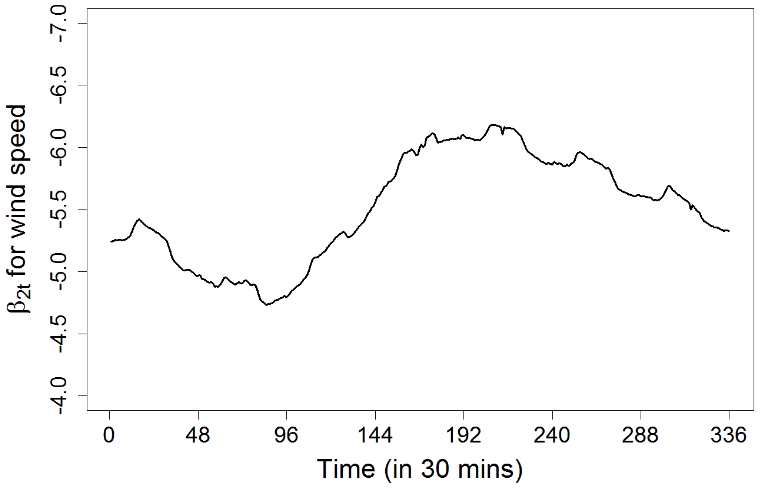
Figure 5.31 Time series of estimated mean states of independent variables (thick lines). The black dot is the observation of square-root NO_2 concentration in ug/m^3. The first segment is Thursday.



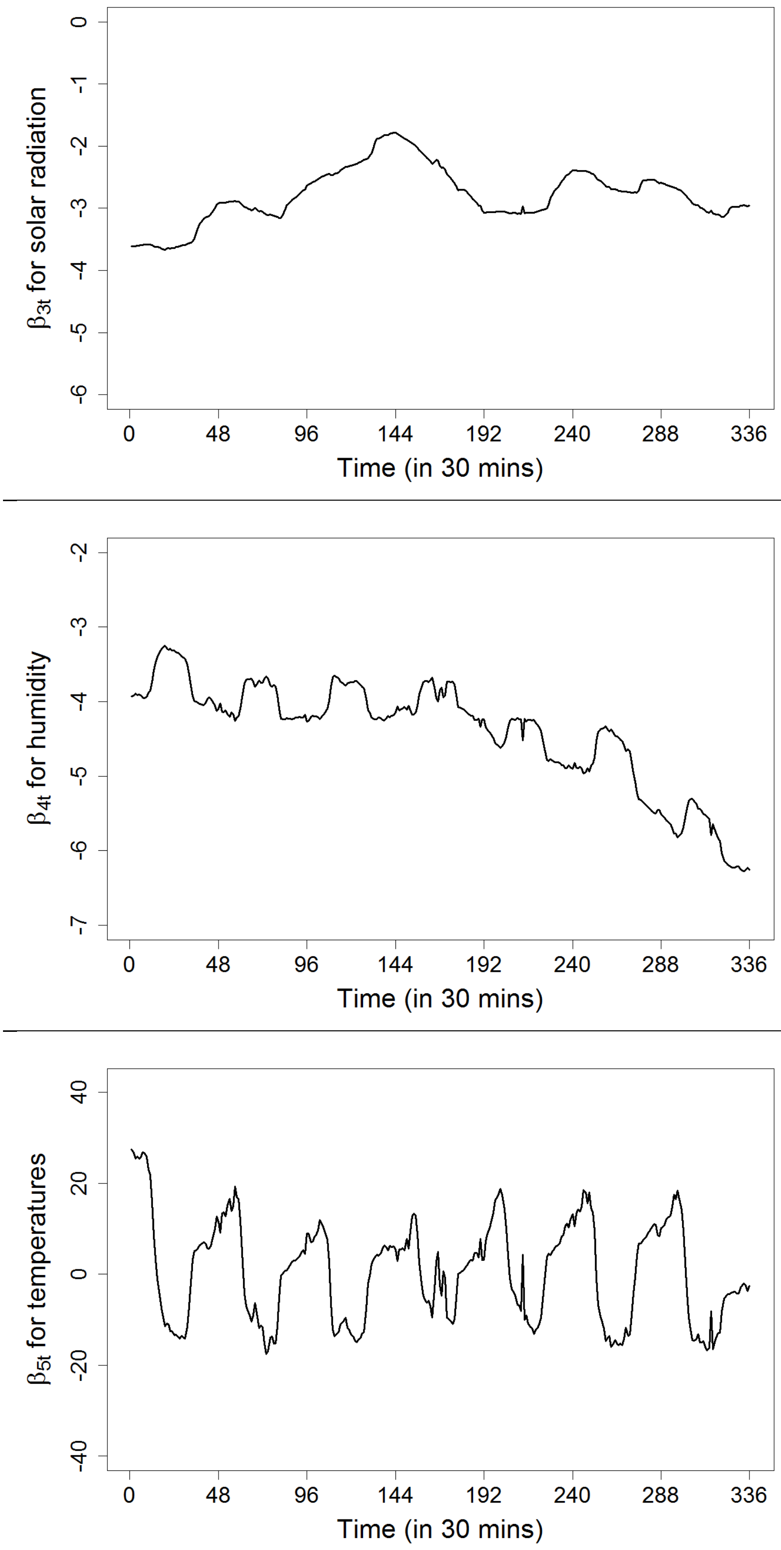
(a) intercept



(b) wind direction



(c) wind speed



(d) solar radiation

(e) humidity

(f) temperatures

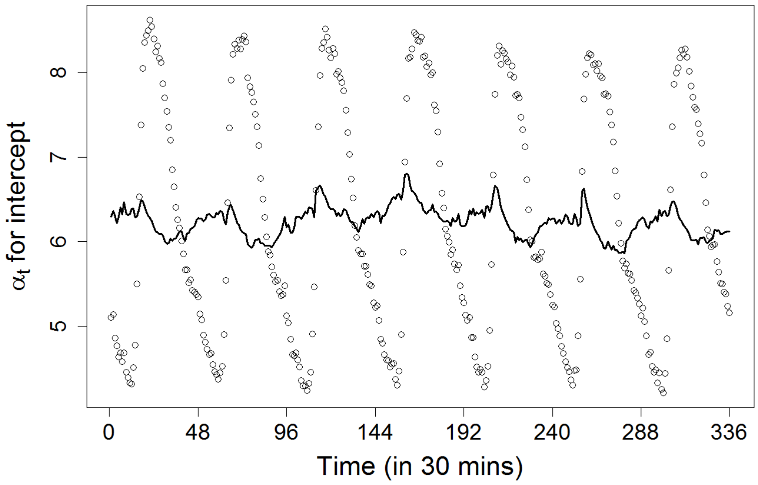
Figure 5.32 PM_{10} Time series of estimated mean states of independent variables (thick lines). The black dot is the observation of square-root PM_{10} concentration in ug/m^3. The first segment is Thursday.

Figure 5.33 displays parameter estimates for O_3. We can observe that the intercept parameter is not as dynamic as the intercept for NO_2 and PM_{10}. It is logical because O_3 is formed due to a series of photochemical reactions combined with the association of solar radiation as shown by Figure 5.33d. There was a dynamic change in the solar radiation parameter estimate. During the night, the solar radiation sensitivity factor was relatively low to accommodate the fact that at night there was no solar radiation, whereas, in the absence of radiation, the concentration of O_3 persisted. This also shows that even though solar radiation is one factor affecting O_3, it is not the most important. O_3 is formed with the presence of precursors such as NO and NO_2. This result also depicted the increase of O_3 during morning due to the increase of NO_2 and NO rate.

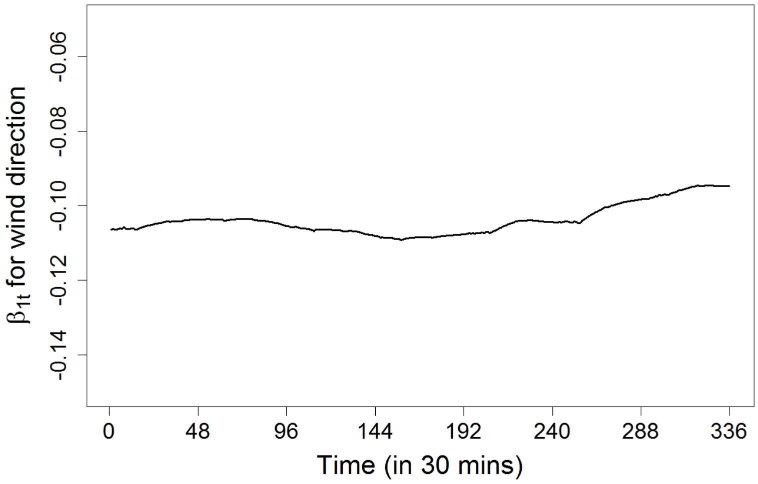
As can be seen, the important factors for O_3 had been solar radiation, humidity and temperatures. However, solar radiation was the one that has most influential factors. Humidity tended to decrease O_3 concentrations, especially during working days. This is in agreement with the fact that when the presence of NO_2 and NO is relatively higher, the O_3 concentrations tends to be lower than if the opposite environment occurs, suggesting the reaction rate of O_3 is complex. Furthermore, in the morning, the solar radiation will increase O_3 levels. Therefore, it is important to highlight the importance of air quality management especially during morning session. On the other hand, higher temperature will lead to the increase of O_3 levels, which is again in agreement with previous result. The effect of temperatures during working days slightly diminishes though.

Figure 5.34 shows temporal dynamic factors for NO which is similar with O_3 (Figure 5.32), and CO intercept (Figure 5.35) was not as dynamic as the intercept of NO_2 and PM_{10} suggesting the lower effect may involve factors other than meteorological variables. However, additional information is needed. The estimate of intercept was lower on Sunday, similar to NO_2. This similar pattern is because NO and NO_2 came from a similar source, which is vehicular or transport related. In the meantime, the fluctuation of wind direction estimates explains that in the morning the wind direction had significant impact on the concentration of NO. The estimate of solar radiation is similar to NO_2 and $PM_{10,}$ explaining that solar radiation may reduce NO concentration, indicated by negative signs of estimates. Parameters, such as wind speed, solar radiation, humidity and temperatures, were not highly dynamic. The important parameter in this case is intercept, indicating the significance of other variables besides meteorological factors.

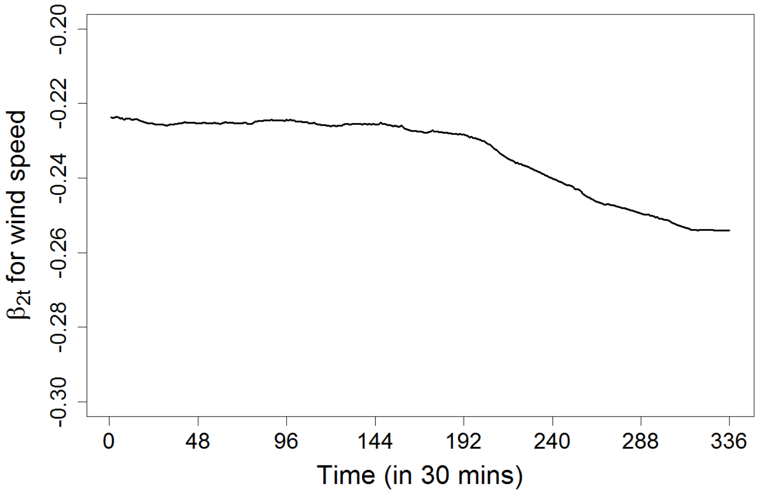
The two-wave peaks as shown on Figure 5.34b shows that there is a certain direction of which contributed the most of the pollution. This direction may came from transportation and from previous observation, it can be estimated that these peaks were contributed from the road connecting between city center to the port harbor. From the results, there is one different noted between NO and NO_2, that is wind speed. Although over time the parameter decreased, but it showed positive value, on the contrary the wind speed parameter for NO_2 had negative value. This positive value indicated that higher wind speed may increase the concentration of NO. These results are also in agreement with the result from multilevel model. Solar radiation had negative impact towards the NO concentrations but on the opposite, the parameter estimate for humidity is positive, indicating higher humidity may contributed to the increase of NO concentrations. The results of humidity and temperatures effects toward NO were different with their effects towards NO_2, suggesting different chemical properties as well as reaction in the atmospheric layer.



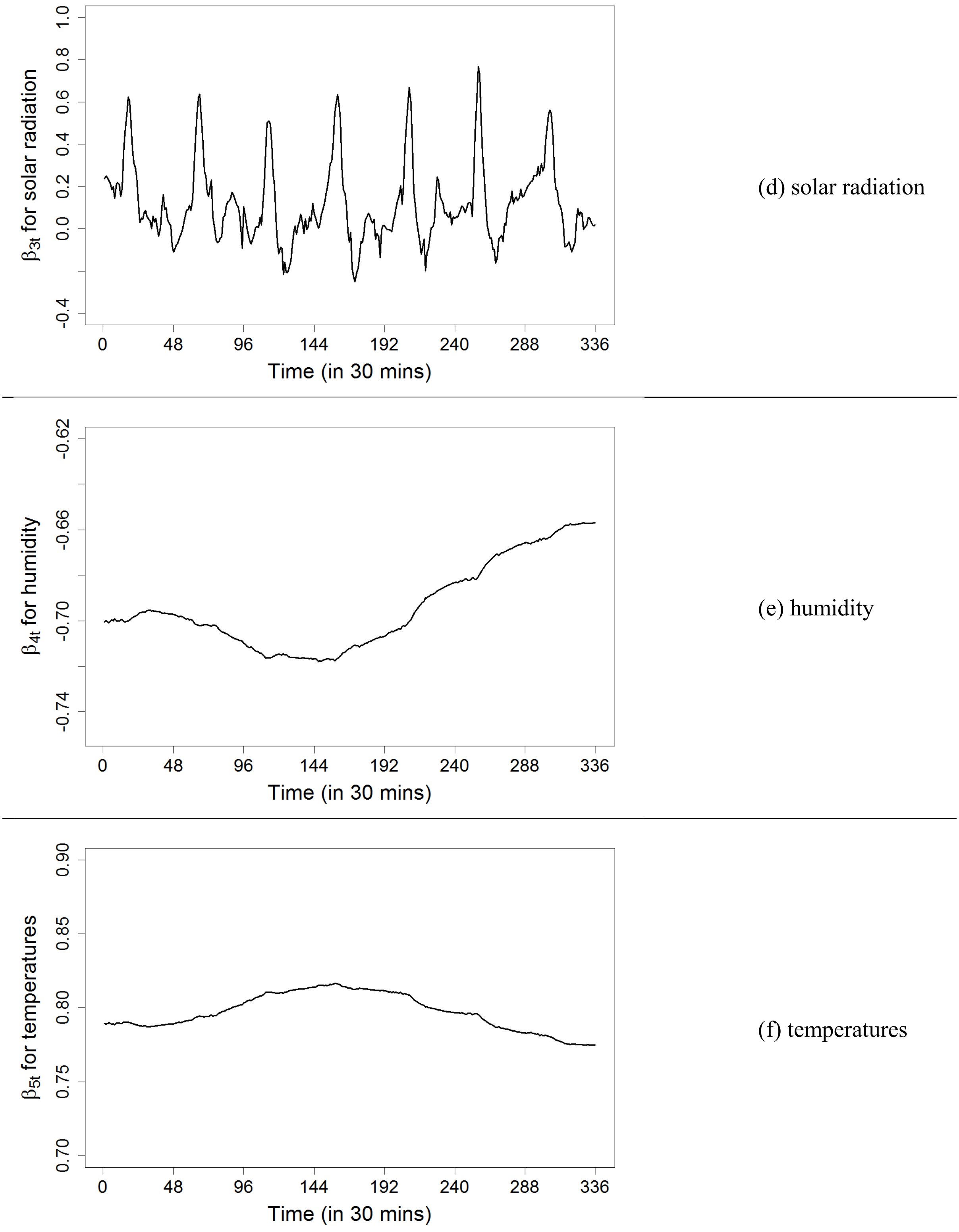
(a) intercept



(b) wind direction



(c) wind speed

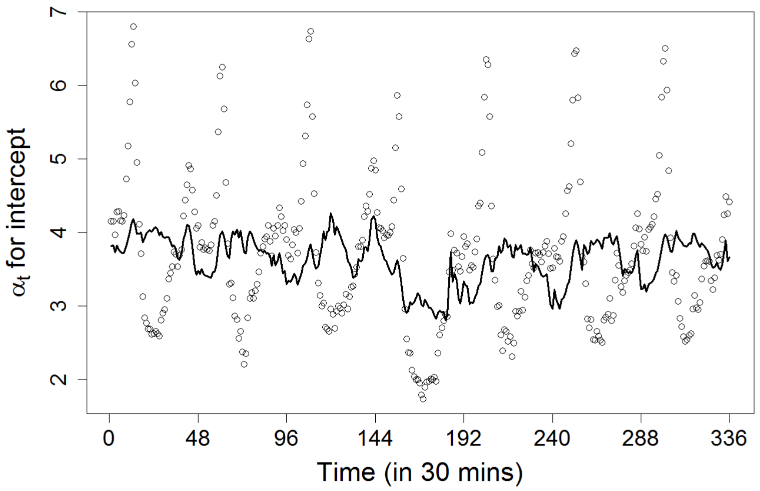


(d) solar radiation

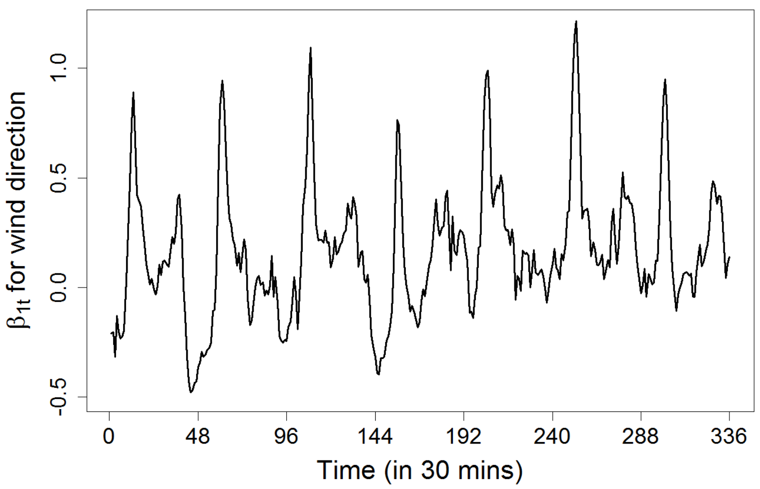
(e) humidity

(f) temperatures

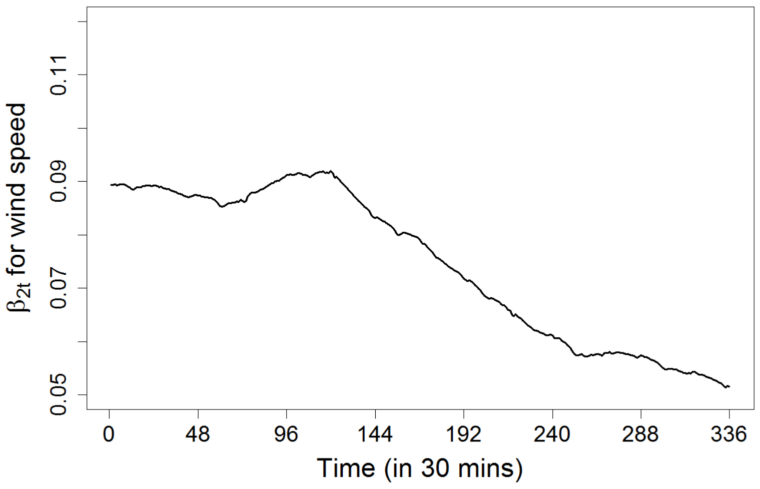
Figure 5.33 O_3 Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root O_3 concentration in ug/m^3. The first segment is Thursday.



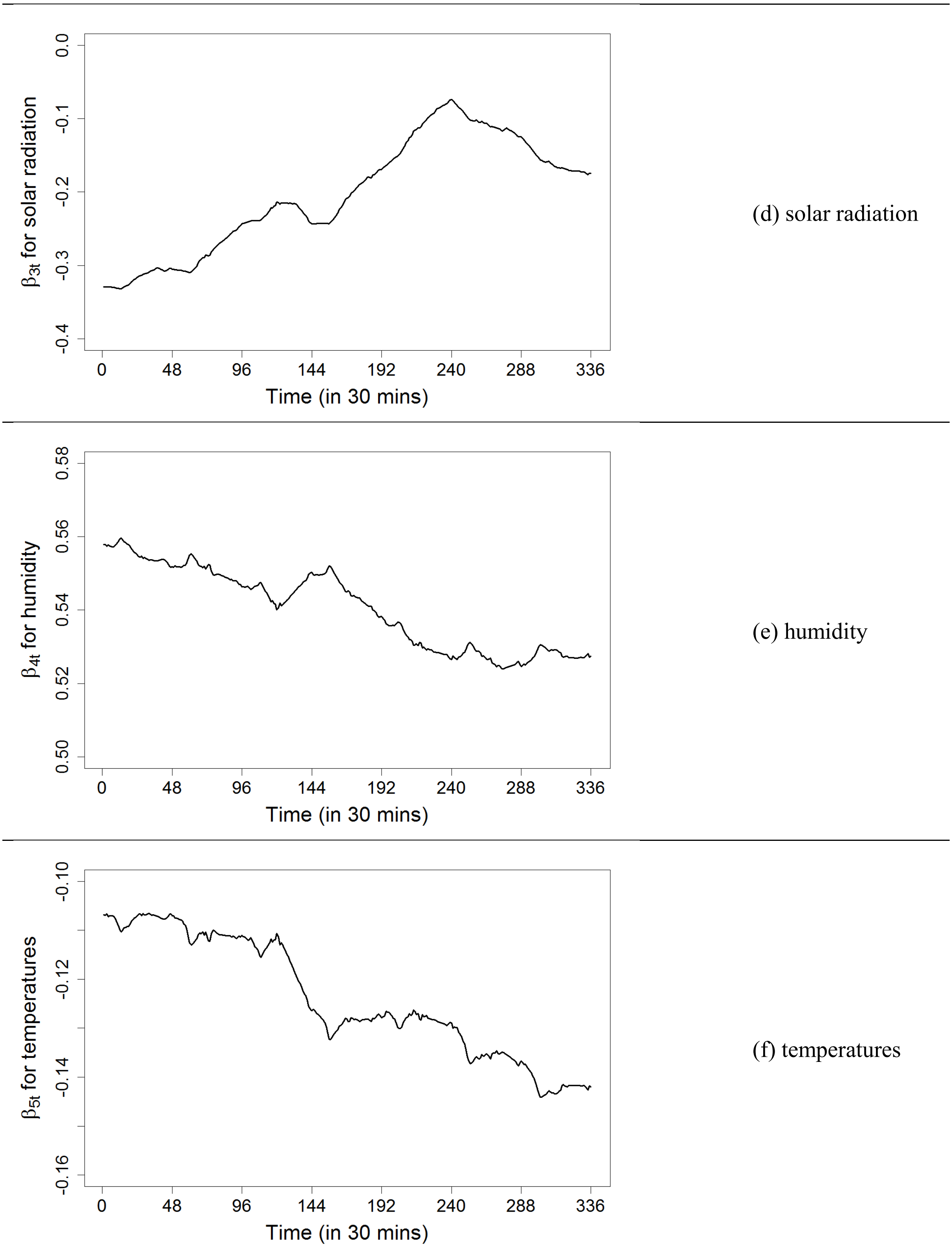
(a) intercept



(b) wind direction



(c) wind speed



(d) solar radiation

(e) humidity

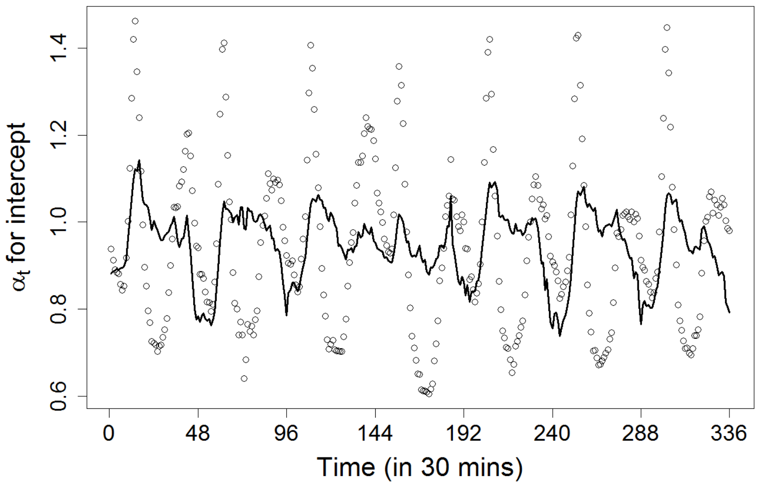
(f) temperatures

Figure 5.34 NO Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root NO concentration in ug/m^3. The first segment is Thursday.

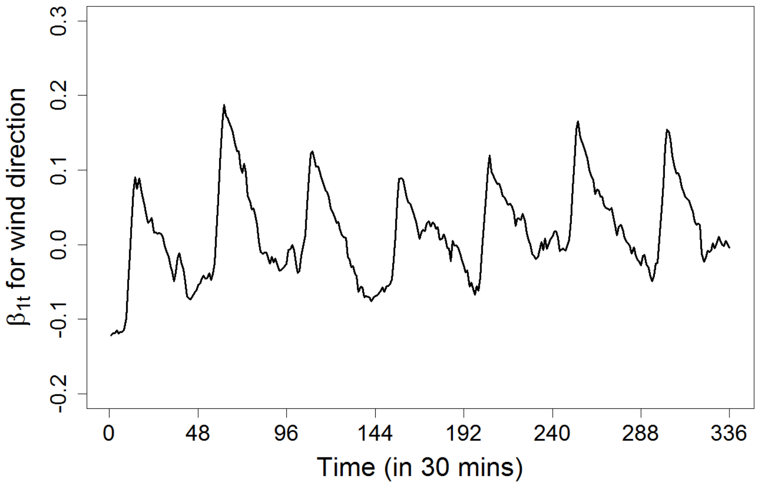
Figure 5.35 shows the dynamic of meteorological factors towards the CO concentrations. It can be shown that meteorological factors such as wind speed, solar radiation, humidity, and temperatures had small effect for CO concentrations because their fluctuation interval was small. However, we have noted several points as follows. Like NO, wind direction played big role in forming the CO concentrations. In the morning the effect of wind direction was seen, showing an indication that traffic in the morning on Perak Barat and Perak Timur Rd was significant. Furthermore, the parameter for wind speed is negative, therefore leading to a conclusion that higher wind speed increased the concentration of CO. However, the effect of wind speed on Sunday had been low. The value of solar radiation parameter was small and stayed relatively flat, not as dynamic as intercept. Higher humidity tended to increase the CO concentrations. The result matches with the result from the multilevel model although the estimate from multilevel model is small but significant. The estimate of humidity on working days tended to be smaller than weekends. Therefore, during weekends, if the humidity was high, the CO concentrations were likely to be higher than working days as well.

SO_2 comes mainly from the combustion process, and one major source is from power plants. SO_2 is also formed by a reaction between hydrogen sulfide (H_2S) and Oxygen (O_2). However, its oxidation is slow in a mixture of pure gases (Vallero, 2008), but the rate will increase because of light, which may explain Figure 5.36a. The intercept decreased as it approached mid-day. The rate of oxidation is also affected by NO_2, oxidants, and metallic oxides. Metallic oxides serve as catalysts. Since this process is highly affected by the presence of other gases, the effect of meteorological factors seems not to be too dynamic.

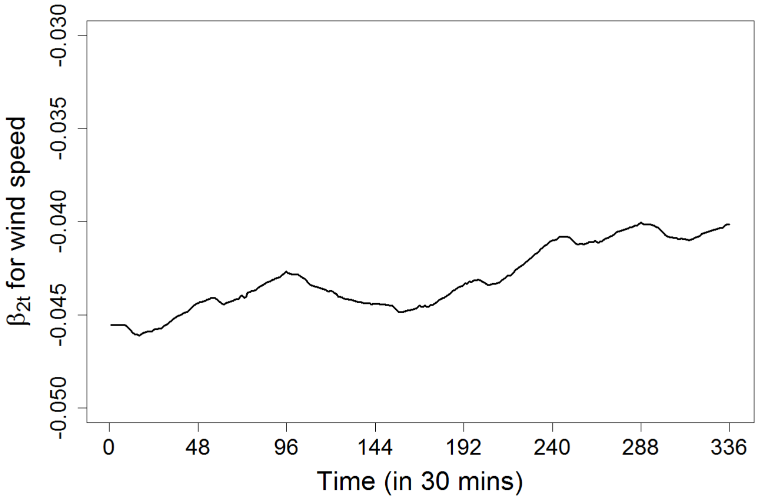
From the figure as well, we noted that humidity and temperatures were relatively dynamic suggesting their significant role towards the concentration of SO_2. In the night, the parameter was higher than during day time. However, the positive sign shows that higher temperature and humidity may lead to the increase of SO_2 levels.



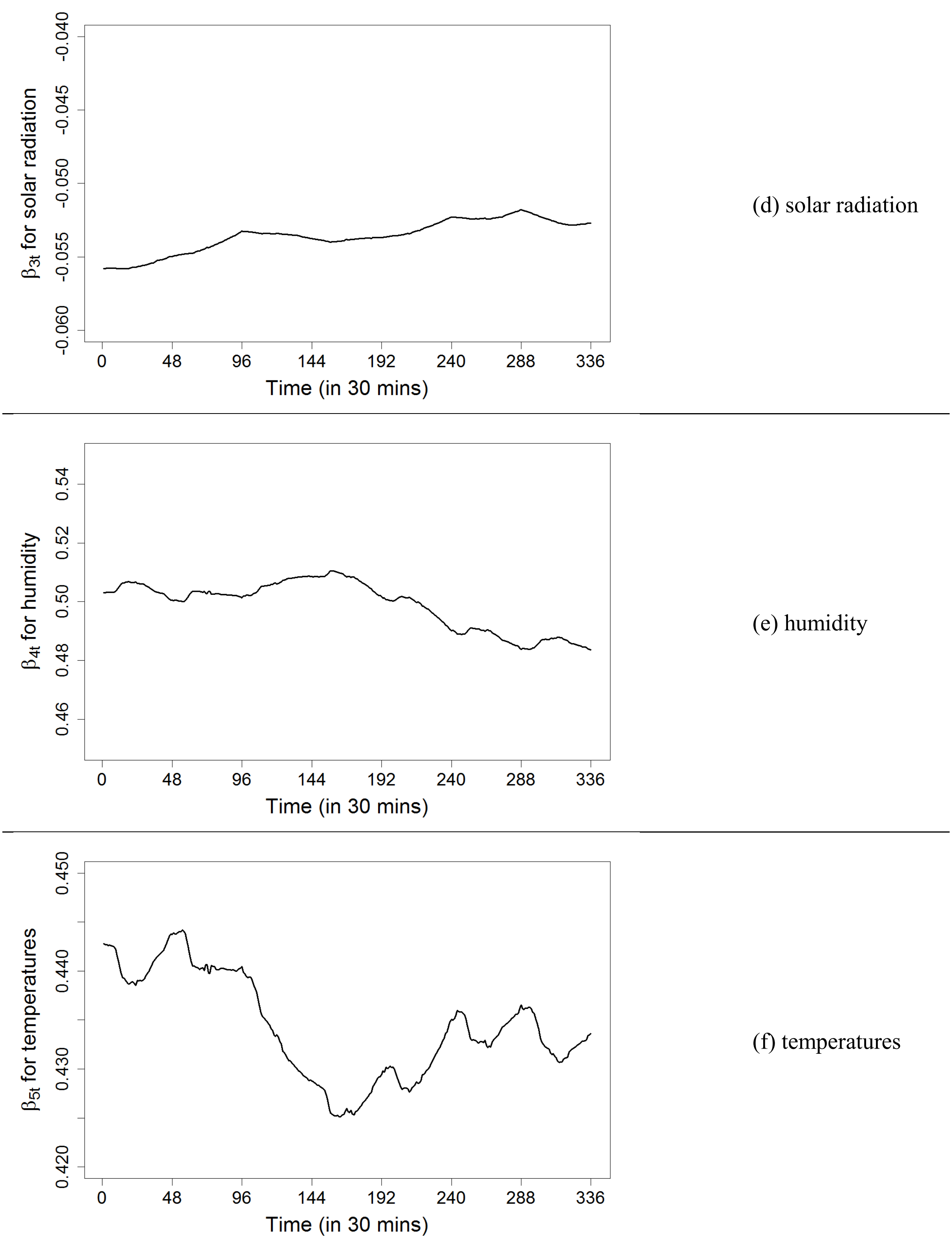
(a) intercept



(b) wind direction



(c) wind speed

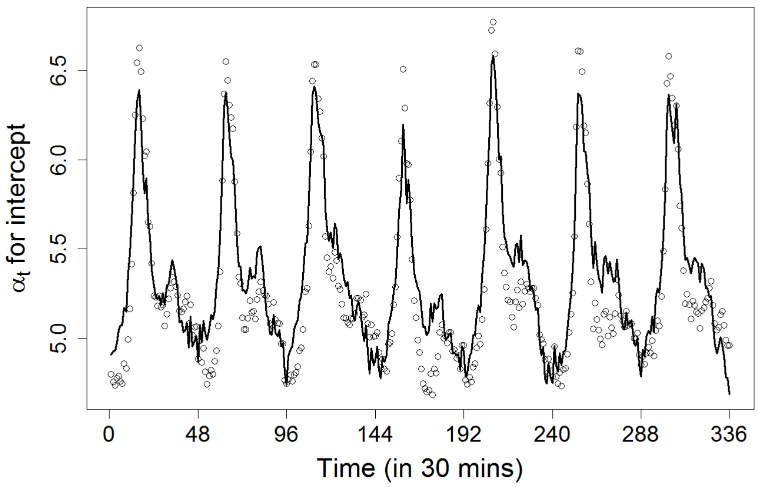


(d) solar radiation

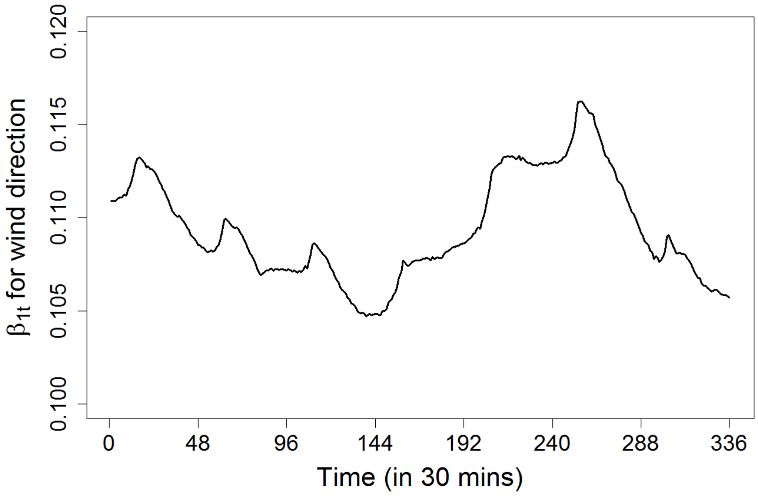
(e) humidity

(f) temperatures

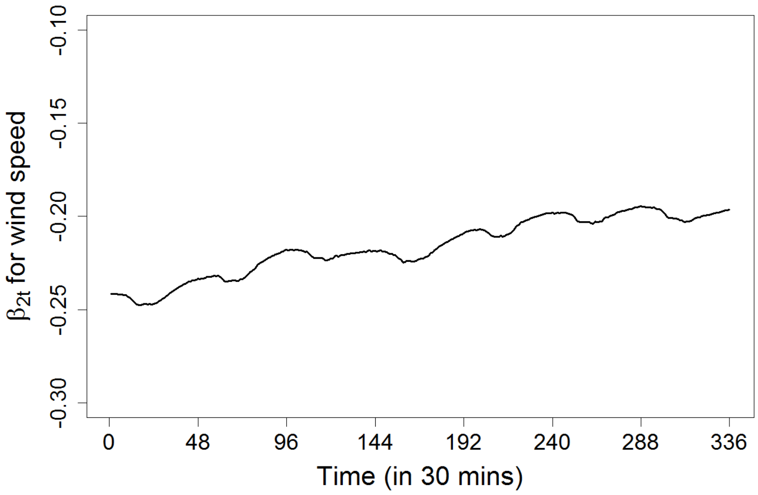
Figure 5.35 CO Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root CO concentration in ug/m^3. The first segment is Thursday.



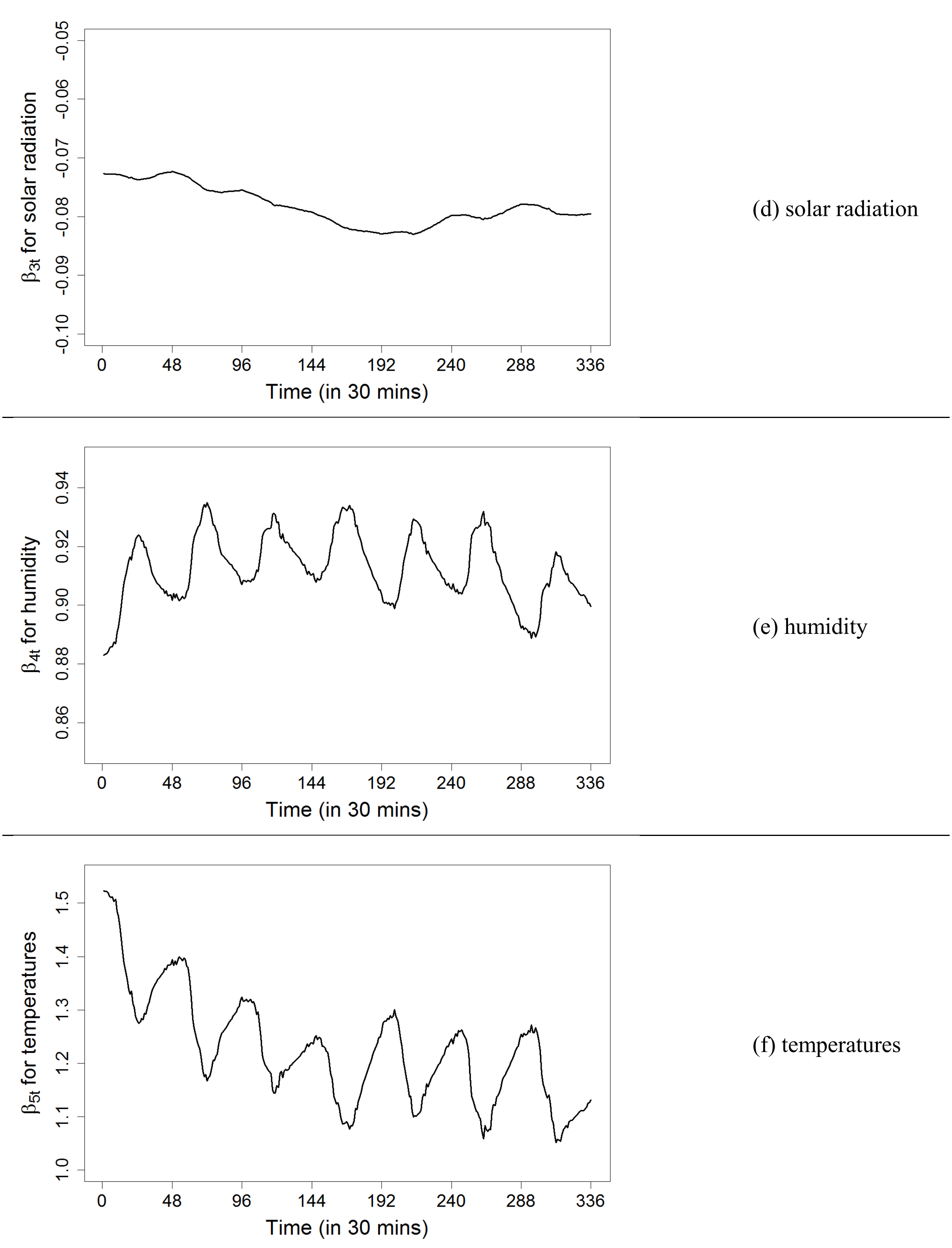
(a) intercept



(b) wind direction



(c) wind speed



(d) solar radiation

(e) humidity

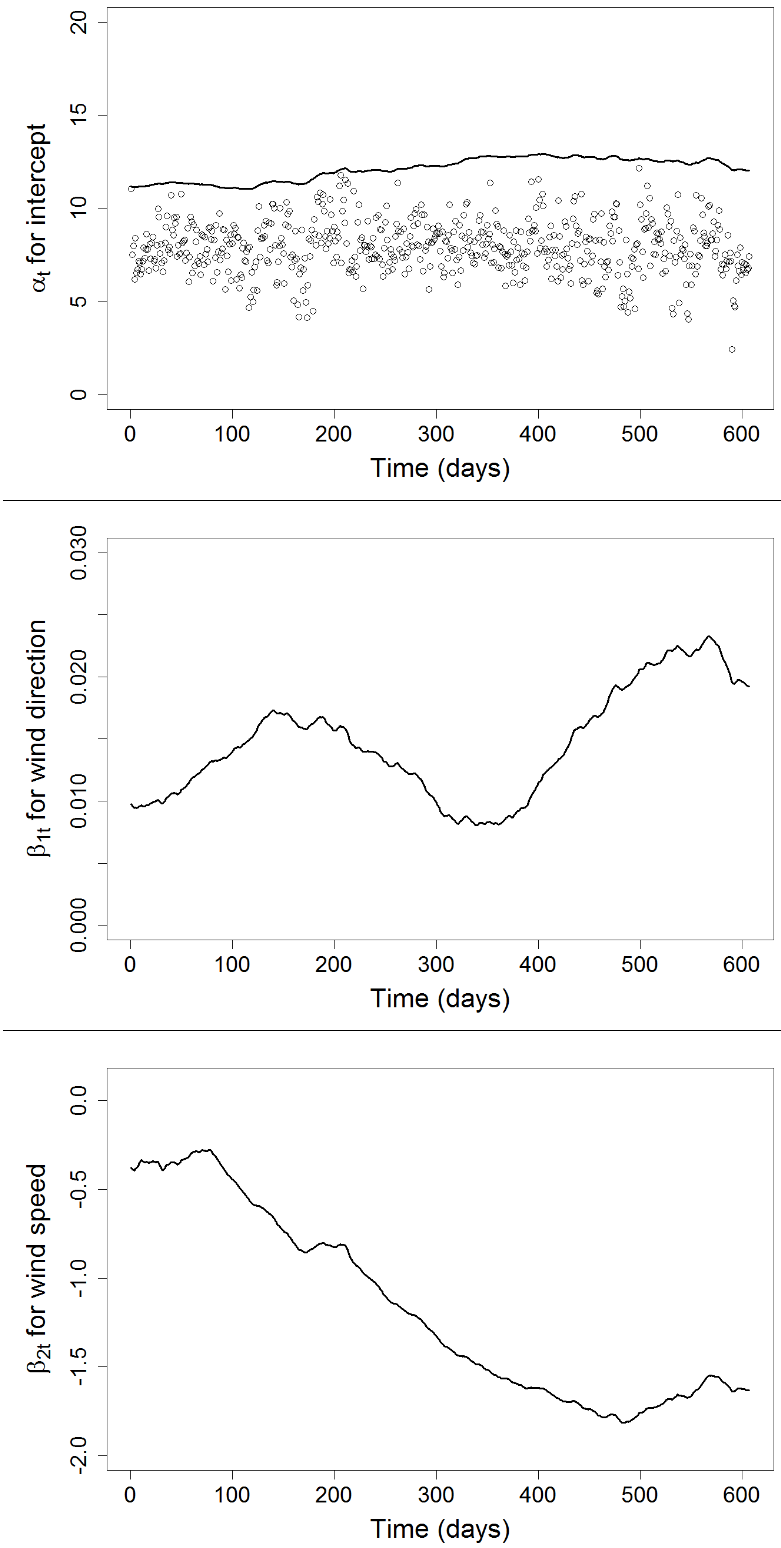
(f) temperatures

Figure 5.36 SO_2 Time series of estimated mean states of independent variables (thick lines). The black dot is an observation of square-root SO_2 concentration in ug/m^3. The first segment is Thursday.

Based on the evaluation above, we tried to find the "best" parameters for policy implementation. For that purpose, we determined and filtered pollutants that are emitted directly from the source (and also reactive), so that we can observe the temporal change more clearly. Therefore, we focused on NO_2 and PM_{10} because both are emitted by the transport sector, which change in minutes and in hours despite that PM_{10} sources are more complex. However, to promote and suggest a policy, we gained the benefit from the patterns obtained from these two pollutants in terms of pollutant profile. Based on the Figures 5.30 and 5.31, the temporal peak time frame on this zone that must be taken into account is between 7am to 10pm because the estimates are high during these hours. We must take this fact as indicating a daily pattern, not just considering working days.

5.5.2 Seasonal Change dynamically of air quality

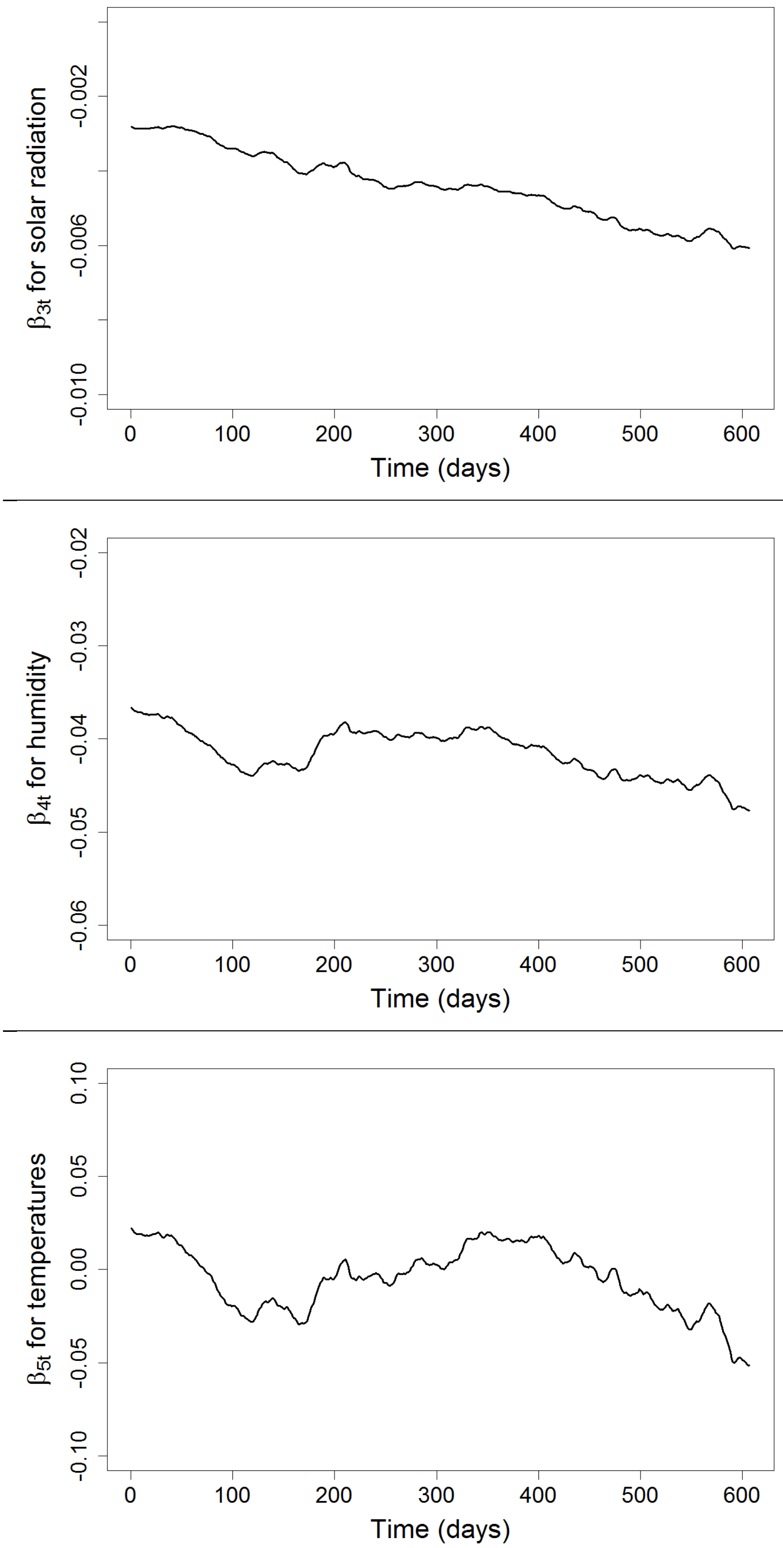
To evaluate seasonally and dynamically changes in air quality, we used daily average data as displayed in Figure 5.37 (NO_2) and Figure 5.38 (PM_{10}). We observed that the dynamics of NO_2 and PM_{10} concentration were high by season, for wind direction, and for wind speed. Wind direction highly affects both pollutants, especially during the dry season. In addition, as time goes by, we noted that the effect of wind speed increases with the following understanding: the longer the time runs, the more negative the estimated parameter for wind speed. This means the faster the wind speed, the better the self-purification of pollutants. In addition, regarding PM_{10}, during the wet season, humidity helped to significantly reduce the pollutant concentration levels. This is in agreement with the theory of collision between particulate matter with droplets contained inside the air parcel, and then deposited on the ground. During the wet season, the concentration of NO_2 tended to be higher, suggesting a change of travel behavior. Whereas for PM_{10}, the concentrations tended to be lower during the rainy season.



(a) intercept

(b) wind direction

(c) wind speed

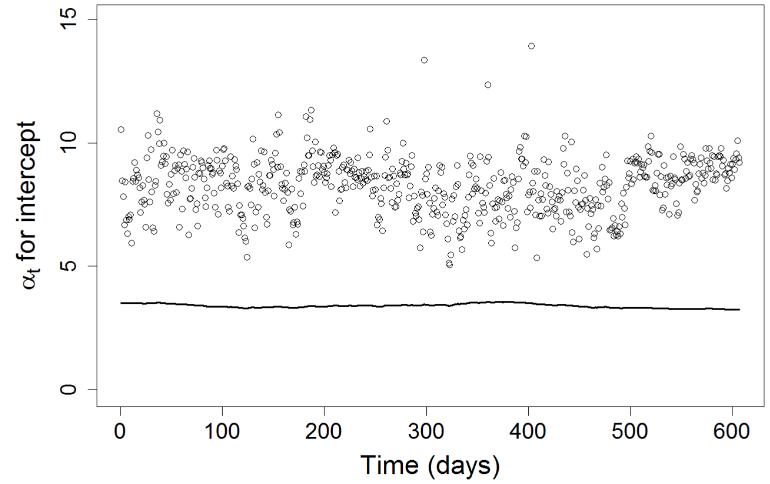


(d) solar radiation

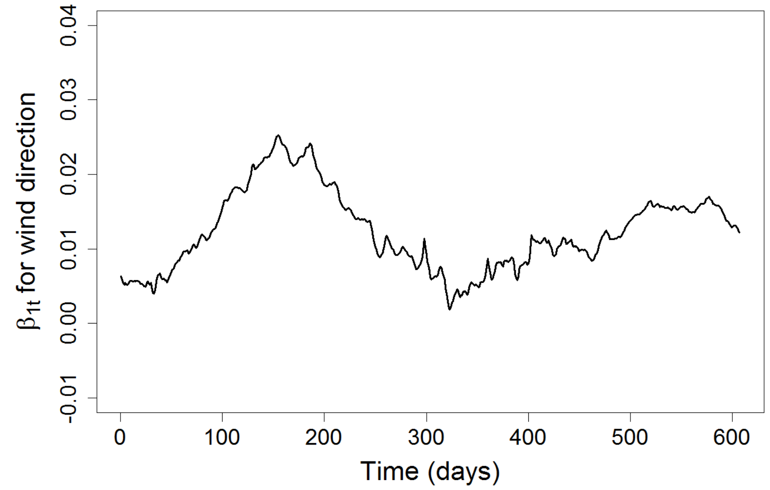
(e) humidity

(f) temperatures

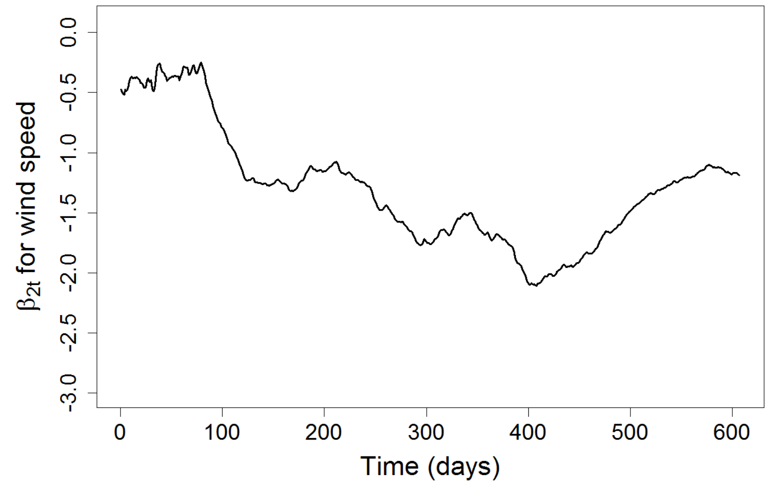
Figure 5.37 Estimated meteorological parameter coefficients for NO_2.The black dot is an observation of square-root NO_2 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry



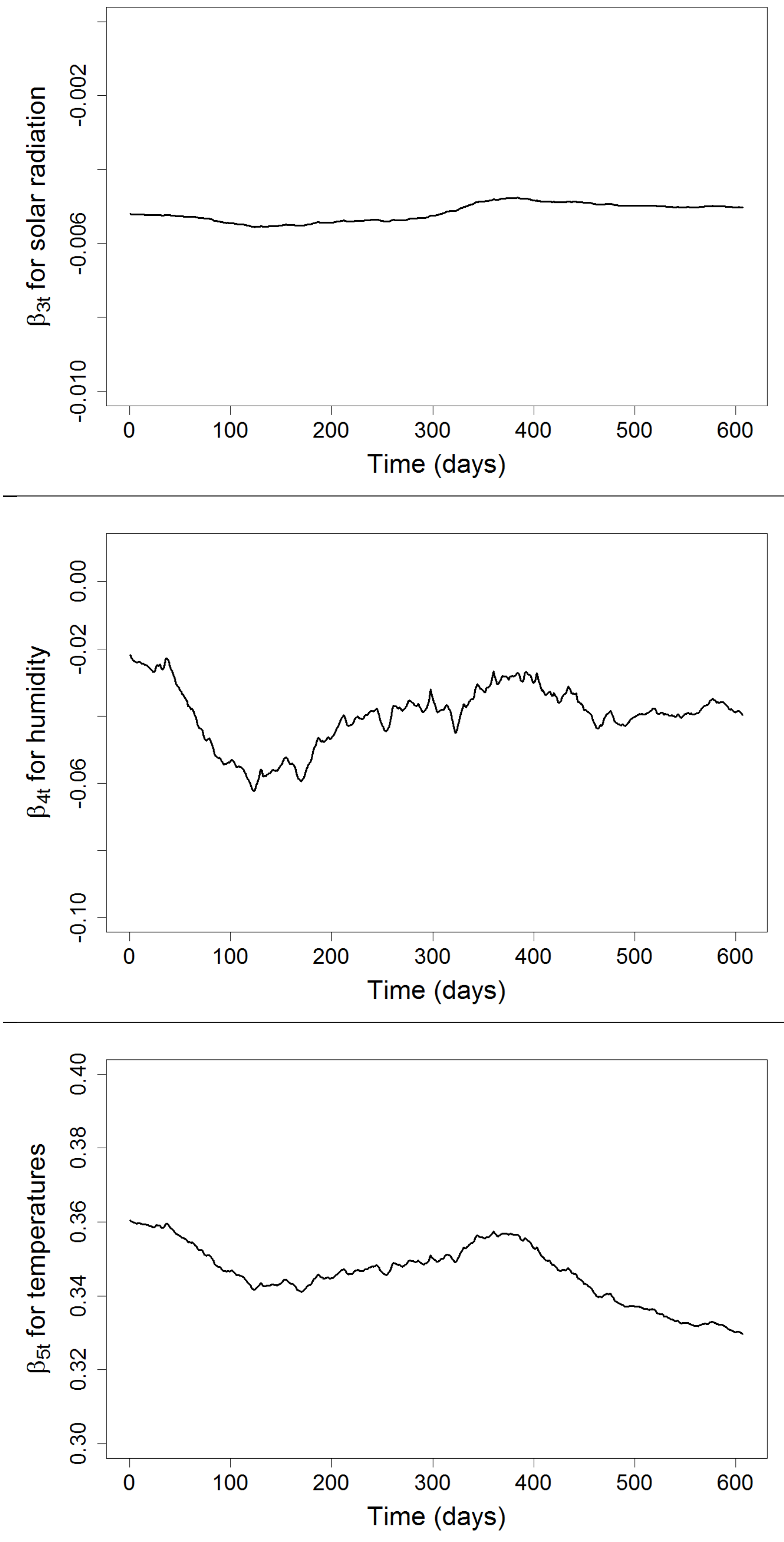
(a) intercept



(b) wind direction



(c) wind speed



(d) solar radiation

(e) humidity

(f) temperatures

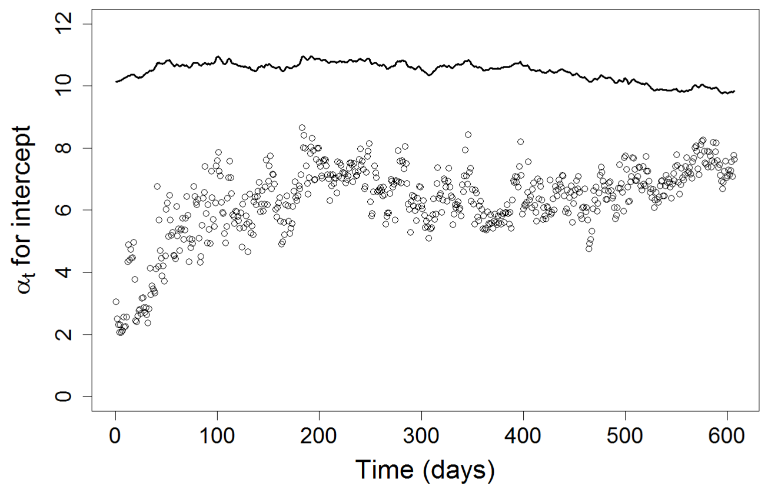
Figure |5.38 Estimated meteorological parameter coefficients for PM_{10}. The black dot is an observation of square-root PM_{10} concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry

For O_3, despite the not-so-dynamic parameter coefficients for intercept and wind direction, we observed a high dynamic of wind speed, especially in the wet season, which forms a plateau (Figure 5.39). During this season, the higher wind speed causes the lower concentration of O_3. This reaction occurs in the tropospheric layer indicating a complex purification process in the atmospheric layer. We didn’t see such a pattern in the dry season. Therefore, we may conclude that there was a tendency that during the wet season the probability of O_3 exceeding threshold value was lower than in the dry season. This result is in agreement with the result from multilevel model that during wet season higher wind speed will reduce the O_3 levels. The effect of solar radiation appeared to decline over time, but during the wet season, we saw a slight decrease of the parameter. This means that there was tendency that higher solar radiation improved O_3 in the atmospheric layer. Moreover, the humidity parameter was dynamic over time with tendency to increase. The power of pollutant purification was stronger during dry season than wet season, as indicated that during dry season the estimate was more negative.

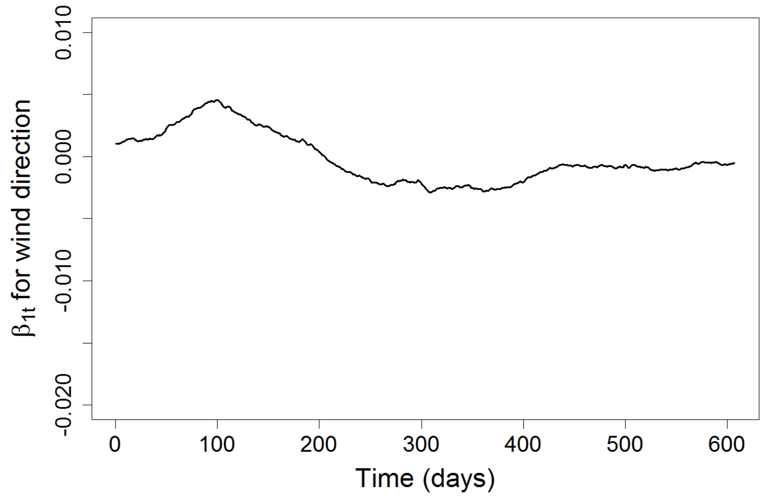
Overall, judging from annual data, we observed the significance of meteorological factors towards the concentration of O_3. This was due to O_3 production was highly dominated by a series of chemical reaction in the atmospheric layer, that was completely different from NO_2 and PM_{10} which were emitted directly from the sources e.g., vehicles.

Observing NO seasonally, the intercept was not too dynamic, and daily average wind direction affects daily average concentrations of NO (Figure 5.40). The effect of wind direction was higher during the dry season than in the wet season, indicating the factor of wind direction contributed more towards the concentration than during the wet season. In the meantime, the increasing wind speed will reduce the ambient concentration of NO, whereas humidity and temperatures affected the concentrations seasonally. The wet season contributed towards lower NO concentrations. On the other hand, Figure 5.41a shows that the intercept was not dynamic compared with wind speed, indicating that for CO, the role of wind speed was crucial. The faster wind speed had the largest the effect on CO purification.

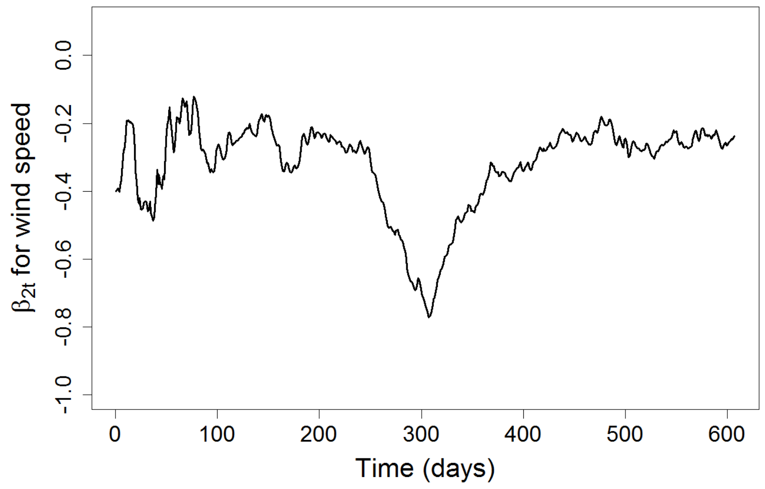
Seasonally, we noted that the wind speed parameter tended to reduce SO_2 during the dry season (Figure 5.42). This means that in the dry season, the faster wind speed led to further reduction in pollution from SO_2. There is no significant difference between seasons for solar radiation. However, we noted an increasing coefficient over time. Interestingly, the parameter estimate for temperature during the dry season was lower than during the wet season, which indicates that the tendency to reduce SO_2 pollution was better in the dry season than during the wet season for the same temperature.



(a) intercept



(b) wind direction



(c) wind speed

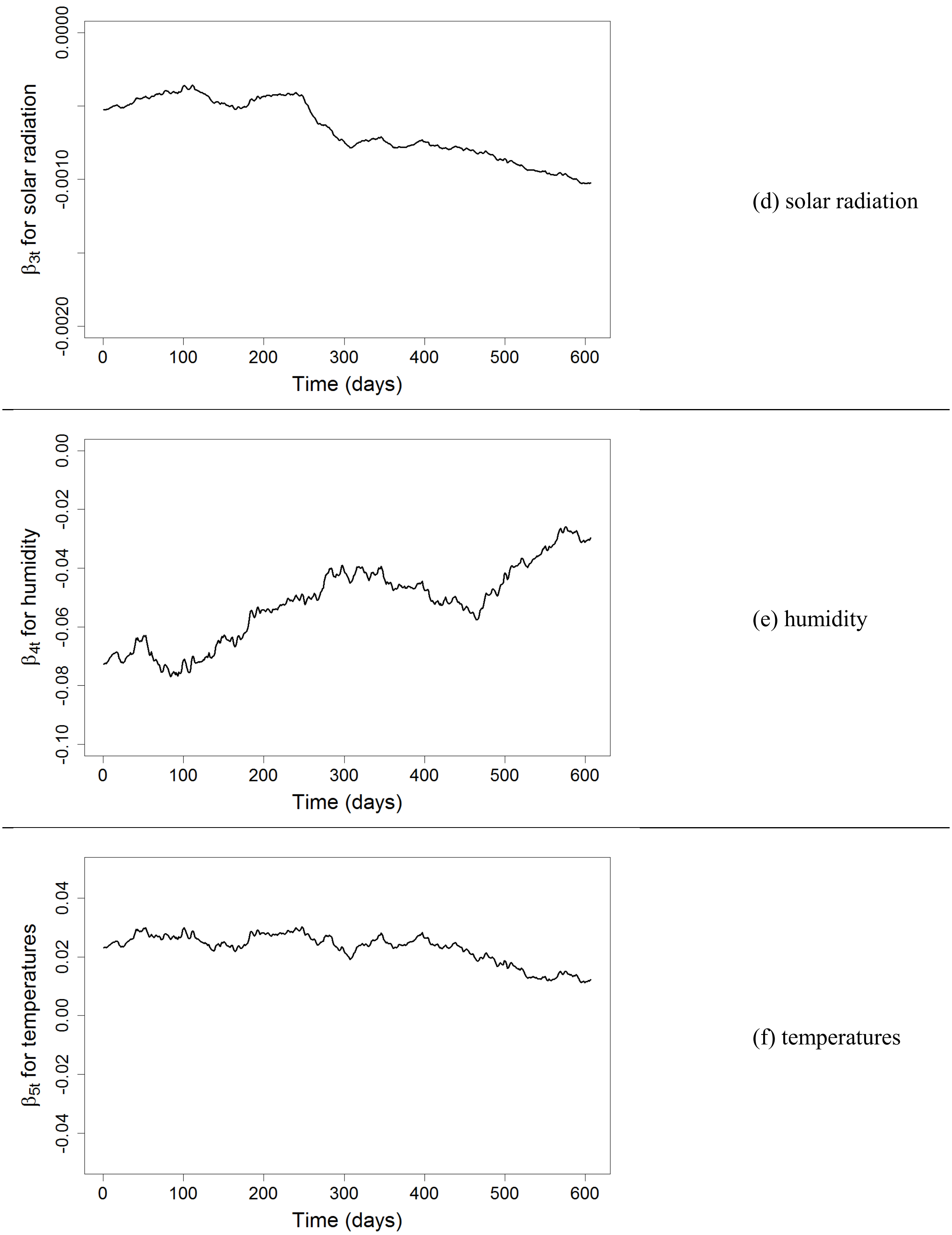
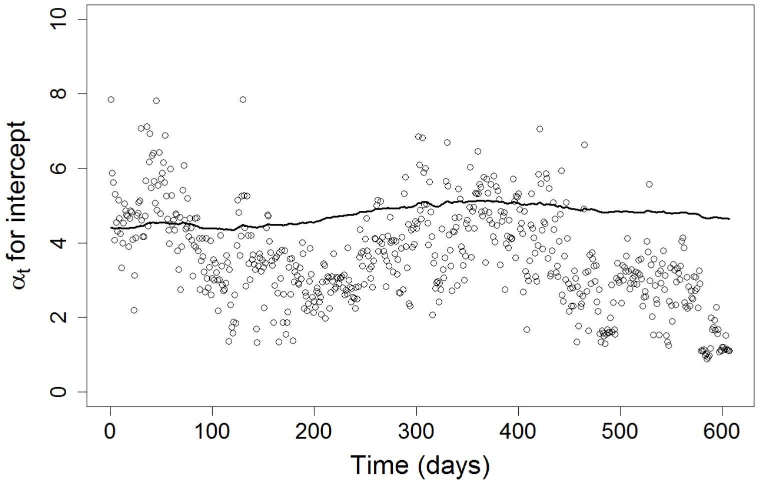
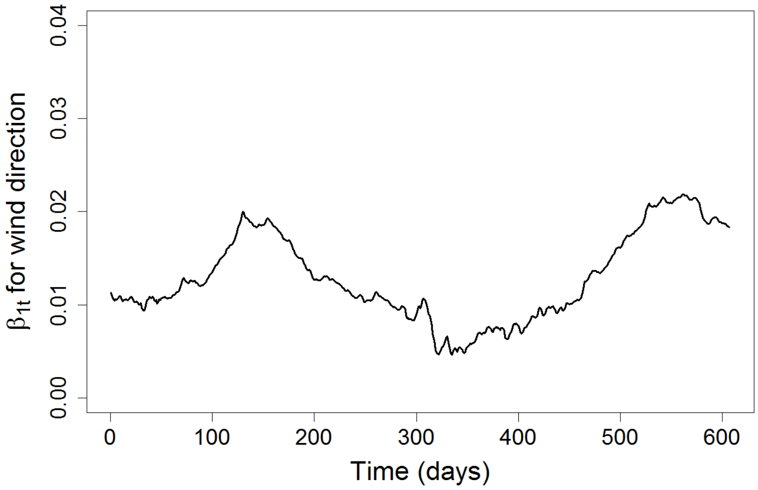


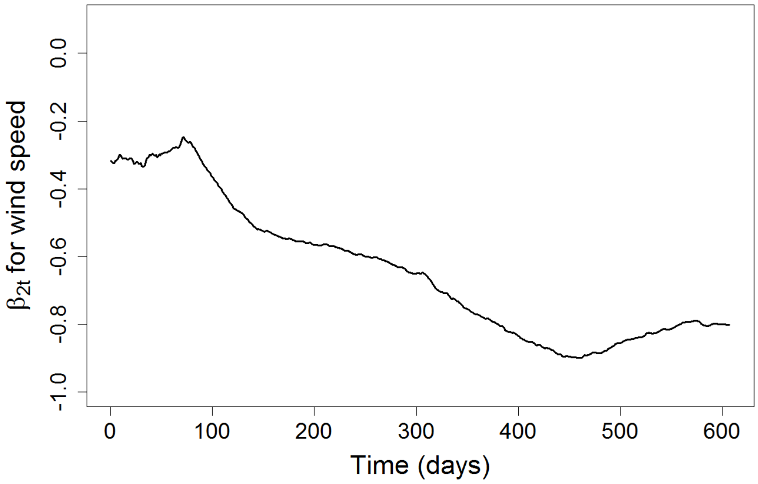
Figure [5.39 Estimated meteorological parameter coefficients for O_3. The black dot is an observation of square-root O_3 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry



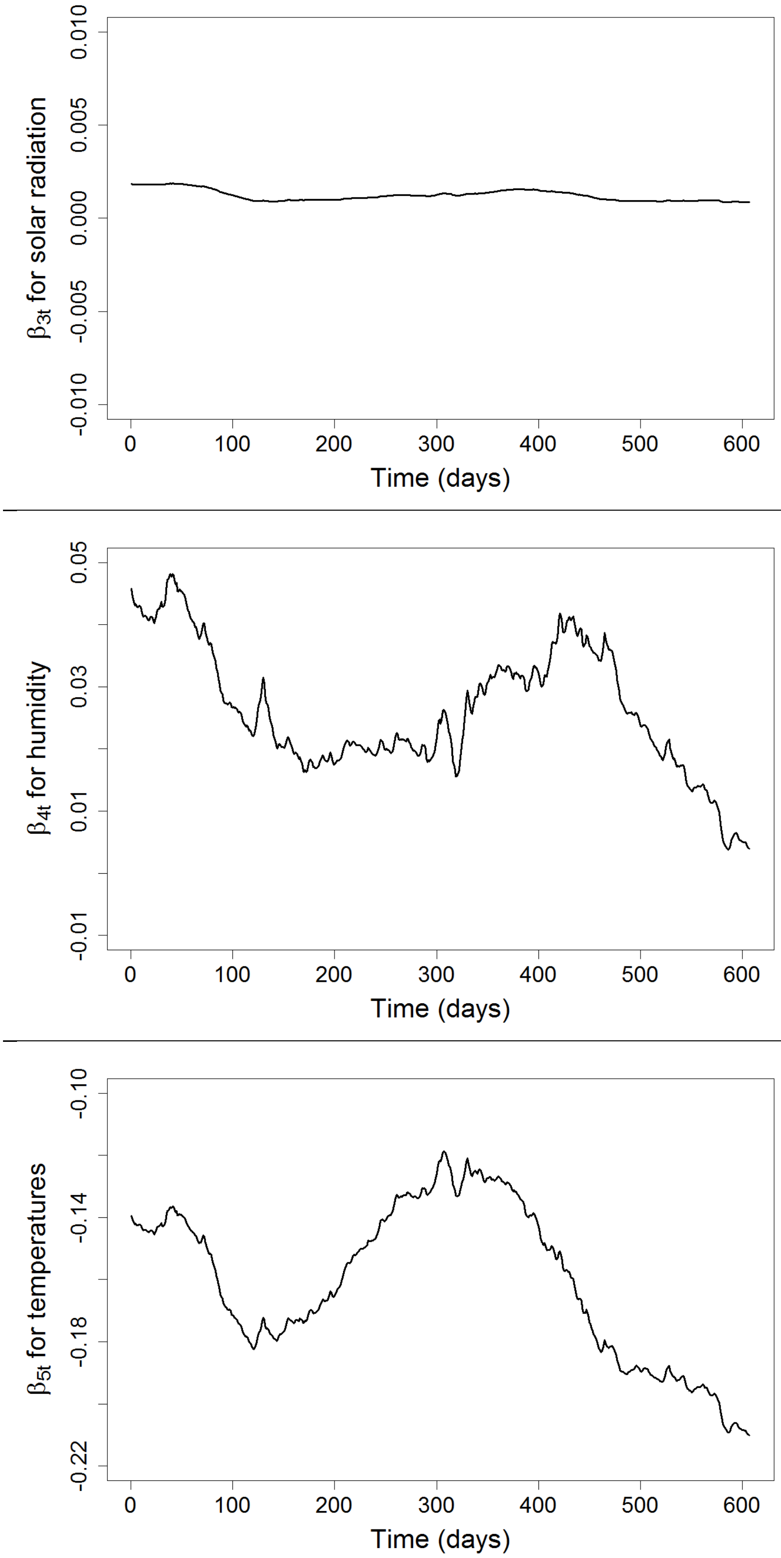
(a) intercept



(b) wind direction



(c) wind speed

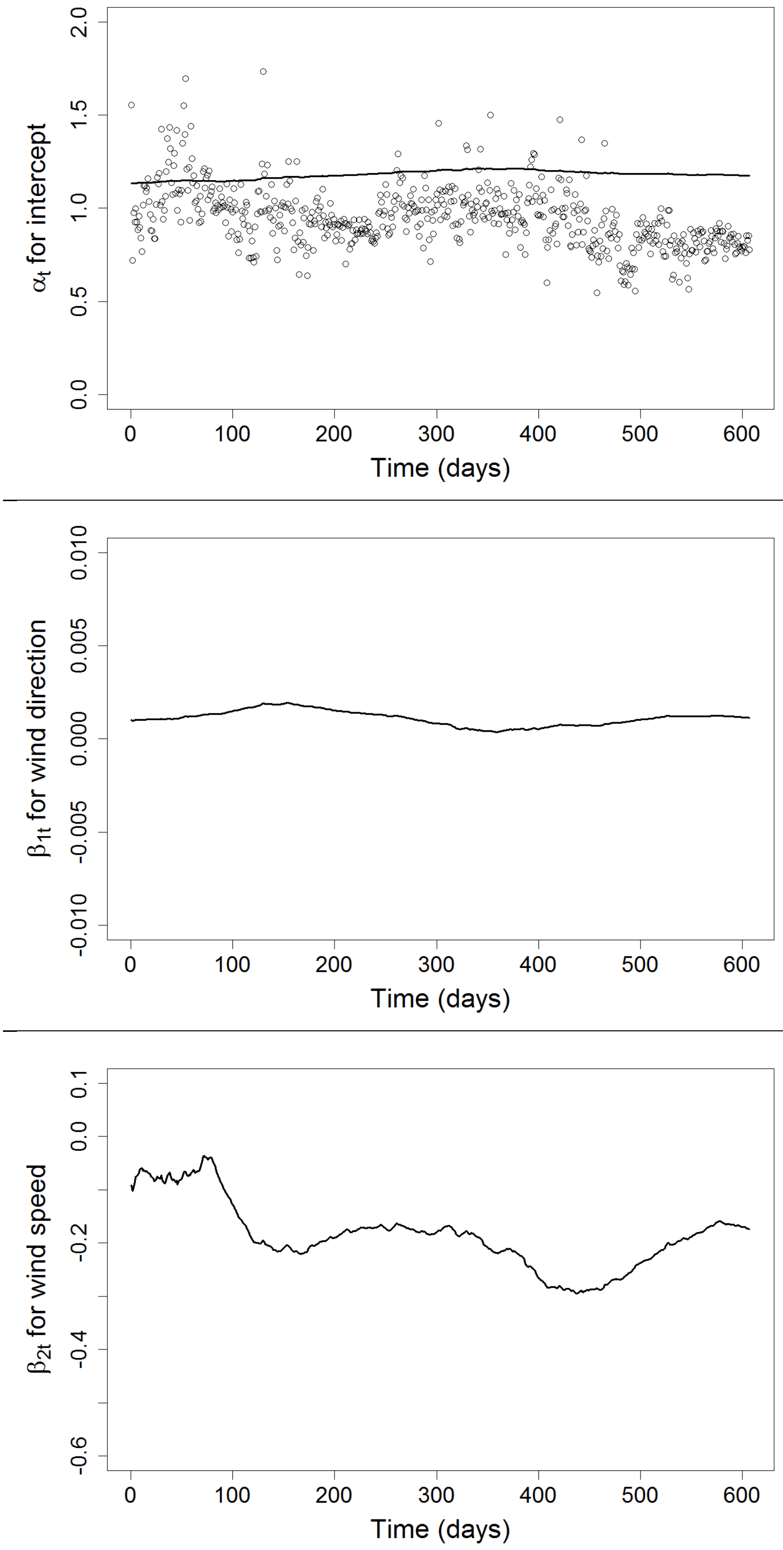


(d) solar radiation

(e) humidity

(f) temperatures

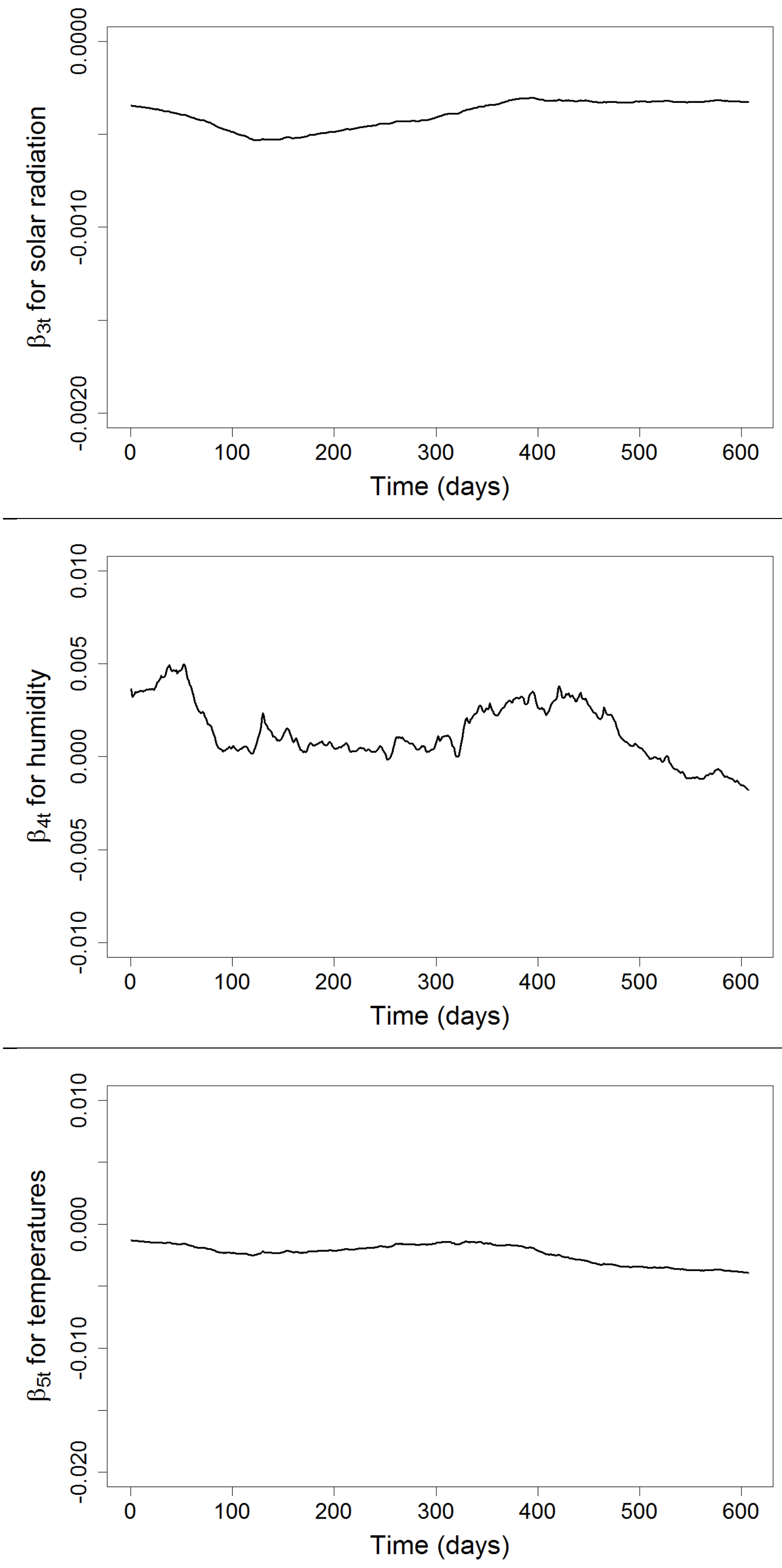
Figure 5.40 Estimated meteorological parameter coefficients for NO. The black dot is an observation of square-root NO concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry



(a) intercept

(b) wind direction

(c) wind speed

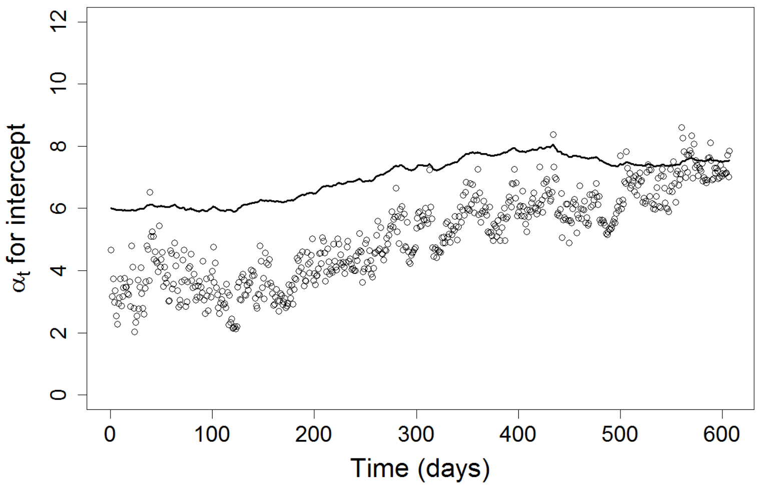


(d) solar radiation

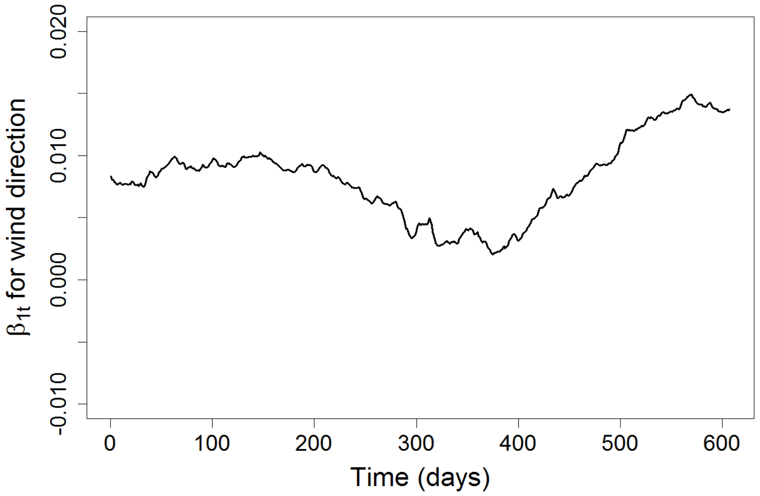
(e) humidity

(f) temperatures

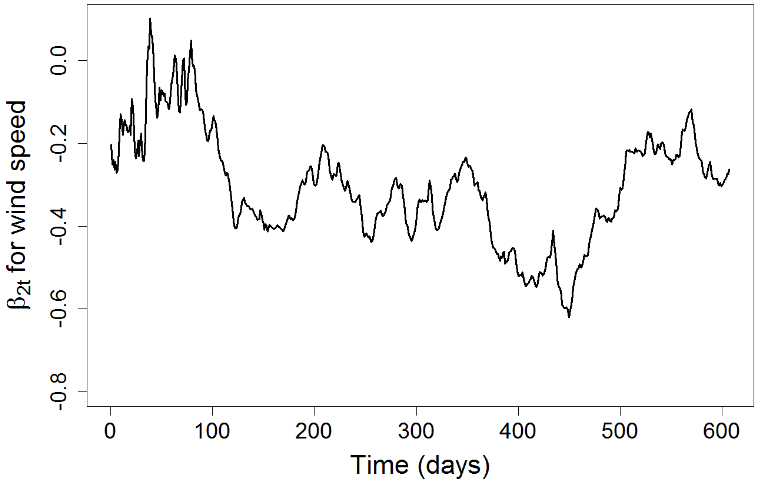
Figure 5.41 Estimated meteorological parameter coefficients for CO. The black dot is an observation of square-root CO concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry



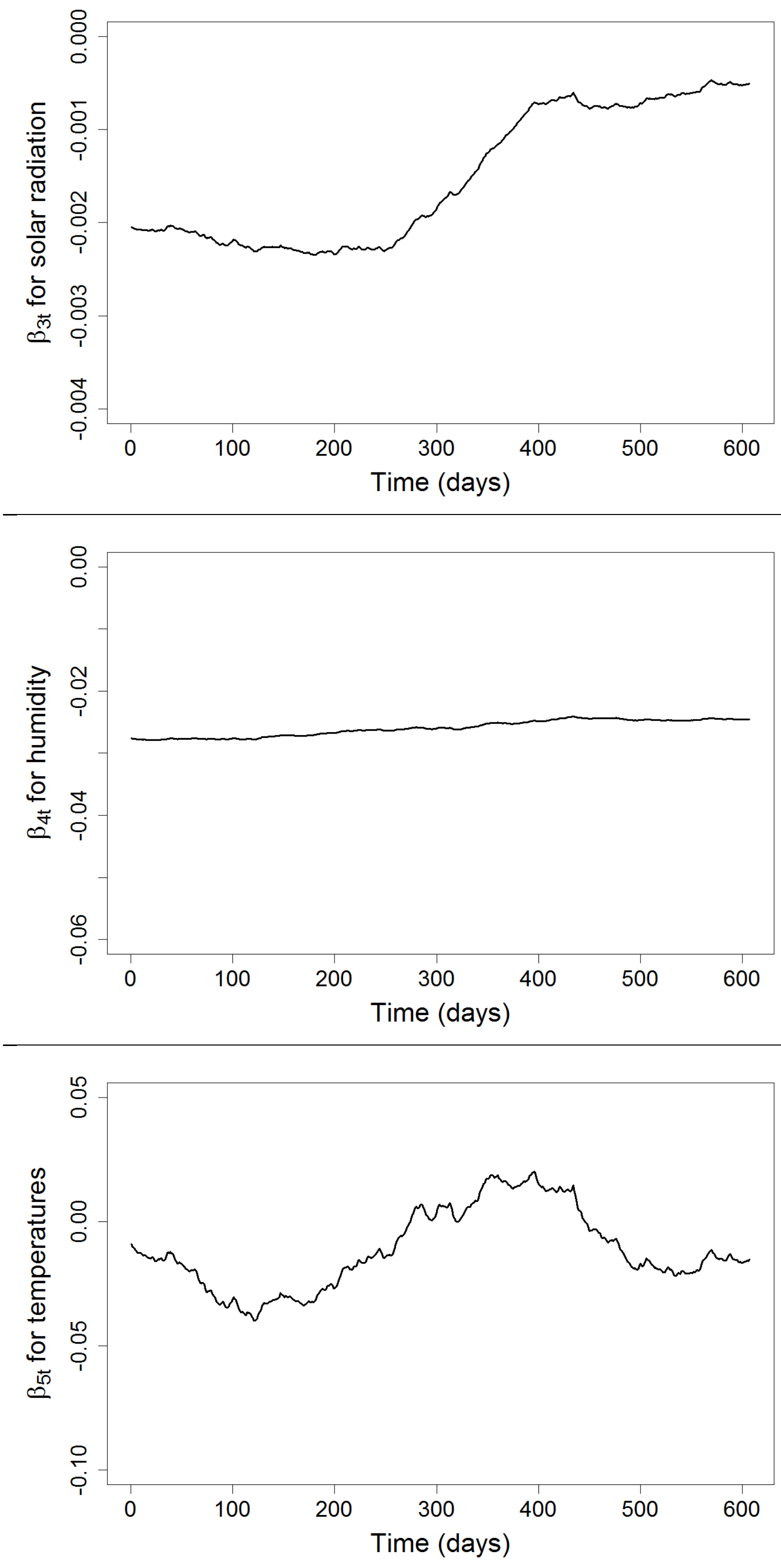
(a) intercept



(b) wind direction



(c) wind speed



(d) solar radiation

(e) humidity

(f) temperatures

Figure |5.42 Estimated meteorological parameter coefficients for SO_2. The black dot is an observation of square-root SO_2 concentration in ug/m^3. Season segment 1-59: wet, 59-242: dry. 242-424: wet, 424-607: dry

Overall, except for O_3, the effect of meteorological factors towards the concentration was small although significant as shown by the dynamic changes of estimates. The results shown by state-space models have been in agreement largely with the results obtained from multilevel model.

5.5.3 Pollutants Concentrations Subpatterns

In the present study, one of our objectives is to evaluate several different concentration subpatterns concerning spatial pattern that determine the NO_2 temporal pattern using state-space model. However, in order to obtain optimum set up and best subpattern composition we have set up several hypotheses. The hypotheses are:

Uniform (m = 1): all NO_2 concentrations are similar regardless of the site locations. There is only one single concentration trajectory, and all observation values are independent measurements of this trajectory.

Zones (m = 4): Since monitoring station 3 and 5 are located on suburban area, we group them as one, representing one that resides in suburban area. Other monitoring sites are representative by them alone based on their zone: city center, trading zone, and near highway zone.

Traffic (m = 3): We group station in city center and suburban sites become one because we assume that in the city center, the contribution of NO_2 are mainly from traffic related to school and work of people from all directions. This behavior may be represented by people in the suburban areas. This pattern might be slightly different in the trading zone and near highway zone. In the trading zone, the traffics are dominantly affected by business activity hours. On the highway zone, we note that the increase starts earlier than other zones, due to the highway connects Surabaya City with other neighbor cities.

Independent (m = 5): each monitoring site represents its own unique NO_2 temporal patterns and the emission source of the pollutants is completely independent between one and another.

To obtain best model based on several scenarios explained above, we tested several alternative for model input. We allow process variation to be equal or different across different concentration subpattern, and to be independent or correlated, represented in the matrix input **Q**. We also try several variation of observation variation, denoted as **R**, to be different of equal across subpattern.

To estimate parameters and observe models, we used the Kalman Filter and EM algorithm explained by Shumway and Stoffer (2006), which is implemented in the package MARSS (Holmes, Ward, and Wills, 2012) which is run through R open source platform. To select best model for subpattern study, we use AIC*b*, described as Akaike's Information Criterion based on bootstrapping method (Cavanaugh and Shumway, 1997). This method uses bootstrapping to estimate the penalty term. Since in the present study we try variation of **Z**, we use AIC*b* for model selection. The AIC*b* values generated were used to judge the model with lower AIC*b* value means that the model has more support over larger value.

A time series state-space model framework is useful to determine the subpattern of pollutant's concentrations among zones, and the correlation helps us infer which factors or mechanisms that causes the similarity in the patterns. In this section we focus on NO_2 only as example of state-space model and also because of the analysis from previous chapter and due to the main focus is traffic-related pollutant.

In regard of NO_2, the pattern of NO_2 can be described as follows. We observed two-peaked cycle every day that reflected peak concentration of NO_2 on those days. NO_2 are mainly from traffic emission, and therefore it may indicate the traffic congestion level. During the congestion level, people are exposed to the risk of health problems due to high possibility of NO_2 concentration exceeding the permissible level, 92.5 (ug/m^3). It is therefore mandatory to observe daily as well as weekly pattern of high concentrations.

In terms of policy related to the emission, we note that the presence of the monitoring stations is essential. Its existence is crucial for a number of vital reasons. By investigating the pattern of the concentration, in particular NO_2, we can issue preventive action in the future. However, one of the problems regarding of policy related to monitoring stations is that one policy may be not feasible to be implemented to other zones. In this instance, the location of monitoring stations depends on the nature of each zone. There are five zones represented by each station e.g, city center, trading zone, suburban in the west side of Surabaya, near highway zone, and suburban area in the east side of Surabaya. Each has its own policy implication.

By realizing the fact above, it is necessary to first doing preliminary study to obtain clear pattern of NO_2 concentrations. In one of our hypotheses, we try a scenario where we assume that the pattern is global, by indicating $m = 1$ in the MARSS framework model. This set up means that the NO_2

concentrations on those five zones are alike and similar. These may refer that the nature of fluctuation is similar, probably due to similar emission sources, or due to global atmospheric reaction because of the sensor location. If the model based on this scenario is chosen based on the selected AIC*b*, then we could easily derive draft policy that can be implemented to all zones. Unfortunately this is not the case in this study.

The best model to observe the NO_2 concentrations subpattern across zones was obtained with *m* = 4, based on the zone characters (Table 5.24). The division of zones was based on the goal of the installation of the sensors. There were two sensors were installed in the suburban area, and the grouping of these two zones yielded the lowest value of AIC*b*. We can conclude that the patterns of NO_2 concentrations on these two zones are similar and alike. The best model allows each structure (group) to have a unique concentration fluctuation. However, the best model for PM_{10} was obtained with *m* = 5 (Table 5.25) suggesting each profile represents unique pattern on each location. Also, the result suggested complex emission sources of PM_{10} that were unique on each station.

Table 5.24 Model performance based on Akaike's Information Criterion (AIC) *b*-value from the five hypotheses for the NO_2 concentration subpattern

Scenario	Parameters			Hypotheses (*m* = number of subpattern)			
	U	Q	R	Atmospheric (*m* = 1)	Zones (*m* = 4)	Traffic (m = 3)	Independent (*m* = 5)
1	Same	Same	Same	2263.074	2930.414	2680.071	736.812
2	Same	Same	Unique	1901.79	1408.759	1487.6	781.455
3	Same	Unique	Same		2936.534	2671.132	710.183
4	Same	Unique	Unique		1430.857	1460.005	681.419
5	Unique	Unique	Same		2935.051	2670.7	698.003
6	Unique	Unique	Unique		1437.114	1472.085	707.339
7	Unique	Same	Unique		1499.306	1496.517	777.7477
8	Unique	Same	Same		2938.922	2680.87	765.8606
9	Same	Correlated	Same		2147.435	2191.231	-787.492
10	Same	Correlated	Unique		**194.7197**	595.911	-815.400
11	Unique	Correlated	Same		2165.386	2223.45	-786.765
12	Unique	Correlated	Unique		202.8271	593.179	-797.328

Table 5.25 Model performance based on Akaike's Information Criterion (AIC) *b*-value from the five hypotheses for the PM_{10} concentration subpattern

Scenario	Parameters			Hypotheses (*m* = number of subpattern)			
	U	Q	R	Atmospheric (*m* = 1)	Zones (*m* = 4)	Traffic (m = 3)	Independent (*m* = 5)
1	Same	Same	Same	12544.17	12417.13	12223.34	11268.52
2	Same	Same	Unique	11876.08	11572.64	11506.67	11134.42
3	Same	Unique	Same		12433.46	12146.2	11053.97
4	Same	Unique	Unique		11600.42	11509.95	11087.05
5	Unique	Unique	Same		12423.16	12148.53	11105.4
6	Unique	Unique	Unique		11621.89	11515.87	11127.08
7	Unique	Same	Unique		11758.86	11659.17	11294.77
8	Unique	Same	Same		12509.72	12228.82	11297.62
9	Same	Correlated	Same		11762.24	11731.25	10300.2
10	Same	Correlated	Unique		11012.24	11085.38	**10292.33**
11	Unique	Correlated	Same		11928.71	11724.88	10304.19
12	Unique	Correlated	Unique		11359.52	11086.51	10313.44

The best model obtained has unconstrained variance-covariance matrix **Q** for the process variation. Each concentration subpattern has different process variance whereas it also allows correlation in the temporal variation between subpatterns (Table 5.26 and Table 5.27). We can observe that among zones there are high correlations, and all of them are positive.

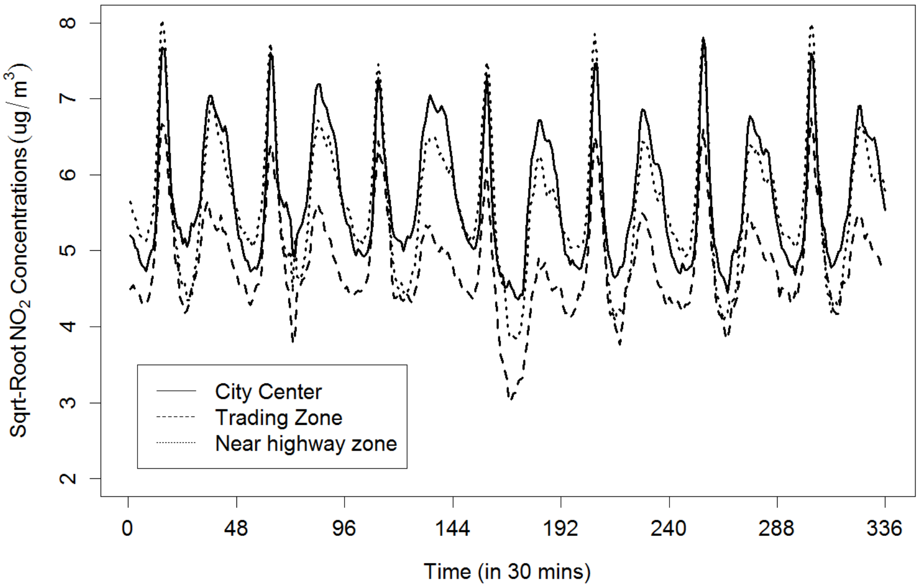
Table 5.26 Process variation and correlation among zones

Subpattern	Process variation (Q)	Correlation (Q)			
		City center	Trading	Suburban	Near highway
City center	0.093	1.000	0.787	0.781	0.861
Trading zone	0.052	0.787	1.000	0.671	0.798
Suburban	0.063	0.781	0.671	1.000	0.723
Near highway	0.098	0.861	0.798	0.723	1.000

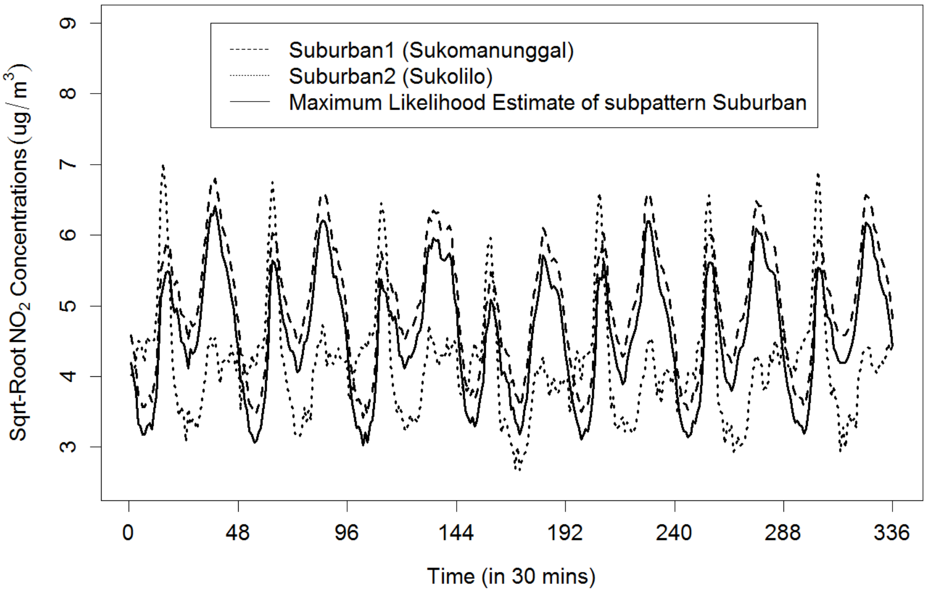
Table 5.27 Process variation and correlation among zones

Subpattern	Process variation (Q)	Correlation (Q)				
		City center	Trading	Suburban1	Near highway	Suburban2
City center	22.525	1.000	0.611	0.614	0.751	0.778
Trading zone	42.135	0.611	1.000	0.393	0.563	0.598
Suburban1	22.158	0.614	0.393	1.000	0.484	0.497
Near highway	78.542	0.751	0.563	0.484	1.000	0.754
Suburban2	35.067	0.778	0.598	0.497	0.754	1.000

The results show that NO_2 concentrations subpattern consisted of 4 zones (m = 4) with a parameter set up unique concentration fluctuation, **u**, correlated and unconstrained variance and covariance matrix, **Q**, and unique observation (measurement) variation error which means the variation is restricted to site-specific. The configuration result of **Q** allow us to derive conclusion that each subpattern of NO_2 concentration has their own characteristic, however, the main limitation of this study was we may be able to see major usefulness of this distinctness (Figure 5.43). We only have five stations whereas the best model was with m = 4, therefore allowing other 3 (three) hidden states to be equal with the measurement value of NO_2 concentration in city center, trading zone, and near highway zone. The power of the time-series state-space analysis will be more observable and useful if we have more sites to observe. However, this problem is common that developing countries are facing.



(a) City center, trading zone, near highway zone



(b) suburban

Figure 5.43 Maximum likelihood (ML) estimates. Note that the estimates in (a) are equal with actual NO_2 measurement, whereas in (b) the continuous line represents the ML estimates with other lines as the actual NO_2 measurement. The time interval starts from Thursday (0-48) in the x-axis

5.6 Conclusions

This study investigated the phenomena and the events where specific pollutants exceeded permissible levels during 2001-2002 in Surabaya. ICA has been very helpful in understanding the characteristics of pollutants. Its main usefulness is to draw daily and seasonal patterns from data of all pollutants. We were able to draw a seasonal pattern that occured every day, only workdays, only weekends, the Monday to Saturday pattern, and certain days-only pattern using ICA. By using ICA, we were also able to confirm the hypothesis we built up that each station may represent a unique daily pattern. Despite the fact that some timeframes occured on several locations, we still found pattern uniqueness between stations. The temporal patterns of pollutants, e.g., NO_2 and $PM_{10,}$ that we found were characterized by similar temporal patterns in the morning for Ketabang Kali (city center) and Sukolilo (suburban), whereas we found a wider peak morning time interval in Gayungsari. In the region where it is near the highway, one can suspect an earlier rise of NO_2 and PM_{10} concentration due to transport from/to and within Surabaya, leading to the conclusion that regions, which are passed by a highway may suffer an exposure to the high concentration of NO_2 and PM_{10}. In Perak (trading zone), the occurrence of morning peak time was delayed then in other zones, as the main driving force of transport activities were dominated by business and/or goods delivery. This fact was also true for the evening session. However, in the evening, the temporal pattern in the trading zone was similar to Suburban1, whereas on near highway zone, the peak concentration occured earlier than those two sites, showing a much segmented interval during which traffic was the highest. Besides what can be drawn from the present study, ICA can be further expanded into the air quality management system, such as evaluating the number of monitoring stations, as well as improving the prediction method.

The use of daily average NO_2 and PM_{10} concentration helped us to confirm seasonal effect on concentrations. We confirmed that for NO_2, the concentrations during the dry season were higher than for the wet season. To the contrary, the PM_{10} concentrations during the dry season were lower than in the wet season. It clearly can be seen that the deposition of PM_{10} was much higher in the wet season. On the other hand, we estimated that the higher concentrations of NO_2 during the wet season were probably due to travel behavior of people that causes more NO_2 emission.

In the present study, we also estimated sources of pollution, in particular for NO_2 and $PM_{10,}$ based on wind speed and wind direction information. In the city center, traffic contributed the most to the area where NO_2 concentrations exceeded the ambient threshold level, whereas activities from offices, industries, and the most important building, that is, a shopping mall, were responsible for the high PM_{10} concentrations. In the trading zone, the cause of NO_2 events were due to traffic in the SE direction, where there are several protocol roads heading to the Tanjung Perak harbor port, while sources that contributed to high concentration of PM_{10} were from houses, religion-related buildings, and traffic in the Tanjung Perak Barat St. area. One option to minimize the events of high concentration is by managing the time frame when motorcycles, big vehicles (trucks, container trucks) may pass, to better manage the distribution of vehicles. In Sukomanunggal, the source of NO_2 events was traffic flow, and the source of PM_{10} events were from activities of offices, a shopping mall, and a resident tower. In the near highway zone (Gayungsari) site, the time interval of the events of PM_{10} exceeded the ambient threshold level had more variation. In this location, the source of pollution came from traffic on the highway. Chapter 7 summarizes recommended action based on these results.

To confirm temporal time frame patterns and investigate the effects of meteorological variables, we employed a state-space approach on the trading zone. We were able to confirm and extract the important time frames during which pollutant concentrations may be high and potentially that during these time frames it is likely that the concentrations may breach the permissible level. Determining the timeframe is one important factor that can be kept into the basket to support policy makers in deciding policies to reduce the impact of pollutants. In this case, we focused on NO_2 and PM_{10} in the trading zone, as they are the most dynamic pollutants as well as the fact that they are emitted from transport-related activities. Moreover, the implementation of short-term policy may be improved in this zone compared with other zones because within this zone, we may propose a policy to regulate company's big vehicles, which are heading to the harbor port of Tanjung Perak in Surabaya. The time interval between 7am to 10am was selected as the most important time interval, during which we will recommend that at these times, a planned distribution of vehicles be implemented, e.g., we suggest limiting the volume of those big vehicles (trucks) on the road near the sensor, heading to the harbor port. We discussed this on Chapter 7. By imposing such a policy, it is expected that the peak concentration of pollutants will be lower, and will therefore reduce the events when NO_2 concentrations exceed the permissible level. This action is required to minimize the adverse effect on human health and the environment. The results reported suggest that the model was practical and provides an alternative way to explain the pattern of air pollutants better than from using the conventional approach.

By using a dynamic linear model, we found that other factors may have a bigger influence than meteorological variables, as indicated by a more dynamic sensitivity factor of the intercept variable, as shown by NO_2, PM_{10}, CO, and SO_2. Besides that, some meteorological factors had the most significant contribution to each pollutant. For example, with NO_2, solar radiation appeared to have less dynamic compared with other variables, whereas for PM_{10}, we noted all meteorological factors were dynamic, indicating contributions of concentration fluctuation. The dynamics of meteorological factors were helpful in understanding the phenomena of air quality, and we had shown the application of time series state-space to achieve such a goal.

We were also able to specify temporally the effect of season on the concentration of pollutants by observing dynamically changing meteorological parameters for both short-term and mid-term, using a daily average. For example, with NO_2, the patterns found during the wet season indicated the change of travel or driving pattern that causes the NO_2 concentrations to be high in this particular zone. From these results, we obtained data showing that wind direction and wind speed had important roles in the concentration of pollutants. We noted that the faster wind speed, the faster the rate of emission reduction. Moreover, humidity had a contribution to the formation of pollutants.

This study tries to explore the use of time series state-space model to identify subpattern of NO_2 concentration in Surabaya. We have identified that the patterns of NO_2 concentration represented the location of which these sensors (monitoring stations) reside. The monitoring stations in the two suburban areas were found to form a certain subpattern so that under this configuration the best state-space model was formulated. Therefore we can conclude that each of the remaining zones best represent characteristics of emission source of each zone. In the trading zone, the subpattern may indicate uniqueness of the subpattern among other zones.

Chapter 6 Prediction of Air Quality Concentrations Based on Time Series Models and Generalized Least Square Models

6.1 Introduction

One part of exploiting the usefulness of monitoring data is we can use the data to predict future concentrations. The purpose in prediction is to estimate next pollutant levels so that this information may be shared to the users for preventive actions. These preventive actions are necessary to avoid people get exposed to the risk of pollutants which may exceed the permissible level. Pollution on which it exceeds the permissible level will affect human health. Not only human health but also ecosystem is affected by pollution. To increase awareness of such events, information of future concentration is crucial and vital.

There are several types of prediction model. One is time series model and other type is model that includes exogenous variables. One example of ordinary time series model is autoregression. However, many types of time series model have been developed, as explained in Chapter 2. The challenges emerge when sometimes some air quality data cannot be predicted very well with only time series model due to the nature of air quality, instead it requires additional exogenous variables (e.g., meteorological variables) that may highly represent the variation of concentrations. Therefore, it is one of major motivation why the prediction model is done in the present study.

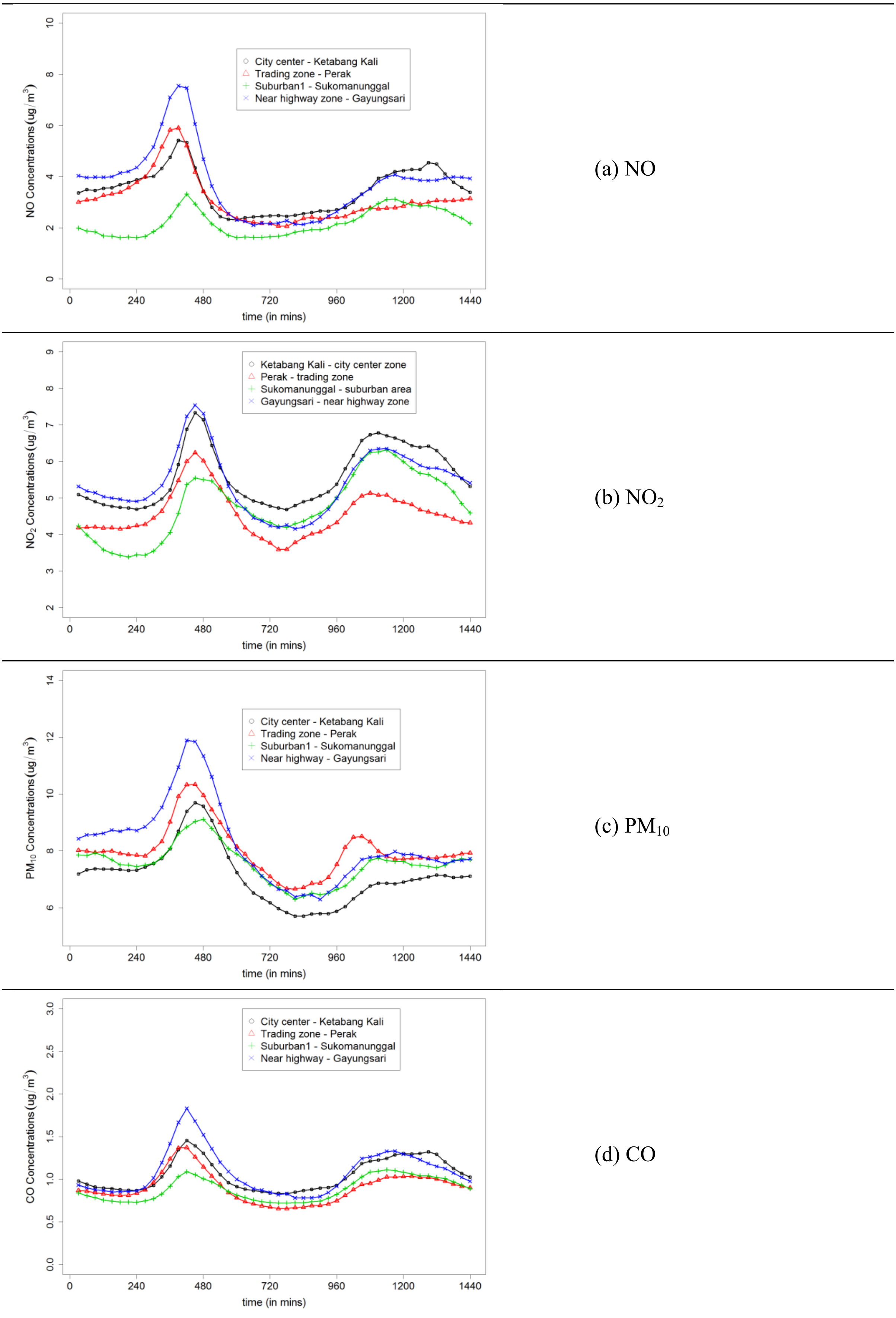
To compare the performance between time series model and model with exogenous variables we investigated several models. For time series model, we tested and compared several models: autoregression, Bayesian Markov Switching model, Modified Bayesian Switching model, as well as ARIMA to predict pollutant's concentrations and true distribution air quality (independent components). These time series models that predict pollutant concentrations were compared with models utilizing exogenous variables. For this purpose, we employed a Generalized Least Square (GLS) model because this model allows can handle time series residual autocorrelation. However, sometimes on air quality data, there is multicollinearity among variables. Moreover, there are noises inside the variables. Because of multicollinearity and noises we employed Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to obtain variables that are relatively clean from noises and multicollinearity. Therefore, in this chapter we aimed to investigate time series models and GLS models.

6.2 Materials and Data

In this chapter we presented two different concepts of prediction models each using different time scale. However, the performance is comparable since the difference of time scale is not big, between hourly data and 30-mins data. The details of data and validation are described in the following session.

6.2.1 Data for Time series models

We utilized hourly mean NO, NO_2, PM_{10}, and CO square-root concentrations data, taken from four stations on Surabaya City, starting from 1st February 2001 to 15th February 2001 (15 days). Hourly mean concentrations were considered sufficient to capture temporal changes due to the influence of traffic flows in four stations. Figure 6.1 displays daily diurnal NO, NO_2, PM_{10} and CO concentrations, averaged of 607 days starting from 1st February 2001. It can be observed that peak concentrations were achieved between 6:00 am (i.e., 360 mins from 0:00 am) to 8:00 am (i.e., 480 mins). On the other hand, in the evening, the peak concentrations were observed between 6:00 pm (i.e., 1080 mins) and midnight; however the concentrations did not appear obviously as in the morning peak.



(a) NO

(b) NO$_2$

(c) PM$_{10}$

(d) CO

Figure 6.1 Daily diurnal pollutants concentrations (square-root) on four stations

The data sets used are from Station one, located at Ketabang Kali (the city center), Station two – Perak (trading zone), Station three – Sukomanunggal (suburban area), and Station four – Gayungsari (highway zone). We utilized an Expectation-Maximization based algorithm to impute missing values as described in the Chapter 3.

Next 24-hr ahead hourly NO, NO_2, PM_{10}, and CO concentrations were predicted based on the information of previous hourly concentrations. The resulting data sets for parameter estimation and evaluation are shown in Table 6.1. Each of data set consists of 336 samples of hourly mean data for parameter estimation, and 24 samples for evaluating the differences between observed and predicted values. To clarify the influence of sample size, we also compared the performance of a model using more data with 720 observations. Figure 6.2 shows 336 data of hourly mean data, starting from 1 February 2001 for all four locations after transformation for NO concentrations (as example). Figure 6.3 illustrates the differences of data used for parameter estimation and evaluation (forecasting) at the Gayungsari station. Table 6.2 shows statistical properties of transformed data at each station. We can observe that the mean value was found to be the lowest at Sukomanunggal, which makes sense because the station is located at a suburban area. Furthermore, the characteristics at Perak and Gayungsari stations are similar in terms of standard deviations, although the range is bigger in Gayungsari.

Table 6.1 Data for parameter estimation, evaluation and forecasting for each station

Monitoring Station	Data for parameter estimation	Data for evaluation and forecasting
Ketabang Kali (city center)	336	24
Perak (trading zone)	336	24
Sukomanunggal (suburban area)	336	24
Gayungsari (highway zone)	336	24
Ketabang Kali – B	720	744

In this part, not only we used mean-hourly, but also 4-hour data to obtain best time interval for prediction. Furthermore, to improve the models, we aimed to predict true distribution of air quality instead of pollutant's concentrations. The true distribution patterns were extracted from ICA method.

6.2.2 Data for GLS models

We made use of 30-mins interval concentrations of NO, NO_2, O_3, SO_2, CO, and PM_{10} as well as meteorological factors that consist of wind direction, wind speed (m/s), solar gradiation (W/m^2), humidity (%), and temperatures (^oC) as independent variables. For wind direction we created eight variables representing direction as dummy variables. They are north, northeast, east, southeast, east, south, southwest, west, and northwest, with north as base reference. These variable values were different for each station, thus creating more 35 wind direction variables for model input.

We attempted to predict next-30mins ahead of NO_2, PM_{10}, and O_3 using GLS model using the data taken from March 2002 to April 2002, with total 14635 observations (five stations), as training set, while the test set, which was not used for parameter estimation, was taken from May 2002. Table 6.2 shows shows each mean value, standard deviation, and median of each pollutant concentrations (in ug/m^3) from five stations in Surabaya. It can be seen that average emission that is related to traffic (NO_2) was high in city center and low in suburban2 on the east side of Surabaya. With regards of PM_{10}, high average concentration was found on suburban1 on west side of Surabaya. Interestingly, mean value of O_3 concentration was high on highway zone suggesting high reaction rate between NO and NO_2.

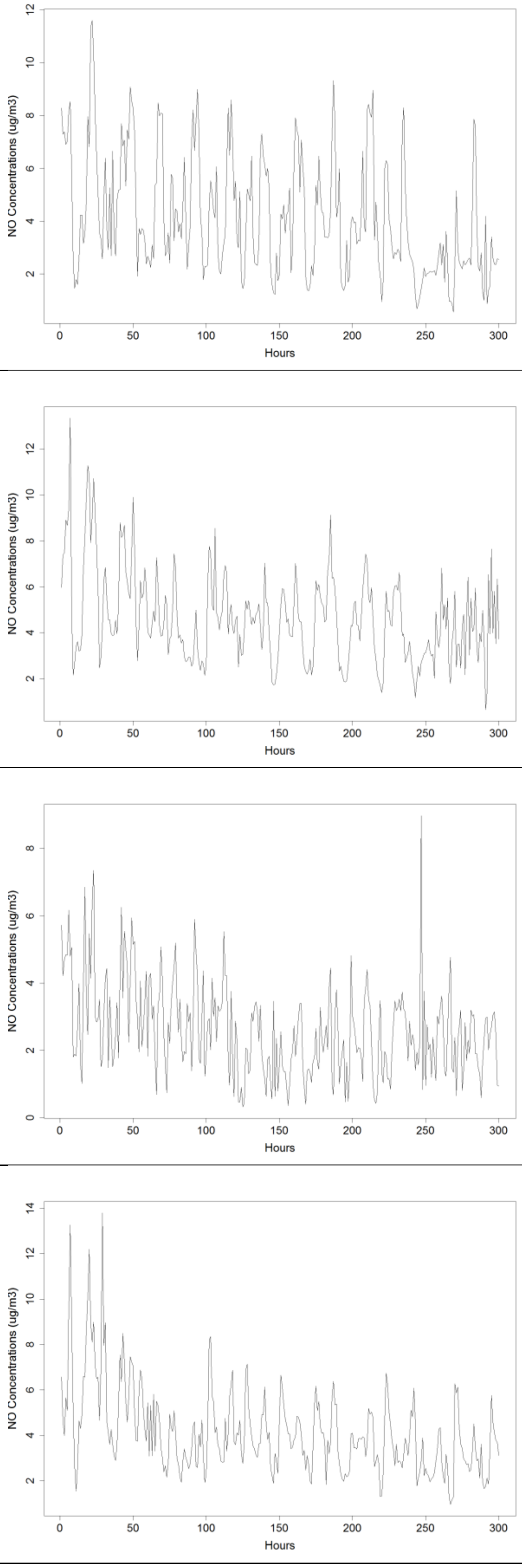
Table 6.2 Descriptive statistics of levels of pollutants in Surabaya from five monitoring stations (ug/m^3)

	Description	City Center	Trading	Suburban1	Near Highway	Suburban2
NO_2	Min	1.603	0.335	1.175	0.48	0.055
	Max	12.309	11.577	10.013	16.024	8.538
	Mean	5.487	4.745	4.85	5.266	4.151
	Standard Deviation	1.518	1.703	1.503	1.758	1.62
PM_{10}	Min	0.656	0.1	0.317	0.541	0.117
	Max	17.38	48.799	17.689	48.99	14.979
	Mean	7.137	7.673	7.857	7.488	6.79
	Standard Deviation	2.437	2.768	2.527	3.122	2.38
O_3	Min	1.558	0.042	0.01	0.042	0.174
	Max	15.612	13.613	13.976	26.088	14.453
	Mean	6.067	6.305	5.614	4.372	6.864
	Standard Deviation	1.984	1.579	1.941	2.21	2.232

A GLSs were then fitted, with square-root transformed of each pollutant (NO_2, O_3, or PM_{10}) as the dependent variable. The explanatory variables were functions of 30-mins interval of pollutant levels: NO_2, NO, O_3, SO_2, CO, PM_{10}, wind speed, solar gradiation, humidity, temperatures, status of day (weekends, workdays as base reference), peak time of morning and afternoon session (non-peak time as base reference with peak time morning is between 6:30am to 9am and in the evening between 4:30pm to 7pm), holidays, spatial covariates of zones: trading, suburban1, highway, suburban2, with city center as base reference, and wind direction. All air pollutants and wind speed were all square-root transformed as standard procedures to stabilize the variance. Each dependent variable was predicted by the interaction of other pollutants one-step backward (last 30-mins concentration). The three pollutants were considered in three separate GLS models because of their substantial correlation. On a second and third model, the variables of six air pollutants and four meteorological factors were replaced by components extracted from an ICA and PCA. Total there were 10 ICs and 10 PCs were obtained and used as predictor variabels along with other independent variables as described above.

6.3 Time series models

Here, three models will be described, including a first-order autoregressive model (AR(1)), a Bayesian Markov Switching (BMS) model, and a modified BMS (MBMS) model. In the AR(1) model, the current pollutant's concentrations were calculated by using a constant parameter and a prior information of concentrations (here, the concentration observed one hour ago is used to compute the current concentrations). In the BMS model, we assumed that time series data can be divided into two types, each of which shared similar data pattern related to the relationship between a prior concentration and the current concentration. The two types were captured by two latent states. In other words, a state refers to a series of data with similar attributes or patterns. When a state is built along with its parameters, values can be fitted. The parameters inside a state consist of a constant parameter and a parameter describing a prior information of concentration, which is defined in the AR(1) model. The MBMS model attempts to make full use of forecasting errors obtained in the BMS model in order to improve the forecasting ability, where the errors the BMS model are divided into different intervals.



(a) City center

(b) Trading zone

(c) Suburban1

(d) Near highway

Figure 6.2 Hourly mean NO concentrations on four sites

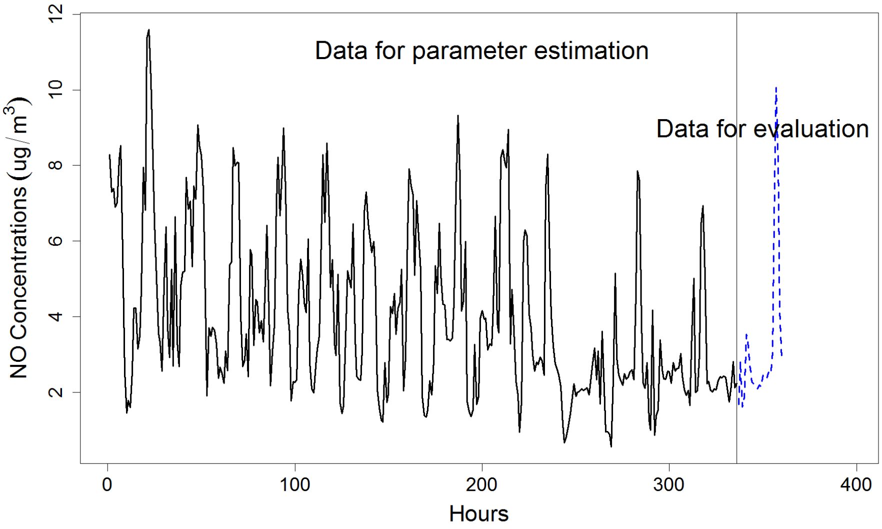


Figure 6.3 Data for parameter estimation and evaluation in near highway station (as example)

Table 6.3 Statistical properties of hourly NO, NO_2, PM_{10}, and CO concentrations

Parame-ters	Statistical measures (ug/m^3)	Ketabang Kali	Perak	Sukomanunggal	Gayungsari
NO	Minimum	0.556	0.648	0.323	0.958
	Maximum	11.60	13.35	8.966	13.80
	Mean	4.154	4.630	2.607	4.154
	Standard deviation	2.218	1.913	1.389	1.860
NO_2	Minimum	2.339	1.921	0.173	0.667
	Maximum	9.879	9.819	8.642	9.879
	Mean	4.974	4.924	3.614	4.241
	Standard deviation	1.392	1.369	1.603	1.396
PM_{10}	Minimum	1.103	2.451	0.075	3.014
	Maximum	16.899	16.586	28.768	37.884
	Mean	6.497	7.440	6.363	7.809
	Standard deviation	1.716	1.956	3.271	3.626
CO	Minimum	0.462	0.204	0.229	0.480
	Maximum	2.668	2.467	1.960	2.493
	Mean	0.992	0.930	0.767	0.927
	Standard deviation	0.328	0.309	0.256	0.322

6.3.1 A first-order autoregressive (AR) model

Concentrations of air pollutants are often related to previous concentrations, often explained as *lagged* data. In other words, data in the present may be closely linked with those in the past. Therefore we explored the use of a simple time series model with lag-1. We used a first-order autoregressive (AR(1)) as a start off to fit and model NO concentrations. AR model has been widely applied because of its fast and robust implementation. Autoregressive model of order p, AR(p=1) has the form:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \qquad (1)$$

Where $\alpha = \mu(1-\phi)$, therefore the equation used in this study become

$$y_t = \mu(1-\phi) + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \qquad (2)$$

where $\emptyset_i$indicates the ith unknown autoregressive parameter, α is the intercept term, μ is the mean term, and ε_t is an error term. Each AR term corresponds to the use of a lagged concentration in the equation. The term autoregressive refers to the fact that the model computes the current concentration values y_t as a function of past concentration values $y_{t-1}, y_{t-2},\dots y_{t-p}$.

6.3.2 Autoregressive Integrated Moving Average (ARIMA) model

ARIMA is a combination of two processes AR and MA (moving average). AR has been explained previously, however in this step we determine the AR(p) where the p indicates the number of lag or previous steps needed to forecast the current value. The moving average of the order q, as shortened with MA(q) assumes the white noise w_t up to lags q are integrated linearly to build observed data. A time series ARMA(p,q) model can be represented as (Shumway and Stoffer, 2006; Kumar and Jain, 2009):

$$y_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \tag{3}$$

with $\phi_p \neq 0$, $\phi_q \neq 0$, and $\sigma_w^2 > 0$ and the parameters p and q refer to the autoregressive and the moving average orders. Unless state otherwise, $\{w_t; t = 0 \pm 1, \pm 2, \dots\}$ is a Gaussian white noise sequence. The ARIMA(p,d,q) is used when applying the stationary differenced time series where d denotes the order of differencing of the data. The estimation of ARIMA coefficients (ϕ, θ) for time series data can be determined using conditional-sum-of-squares to find starting values followed by maximum likelihood (Hyndman *et al.*, 2014).

In constructing an ARIMA(p,d,q) model both autocorrelation function (ACF) and partial autocorrelation function (PACF) might be useful. However, we used objectively defined criterions such as AIC, AICc, and BIC. These information are statistical model fit measures and applied in the algorithm inside a package *forecast* (Hyndman *et al.*, 2014) run under R platform.

6.3.3 A Bayesian markov switching model

Here, we build a BMS model based on the autoregressive lag-1 concept, where each state constitutes a parameter explaining the effect of lag-1 concentration. Consider the following equation:

$$y_t = \mu_{S_t} + \beta_{S_t} y_{t-1} + e_t \tag{4}$$

where S_t refers to a state (1…k), and e_t follows a normal distribution with zero mean and variance $\sigma_{S_t}^2$. The intercept μ_{S_t} represents switching states with respect to state S_t. And k states lead to k values for μ_{S_t} and $\sigma_{S_t}^2$. y_{t-1} represents lagged NO concentration with β coefficient corresponding to each state k.

The transition of states for a markov switching regime is stochastic, where dynamics of the switching process is known and driven by the following transition matrix and p_{ij} refers to the probability of a switch from state i to state j.

$$\begin{bmatrix} p_{ii} & \cdots & p_{ik} \\ \vdots & \ddots & \vdots \\ p_{kj} & \cdots & p_{ij} \end{bmatrix} \tag{5}$$

where i and j indicates states 1,…, k.

This study only considers two states (k = 2) and equation (3) is re-written as,

$$y_t = \mu_1 + \beta_1 y_{t-1} + e_t \text{ for state 1} \tag{6}$$
$$y_t = \mu_2 + \beta_2 y_{t-1} + e_t \text{ for state 2} \tag{7}$$

where:

$$e_t \sim (0, \sigma_1^2) \text{ for state 1} \tag{7}$$
$$e_t \sim (0, \sigma_2^2) \text{ for state 2} \tag{8}$$

The probability of a switch from state 1 to state 2 between time t-1 and t is given by p_{12}. Analoguely, the probability of staying on state 2 is determined by p_{22}. The model is estimated based on the standard maximum likelihood method (Perlin, 2010). The log-likelihood of the model is given by:

$$logL = \sum_{t=1}^{T} log\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{y_t-(\beta_t X)}{2\sigma^2}\right)\right) \tag{9}$$

where X is variable composed of integer 1 to form intercept value (μ) and NO concentration (y_t). For equation (7), if all of the states were known, the values of St are available, the estimating parameters is straightforward, by maximizing equation (9) as a function of parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. In case of a markov switching model where the states are unknown, we use the following equation:

$$logL = \sum_{t=1}^{T} log \sum_{j=1}^{2} \big(f(y_t|S_t = j, \Theta) Pr(S_t = j)\big) \tag{10}$$

where $f(y_t|S_t = j, \Theta)$ indicates the likelihood function for state j conditional on a parameter set (Θ). Equation (10) is a weighted average of likelihood functions for all the states, where $Pr(S_t = j)$ describes the weight. The weights are determined by state probabilities. When these probabilities are not observed, we make inferences on the probabilities based on the available information, using Hamilton's filter. Detailed calculation steps are given by Perlin (2010). Consider ψ_{t-1} as the matrix of information (intercept with value of 1 and concentration) at time t-1, we can estimate $Pr(S_t = j)$.We set up an initial guess for the starting probabilities $Pr(S_0 = j)$ for j = 1, 2 at t = 0. The unconditional probabilities of S_0 are given by:

$$Pr(S_0 = 1|\psi_0) = \frac{1-p_{11}}{2-p_{11}-p_{22}} \tag{11}$$
$$Pr(S_0 = 2|\psi_0) = 1 - \frac{1-p_{11}}{2-p_{11}-p_{22}} \tag{12}$$

Prior probabilities of each state given information on t – 1 is:

$$Pr(S_t = j|\psi_{t-1}) = \sum_{i=1}^{2} p_{ji}\big(Pr(S_{t-1} = i|\psi_{t-1})\big) \tag{13}$$

p_{ji} is a transition probability from state j to state i. In this study, we assume that the transition probability between states is 0.2, which indicates $p_{11} = p_{22} = 0.22$, and $p_{12} = p_{21} = 0.8$. The first value of equation $\big(Pr(S_{t-1} = i|\psi_{t-1})\big)$ in Equation (13) is obtained from equation (11) and (12). The value of $Pr(S_t = j|\psi_{t-1})$ in (13) will be used as input in the equation (14).

Posterior probability of each state is updated using new information from time t using Bayes' rule:

$$Pr(S_t = j|\psi_t) = \frac{f(y_t|S_t=j,\psi_{t-1})Pr(S_t=j|\psi_{t-1})}{\sum_{j=1}^{2} f(y_t|S_t=j,\psi_{t-1})Pr(S_t=j|\psi_{t-1})} \tag{14}$$

The value of $Pr(S_t = j|\psi_t)$ in the equation (14) will be used as input $\big(Pr(S_{t-1} = i|\psi_{t-1})\big)$ in the equation (13). Repeating the above process for all observations, one can obtain the resulting log-likelihood function. The value of $Pr(S_t = j|\psi_t)$ from equation (15) come from $Pr(S_t = j|\psi_t)$ equation (14).

$$logL = \sum_{t=1}^{T} log \sum_{j=1}^{2} \big(f(y_t|S_t = j, \Theta) Pr(S_t = j|\psi_t)\big) \tag{15}$$

As seen above, the BMS model has flexibility in handling processes driven by heterogeneous states of the world. Details of the models are explained in existing literature (e.g., Goldfeld and Quantd, 2005; Perlin, 2010).

6.3.4 A modified Bayesian Markov Switching Model

Here, we build an MBMS model, which concept was originally proposed by Ji *et al.* (2013) for the case of wind speed. First, we run the above BMS model with lag-1 information coefficient and calculate the fitted values by the model to all observations on the training data set. Second, we obtain forecasting errors by subtracting the fitted values from the observed values. Third, we divide these forecasting errors into four states (denoted by m), each of which indicates a particular interval of forecasting error. We adopt four states, following the suggestion by Ji *et al.* (2013), but with different interval values. It must be noted that the states here differ with the states in the BMS model. The term states here represent intervals which are built to improve forecasting errors. In this case,

$$\otimes\, i = \left[\widetilde{\otimes}_{1i}\widetilde{\otimes}_{2i}\right], \widetilde{\otimes}_{1i} = \hat{X}_t + M_i, \widetilde{\otimes}_{2i} = \hat{X}_t + N_i \tag{16}$$

where, $\hat{X}_t$ is a vector of forecasted values from the BMS model, M_i is a vector of minimum forecasting errors in state i, and N_i is a vector of maximum forecasting error in state i.

As stated, there are four states, which consists of two states represent positive errors and other two states represent negative errors. We collect minimum and maximum value of errors and divide these to half. Forecasted errors belong to State 1 if the error is between 0 and ½ min(e_t), with e_t is the error from the training data set. Analogously, State 2, denoted by $\otimes_2$ is for interval e_t < ½ min(e_t), further $\otimes_3$ for interval 0 > e_t ≥ ½ max(e_t), and finally $\otimes_4$ for interval e_t > ½ max(e_t). Min and max refers to the minimum and maximum value of errors in the training data set. We determine the states of forecasted value by calculating the error between fitted values from BMS model with observation value from the last position.

Finally, we revise the forecasted values based on the following equation, where $\hat{X}_t$ is the revised value.

$$\hat{X}_t = X_t + \frac{1}{2}(\otimes_{i1} + \otimes_{i2}) = X_t + \frac{1}{2}(M_i + N_i) \tag{17}$$

$\hat{X}_t$ is a vector of forecasted values from the BMS model, M_i is a vector of minimum forecasting errors in state i, and N_i is a vector of maximum forecasting error in state i.

6.4 GLS Models

6.4.1 Generalized Least Square

We employed a Generalized Least Squares (GLS) model to formula the mixed linear effect of predictor variables towards the concentration of pollutants, following the equation:

$$y = X\beta + \varepsilon \tag{18}$$

Where y is n x 1 response variable (pollutant) and X is an n x p matrix, β is a p x 1 vector of estimated parameters, and ε is n x 1 vector of errors. With the assumption that $\varepsilon \sim N_n(0, \sigma^2 I_n)$, we can estimate ordinary least square estimator of β:

$$\beta_{GLS} = (X'X)^{-1}X'y \tag{19}$$

With covariance matrix

$$V(\beta_{GLS}) = \sigma^2(X'X)^{-1} \tag{20}$$

When the error covariance Σ positive-definite and symmetric and its diagonal entries Σ correspond to non-constant error variances, and nonzero off-diagonal entries are associated with correlated errors, we can estimate the log-likelihood of the model, given that Σ is known:

$$Log_e\, L(\beta) = -\frac{n}{2} log_e(det\Sigma) - \frac{1}{2}(y - X\beta)'(y - X\beta) \qquad (21)$$

The function is maximized by the GLS estimator of β:

$$\beta_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \qquad (22)$$

With covariance matrix:

$$V(\beta_{GLS}) = (X'\Sigma^{-1}X)^{-1} \qquad (23)$$

However, in the application, the matrix of Σ is not known and therefore must be estimated from the data with the regression coefficients, β. In time series data, though, there is a concern of error correlation. Assuming that all errors have same expectation and same variance, the covariance of two errors depends on their separation *s* in time:

$$C(\varepsilon_t, \varepsilon_{t+s}) = C(\varepsilon_t, \varepsilon_{t-s}) = \sigma^2\rho_s \qquad (24)$$

Where $\rho_s$is the error autocorrelation at lag *s*. The error-covariance matrix will become:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & & \rho_{n-3} \\ & \vdots & & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma^2 P \qquad (25)$$

For stationary time-series, we apply first-order auto-regressive process, AR(1) for autocorrelated regression errors:

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t \qquad (26)$$

Under this model, the v_t is assumed to be Gaussian white noise, $\rho_1 = \phi\varepsilon_{t-1}$, $\rho_s = \phi^s$, and $\sigma^2 = \sigma_v^2/(1-\phi^2)$, along with the time run, the error autocorrelations ρ_s will decay exponentially as *s* increases to 0. A GLS model is run through a *gls* command under *nlme* library package within R open source software.

6.4.2 Independent Component Analysis

In ICA, the input variables are regarded as linear combinations of latent variables which are considered independent and non-Gaussian. Along with Principal component analysis (PCA), ICA is considered linear representation models. ICA establishes independent components from original variables. The concept of ICA is regarded to be able in explaining more for variable relationship because independence is a high-order statistic that is in favor over orthogonality (Shao *et al.*, 2006). Multilevel IC regression forms relationship between response variable (*y*) and the ICs from ICA along with other explanatory variables (e.g., days in week, season)

ICA uses higher-order statistic to obtain statistically independent components (Westra *et al.*, 2009). A typical ICA model is expressed as:

$$X = SA \qquad (27)$$

X is observation matrix, derived through the mixing of an n-dimensional source matrix, S = $(s_1, \ldots, s_n)^T$, with temporal dimension of l, referred to ICs, with n is independent components extracted. A is the mixing matrix of dimension n x n or m x n where $m \leq n$. The objective of ICA is to estimate A and S, knowing only the observations matrix X. The present study uses *FastICA* algorithm to estimate A and S from observations X (Hyvärinen and Oja, 2000). The S components will be used as input variables in multilevel model. A *fastICA* function within R program was used to obtain ICs.

For forecasting purpose, the following formula was used to obtain ICs for input to the prediction model:

$$XA^{-1} = S \tag{28}$$

X is lag-1 independent variables while A is the inverse of loading matrix obtained from training set data.

6.4.3 Principal Component Analysis

PCA creates principal components (PCs) that are orthogonal and uncorrelated and linear combinations of the original variables. The first PC is the one that has the largest portion of original data variability. A varimax rotation is commonly used to obtain rotated factor weight loadings that represent effect of each each variable in one particular PC. MPC regression examines a relationship between the output variable (y) and the PCs obtained from explanatory variables (air pollutants: NO, NO_2, O_3, SO_2, CO, and PM_{10}, and meteorological factors: wind speed, solar gradiation, humidity, and temperatures). The estimation procedure is given in the following equations:

$$PC_{ij} = \sum_{k=1}^{m} w_{ik} x_{kj} \tag{29}$$

Where PC_{ij} is the PC score for ith component and j-th object. The loading weight is represented by w_{ik} for k-th variable variable on the i-th component, and x_{kj} is the standardized value of k-th variable for the j-th observation (Verbeke *et al.*, 1984). A PCA is run using *prcomp* function within R open source program. To obtain PCs for prediction purposes, the x_{kj} were simply lag-1 independent variables.

6.4.4 Evaluation indicators

We implemented the forecasting and consequently we can obtain a fitted value F_t for each observation. Here we calculate the following three types of evaluation indicators for the forecasting: mean error (ME), mean absolute error (MAE), and root mean square error (RMSE). MAE calculates the average magnitude of the errors in a set of fitted values without considering direction. On the other hand, RMSE measures the average magnitude of the error. The MAE and RMSE can be used together to diagnose the variation of errors in a set of fitted or predicted values. The greater the difference between RMSE and MAE is the greater the variance in the individual errors in the sample. Lower values of ME, MAE, and RMSE are better. In addition, we also compute R^2 (or R-squared).

6.5 Results and Discussion for time series models

Each model is evaluated by first focusing on its internal validity and then its external validity. The internal validity indicates how each model fits data used to estimate it, i.e., the model accuracy itself. The external validity refers to the forecasting ability of each model, i.e., it reveals how well the model can forecast data that are not used to estimate it.

6.5.1 Internal validity: Model accuracy

Table 6.4 shows the estimation results of AR(1) model for air quality on all locations, all coefficients appear to be statistically significant.

Table 6.4 Estimation results of AR(1) models for all air quality parameters: NO, NO_2, PM_{10}, and CO

Parameter	Monitoring station	Mean	Lag-1 coefficient
NO	City center A	4.024 (12.39)	0.759 (21.26)
	Trading	0.693 (17.733)	4.606 (18.974)
	Suburban1	0.456 (9.295)	2.622 (21.159)
	Near highway	0.749 (20.738)	4.053 (15.257)
	City center B	0.7615 (31.408)	3.5875 (16.766)
NO_2	City center A	0.842 (27.662)	5.053 (19.268)
	Trading	0.680 (17.077)	4.906 (27.971)
	Suburban1	0.584 (13.162)	3.5573 (20.193)
	Near highway	0.678 (16.970)	4.248 (24.660)
	City center B	0.827 (38.753)	5.002 (31.508)
PM_{10}	City center A	0.772 (21.245)	6.585 (24.719)
	Trading	0.647 (15.624)	7.442 (32.532)
	Suburban1	0.475 (9.919)	6.3715 (21.403)
	Near highway	0.615 (14.389)	7.8094 (19.421)
	City center B	0.772 (31.836)	6.586 (37.984)
CO	City center A	0.860 (30.249)	1.006 (15.371)
	Trading	0.685 (17.327)	0.932 (24.178)
	Suburban1	0.607 (13.955)	0.7704 (27.365)
	Near highway	0.726 (19.521)	0.9275 (21.291)
	City center B	0.829 (39.338)	0.989 (25.390)

Note: t-value is shown inside bracket for each parameter.

Table 6.5 Estimation results of BMS models for all parameters NO, NO_2, PM_{10}, and CO

Paramater	Stations	State 1		State 2	
		Intercept	Lag-1 coefficient	Intercept	Lag-1 coefficient
NO	City center A	0.938 (4.119)	0.841 (13.820)	0.938 (4.119)	0.685 (11.261)
	Trading	1.335 (4.711)	0.799 (11.707)	1.335 (4.711)	0.622 (9.107)
	Suburban1	1.264 (6.127)	0.690 (7.395)	1.264 (6.127)	0.355 (3.804)
	Near highway	1.591 (8.96)	0.537 (14.531)	-0.115 (-0.648)	1.111 (30.052)
	City center B	0.812 (5.892)	0.885 (21.974)	0.812 (5.892)	0.663 (16.453)
NO_2	City center A	1.127 (5.700)	0.756 (20.495)	0.592 (2.998)	0.895 (24.260)
	Trading	2.311 (10.187)	0.530 (12.072)	0.624 (2.752)	0.867 (19.744)
	Suburban1	1.609 (7.305)	0.639 (9.417)	1.225 (5.561)	0.562 (8.292)
	Near highway	1.967 (9.269)	0.505 (11.209)	0.50 (2.356)	0.917 (20.354)
	City center B	1.178 (8.688)	0.749 (29.309)	0.603 (4.449)	0.891 (34.903)
PM_{10}	City center A	3.430 14.495)	0.456 (13.255)	-0.880 (-3.717)	1.146 (33.297)
	Trading	3.411 (8.620)	0.495 (10.342)	1.444 (3.650)	0.857 (17.906)
	Suburban1	5.192 13.055)	0.167 (2.980)	0.585 (1.471)	0.936 (16.682)
	Near highway	2.567 (3.517)	0.807 (7.779)	2.567 (3.517)	0.534 (5.148)
	City center B	3.153 18.583)	0.503 (20.610)	-0.570 (-3.359)	1.102 (45.179)
CO	City center A	0.273 (8.203)	0.681 (22.354)	-0.036 (-1.087)	1.083 (35.536)
	Trading	0.428 (9.566)	0.484 (11.383)	0.10 (2.233)	0.956 (22.476)
	Suburban1	0.416 10.295)	0.422 (8.895)	0.138 (3.411)	0.861 (18.161)
	Near highway	0.416 10.502)	0.521 (13.158)	-0.020 (-0.512)	1.061 (26.799)
	City center B	0.326 (13.701)	0.625 (28.554)	-0.017 (-0.728)	1.064 (48.613)

Note: t-value is shown inside bracket for each parameter

Our final model is the MBMS model, where we attempted to improve the forecasting by incorporating the information of forecasting errors from the BMS model. We divided forecasting errors from the BMS model into four states (intervals) (Figure 6.4a – e, and Figure 6.5a – e for increased sample size), representing each parameter NO, NO_2, PM_{10}, and CO in City center site, other sites are not shown. $\otimes_1$ stands for interval 1/2 min(e_t) ≤ e_t < 0, $\otimes_2$ for interval e_t < ½ min(e_t), $\otimes_3$ for interval 0 > e_t ≥ ½ max(e_t), and finally $\otimes_4$ for interval e_t > ½ max(e_t). Min and max refers to the minimum and maximum value of errors. The state intervals for all stations for all quality parameters are shown below in Table 6.6.

Table 6.6 Min and Max Errors of BMS model for city center, trading zone, Suburban1, near highway, and Suburban2 stations for all quality parameters

Parameters	States ($\otimes$)	City center A		Trading zone		Suburban1	
		Max: M_i	*Min: N_i*	*Max: M_i*	*Min: N_i*	*Max: M_i*	*Min: N_i*
NO	$\otimes_1$	-0.135	-2.666	-0.029	-2.982	-0.027	-1.88
	$\otimes_2$	-2.696	-5.352	-3.007	-5.998	-1.903	-3.760
	$\otimes_3$	2.684	0.071	3.371	0.003	2.308	0.037
	$\otimes_4$	5.671	3.05	7.224	4.093	4.883	2.756
NO_2	$\otimes_1$	-0.0002	-1.318	-0.007	-2.420	-0.0002	-2.140
	$\otimes_2$	-1.341	-2.662	-2.503	-4.920	-2.191	-4.353
	$\otimes_3$	2.180	0.003	2.433	0.004	2.076	0.022
	$\otimes_4$	4.466	2.247	4.898	2.538	4.189	2.102
PM_{10}	$\otimes_1$	-0.016	-3.003	-0.0006	-2.063	-0.001	-3.069
	$\otimes_2$	-3.164	-6.053	-2.195	-4.302	-3.094	-6.159
	$\otimes_3$	5.293	0.006	4.667	0.013	11.233	0.031
	$\otimes_4$	10.590	6.984	9.833	5.307	22.534	11.424
CO	$\otimes_1$	-0.001	-0.281	-0.001	-0.310	-0.001	-0.245
	$\otimes_2$	-0.285	-0.563	-0.315	-0.625	-0.246	-0.491
	$\otimes_3$	0.871	0.002	0.796	0.004	0.618	0.001
	$\otimes_4$	1.813	0.950	1.633	0.893	1.241	0.621

Table 6.6 Min and Max Errors of BMS model for Trading zone, Suburban1, and Near highway stations for all quality parameters (Contd)

Parameters	States ($\otimes$)	Near highway		City center B	
		Max: M_i	*Min: N_i*	*Max: M_i*	*Min: N_i*
NO	$\otimes_1$	-0.007	-1.255	-0.007	-3.260
	$\otimes_2$	-1.266	-2.524	-3.269	-6.526
	$\otimes_3$	5.159	0.005	2.250	0.030
	$\otimes_4$	10.361	5.221	4.587	2.469
NO_2	$\otimes_1$	-0.009	-1.652	-0.004	-1.304
	$\otimes_2$	-1.659	-3.308	-1.328	-2.641
	$\otimes_3$	2.947	0.004	2.214	0.010
	$\otimes_4$	5.904	3.061	4.458	2.231
PM_{10}	$\otimes_1$	-0.179	-5.129	-0.003	-2.790
	$\otimes_2$	-5.150	-10.293	-2.969	-5.791
	$\otimes_3$	9.504	0.304	5.263	0.007
	$\otimes_4$	24.581	14.886	10.560	5.498
CO	$\otimes_1$	-0.002	-0.193	-0.001	-0.270
	$\otimes_2$	-0.195	-0.388	-0.274	-0.546
	$\otimes_3$	0.783	0.003	0.881	0.001
	$\otimes_4$	1.624	0.918	1.798	0.909

Based on internal validation in City center we can conclude the followings. BMS didn't perform well in terms of all performance indicators for all sites. For instance in City center site for sample size 336 hours, the ME, MAE, and RMSE are -1.932, 2.606, and 2.923, showing higher bias compared with AR(1), however, the R^2 improves. The ME, MAE, and RMSE for AR(1) are -0.076, 1.799, and 2.174, respectively. However, MBMS produced better result than both AR(1) and BMS, as indicated by the ME, MAE, and RMSE. The R^2 also improves to almost 90% (Table 6.7). In terms of the effect of the increased sample size, the ME, MAE, and RMSE of models using 720 data were higher than fewer sample size for BMS. This indicated that the increase of sample size did not help the internal validation performance for NO. We also noted consistent performance of MBMS which is better than AR(1).

Table 6.7 Model performances of internal validation based ME, MAE, and RMSE values for all site locations for all quality parameters

Parameter	**Stations**	**Model**	**ME**	**MAE**	**RMSE**	**R^2**
NO	City center A	AR(1)	-0.076	1.799	2.174	0.032
		BMS	-1.932	2.606	2.923	0.035
		MBMS	0.046	0.627	0.729	0.899
	Trading zone	AR(1)	-0.019	1.464	1.898	0.047
		BMS	-2.027	2.453	2.814	0.061
		MBMS	-0.049	0.783	0.906	0.81
	Suburban1	AR(1)	-0.031	1.084	1.371	0.015
		BMS	-1.496	1.756	2.026	0.032
		MBMS	-0.058	0.471	0.552	0.859
	Near highway	AR(1)	-0.036	1.365	1.841	0.123
		BMS	0.579	1.328	1.945	0.003
		MBMS	-0.570	0.980	1.274	0.700
	City center B	AR(1)	-0.055	1.645	2.089	0.022
		BMS	-3.519	3.772	4.108	0.028
		MBMS	0.039	0.647	0.794	0.863
NO_2	City center A	AR(1)	-0.158	1.086	1.353	0.044
		BMS	0.291	1.085	1.402	0.017
		MBMS	-0.066	0.480	0.571	0.85
	Trading zone	AR(1)	-0.019	1.097	1.413	0.007
		BMS	-0.031	1.100	1.415	0.003
		MBMS	-0.038	0.645	0.738	0.767
	Suburban1	AR(1)	-0.030	1.328	1.656	0.007
		BMS	-0.916	1.556	1.897	0.009
		MBMS	-0.034	0.520	0.609	0.878
	Near highway	AR(1)	-0.020	1.103	1.389	0.008
		BMS	0.251	1.112	1.414	0.003
		MBMS	-0.136	0.590	0.717	0.79
	City center B	AR(1)	-0.068	1.047	1.283	0.019
		BMS	0.250	1.040	1.317	0.008
		MBMS	-0.046	0.454	0.548	0.841
PM_{10}	City center A	AR(1)	-0.170	1.213	1.668	0.030
		BMS	0.127	1.218	1.711	0.007
		MBMS	-0.457	1.207	1.404	0.667
	Trading zone	AR(1)	-0.015	1.404	1.945	0.020
		BMS	0.675	1.494	2.062	0.009
		MBMS	-0.557	1.086	1.283	0.714
	Suburban1	AR(1)	-0.012	1.858	3.270	0.001
		BMS	0.124	1.829	3.273	0.0001
		MBMS	-1.340	2.172	2.909	0.638
	Near highway	AR(1)	0.005	2.124	3.629	0.018
		BMS	-5.389	6.133	6.602	0.047
		MBMS	0.229	1.303	1.570	0.846
	City center B	AR(1)	-0.078	1.189	1.630	0.014
		BMS	0.178	1.182	1.658	0.004
		MBMS	-0.455	1.160	1.374	0.646

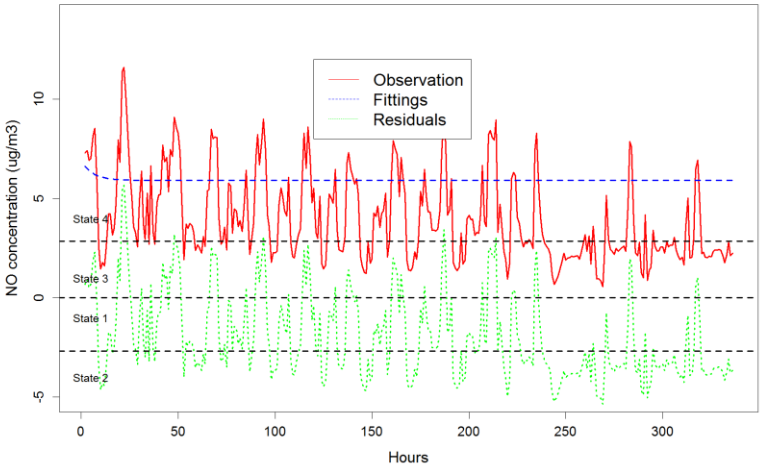
Observing the performance of BMS on other sites, the performances were observed to be worse than AR(1). However, it was able to improve the R^2 from the AR(1) model. Additional action by involving prediction error didn't help the performance as well, in terms of RMSE, with additional information that the MBMS also increased the R^2 values, which means that the forecasted values had more capability in explaining the variance of the observation values compared with AR(1) and BMS. However, we observed that in terms of ME, the performance of MBMS were underperformed by AR(1) model.

The coefficients of AR(1) for NO_2 are shown on Table 6.4, while the estimation parameters for BMS model are shown on Table 6.5. All coefficients were statistically significant. Table 6.6 shows the forecasting error intervals which were used to forecast NO_2. For NO_2, BMS model did not perform better than AR(1) for all site locations in terms of RMSE. Moreover, the model only improved the R^2 in the Suburban1 site. In regards to ME and MAE, the BMS model also shows that its performance was slightly under the performance of AR(1) model. In terms of RMSE, similar results with NO were obtained. The performances (ME, MAE, and RMSE) of MBMS were better for all sites compared with AR(1), the model also improved the R^2 significantly. Another fact we observed was that increasing sample size did not boost improvement of the validation. As for PM_{10}, estimation parameters for PM_{10} data for AR(1), BMS model are shown on Table 6.4 and Table 6.5, while the intervals for MBMS model are shown on Table 6.6. Based on all performance indicators, the performances of AR(1) were better than BMS and MBMS model. However, the MBMS model also improved the R^2.

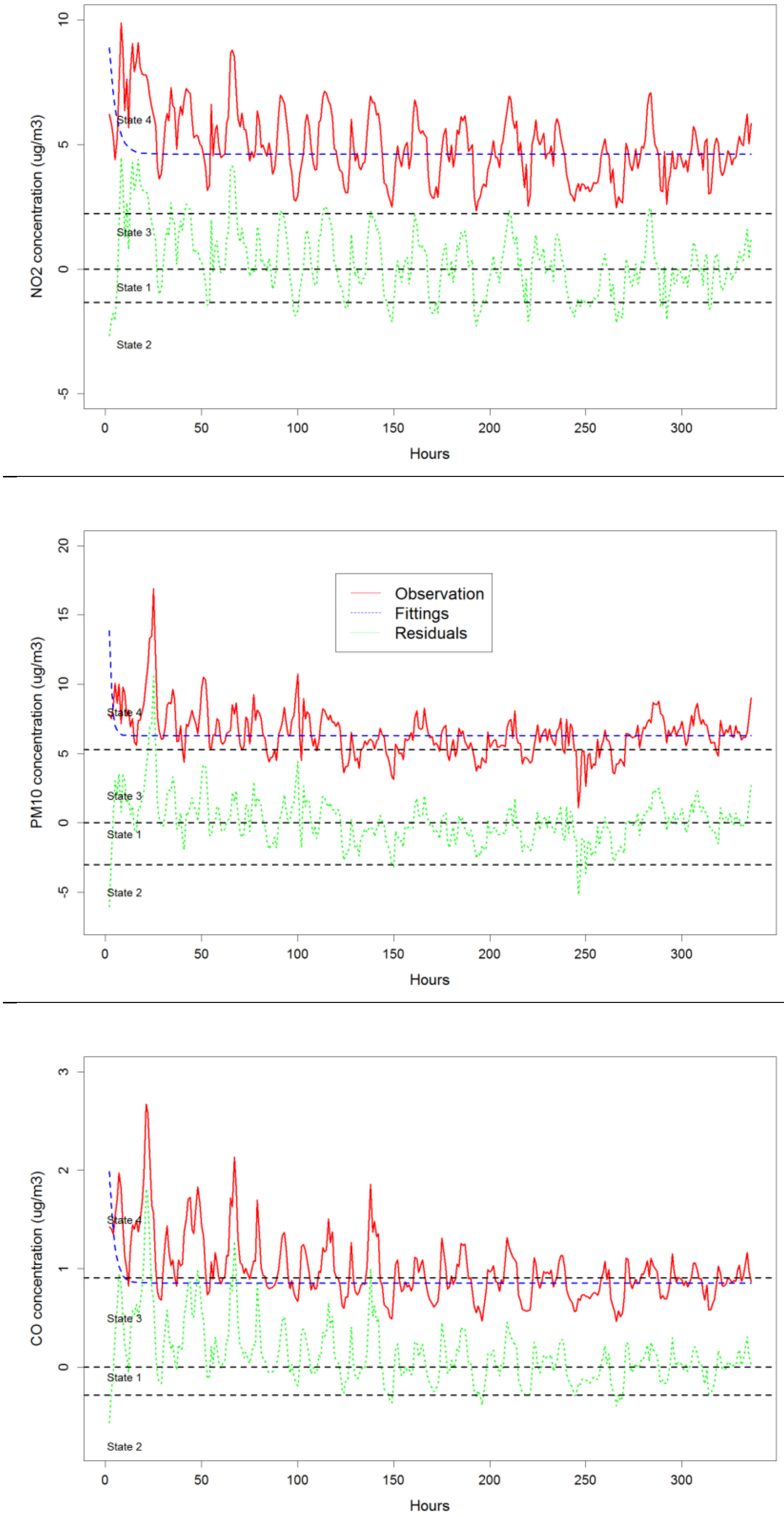
Table 6.6 Model performances of internal validation based ME, MAE, and RMSE values for all site locations for all quality parameters (Contd)

Parameter	**Stations**	**Model**	**ME**	**MAE**	**RMSE**	**R^2**
CO	City center A	AR(1)	-0.032	0.231	0.311	0.091
		BMS	0.124	0.229	0.343	0.034
		MBMS	-0.105	0.193	0.236	0.659
	Trading zone	AR(1)	-0.004	0.229	0.306	0.046
		BMS	0.100	0.235	0.323	0.018
		MBMS	-0.080	0.164	0.203	0.729
	Suburban1	AR(1)	-0.006	0.199	0.255	0.009
		BMS	0.045	0.195	0.259	0.003
		MBMS	-0.047	0.117	0.146	0.760
	Near highway	AR(1)	-0.002	0.244	0.321	0.004
		BMS	0.057	0.238	0.327	0.000
		MBMS	-0.067	0.130	0.172	0.790
	City center B	AR(1)	-0.015	0.224	0.311	0.037
		BMS	0.106	0.217	0.332	0.014
		MBMS	-0.114	0.189	0.236	0.691

For CO, based on the performance indicators of each model, as shown on Table 6.7, it can be seen that the performance of MBMS outperformed BMS and AR(1) model for RMSE, MAE, and R^2. For example in Trading zone, we observed that the MAE and RMSE of MBMS model were 0.203, and 0.854, respectively, while for AR(1) were 0.229, and 0.309, respectively. MBMS model worked very well for interval validation for this parameter with slight improvement from AR(1) model. We also found that there was no difference between increasing sample size with original data sets, suggesting that the parameters estimation were not highly affected by the sample size.



(a) NO

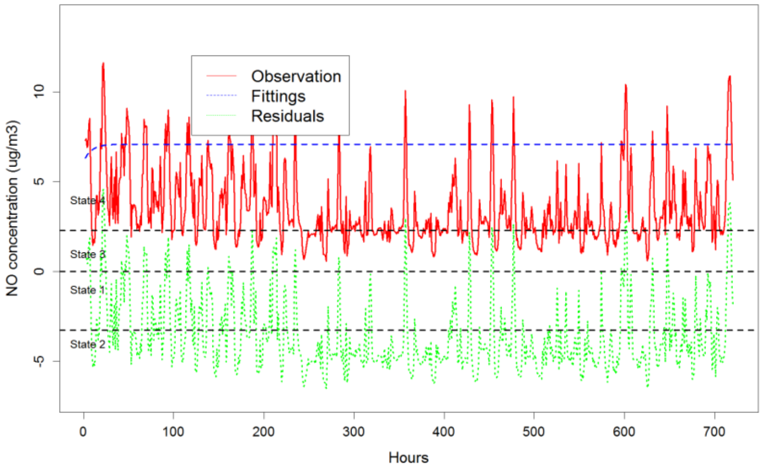


(b) NO_2

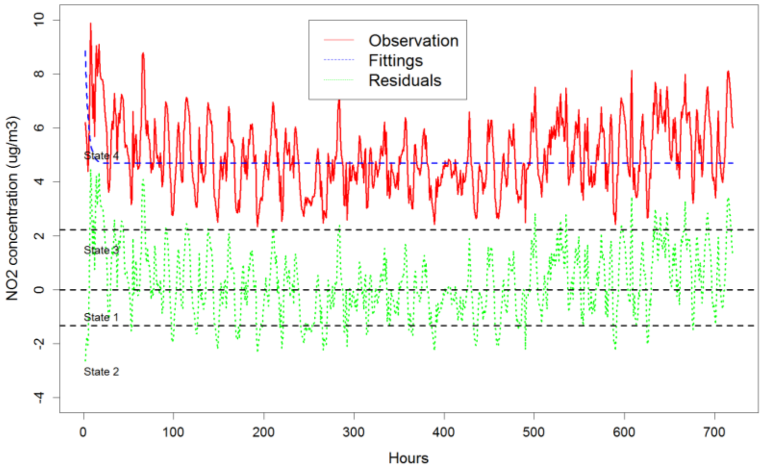
(c) PM_{10}

(d) CO

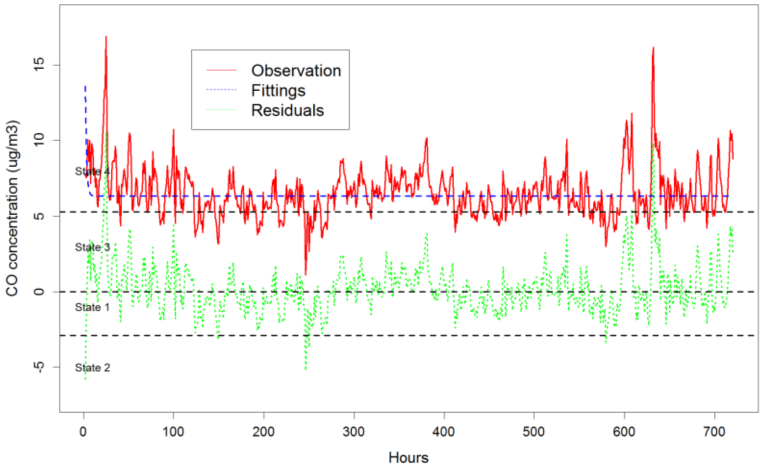
Figure 6.4 Observed values, fitted values and residuals from the BMS model at City center (city center) station



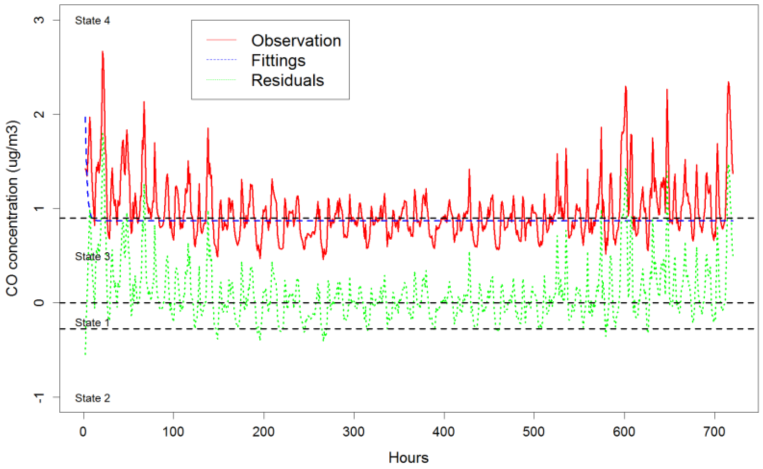
(a) NO



(b) NO_2



(c) PM_{10}



(d) CO

Figure 6.5 Observed values, fitted values and residuals from the BMS model at City center station (city center) for increased sample size

6.5.2 External validity: Forecasting ability

Figure 6.6 – 6.9 show the observation values starting from Hour 300 to Hour 360 with the forecasted values from hour 337 to 360, we forecasted 24-h values of NO, NO_2, PM_{10}, and CO squared concentrations. The performance for each model based on several performance criteria are shown on Table 6.7 for all air quality parameters NO, NO_2, PM_{10}, and CO.

We noted that the performance of BMS model was better than AR(1) in City center based on ME, MAE, and RMSE (-0.401, 1.085, and 1.451, respectively). The BMS model in this site also improved R^2. Similar performance in terms of ME was also observed on near highway, however, its RMSE was worse than AR(1) although the R^2 improved. Overall, we observed that the performance of BMS model was not better than AR(1) in terms of RMSE. In terms of MBMS model, we only found that this model only performed well in city center site based on ME and RMSE. Its performances were even worse compared with Trading zone and Suburban1 in terms of RMSE. This indicated the model suffered from high fluctuation of the concentration on Trading zone and Suburban1. Moreover, we noted that increasing sample size for parameter estimation did not help improving the performance. However, one particular note was that the MBMS model was able to enhance the R^2 index because it incorporated the addition of prediction error intervals. Interesting fact was the BMS model performed well in City center and slight worse result in Near highway, meaning if the concentrations are not very high fluctuations, the performance is similar with AR(1).

For NO_2, based on RMSE, the performances of BMS were better for these sites: City center, Trading zone, and Suburban1 compared with AR(1). However, the performances of BMS only better than AR(1) only in City center and Suburban1, based on ME. In Trading zone, the ME was slightly higher than AR(1). Another observation for MBMS model was that the performance of MBMS model was better on these sites: Suburban1 and Near highway based on ME, MAE, and RMSE. This result suggested that the performance was highly affected by the pattern of concentrations used to estimate parameters. The difference of performance of each site suggests that each site poses different air quality patterns. We also note that the increased sample size could not improve the performance of the forecasted values.

Table 6.8 Model performances of external validation based ME, MAE, and RMSE values for all site locations

Parameter	**Stations**	**Model**	**ME**	**MAE**	**RMSE**	**R^2**
NO	City center A	AR(1)	-0.542	1.621	2.040	0.093
		BMS	-0.401	1.085	1.451	0.546
		MBMS	-0.297	1.569	1.926	0.661
	Trading zone	AR(1)	-0.357	1.005	1.219	0.062
		BMS	-2.023	2.082	2.359	0.033
		MBMS	-0.518	1.060	1.318	0.033
	Suburban1	AR(1)	-0.420	1.029	1.178	0.006
		BMS	-1.885	1.938	2.186	0.0002
		MBMS	-3.058	3.058	3.252	0.0002
	Near highway	AR(1)	-0.235	0.903	1.037	0.271
		BMS	0.212	0.927	1.094	0.311
		MBMS	0.843	1.093	1.365	0.311
	City center B	AR(1)	0.289	1.604	1.907	0.102
		BMS	-2.123	2.289	2.674	0.219
		MBMS	-0.490	1.522	1.698	0.219
NO_2	City center A	AR(1)	-0.666	0.939	1.133	0.341
		BMS	-0.217	0.802	0.969	0.257
		MBMS	-1.309	1.410	1.614	0.257
	Trading zone	AR(1)	0.045	1.076	1.30	0.058
		BMS	0.063	1.067	1.287	0.061
		MBMS	-1.155	1.494	1.728	0.061
	Suburban1	AR(1)	9.591	12.189	15.473	0.158
		BMS	8.350	11.667	14.977	0.174
		MBMS	5.205	10.596	13.478	0.174
	Near highway	AR(1)	18.827	18.904	21.939	0.036
		BMS	19.587	19.587	22.532	0.058
		MBMS	15.105	15.698	18.767	0.058
	City center B	AR(1)	17.995	18.505	22.539	0.142
		BMS	20.523	20.693	24.012	0.083
		MBMS	17.179	17.709	21.224	0.083
PM_{10}	City center A	AR(1)	0.363	0.625	0.800	0.060
		BMS	0.884	0.933	1.074	0.124
		MBMS	-1.766	1.766	1.868	0.124
	Trading zone	AR(1)	0.568	1.650	1.908	0.021
		BMS	1.183	1.872	2.182	0.031
		MBMS	2.215	2.403	2.876	0.031
	Suburban1	AR(1)	-0.719	1.289	1.650	0.001
		BMS	-0.510	1.186	1.549	0.004
		MBMS	-6.143	6.143	6.314	0.004
	Near highway	AR(1)	-0.315	0.897	1.019	0.0004
		BMS	-4.946	4.946	5.128	0.005
		MBMS	-2.291	2.357	2.662	0.005
	City center B	AR(1)	0.177	1.292	1.717	0.190
		BMS	0.632	1.383	1.919	0.055
		MBMS	-2.004	2.396	2.701	0.055

On the other hand, for PM_{10}, it appears that the BMS model didn't perform quite well. Its performance only better than AR(1) in Suburban1 with ME, MAE, and RMSE are -0.510, 1.186, 1.549, respectively. The forecasted values were overestimated. The addition of interval values from testing data also did not improve the performance. Its performance was worse on all sites. This result shows that the performance of the model was highly affected by the character of the testing data, which for PM_{10}, the testing data behaviors were not in favor for forecasting. The increase of sample size did not aid the performance as well.

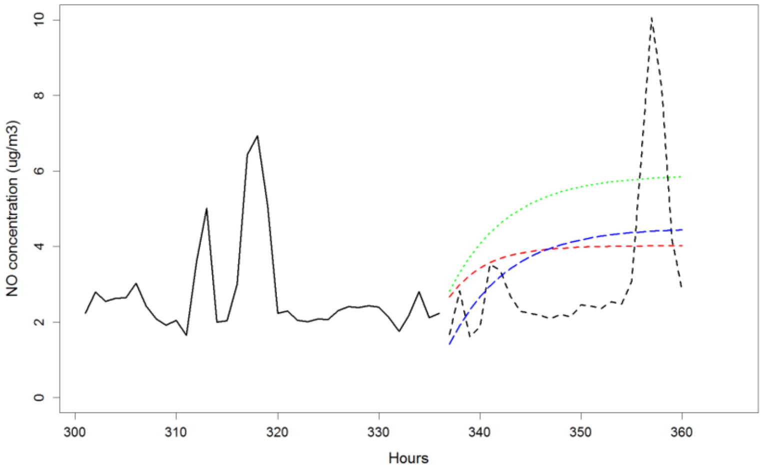
Different performance was observed for CO, where the BMS model was better than AR(1) model in these sites: City center, Suburban1, and Near highway in terms of ME, MAE, and RMSE. The ME, MAE, RMSE on Suburban1 for both BMS mode are -0.039, 0.193, 0.236, respectively, and for AR(1) model were -0.099, 0.213, and 0.255 respectively. Interestingly, overall, the MBMS for CO did not perform well, its performance were generally worse than AR(1) and BMS model. Increasing sample size did not help the forecast either.

Table 6.7 Model performances of external validation based ME, MAE, and RMSE values for all site locations (Contd)

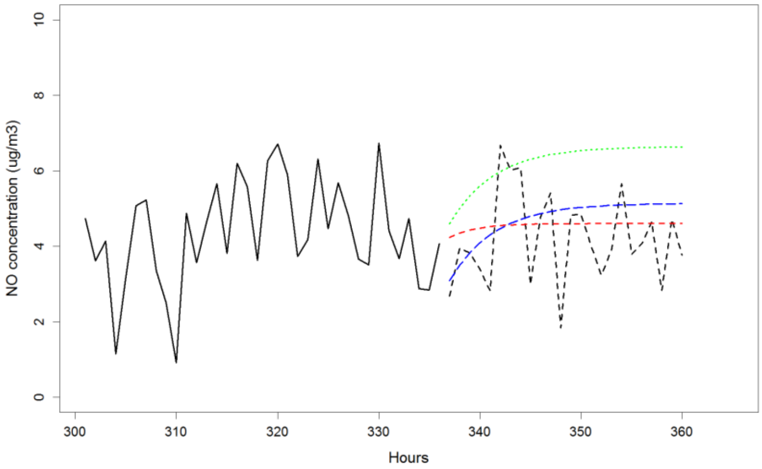
CO	City center A	AR(1)	-0.103	0.126	0.151	0.561
		BMS	0.0124	0.114	0.136	0.449
		MBMS	-0.424	0.424	0.445	0.449
	Trading zone	AR(1)	-0.013	0.232	0.287	0.207
		BMS	0.090	0.256	0.303	0.158
		MBMS	-0.309	0.337	0.423	0.158
	Suburban1	AR(1)	-0.099	0.213	0.255	0.005
		BMS	-0.039	0.193	0.236	0.006
		MBMS	-0.349	0.367	0.419	0.006
	Near highway	AR(1)	-0.079	0.132	0.151	0.248
		BMS	-0.031	0.124	0.140	0.207
		MBMS	0.067	0.118	0.152	0.207
	City center B	AR(1)	0.015	0.189	0.236	0.024
		BMS	0.176	0.214	0.286	0.011
		MBMS	-0.266	0.303	0.348	0.011

Short term training data set used for parameters estimation may have high fluctuation, so we hope that increasing sample size would help in producing more reliable parameters. However, the results showed that the performances of BMS model and MBMS model using parameters estimated from increased sample size were relatively not better than AR(1). This may be due to that over the period of the sample size, the concentrations remained fluctuate based on their daily diurnal pattern, therefore this result strengthen the previous finding that reliable parameters would be obtained for sites where the concentrations are not so fluctuate. Determining the level of fluctuation level would be outside the scope of the present exploration.

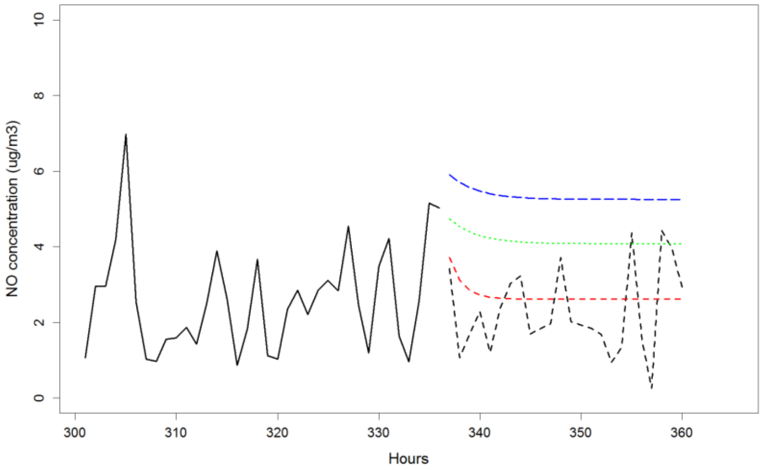
Based on the findings above, the BMS model and MBMS model show a promising if the training data set used to estimate parameters do not highly fluctuate. This condition might be good for sites that the traffic flow between hours are not so big difference, e.g., predicting air quality in suburban areas where the traffic flow is not so fluctuates compared with trading zone, may produce better performance compared with other AR(1) model.



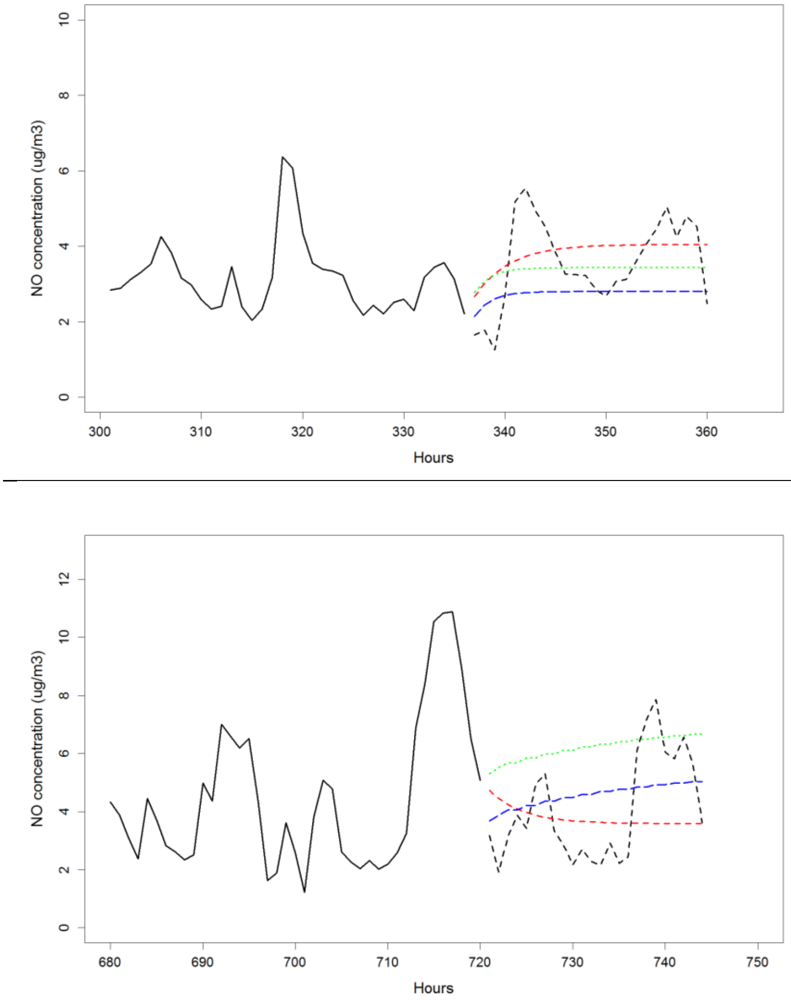
(a) City center



(b) Trading zone



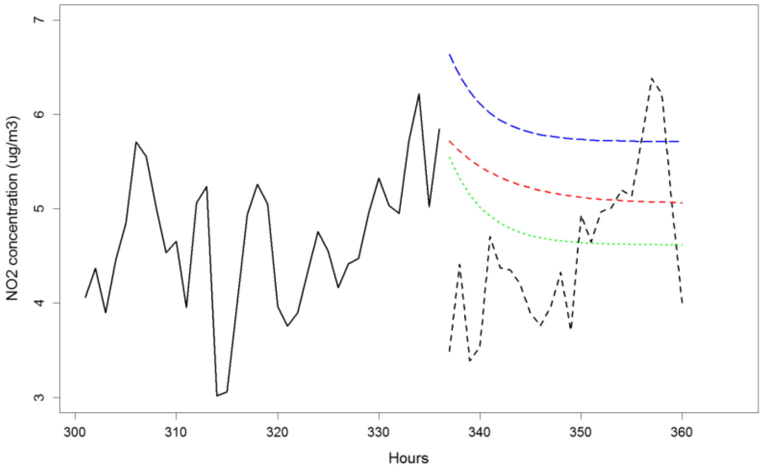
(c) Suburban1



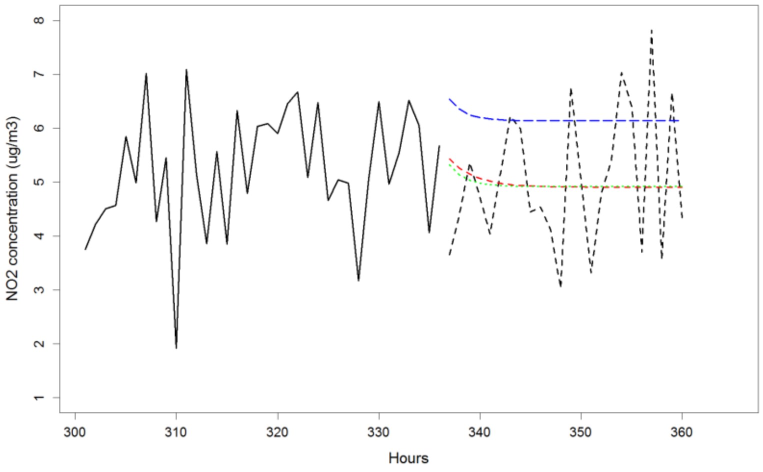
(d) Near highway zone

(e) City center B

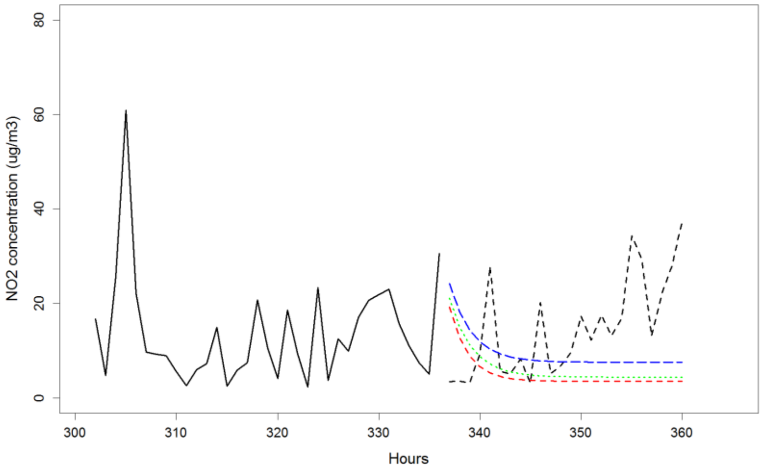
Figure 6.6 Forecasting by AR(1), BMS, and MBMS models for NO on all sites



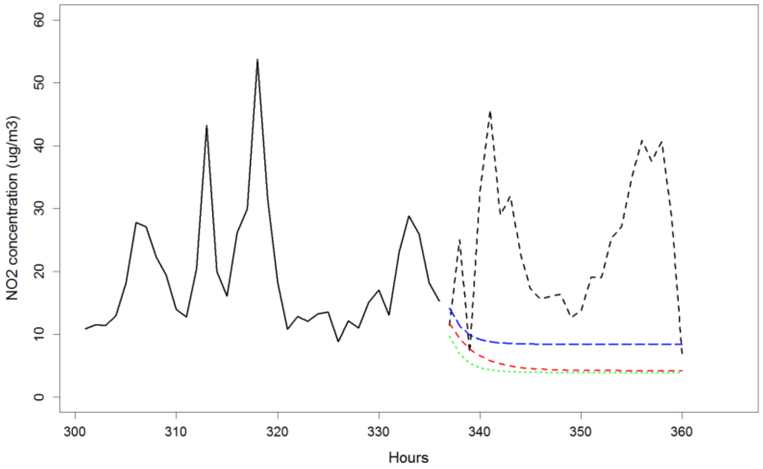
(a) City center A



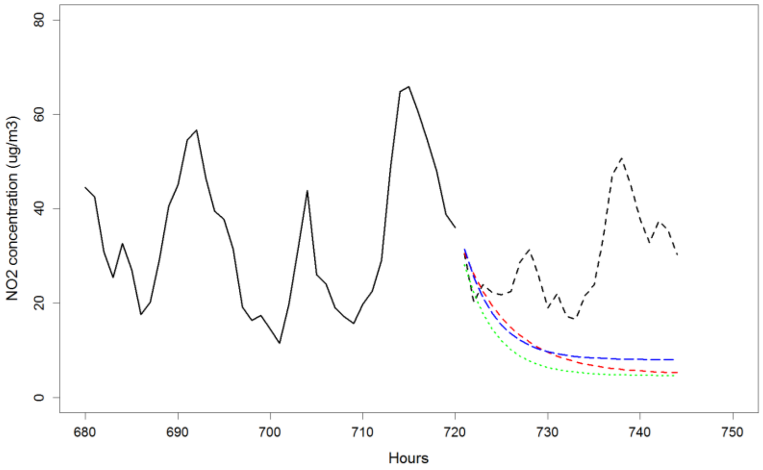
(b) Trading zone



(c) Suburban1

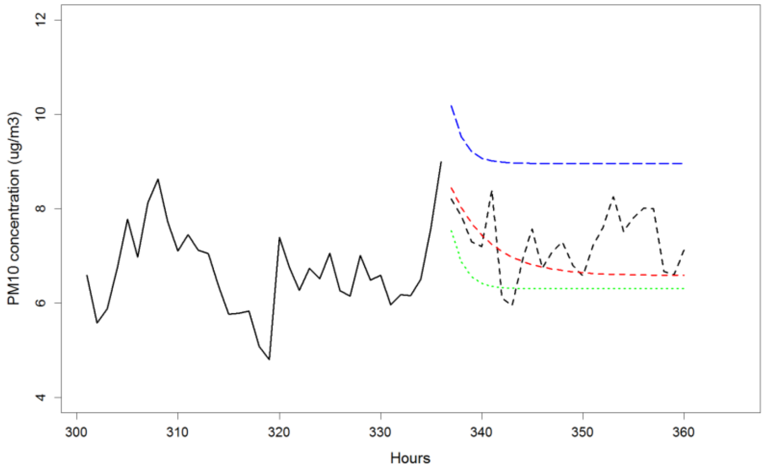


(d) Near highway zone

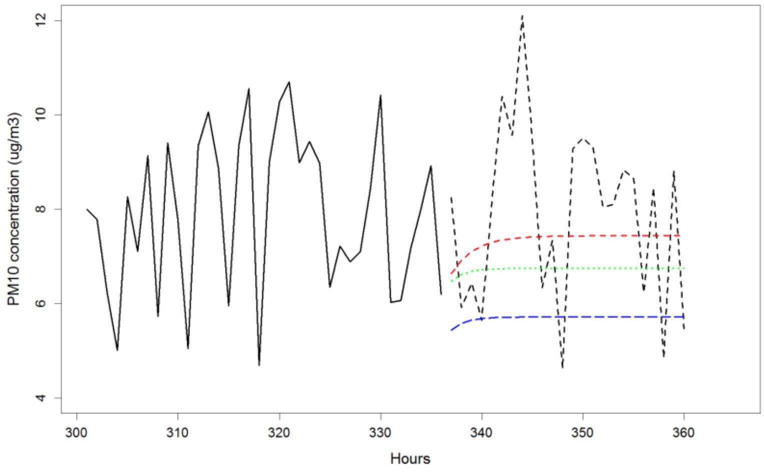


(e) City center B

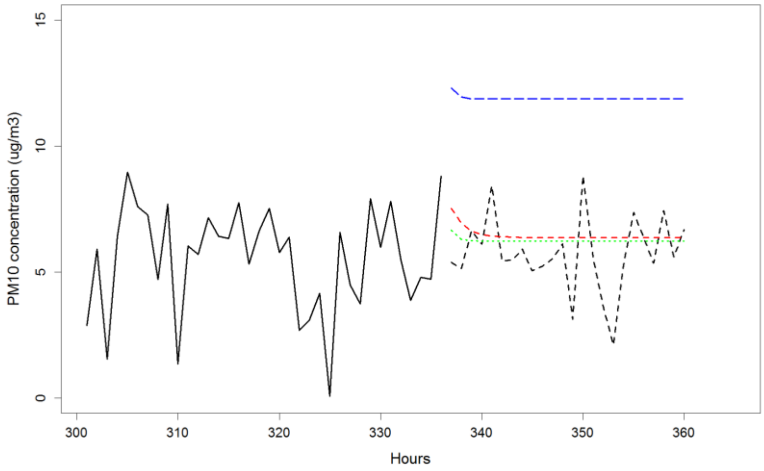
Figure 6.7 Forecasting by AR(1), BMS, and MBMS models for NO_2 on all sites



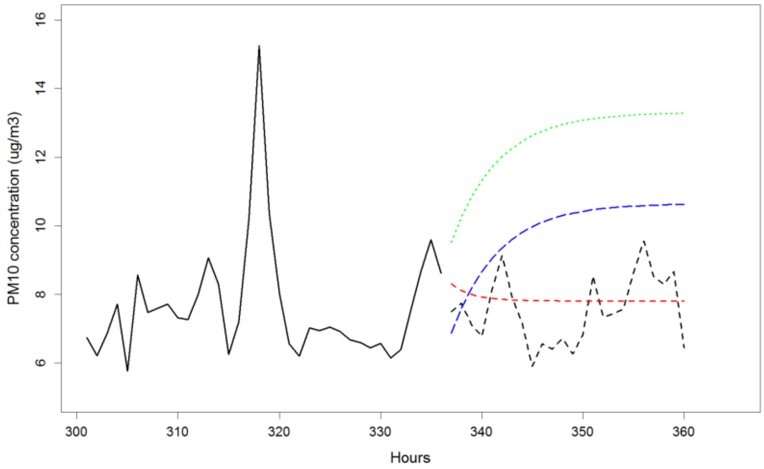
(a) City center A



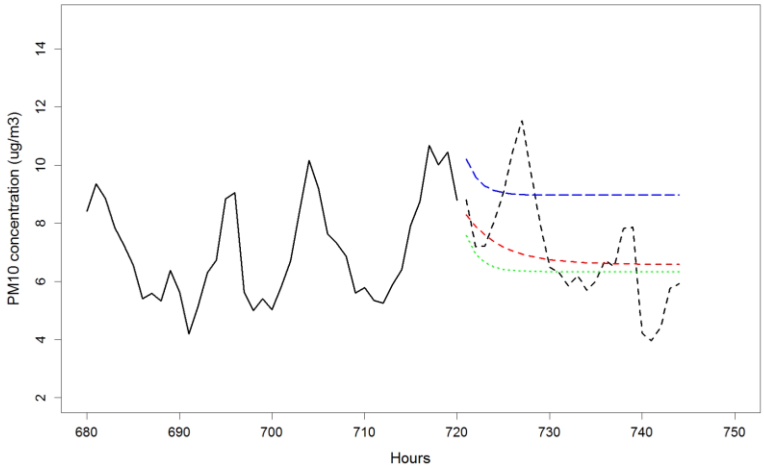
(b) Trading zone



(c) Suburban1



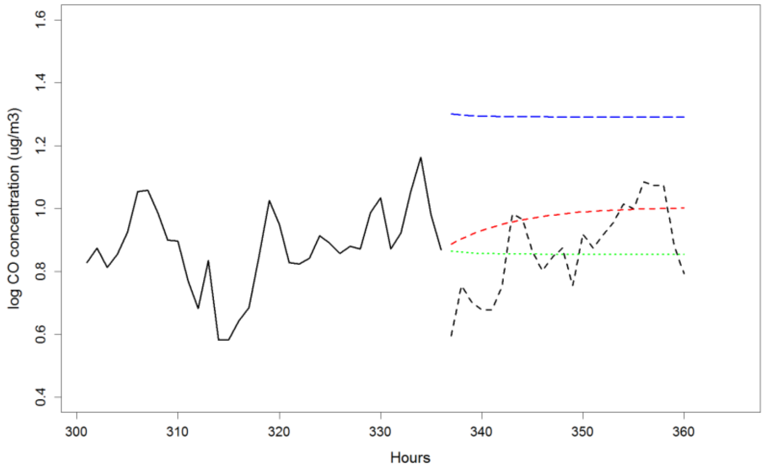
(d) Near highway zone



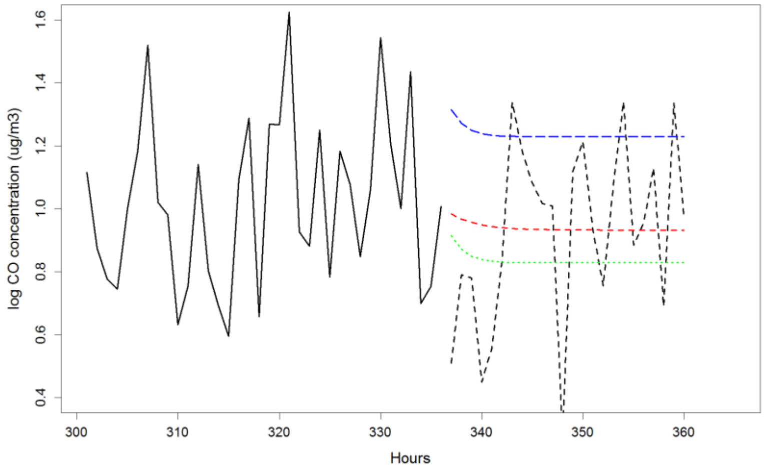
(e) City center B

Figure 6.8 Forecasting by AR(1), BMS, and MBMS models for PM_{10} on all sites

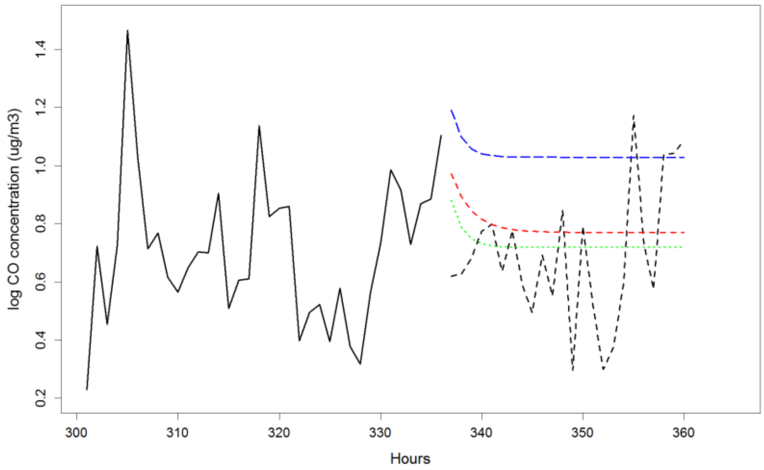
Because of poor performance of time series models demonstrated in the present chapter, there are possible solutions to improve the performance. One way is to determine and check for optimum previous concentration that has most effect towards the current concentration. Another attempt can also be made by using different time interval instead of 1-hour data e.g., 4-hrs or daily.



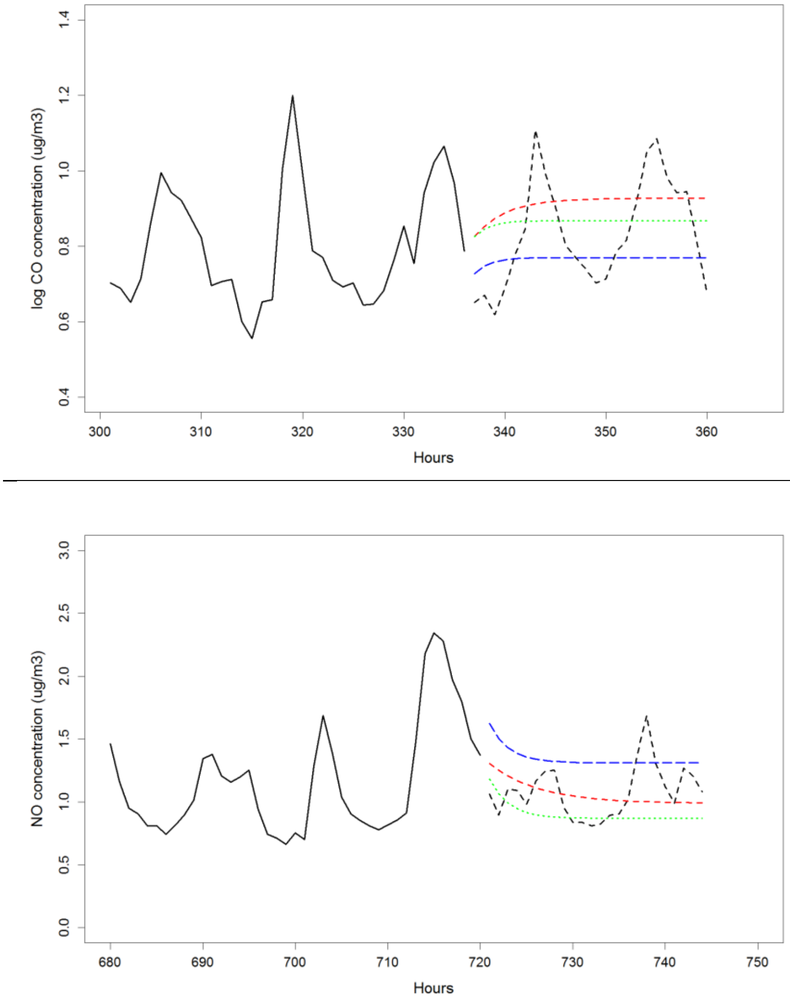
(a) City center A



(b) Trading zone



(c) Suburban1



(d) Near highway zone

(e) City center B

Figure 6.9 Forecasting by AR(1), BMS, and MBMS models for CO on all sites

6.6 Time Series Model with Optimum Order of Parameters

In this section we focus on the use of ARIMA and AR models. Before each attempt, we determined optimum orders for ARIMA and AR by using *auto.arima* function on *forecast* package run under R platform (Hyndman *et al.*, 2014). Table 6.9 below shows the performance result of internal validation and Table 6.10 shows the performance result of forecasting obtained from ARIMA and AR models with their optimum order. Figure 6.10 shows forecasting graphics for NO_2 on all four locations.

We have observed R^2 for internal validation which are ranging between 0.4 up to 0.76 for NO_2. The R^2 are much better than the previous AR(1) and BMS models we used. In the same time, the performances of internal validation of PM_{10} were also better than the previous models as explained in the previous section.

The results for NO_2 are we have seen improvement in the forecasting performance. For example in the city center, after including optimum order of lag-23 concentration, the R^2 of AR becomes 0.398, whereas for ARIMA the R^2 is 0.235. We observed another increase of R^2 in the suburban1 and near highway zone. Especially for near highway zone, only AR with 24-lag concentrations yields significant improvement with R^2 0.259, ARIMA didn't improve the performance. On the other hand, even though there was improvement of PM_{10} forecasting accuracy but the R^2 did not improve by using ARIMA and AR with the optimum order.

Table 6.9 Internal validation of time series models with optimum order parameters on all sites

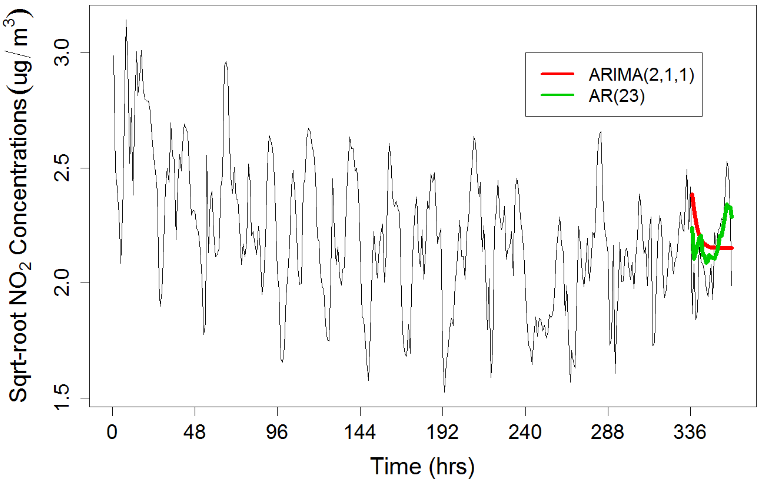
Pollutants	Zone	Model	ME	MAE	RMSE	R^2	R
NO_2	City center	ARIMA(2,1,1)	-0.015	0.127	0.168	0.708	0.842
		AR(23)	-0.014	0.115	0.151	0.765	0.875
	Trading	ARIMA(3,1,3)	-0.067	0.897	1.155	0.454	0.674
		AR(2)	-0.007	0.914	1.172	0.432	0.657
	Suburban1	ARIMA(1,1,2)	-0.067	0.897	1.155	0.454	0.674
		AR(2)	-0.007	0.914	1.172	0.432	0.657
	Near highway	ARIMA(2,1,1)	-0.065	0.765	0.994	0.488	0.698
		AR(24)	-0.039	0.722	0.925	0.553	0.743
PM_{10}	City center	ARIMA(4,1,4)	-0.004	0.163	0.225	0.552	0.743
		AR(3)	-0.005	0.165	0.224	0.555	0.745
	Trading	ARIMA(1,1,2)	-0.001	0.191	0.26	0.482	0.694
		AR(2)	-0.001	0.191	0.259	0.485	0.696
	Suburban1	ARIMA(2,1,5)	-0.001	0.301	0.468	0.263	0.513
		AR(19)	-0.007	0.288	0.427	0.386	0.621
	Near highway	ARIMA(1,1,2)	0	0.248	0.389	0.473	0.687
		AR(23)	-0.013	0.24	0.341	0.594	0.771

Table 6.10 External validation of time series models with optimum order parameters on all sites

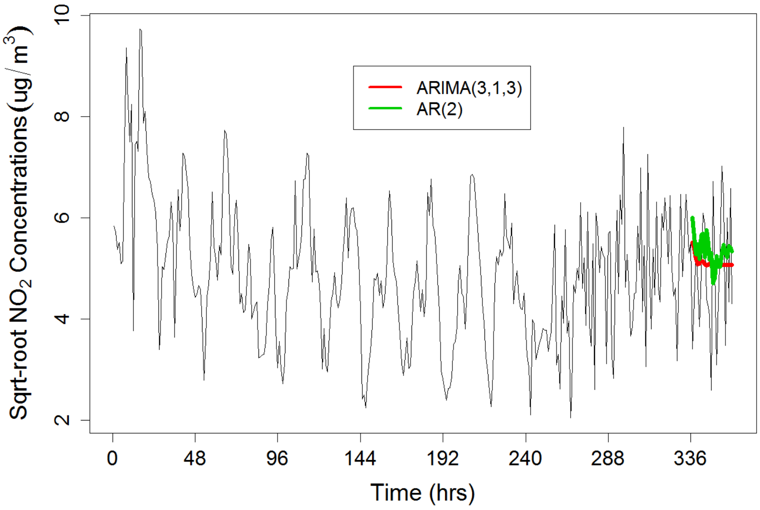
			ME	MAE	RMSE	R^2	R
NO_2	**City center**	**ARIMA(2,1,1)**	**-0.059**	**0.186**	**0.227**	**0.235**	**0.485**
		AR(23)	**-0.059**	**0.126**	**0.159**	**0.398**	**0.631**
	Trading	ARIMA(3,1,3)	-0.261	1.062	1.239	0.051	0.225
		AR(2)	-0.483	1.175	1.362	0.045	0.212
	Suburban1	**ARIMA(1,1,2)**	**-0.063**	**1.189**	**1.426**	**0.203**	**0.45**
		AR(2)	**-0.273**	**1.245**	**1.481**	**0.284**	**0.533**
	Near highway	ARIMA(2,1,1)	0.631	1.04	1.284	0.064	0.253
		AR(24)	**0.436**	**0.868**	**1.089**	**0.259**	**0.509**
PM_{10}	City center	ARIMA(4,1,4)	0.105	0.134	0.166	0.078	0.28
		AR(3)	0.081	0.12	0.156	0.036	0.189
	Trading	ARIMA(1,1,2)	0.066	0.282	0.353	0.084	0.29
		AR(2)	0.065	0.283	0.353	0.092	0.303
	Suburban1	ARIMA(2,1,5)	0.079	0.266	0.331	0.081	0.284
		AR(19)	0.169	0.281	0.347	0.085	0.291
	Near highway	ARIMA(1,1,2)	-0.015	0.152	0.176	0	0.003
		AR(23)	-0.051	0.154	0.184	0.028	0.167

Compared with previous results with only lag-1 information, the results obtained using optimum orders improved a lot. This means that the determination of parameters should be based on which lag information is important that affect current concentrations. In Surabaya case, the time series models with optimum order did not have good performance to predict PM_{10}, as shown by R^2. Even though the R^2 is better than lag-1 models, it seems the dynamic of PM_{10} cannot be captured very well. Therefore, Surabaya Government should focus on the development of NO_2 prediction model using time series.

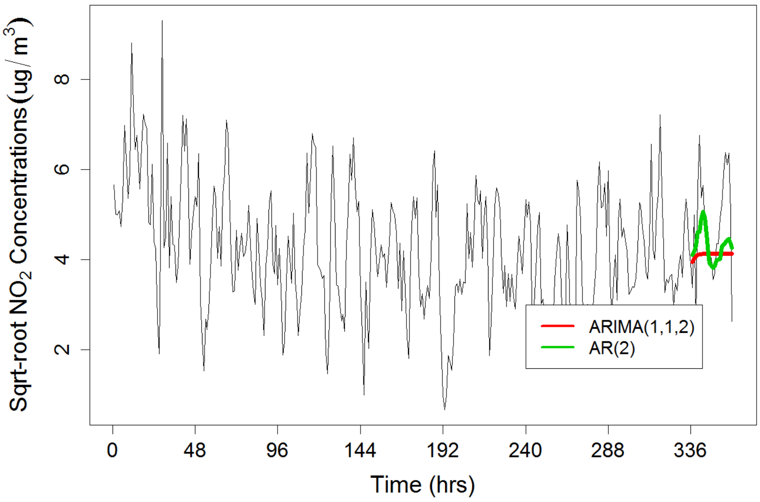
In general, PM_{10} emitted from mixed sources that contribute towards the fluctuation. Therefore, in regions where the land uses are not so diverse, time series models may be applied. Further, to improve these models, we attempted to use longer time interval data to obtain better performance results as well as to determine optimum time interval for time series models application.



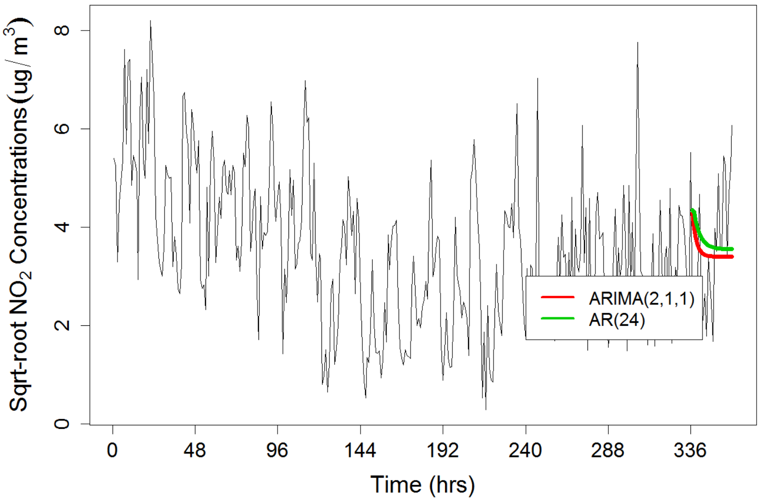
(a) City center



(b) Trading zone



(c) Suburban1



(d) Near highway

Figure 6.10 Forecasting by ARIMA and AR for NO_2 concentration on four stations

6.7 Results and Discussions of Time Series Models for Air Quality Data using 4 hours interval data

In this section we used 4 hours data as initial attempt to use longer time interval for prediction. In the future direction we will expand time interval to obtain best time interval for Surabaya application. This way, we want to prove that longer time interval may produce better performance results. Table 6.11 shows the internal validation and compared with previous results, they show significant improvement for both NO_2 and PM_{10}. Moreover, external validation (Table 6.12) also shows better performance from 4 hours data than if we use one hour data.

Table 6.11 Internal validation of ARIMA and AR using 4 hours data for NO_2 and PM_{10}

Parameters	Zones	Models	ME	MAE	RMSE	R^2	R
NO_2	City center	ARIMA(1,1,4)	-0.025	0.158	0.198	0.501	0.708
		AR(11)	-0.018	0.099	0.126	0.787	0.887
	Trading	ARIMA(2,1,3)	-0.011	0.154	0.19	0.485	0.697
		AR(9)	-0.011	0.13	0.167	0.592	0.769
	Suburban1	ARIMA(2,1,2)	-0.019	0.199	0.25	0.477	0.691
		AR(7)	-0.022	0.183	0.224	0.572	0.756
	Highway	ARIMA(2,1,3)	-0.021	0.175	0.225	0.356	0.597
		AR(6)	-0.015	0.189	0.236	0.277	0.526
PM_{10}	City center	ARIMA(1,1,1)	-0.026	0.172	0.226	0.408	0.639
		AR(4)	-0.012	0.172	0.222	0.382	0.618
	Trading	ARIMA(1,0,0)	-0.002	0.186	0.254	0.293	0.541
		AR(1)	-0.002	0.186	0.254	0.293	0.541
	Suburban1	ARIMA(1,0,0)	-0.001	0.278	0.427	0.11	0.332
		AR(4)	-0.01	0.28	0.405	0.198	0.445
	Highway	ARIMA(1,1,4)	-0.03	0.265	0.346	0.463	0.68
		AR(6)	-0.03	0.26	0.346	0.433	0.658

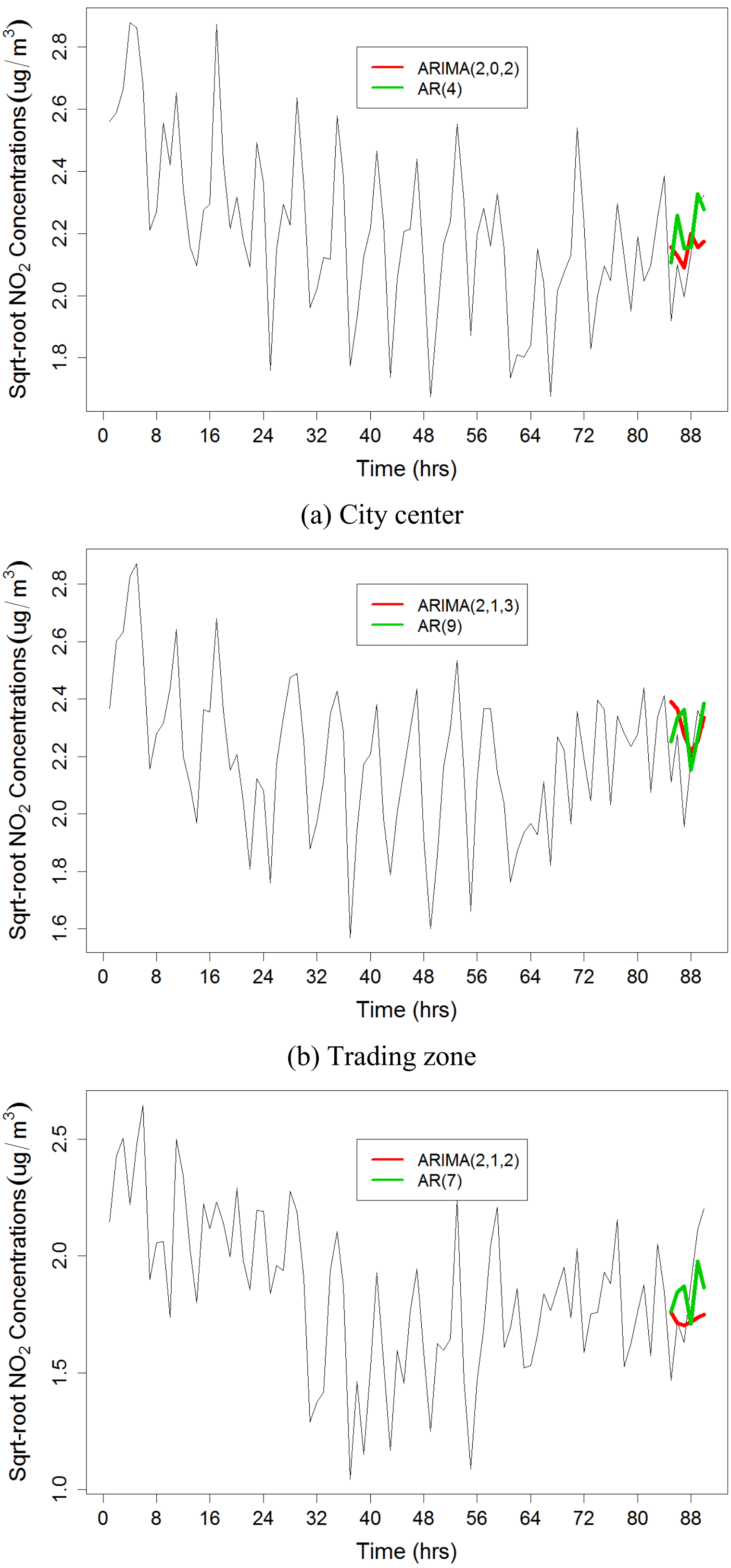
There are several conclusions can be drawn below. First, changing the interval data improves the forecasting result, therefore it is recommended in the future direction to obtain best time interval for time series application. Second, there was no consistent result of the performance. For instance, for NO_2 best model for city center is AR while for highway is ARIMA. However, we found that the models consistently produced better result for city center and highway for NO_2. This might be due to there was a certain pattern fixed in these areas related to working and school activities therefore the patterns might be clearer than other two zones. Therefore in the practical purposes for Surabaya, time series models might be useful to predict NO_2 concentrations in city center and highway. For PM_{10}, we did not find consistent result with previous models, however for 4 hours data, the prediction in the trading was good.

In this section, we have shown the prediction model for next 24-hrs ahead. We would like to extend the prediction up to one (1) week. The common problem with time series models is when we predict future values, the future values will eventually converge to its mean value because of its stationarity characteristic. Therefore the accuracy of longer term prediction is significantly reduced. The power of longer term prediction significantly is reduced because we use exactly same parameters to predict future values or concentrations. As example for this case we use 14 days training data and we predict next-24 hrs, when we try to predict the second day up to one week, we still use same parameters that were obtained from previous original 14 days training data and therefore the power of the prediction will decrease.

In this case we attempted to improve such limitation by using updating approach to predict up to next 7 days. The concept is we predict next 24-hrs using previous 14-days training data and for the next day we repeat again the algorithm using the forecasted values of the 24-hrs data plus previous 13-days training data, we will obtain new regression parameters with new optimum of ARIMA and AR models. These new parameters are used to forecast the second day. This algorithm will be repeated again until we obtain concentrations for consecutive 7 days which will be discussed in the next section.

Table 6.12 External validation of ARIMA and AR using 4 hours data for NO_2 and PM_{10}

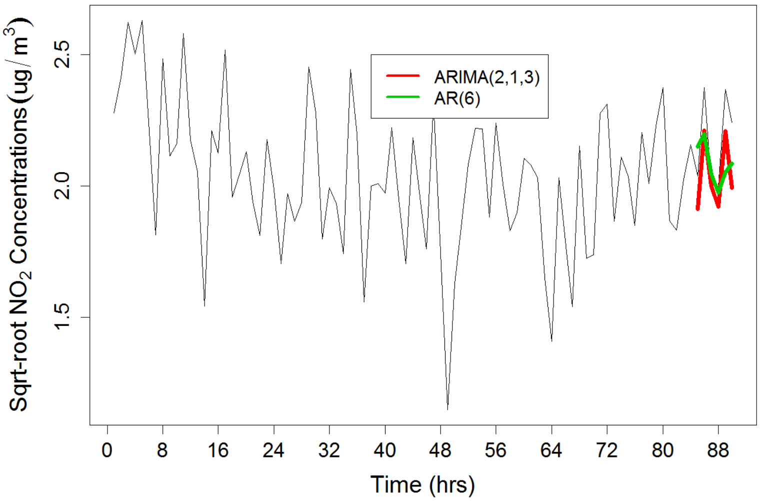
Parameters	Zones	Models	ME	MAE	RMSE	R^2	R
NO_2	City center	ARIMA(1,1,4)	-0.022	0.119	0.136	0.2	0.447
		AR(11)	**-0.085**	**0.1**	**0.121**	**0.757**	**0.87**
	Trading	ARIMA(2,1,3)	-0.107	0.143	0.184	0	0.007
		AR(9)	-0.094	0.136	0.185	0.002	0.049
	Suburban1	ARIMA(2,1,2)	0.106	0.228	0.278	0.024	0.153
		AR(7)	-0.003	0.219	0.232	0.225	0.474
	Highway	**ARIMA(2,1,3)**	**0.131**	**0.131**	**0.151**	**0.81**	**0.9**
		AR(6)	0.088	0.137	0.169	0.215	0.464
PM_{10}	City center	ARIMA(1,1,4)	0.105	0.105	0.123	0.063	0.251
		AR(11)	0.096	0.096	0.115	0.137	0.37
	Trading	**ARIMA(1,0,0)**	**0.101**	**0.146**	**0.18**	**0.588**	**0.767**
		AR(1)	**0.101**	**0.146**	**0.18**	**0.588**	**0.767**
	Suburban1	ARIMA(1,0,0)	0.071	0.219	0.223	0.136	0.369
		AR(4)	0.172	0.24	0.266	0.044	0.211
	Highway	ARIMA(1,1,4)	0.011	0.107	0.114	0.214	0.462
		AR(6)	-0.048	0.107	0.132	0.049	0.222



(a) City center

(b) Trading zone

(c) Suburban1



(d) Near highway

Figure 6.11 Forecasting by ARIMA and AR for 4-hour NO_2 concentration on four stations

6.8 Result and Discussions for Updated Algorithms Time Series Models

In this section we employed the updating algorithm of time series models as explained in the previous section. Figure 6.12 below shows the forecasted values between models using updating algorithms and models without updating algorithms.

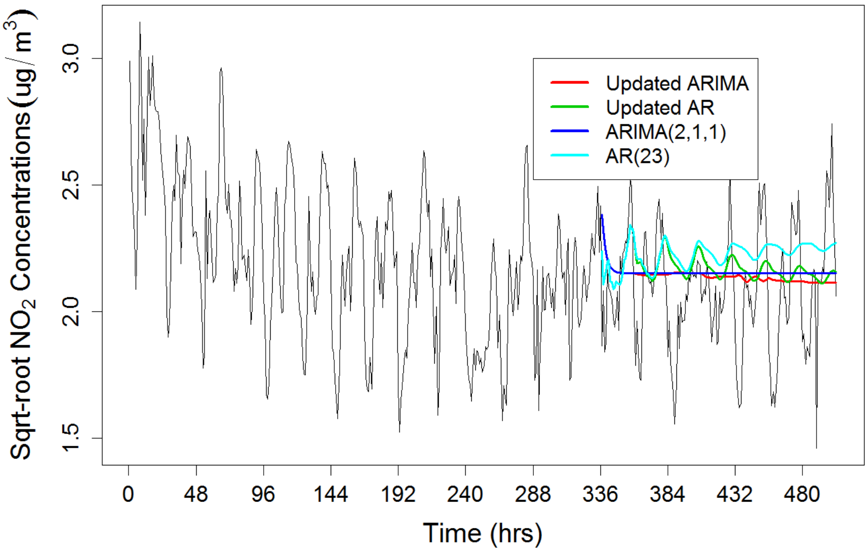


Figure 6.12 Forecasting by updating algorithm of ARIMA and AR compared with ordinary ARIMA(2,1,1) and AR(23) for NO_2 in the city center

Internal validation (Table 6.13) shows that the performance of the models in particular the updated time series models were quite good as indicated by R^2 scores which are higher than 0.5. The updated autoregression (AR) model had better R^2 than the updated ARIMA model. The seven values of internal validation from each updated model show that we check the accuracy once the parameters were updated seven times. The R^2 of internal validation fluctuated. The remaining two rows inside the Table 6.14 show the performance of time series models without the updating process. The results were comparable between the updated time series models and non-updated models. However, for long term in this case one week, the updated models yielded better performance than models without updating (Table 6.14). The results show that the updating models with updated parameters with training data set with the inclusion of forecasted values performed better than without the updating algorithm, although we have to improve the model because the R^2 are low.

Table 6.13 Internal validation of updated ARIMA and AR in the city center for NO_2

Models	ME	MAE	RMSE	R^2	R
Updated ARIMA	-0.015	0.127	0.168	0.708	0.842
	-0.01	0.118	0.155	0.669	0.818
	0.003	0.112	0.151	0.648	0.805
	0.006	0.105	0.139	0.653	0.808
	0.009	0.097	0.131	0.672	0.82
	0.011	0.096	0.132	0.613	0.783
	0.014	0.09	0.126	0.601	0.775
Updated AR	-0.014	0.115	0.151	0.765	0.875
	-0.006	0.102	0.137	0.743	0.862
	-0.004	0.094	0.131	0.735	0.858
	-0.003	0.083	0.118	0.752	0.867
	-0.002	0.076	0.112	0.76	0.872
	-0.002	0.07	0.106	0.749	0.865
	0	0.063	0.101	0.74	0.86
ARIMA(2,1,1)	-0.015	0.127	0.168	0.708	0.842
AR(23)	-0.014	0.115	0.151	0.765	0.875

Table 6.14 External validation of updated ARIMA and AR in the city center for NO_2

Models	ME	MAE	RMSE	R^2	R
Updated ARIMA	-0.053	0.19	0.244	0.025	0.157
Updated AR	-0.083	0.189	0.244	0.018	0.133
ARIMA(2,1,1)	-0.068	0.191	0.244	0.01	0.101
AR(23)	-0.137	0.212	0.269	0.008	0.089

6.9 Results and Discussion of Time Series Models to Predict True Distribution of Air Quality

6.9.1 Prediction of ICs using 1-hour NO_2 data

On the previous results, we have discussed that the forecasting result performances were not good might be because of the fluctuation of air quality, in this case because of emission of air quality that came from mixed sources. So in this section we want to improve the quality of prediction by using Independent Component data. For this purpose, we used NO_2 as example. Our hypothesis was since the ICs were statistically independent patterns therefore they may represented the true distribution of air quality, therefore it was expected to obtain better performance results. Not only that, the true distribution information from ICs may represented the pattern of each location e.g., transport, business activities. Therefore, in this case indirectly we predicted the transport or business-related activities pattern. To obtain conclusions we attempted predicting ICs using NO_2 because NO_2 is primarily emitted from vehicles. The tranport character between zones differed, so these different patterns captured by IC represented different activities of each zone.

Table 6.15 Weght loadings of ICA for NO_2 using 1-hour data

Zones	IC1	IC2	IC3	IC4
City center	**13.904**	-3.859	-3.472	1.065
Trading zone	**11.987**	-3.797	**6.732**	0.304
Suburban1	4.323	**-11.338**	-0.429	0.960
Near highway	7.783	-2.880	-1.615	**-9.064**

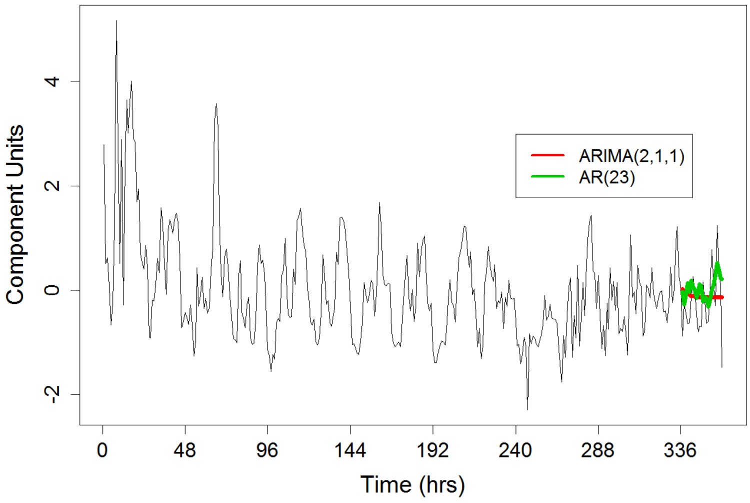
Table 6.16 Internal validation of ARIMA and AR for each ICs on 1-hour NO_2 data

IC	Models	ME	MAE	RMSE	R^2	R
IC1	ARIMA(2,1,1)	-0.057	0.478	0.66	0.593	0.77
	AR(23)	-0.047	0.443	0.604	0.655	0.809
IC2	ARIMA(1,1,1)	0.039	0.653	0.881	0.239	0.489
	AR(2)	0	0.655	0.897	0.206	0.454
IC3	ARIMA(2,1,1)	0.083	0.655	0.903	0.151	0.389
	AR(2)	0.006	0.644	0.907	0.128	0.357
IC4	ARIMA(2,0,1)	-0.001	0.63	0.894	0.185	0.43
	AR(2)	-0.002	0.631	0.898	0.179	0.423

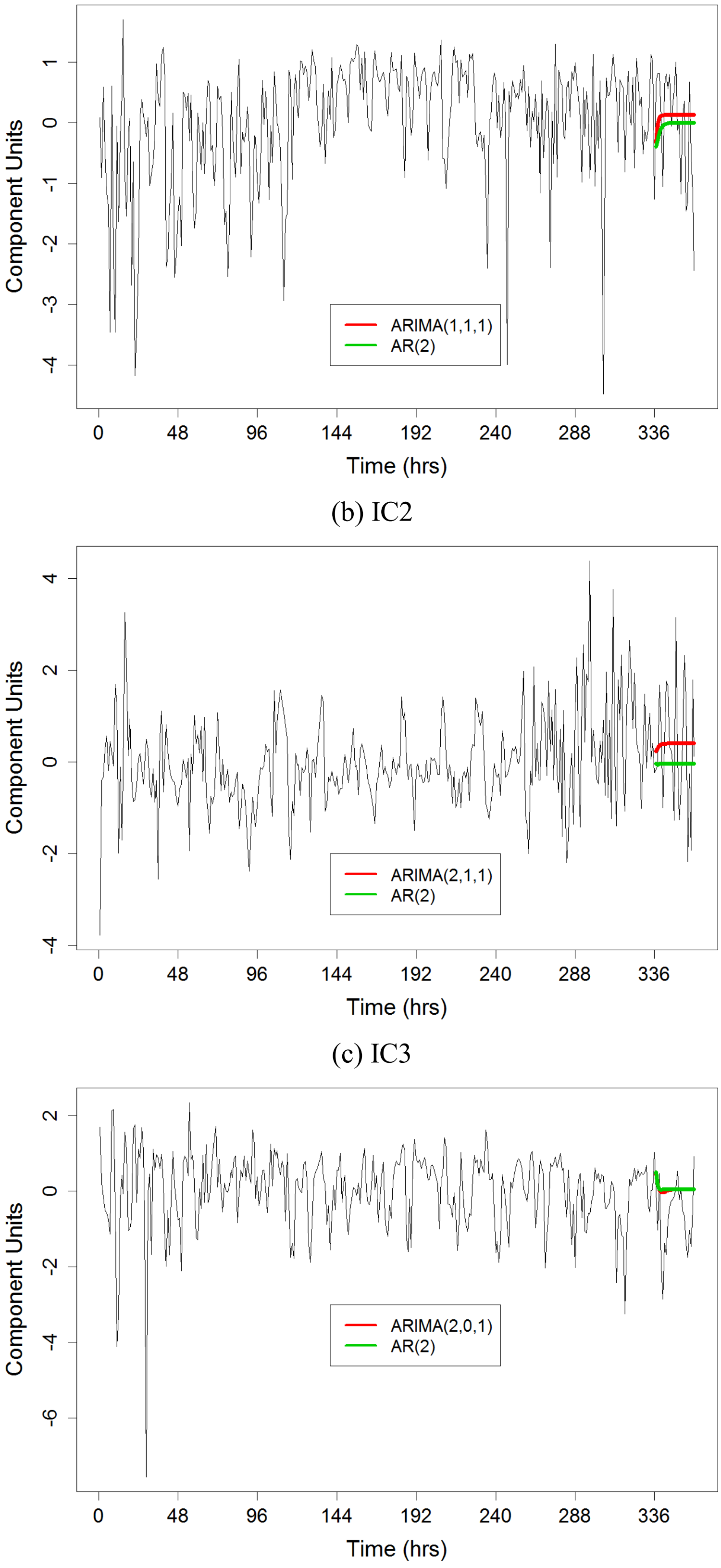
The weight loadings are shown in the Table 6.15. Internal validation (Table 6.16) shows high goodness of fit in terms of R^2 for IC1. IC1 represented pattern in city center and trading zone. Unfortunately, the forecasting performances (Table 6.17) were not good. As shown by Figure 6.13, the prediction was not able to provide accurate future distribution related to traffic pattern in the city center and trading zone as depicted weight loadings. This results show several things. Firstly we expected that the use of true distribution of air quality will be better than using air quality data. However, the results show the opposite. The R^2 for all ICs are lower than 0.5. This might be due to the fluctuation of IC components as shown in the Figure 6.13. The distribution of ICs highly fluctuated. To overcome this, we attempted to use longer time interval which might produce better prediction results.

Table 6.17 External validation of ARIMA and AR for each ICs on 1-hour NO_2 data

IC	Models	ME	MAE	RMSE	R^2	R
IC1	ARIMA(2,1,1)	-0.126	0.48	0.594	0.076	0.276
	AR(23)	-0.27	0.467	0.602	0.103	0.321
IC2	ARIMA(1,1,1)	-0.153	0.735	0.943	0.059	0.242
	AR(2)	-0.011	0.763	0.939	0.101	0.318
IC3	ARIMA(2,1,1)	-0.081	1.029	1.309	0.002	0.045
	AR(2)	0.343	1.044	1.352	0	0.005
IC4	ARIMA(2,0,1)	-0.734	0.885	1.124	0.089	0.298
	AR(2)	-0.75	0.899	1.145	0.023	0.153



(a) IC1



(b) IC2

(c) IC3

(d) IC4

Figure 6.13 Forecasting ICs by ARIMA and AR using 1-hour NO_2 data

6.9.2 Prediction of ICs using 4-hour NO_2 data

The weight loadings from ICA suggested that IC1 represented highway zones whereas city center was represented by IC2, as shown in Table 6.18. The internal validation shows few improvement compared with 1-hour data. Showing that the use of 4-hour data improved the model performances. Furthermore, there was also improvement over the prediction power as shown in the Table 6.20 which was in agreement with improvement of internal validation (Table 6.19). However, these performances were not better than if we use air quality data.

Table 6.18 Weght loadings of ICA for NO_2 using 4-hour data

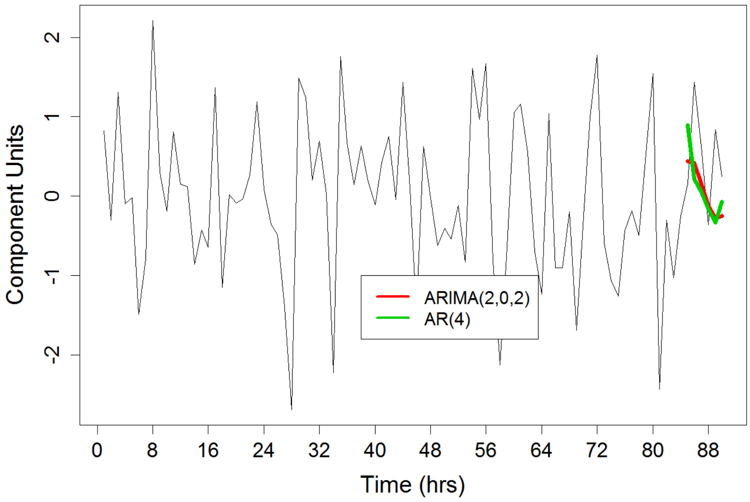
Zones	IC1	IC2	IC3	IC4
City center	2.663	**-11.714**	-3.937	-2.571
Trading zone	-0.847	**-9.657**	-0.328	**-6.257**
Suburban1	-0.836	-4.665	**-6.626**	-4.863
Near highway	**5.503**	-5.464	-0.978	-5.824

Table 6.19 Internal validation of ARIMA and AR for each ICs on 4-hour NO_2 data

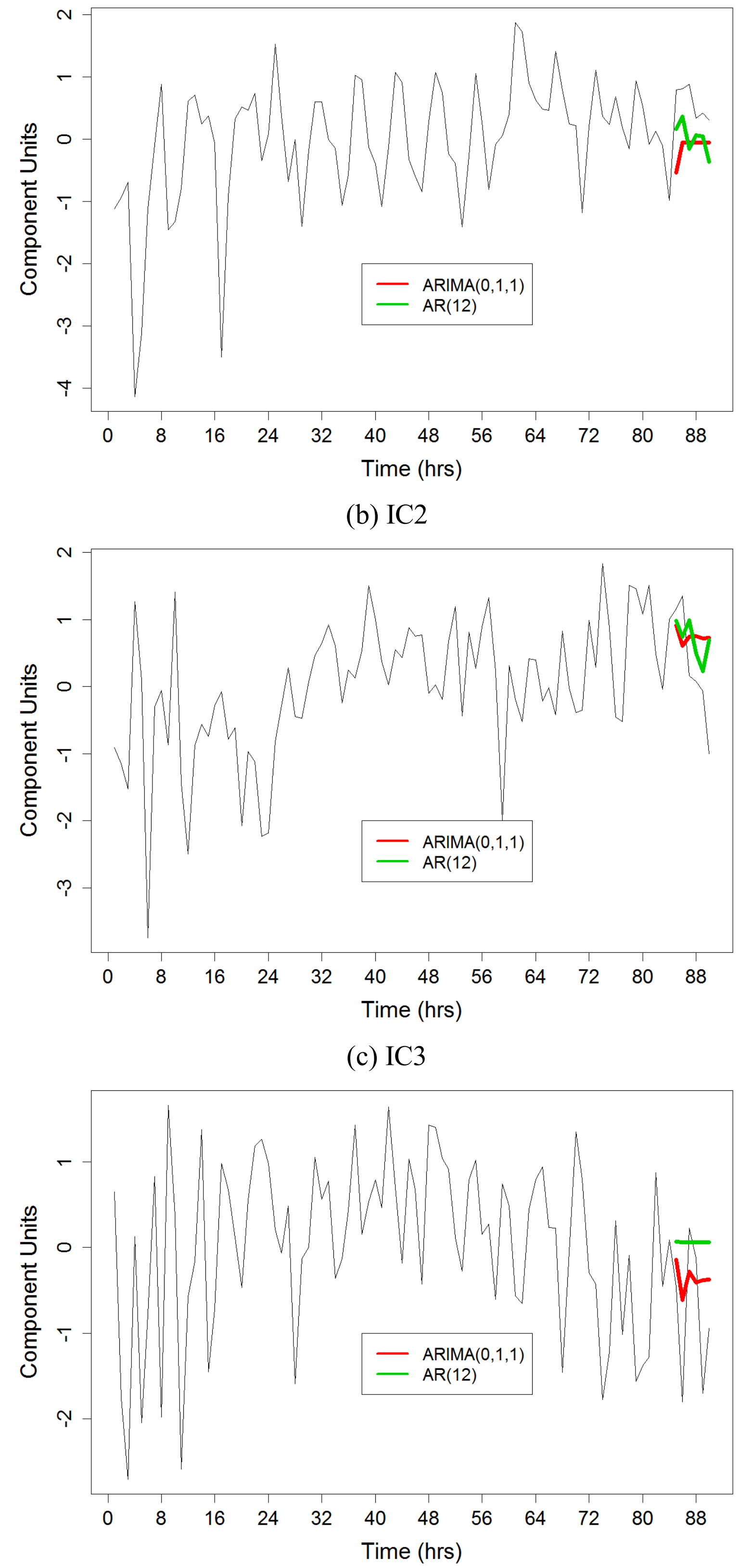
IC	Models	ME	MAE	RMSE	R^2	R
IC1	ARIMA(2,0,2)	0.002	0.613	0.865	0.282	0.531
	AR(4)	0.083	0.494	0.674	0.572	0.756
IC2	ARIMA(0,1,1)	0.002	0.613	0.865	0.282	0.531
	AR(12)	0.083	0.494	0.674	0.572	0.756
IC3	ARIMA(2,1,2)	0.1	0.649	0.855	0.304	0.551
	AR(6)	0.056	0.61	0.786	0.399	0.631
IC4	ARIMA(0,0,1)	0.017	0.732	0.936	0.137	0.371
	AR(1)	-0.002	0.756	0.969	0.043	0.207

Table 6.20 External validation of ARIMA and AR for each ICs on 4-hour NO_2 data

IC	Models	ME	MAE	RMSE	R^2	R
IC1	ARIMA(2,0,2)	0.725	0.725	0.804	0.136	0.369
	AR(4)	0.574	0.574	0.626	0.19	0.436
IC2	ARIMA(0,1,1)	0.725	0.725	0.804	0.136	0.369
	AR(12)	0.574	0.574	0.626	0.19	0.436
IC3	ARIMA(2,1,2)	-0.463	0.795	0.916	0.002	0.046
	AR(6)	-0.408	0.671	0.84	0.138	0.371
IC4	ARIMA(0,0,1)	-0.433	0.697	0.807	0.383	0.619
	AR(1)	-0.861	0.916	1.147	0.012	0.111



(a) IC1



(b) IC2

(c) IC3

(d) IC4

Figure 6.14 Forecasting ICs by ARIMA and AR using 4-hour NO_2 data

6.10 Results and Discussions for GLS models

From some previous results, it seemed that the prediction for air quality could not rely solely based on the intrinsic nature of the concentrations itself. In this case the inclusion of exogenous variables such as meteorological factors was hypothesized to significantly improve the prediction accuracy and this has became our motivation to conduct GLS models.

The correlation analysis of the raw concentration between pollutants and meteorological factors were displayed in Table 6.21. The Pearson correlation reported that there was a linear correlation between the predicted pollutant concentrations and the independent variables. To evaluate correlation coefficients, the significance value was calculated with a significance level of 0.05, marked by values in bold.

Table 6.21 Pearson correlation of air pollutants and meteorological factors over five zones of monitoring stations

	NO	NO_2	O_3	SO_2	PM_{10}	CO	WS	Grad	Hum	Temp
NO	1.00	**0.68**	**-0.67**	0.54	0.49	**0.94**	**-0.73**	**-0.70**	**0.70**	**-0.69**
NO_2	**0.68**	1.00	-0.50	0.41	0.60	**0.80**	**-0.66**	**-0.66**	0.51	-0.50
O_3	**-0.67**	-0.50	1.00	-0.23	-0.50	-0.60	**0.68**	**0.82**	**-0.90**	**0.90**
SO_2	0.54	0.41	-0.23	1.00	0.18	0.51	-0.47	-0.30	0.26	-0.24
PM_{10}	0.49	0.60	-0.50	0.18	1.00	0.59	-0.61	-0.55	0.55	-0.55
CO	**0.94**	**0.80**	-0.60	0.51	0.59	1.00	**-0.75**	**-0.72**	**0.67**	**-0.66**
WS	**-0.73**	**-0.66**	**0.68**	-0.47	**-0.61**	**-0.75**	1.00	**0.64**	**-0.78**	**0.75**
Grad	**-0.70**	**-0.66**	**0.82**	-0.30	-0.55	**-0.72**	**0.64**	1.00	**-0.89**	**0.89**
Hum	**0.70**	0.51	**-0.90**	0.26	0.55	**0.67**	**-0.78**	**-0.89**	1.00	-1.00
Temp	**-0.69**	-0.50	**0.90**	-0.24	-0.55	**-0.66**	**0.75**	**0.89**	-1.00	1.00

It was noted that NO, CO, wind speed, solar gradiation had strong correlation with concentrations of NO_2. NO and CO has positive correlation while wind speed and solar gradiation had negative correlation. Higher wind speed and solar radiation allows for gas purification in the lower layer of atmosphere. It was noted that NO, wind speed, solar gradiation, humidity, and temperatures had statistically significant correlation values with O_3. This was due to that the presence of O_3 are highly due to chemical reaction in the atmospheric layer. The negative correlation of NO and O_3 indicated that higher NO concentration along with NO_2 reaction caused O_3 depletion This fact is in agreement with the result drawn from Chapter 4. On the other hand, wind speed and solar radiation had positive correlation with O_3 despite of opposite coefficient for humidity, lower humidity tended to increase O_3 presence. Since PM_{10} is an inert gas particle, there was no significant variables which affect the concentration of PM_{10} although different results were indicated by Pires *et al.* (2008).

The performance of GLS model with original variables perform the best compared with ICR and PCR model for both internal validation (Table 6.22) and forecasting (Table 6.23) for NO_2. Observing the performance for other pollutants, we also found that the performance of ICR was the worst than PCR and GLS model. Specifically for O_3 prediction, we did not obtain comparable performance as those measured by Al-Alawi *et al.* (2008) which the use of PCR alone yielded R^2 of 0.965, assuming we care less of interval concentration of O_3 they used. However, if we compare the performance of ICR and PCR in the training set (internal validation), ICR performed better than PCR for O_3 and PM_{10} on two stations. This result may indicate better component extraction due to the nature of ICA process. Unfortunately, the performance of ICR in the forecasting was worse than PCR for all pollutants. These facts show the failure of the component analysis method in producing reliable surrogate variables for the model.

Table 6.22 Internal validation of each model for NO$_2$

	Station	Model	ME	MAE	RMSE	R^2	R
NO$_2$	1 (city center)	**1**	**4e-04**	**0.2837**	**0.3658**	**0.9616**	**0.9806**
		2	9e-04	0.898	1.1459	0.5105	0.7145
		3	0.0037	1.0758	1.4043	0.1454	0.3813
	5 (Suburban2)	**1**	**0.0218**	**0.4742**	**0.6437**	**0.8004**	**0.8946**
		2	6.5139	6.5139	6.6138	0.4242	0.6513
		3	-0.2096	0.9462	1.2221	0.3496	0.5913
O$_3$	1 (city center)	**1**	**-2.69e-05**	**0.4805**	**0.7194**	**0.8784**	**0.9372**
		2	-0.0026	1.3244	1.7039	0.3515	0.5928
		3	-0.5805	1.4303	1.8094	0.017	0.1305
	5 (Suburban2)	1	0.7941	2.2566	2.7951	0.0057	0.0752
		2	**0.7941**	**1.9074**	**2.3769**	**0.0134**	**0.1156**
		3	-1.9166	2.125	2.59	1e-04	0.0113
PM$_{10}$	1 (city center)	**1**	**-6e-06**	**0.8834**	**1.2582**	**0.7333**	**0.8563**
		2	0.0021	1.6785	2.1769	0.2104	0.4587
		3	-1.6428	1.9041	2.2173	0.0745	0.2729
	5 (Suburban2)	1	-0.3453	2.247	2.8887	0.0286	0.1691
		2	**-0.3444**	**1.901**	**2.3991**	**0.0376**	**0.1939**
		3	-2.979	3.0094	3.4115	0.0247	0.1572

Observing the result of different stations, we may observe that the performance of ICR was worse than PCR in station 1 (city center) in terms of RMSE. On the other hand, in station 5 (suburban2), the difference of RMSE was not far. This may indicate the sensitivity of performance result highly depend how ICs perform component extraction from the training data set. This may impose a risk where if the particular station has concentrations which highly fluctuate with much noises, this may affect the components extracted, thus affect the prediction severely.

The following factors may contribute to the errors of the models such as: a) errors in data itself (monitoring), b) high fluctuation of concentrations due to short interval time concentrations, c) a linear equation may not be enough to capture the true air quality because of complexity and uncertain system of urban air management.

Table 6.23 Performance of forecasted values in Station 1 and 5 for each model for NO$_2$

	Station	Model	ME	MAE	RMSE	R^2	R
NO$_2$	1 (city center)	**1**	**0.0218**	**0.4742**	**0.6437**	**0.8004**	**0.8946**
		2	6.5139	6.5139	6.6138	0.4242	0.6513
		3	-0.2096	0.9462	1.2221	0.3496	0.5913
	5 (Suburban2)	**1**	**-0.1832**	**0.8289**	**1.0918**	**0.5970**	**0.7727**
		2	0.3821	1.5033	2.0011	0.3327	0.5768
		3	-0.9825	1.8511	2.1211	0.2680	0.5177
O$_3$	1 (city center)	**1**	**-0.0677**	**0.4036**	**0.5672**	**0.8808**	**0.9385**
		2	9.2849	9.2849	9.3918	0.2048	0.4526
		3	-0.3659	1.0223	1.2504	0.6392	0.7995
	5 (Suburban2)	**1**	**-0.0211**	**0.5824**	**0.8565**	**0.7999**	**0.8944**
		2	1.2297	2.1845	2.6100	0.6235	0.7896
		3	-0.4503	1.7747	2.3098	0.5733	0.7572
PM$_{10}$	1 (city center)	**1**	**0.1512**	**0.6419**	**0.8424**	**0.8641**	**0.9296**
		2	-16.8202	16.8202	16.9453	0.1155	0.3398
		3	-0.2522	1.554	2.0382	0.1509	0.3885
	5 (Suburban2)	**1**	**-0.085**	**0.8059**	**1.1339**	**0.8440**	**0.9187**
		2	1.0435	2.1603	3.3041	0.6695	0.8183
		3	-0.9003	2.5138	3.0339	0.4394	0.6629

Table 6.24 shows parameter estimates for the GLS model with NO_2, O_3, and PM_{10} The independent variables were the original square-root transformed variables. For NO_2, among other independent variables, besides lag-1 NO_2, the concentrations of PM_{10} and CO had been the most important predictor for NO_2, which on the opposite, CO also had high contributions to the concentration of PM_{10}. On the other hand, statistically significant as expected, the contribution between NO and NO_2 to the forecasting O_3 concentrations was noticed to be significant.

Furthermore, the increase of wind speed helped mixing process of NO_2, PM_{10}, and O_3 while on the opposite, higher temperature was likely to increase the concentrations. Besides for O_3, solar radiation negatively affected the predicted concentrations of NO_2 and PM_{10}. The congestion level of traffic during peak time affected the emission of NO_2, and as a result, the concentrations during peak times were higher than non-peak time, in particular during morning session, also for PM_{10} and O_3. In the evening time, the concentrations of O_3 tended to be lower than non-peak time. The concentrations were also lower during holidays and weekends for NO_2 and PM_{10}, but the opposite for O_3, indicating photochemical reaction of among contributor gases.

Observing the spatial location, it is interesting that the NO_2 concentrations were found to be higher in suburban1 area (Sukomanunggal, west side of Surabaya) than city center. However, that particular location was highly-densed residentials that emit larger than other zones. Only in east of Surabaya (suburban2), the NO_2 concentrations were lower than other zones. In the trading zone, the predicted PM_{10} concentration was higher in suburban1 (Sukomanunggal) than other zones while we noted higher estimate for O_3 in east side of Surabaya, Sukolilo region, as indicated by suburban2 variable.

The inclusion of wind direction had been proven to be helpful not only for predicting one-step ahead concentrations, but also to identify and estimate possible sources of emission by judging from where the wind dominantly flew. In city center, the direction of southeast was dominant contributor of NO_2 emission because on this direction, there is a big crowded road which includes a governor office. For trading zone, suburban1, highway, and suburban2, the dominant contributors were from northwest, north, southwest, and south, respectively.

The poorer performance of the use of component analysis may become future direction from the current study. One possible problem from component analysis was important values inside the data which were removed by ICA and PCA process. Another possibility is the 10 components extracted from the process may include unnecessary pattern that describe the nature of the data. If the latter is main issue, we should select components that are important based on weight loadings and component figures.

Table |6.24 Parameter estimates of NO_2, PM_{10}, and O_3 with lag-1 independent variables

Variables	**NO_2**	**PM_{10}**	**O_3**
Intercept	1.656 (4.020)	-2.025 (-2.895)	0.872 (1.763)
Pollutants			
NO	-0.006 (-0.916)	0.051 (4.731)	-0.0397 (-5.275)
NO_2	0.685 (94.990)	0.020 (1.642)	0.041 (4.775)
O_3	0.005 (0.845)	0.004 (0.455)	0.744 (107.64)
SO_2	-0.0017 (-0.380)	-0.009 (-1.178)	0.004 (0.852)
PM_{10}	-0.015 (-3.819)	0.628 (93.013)	0.001 (0.198)
CO	0.298 (7.205)	0.601 (8.521)	0.125 (2.532)
Meteorological			
Wind Speed	-0.096 (-4.615)	-0.366 (-10.352)	-0.052 (-2.064)
Solar Gradiation	-0.001 (-11.375)	-0.0005 (-5.440)	0.0002 (3.012)
Humidity	-0.006 (-3.502)	0.008 (3.041)	-0.005 (-2.450)
Temperatures	0.013 (1.221)	0.120 (6.674)	0.027 (2.092)
Peak session time (non-peak time as base)			
Peak morning	0.268 (10.473)	0.110 (2.506)	0.546 (18.086)
Peak afternoon	0.257 (9.404)	-0.015 (-0.325)	-0.218 (-6.724)
Holidays (non-holiday as base)	-0.157 (-5.912)	-0.298 (-6.448)	0.010 (0.335)
Weekends (workdays as base)	-0.012 (-0.689)	0.062 (2.099)	0.014 (0.719)
Zones (city center as base)			
Trading	0.290 (3.543)	0.793 (5.670)	0.100 (1.031)
Suburban1	0.263 (3.312)	1.094 (8.073)	-0.055 (-0.581)
Highway	0.240 (2.843)	-0.037 (-0.258)	-0.409 (-4.05)
Suburban2	-0.157 (-1.216)	0.977 (4.543)	0.253 (1.582)
Wind direction (north as base for each station)			
Northeast1	-0.213 (-2.651)	-0.235 (-1.731)	-0.003 (-0.036)
East1	-0.131 (-1.931)	-0.310 (-2.678)	-0.070 (-0.862)
Southeast1	0.242 (2.929)	0.208 (1.483)	0.085 (0.859)
South1	0.452 (6.073)	1.015 (7.995)	0.025 (0.282)
Southwest1	0.298 (2.933)	0.360 (2.106)	-0.020 (-0.162)
West1	0.123 (1.644)	0.450 (3.518)	0.030 (0.339)
Northwest1	0.328 (3.924)	0.198 (1.402)	-0.061 (-0.605)
Northeast2	-0.493 (-6.690)	-0.465 (-3.705)	0.079 (0.903)
East2	-0.731 (-10.169)	-0.662 (-5.372)	0.004 (0.044)
Southeast2	-0.709 (-9.142)	-0.589 (-4.435)	0.028 (0.305)
South2	-0.100 (-1.192)	-0.032 (-0.228)	-0.099 (-0.990)
Southwest2	-0.039 (-0.536)	0.440 (3.511)	-0.045 (-0.512)
West2	-0.211 (-2.810)	-0.045 (-0.353)	-0.069 (-0.776)
Northwest2	0.254 (2.890)	0.051 (0.340)	0.096 (0.915)
Northeast3	-0.164 (-2.052)	-0.347 (-2.556)	-0.112 (-1.158)
East3	-0.330 (-5.392)	-0.649 (-6.170)	-0.002 (-0.026)
Southeast3	-0.154 (-1.768)	-0.454 (-3.054)	-0.026 (-0.247)
South3	-0.023 (-0.315)	0.425 (3.399)	-0.108 (-1.228)
Southwest3	-0.302 (-4.772)	0.291 (2.687)	0.110 (1.47)
West3	-0.672 (-9.153)	-0.178 (-1.425)	0.100 (1.136)
Northwest3	-0.111 (-1.273)	-0.320 (-2.157)	0.281 (2.674)
Northeast4	-0.135 (-1.443)	-0.068 (-0.430)	0.064 (0.563)
East4	-0.347 (-4.785)	-0.061 (-0.493)	0.016 (0.186)
Southeast4	-0.473 (-6.034)	-0.272 (-2.042)	-0.063 (-0.670)
South4	-0.123 (-1.437)	0.656 (4.531)	0.038 (0.366)
Southwest4	0.031 (0.396)	0.841 (6.257)	0.037 (0.388)
West4	-0.043 (-0.586)	1.121 (9.024)	0.002 (0.028)
Northwest4	-0.171 (-1.974)	0.162 (1.114)	0.077 (0.736)

Variables	NO_2	PM_{10}	O_3
Northeast5	-0.031 (-0.231)	-0.362 (-1.629)	-0.335 (-1.999)
East5	-0.034 (-0.282)	-0.217 (-1.075)	-0.042 (-0.275)
Southeast5	-0.012 (-0.098)	-0.307 (-1.546)	-0.031 (-0.203)
South5	0.126 (1.043)	-0.175 (-0.875)	0.108 (0.714)
Southwest5	0.023 (0.187)	-0.460 (-2.283)	0.202 (1.323)
West5	0.018 (0.145)	-0.165 (-0.796)	0.084 (0.536)
Northwest5	-0.042 (-0.298)	-0.556 (-2.427)	0.149 (0.853)
AR(1) parameter estimates	-0.224	-0.140	-0.364
AIC	44441.81	58558.89	52075.29
BIC	44866.7	58983.79	52500.19
Log likelihood	-22164.9	-29223.44	-25981.65

t-value is listed inside bracket

6.11 Conclusions

Time series data analysis in particular for forecasting air quality concentration is important especially for city where there has been a limited capability of monitoring stations that is able to capture the concentrations reliably. In Surabaya City, there are five monitoring stations, however, the data suffers from missing values, and therefore it is mandatory to have in hand an effective method to forecast 24-h data ahead of air quality. In this study, we attempted to forecast NO, NO_2, PM_{10}, and CO as these are traffic-related gases, associated with the fact of the sharp increase annually of vehicles in Surabaya.

We compared and incorporated three time series model which were autoregressive model, denoted as AR(1), a Bayesian Markov Switching Model, denoted as BMS, and a modified Bayesian Markov Switching Model that incorporated the addition of error intervals to the prediction from BMS model. Judging to the forecasting result, we could not obtain clear performance pattern for BMS and MBMS model. For instance, for NO, the BMS model was better on City center, however, for NO_2, the BMS model showed better performance than AR(1) in City center, Trading zone, and Suburban1. For PM_{10} and CO, the BMS model appeared to be better than AR(1) in Suburban1 (PM_{10}), and City center, Suburban1, and Near highway (CO). This result indicates that the forecasting is highly affected by the training data set which was used to estimate parameters. Moreover, BMS model suffered if the training data set highly fluctuates.

The MBMS model which we expect to improve the performance did not appear to help the forecasted values. For all sites and all air quality parameters, the model's performances were not better than AR(1), only on several sites this model showed better performance. For example, for NO, MBMS model was best on City center, while for NO_2, its performances were better on Suburban1 and near highway. The error intervals which resulted from the errors from the training data set highly affected the forecasting, because of the fluctuation and the range of the training data set, the error intervals might become huge, and therefore will enhance the magnitude of the forecast revision, either become highly positive (overestimate), or become highly negative (underestimate), which further cause high bias. The low performance of time series model led us to improve time series model using optimum order of parameters for AR and tried additional time series model ARIMA.

We observed that the time series models using optimum order of parameters significantly improved the forecasting results both internal validation and for forecasted values. However, we did not see improved forecasting accuracy for PM_{10} indicating that the time series models were not able to predict PM_{10} concentrations. We attempted to increase the power of prediction by increasing the time interval 4-hours. However, only in the trading zone the prediction power of PM_{10} concentration improved. This suggested that time series model can be applied for NO_2 in the case of Surabaya, its application is useful for city center and near highway zone.

Based on above results, we concluded two things. First, the use of time series will be useful if the data source is not complicated. In this case, the model performance is good in the city center and highway because the pattern represents traffic from working and/school pattern. The use of time series models to sites where the emission sources are mixed is not recommended. Second, the use of longer time interval improves prediction power, especially NO_2. However, in the future we will attempt to use longer time interval with different aim of prediciton. One way to improve the model is by using independent component variables. We tried by using both 1-hour and 4-hour data. Unfortunately, the models did not improve the accuracy of prediction. In this case, it is recommended to use air quality data instead of independent components. The use of air quality highly depends on the character of sites. Time series models suit the best with sites where the emission sources are not so mixed.

One particular problem to time series models is the forecasted values will converge to its mean value because of stationarity. For this purpose, we tried using updating algorithm to predict longer time frame in this case one week-ahead concentrations. When applying the updating algorithm for NO_2 in the city center we found that the R^2 of updated time series models were good. It also improved forecasting accuracy however, the R^2 remain low indicating such models cannot be applied to predict longer time scale forecasting values. Furthermore the inclusion of exogenous variables might significantly improve the forecasted values.

Linear models with original variables, ICs, and PCs extracted from six pollutants (NO, NO_2, O_3, SO_2, CO, PM_{10} and meteorological factors (wind speed, solar gradiation, humidity and temperatures) were employed to predict 30-mins ahead of NO_2, PM_{10}, and O_3. In addition, we included serial error correlation computation in the model for better model accuracy. As expected, the presence of NO had positive correlation with NO_2, aside with CO, wind speed and solar gradiation. Furthermore, it was shown that meteorological factors had high role in the formation of O_3. Faster wind speed reduced the concentration of NO_2 while on the opposite increased the concentration of O_3. This pattern was also found for humidity. Since PM_{10} is relatively inert particle gas with less than 10um, using the 30-mins data we obtained, no significant correlation was found with other variables.

During the test step, indicated by internal validation on two zones used for validation, the residual errors from GLS regression were generally lower than ICR and PCR for all pollutants observed (NO_2, PM_{10}, and O_3). Removing the correlation among variables by using predictor variables from ICA and PCA did not improve the performance in the forecasting step. This means the components obtained from these two methods may also remove the necessary pattern of the temporal patterns that affect the prediction.

Using 30-mins interval concentrations of NO_2, PM_{10}, and O_3, we have indicated the effect of other pollutants effect and meteorological factors. The ultimate benefit on using such short interval data is we can exploit the benefit of wind direction instead of taking average if the data were hourly or even daily. Among meteorological factors, we found that the direction of wind highly affected pollutants concentrations, and therefore it is necessary to separate these variables according to their monitoring station location. Holidays and weekends had important role for prediction because the traffic-related emissions tended to be lower during holidays and weekends, due to lower traffic flow. A similar approach in the present study could be extended by incorporating days within week to the data from other stations on other cities to establish a prediction. The information will be very helpful for decision makers in polishing environmental policy related to air quality and human’s health.

Overall in the prediction model, we found that the conventional GLS model with original variable was sufficient to capture and predict future value of pollutant levels. Further in the future direction, we aim to develop more robust time series model by improving the parameter estimation step. For instance, for the time series model we could decide best lag concentration that may effect current concentration and for variables from component analysis, we could pre-filter components by picking the best components leading to less variables to be incorporated inside the GLS model.

Chapter 7 Conclusions, Policy Suggestions, Future Direction, and Limitations

7.1 Summary Findings

This study focuses on exploiting the use of monitoring data to understand the phenomena as well as to predict the air quality levels in a develoing city. First the study aims to obtain factors spatially and temporally that affect pollutant concentrations by using multilevel model. Based on the study, the author summarizes the remarkable findings in chapter 4 as follows:

1. Day-to-day variation and meteorological factors were significant determinants of the air quality. The concentration profiles differed each day because of different magnitude of emission.
2. The fluctuation of air pollutants concentration was mainly contributed by the dynamic changes within station. The variation of different spatial locations was not significant due to the limited number of stations as indicated by the small variation between stations. However, it can be noted that the difference of pollutant concentrations was significant, leading to a conclusion the concentration profile was unique for each station because the emission sources were different each other.
3. Observing the difference between seasons, pollutant concentrations tended to be lower in wet season than in dry season for pollutants SO_2, CO, and PM_{10}. However, for pollutants that were directly emitted by vehicles e.g., NO and NO_2, the concentrations during wet season were relatively higher. Since NO and NO_2 are derived mainly from vehicles, therefore the increase of concentration may be caused by higher car emissions. This higher emission may be triggered by factors such as more traffic congestion occured in wet season or changes of on road users behavior.
4. This study examined the concentration profile on each specified time interval each day when peak concentrations were achieved. It was confirmed that during both morning and evening peak time, the concentrations were significantly higher than non-peak time.
5. By using longer time frame data, it was found that adding more green spaces had been useful to reduce the traffic related emission e.g., NO and PM_{10}.

As a summary of chaper 4, the author extracted an important conclusion and future resposible work. Firstly, it was concluded that the concentrations were significantly different across sites due to different emission sources and land use patterns. This means the concentrations were unique on each location. Secondly, it was required to find out a specific temporal pattern or temporal peak time frame in each location.

Based on the summary, in chapter 5, the author employed ICA that can decompose the compricated mixture discribution into some sourse distributions. Furthermore, state-space model allows us to better understand how meteorological factors affect the dynamic changes in pollutant concentrations. It was found that ICA is very helpful to identify an unique pattern in each location as well as discriminative peak events of pollutants concentrations. In the case of NO_2 mainly emitted from vehicles, the study observed similar a distribution pattern among city center, suburban1, and near highway zone. This pattern may reflect commuting and/or business activities during morning (530am to 10am) and evening peak time (330pm to 12pm). For pollutants that were emitted by more complex sources e.g., PM_{10}, the temporal patterns between different locations were relatively more varied than NO_2.

The emission type of pollutants (NO_2 and PM_{10}) in each location was observed by investigating one by one monitoring station and matching them with wind speed and wind direction. Overall, the concentration of NO_2 primarily was estimated to come from roads and streets. On the other hand, PM_{10} contributors may came from more varied sources e.g., business related buildings, apartments, and government buildings. Therefore, it is crucial for any researchers to first identify the statistically dominant wind direction and what pollutants are most associated with the direction of the wind, then we can obtain the estimated cause of pollution. The observation of the effect of seasons can also be captured by ICA. It is shown that there were different source patterns of pollutants between dry and wet season in particular for NO_2 and PM_{10}. The concentration of NO_2 tended to increase on wet season whereas the concentration of PM_{10} tended to be lower on dry season.

ICA can be used to interpret the phenomena of air quality. Also this study demostrated the use of ICA for predicting pollutant levels. ICA is proven to determine temporal peak time frame, especially for patterns which are unique for each location or stations because its patterns as a function of time. The concentration between stations may differ because the pollutant levels depend on the nature of emission sources. For short time scale, ICA is useful to determine daily cycle, even for weekly and annual cycle. For annual cycle, this study can observe the effect of seasons to the pollutant's concentrations. For prediction purposes, components resulted from ICA, on which these components are statistically independent one another, there are risks posed. If we include IC components into prediction models, there is a possibility some important noises contained within air quality data are removed because of ICA process which reduce the reliability and accuracy of predicted values. One solution is to reduce variables or selecting components which deemed to be representative or which has dominant variant among variables. If ICA produces 10 components, the author can select components used for the models if the patterns are clear by observing weight loadings and component figures. Overall, the future usage of ICA is wide in particular in air quality management.

From ICA, it was found that there were two time peaks (morning and evening) in which the pollutants concentrations were high. However, the peak concentrations in the morning were significant than in the evening. We obtained this conclusion from the result obtained by using time series state-space model, specifically for trading zone. Furthermore, the effect of meteorological factors had been important towards the pollutant levels, however the magnitude was smaller than other factors as indicated by the dynamic of intercept estimate parameter. The intercept parameter of O_3 was particularly not too dynamic as the effect of meteorological factors played an important role for O_3 formation which strengthened the fact that O_3 is formed by a series of reaction between precursor gases in the atmospheric layer.

Prediction cannot be set apart from air quality management because of its ultimate benefit showing the prediction of air quality to expect public will react to this information. The purpose is to increase awareness of people and to reduce the risk of exposed to the pollution especially when the pollution breaches the permissible level. In Chapter 6 this study used two different approaches, the first approach was time series models and the second approach was a linear model, GLS model.

In the time series models, the author employed several steps. At the first step, this study compared three different models of autoregressions to Bayesian approach model. The results showed the models did not have good prediction power to predict future values of air quality. Therefore in the second step, the author employed time series models AR and ARIMA with optimum order or parameters (lag concentrations). Even though the performance was better than models from the first step, we found that the models had a higher prediction power only in the city center and near highway zone for NO_2. Therefore the author can conclude that applying time series models are good for sites with not so mixed emission sources, for instance in this case the time series models had better performance for NO_2 only, not for PM_{10}. Furthermore, the increase of time interval training data to 4-hours had successfully improved the prediction power of the models. However, future works was necessary to find out best time interval data for the prediction.

Because of partially lower performances of the prediction models, the study attempted to predict true distribution pattern obtained from ICA based on 1-hour and 4-hour NO_2 data. However, the results were not better than if we use raw NO_2 data. This might be due to the dynamic flactuation of independent components. Therefore we concluded that the application of time series models are better for site with the emission sources which are not so mixed and complicated. Furthermore, the use of longer interval data such 4-hour will improve the accuracy of predicted values.

The use of time series models have been known of the simplicity. However, we obtained worst result of PM_{10} that might be due to the mixed source of emission sources and the dynamic factors that affect PM_{10}. Therefore we attempted to include additional variables to predict air quality by using GLS. In the GLS model we include meteorological variables e.g., wind direction and wind speed. We compared the performance of conventional model with GLS models with the variables extracted from component analysis. The variation of pollutants can be well explained by the inclusion of meteorological variables. However, the addition of variables from component analysis (ICA and PCA) did not improve the accuracy. Therefore, using a conventional GLS model with original variables was sufficient to predict future pollutants concentrations. The development of time series model in the future has been challenging yet attractive due to the limitation of data obtained from sensors.

7.2 Policy Suggestions

7.2.1 Spatial Policy for each zone

It is noted that the fluctuation of pollutants concentration highly depends on the dynamic within monitoring station. The spatial variability or here we refer to how each monitoring station affect the concentration of other station. In the case of Surabaya City, this kind of effect is minimal as suggested in Chapter 4. Therefore, this study focus on developing policy on each case by case, each location to location. However, it must be noted that based on our previous description that NO_2 and PM_{10} have the most event when their concentrations breach the permissible level, therefore we focus and highlight the recommendations based on the results from these two pollutants.

City center is marked with high traffic volume because it is located in central area of Surabaya. Its location is near with city government office. Not only is that, the location's surroundings full with residential and business activities. The location of the sensor is next to the big river crossing the Surabaya City. There were many sessions when the concentration of NO_2 and PM_{10} exceeded the permissible levels. To prevent or at least to reduce the possibility of the worse events, we propose a policy, specifically for this zone. Based on the analysis from Chapter 4 and Chapter 5, the peak concentration of NO_2 occurs between 6am to 10am in the morning and between 330pm to 10pm. These are resulted mainly from traffic flow, whereas for PM_{10}, the peak concentration occurs between 7am to 11am and in the evening there is a wider interval between 330pm to 12pm. In this zone, our recommendation is a long term one. A reduction of private car usage must be done by promoting shifting to public transport as much as possible. Policy maker could also consider limiting vehicles passing Gubernur Suryo Road on 6am to 10am. The limitation of vehicles may be done in form of restricting big vehicles, or the implementation of 3-in-1 route. The purpose is to reduce traffic flow. However, the implementation may pose high difficulty since the road is the main road where such segmentation may sacrifice people's mobility. The transportation in this road is complex from school activities, working activities, to leisure activities. This type is significantly different with that in the trading zone where the emission source is mainly from a street connecting to a harbor which policy maker can manage with the involvement of many stakeholders and company users.

In the trading zone, the rise of concentrations started from 9pm until 2~3am, which are due to traffic flow vehicles from/to Tanjung Perak Harbor as well as residential and business related activities. Therefore, the plausible policy suggestion is to manage the distribution of vehicles passing the Perak Timur and Perak Barat Road. Based on the study taken from Chapter 4 and 5, what we could propose is to set restriction between 7am to 10am for big vehicles from companies. This policy is for every day. The implementation of such policy will be hard initially, but despite letting people especially in the north of Surabaya, it is one that inevitably must be done. Alternative policy such as car free day or limiting access of small vehicles such as motorbikes or paratransit will not be effective because in this zone, there are many business, offices related activities. There are also medical clinic and hospitals. Therefore, our only option now is to limit the distribution of big vehicles. For this policy implementation, government should coordinate with companies and the ship companies to allow for any adjustment for this scenario. The details of this scenario are outside the scope of this study. However, one could initiate environmental cost-benefit because of the implementation of this proposal.

Sukomanunggal region is located in the west part of Surabaya, we denote as suburban1 in the model. Specifically for this region, to prevent and reduce the events when pollutants, in particular NO_2 and PM_{10} exceeding threshold level, planting more trees is one option that policymaker should consider. However, we also realize that spaces may not be sufficient enough to allow for emission reduction, therefore adding more routes or public transportation are still required in reducing the emission from private cars.

Another complex source is observed in the Gayungsari monitoring station where this site is near with highway. The pollution level is highly related to the traffic in this highway because it connects Surabaya Cities with other neighbor sister cities such as Gresik, Porong, Sidoarjo. In order to reduce emission in this area, we must impose a longer term policy which is asking people's awareness to install catalytic converter into the car's exhaust. This is proven way to reduce the ambient concentration. However, the executive of the government, in this case the province government, should also consider in building more feasible, comfortable public transport between Surabaya to those cities. Even it requires commitment due to its long term implementation, the shift to public transport is the most effective way to reduce pollution, when suppressing the vehicle ownership may be viewed to be difficult. The policy of public transport shift is also recommended to be applied in Sukolilo monitoring station (suburban) on east side of Surabaya. Another recommendation as part of air quality management in the highway zone is to install display, often referred as visual message signboards (VMS) in each gate of highway. The display will show the current pollutants concentration as well as congestion level on each segment. Currently, there are several signs of traffic level e.g., congestion, smooth, within some segments in the highway. If another display of pollutant level could be added, at least two objectives for people can be achieved. First, when the event of both pollutants NO_2 or PM_{10} exceeding the threshold level, the display will warn users and the highway users are expected to take preventive action e.g., taking alternative exit. Furthermore, to reduce health impacts users may wear mask or do other preventive treatment. This is part of changing travel and user's behavior. Observing into a wider scale, users are expected to shift into public transport or with a comprehensive program, government should touch the awareness of car owners to install catalytic converter into their car or at least routinely check for emission values.

The pollutant level shown in the display itself may be obtained from previous values or by using predicted values on which in the present study we propose several methods. Indirectly this policy may encourage users and increase awareness of air pollution. For this, the display should at least show the concentration and the preventive action suggested e.g., to wear mask, or take other exit.

In the Sukolilo (suburban2) site, possible sources of the events were from traffic from/to which passed Arief Rahman Hakim Rd. To reduce further pollutants, it is mandatory to promote the use of public transport to this area. Providing adequate public transport will reduce the use of private vehicles, especially in the peak time between 7am-9am in the morning and around 6pm in the evening. A short term policy such as route adjustment may also be considered.

The tables below contain all possible direct policies that were derived from the temporal variation listed above. The content of the table below (Table 7.1) are self-explanatory based on observation of Table 5.14. From all temporal time frames on which the NO_2 and PM_{10} concentrations are high, we try to list possible direct policies that can be implemented. However, the candidate of policies are short listed and based on assumption that NO_2 pollution is a result of transportation and PM_{10} pollution come from more mixed sources. A more integrated approach even though they seem cliché is still important to be displayed and proposed.

Table 7.1 Possible Direct Policies based on extracted components

Sites	Short term	Long term
City Center (Ketabang Kali)	1) Restriction of vehicles passing Gubernur Suryo Road from 6am to 10am e.g., no big vehicles are allowed (truck) 2) Staggered business hour 3) Traffic route changes	Modal shift from private to public transportation by improving the public transport services
Trading (Perak)	1) Add more green spaces 2) Restriction of vehicles passing Perak Barat and Perak Timur Rd. from 7am to 10am 3) Restriction of vehicles passing Perak Barat and Perak Timur Rd. around 5pm every week days	Modal shift from private to public transportation by improving the public transport services
Suburban1 (Sukomanunggal)	1) Adding more green spaces 2) Staggered business hours of some shopping malls	Modal shift from private to public transportation by improving the public transport services
Near highway (Gayungsari)	Providing air pollutant information on visual message signboards (VMS) to recommend drivers to change their route and wear the masks	Expansion of public transport networks to neighbor cities
Suburban2 (Sukolilo)	Proving air pollutant information on displays to recommend drivers to change their route	Modal shift from private to public transportation by improving the public transport services

7.2.2 Implementing the Prediction of Air Quality

In the present study we have explored the several methods to predict air pollutants. We have shown that the best model is not always the complex one as shown on Chapter 6. We tried to implement two states in differentiating the pattern of pollutants concentration using Bayesian Markov Switching approach. However, the results are in favor for simple AR(1) model. This actually suggests that the concentration of pollutants in Surabaya City, especially for 1-hour interval data, follows a perfect stationary assumption which leads better result in this model. AR(1) can be used in the future by any researchers to train newer data set and use the estimates to forecast up to next 24-h ahead.

In general, we recommend the use of time series models for site that does not have complex land use. Time series models for predicting NO_2 shows good performance than other pollutants such as PM_{10}. Furthermore, using long term interval data for training data is encouraged.

Since the air quality concentration follows a spatio-temporal pattern, many models are used to predict these gases. However, we still found them which are not taking into account time series serial autocorrelation for better prediction. Therefore we use Generalized Least Squared model by taking into account residual serial autocorrelation. In this scenario we include the spatial meteorological variables including the effect of wind direction. We observe all air quality spatio-temporal model should include wind direction because the significant contribution of emission is unique for each location. We found this variable to be highly important in predicting the concentrations. Further finding includes the conclusion of not using independent variables extracted from component analysis. In the present study we show that it is better to use original variables (after transformation) instead of using variables from component analysis. Component analysis produces components that are statistically independent each other and the assumption is that these components represent true pattern of air pollutants. However, since the results in the present study are not in favor of using the variables from both PCA and ICA, we conclude that we must be very cautious when we use these methods as there is high change that they also remove significant patterns intrinsically within the fluctuation of the air pollutants concentration.

7.3 Limitation of Study

This section explains some limitations of the study. There are two major limitations of data availability and prediction models. Firstly, regarding the data source this study originally wishes to investigate spatio-temporal model, however Surabaya City does not have adequate monitoring stations. Therefore this study limits its discussion on temporal prediction. The furthest we can go is to develop a model with spatial indicators as independent variables. Secondly, we suffers missing data information. There are high rate of missing because of sensor disorder. In fact, we do not have at least 50% incomplete data from year 2003 to 2009. Therefore the author was forced to concentrate only two year data between year 2001 and year 2002. There is also a concern about the height of sensors which is 5 m above the ground. This fact must be taken into account for the quality of data.

Regards on the findings from the models, the author attempted to figure out emission sources for pollutants. However, it was difficult to justify the results with evidence. This is because such data for validation is not exist and beyond the scope of this study. It is therefore necessary to focus in the future to collect databases of emission sources as well traffic volume to validate the results obtained in the present study.

In addition for research and model limitation, the author suffers from poor performance of prediction model. Not only because of the highly dynamic of the data nature but also because of selecting the indicators such as time interval and lag of concentrations. This study did not deal with these two points due to the data quality again. In other words, the results of this study highly depends on the specific condition of Surabaya. Therefre, it is noted that the conclusions from this study are not generalized to other developing cities, but should be shared as a lesson.

7.4 Future Directions

There are two main points of our future directions based on our limitations above. Firstly regarding the limitation of data source, more detail studies on spatial model will be started off with different data sources such as remote sensing mode. It is known that we can obtain information of NO_2 pollutants from MODIS (*Moderate Resolution Imaging Spectroradiometer*) image. Therefore, the author will study the step of acquiring image of MODIS and thorough investigation using MODIS as a tool to obtain information about air pollutants over the region of Surabaya.

The second future direction is to strengthen and deepen the methods to predict future concentration of air pollutants, in particular for O_3, because O_3 doesn't come from direct emission instead it is resulted from a series of atmospheric reactions. There are many time series methods that can be applied which will be our next concern.

Our third direction is to combine the air quality research with: 1) transportation mode policy, in particular in Surabaya City, 2) co-benefit analysis, and 3) social and quality of life. More temporal aggregation will be considered into model application and development. Furthermore, we will develop more model e.g., state-space model.

Regarding to the limitation of study, the author would extend to investigate the effect of lag concentrations influence for prediction accuracy, and to explore the performances from different time intervals. As for GLS models with component analysis, tit is required to optimize the selection of variables from component analysis to be included in the GLS models. These works will improve overall performance of forecasted values.

References

Abdul-wahab, S. A., Bakheit, C. S., and Al-alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations, 20, 1263–1271.

Al-Alawi, S. M., Abdul-Wahab, S. A., and Bakheit, C. S. (2008). Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environmental Modelling and Software, 23, 396–403.

Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., and Vieu, P. (2004). Maximum ozone concentration forecasting by functional non-parametric approaches. Environmetrics, 15(7), 675–685. doi:10.1002/env.659

Anonym (2009) Regulation Governor of East Java about Threshold Level of Ambient Air Quality and Stationary Emission in East Java

Arhami, M., Kamali, N., and Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research International*, 20(7), 4777–89

Athanassiadou, M., Baker, J., Carruthers, D., Collins, W., Girnary, S., Hassell, D., Hort, M, Johnson, C., Johnson, K., Jones, R., Thomson, D., Trought, N., Witham, C. (2010). An assessment of the impact of climate change on air quality at two UK sites. *Atmospheric Environment*, 44(15), 1877–1886.

Back A., and Weigend A. (1997) Discovering structure in finance using independent component analysis, Proceedings of 5th International Conference on Neural Networks in Capital Market, Kluwer Academic, 1997, pp. 15–17.

Bigi, A., and Harrison, R. M. (2012). Analysis of the air pollution climate at a central urban background site. *Atmospheric Environment*, 44(16), 2004–2012. doi:10.1016/j.atmosenv.2010.02.028

Cai, M., Yin, Y., and Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment*, 14(1), 32–41.

Cavanaugh, J.E., and Shumway, R.H. (1997) A bootstrap variant of AIC for state-space model selection. *Statistica Sinica,* 7, 473,496

Cesaroni, G., Porta, D., Badaloni, C., Stafoggia, M., Eeftens, M., Meliefste, K., and Forastiere, F. (2012). Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. Environmental Health, 11:48

Chaloulakou A, Grivas G, and Spyrellis N (2012) Neural network and multiple regression models for PM_{10} prediction in Athens : a comparative assessment. *Journal of the Air and Waste Management Association* 53:10, 1183-1190

Chamida. (2004). Policy For Air Pollution Control Strategy By Using The Air Pollutant Dispersion Model (PM_{10}) in Surabaya, URL: cleanairinitiative.org/portal/system/files/59169_chamidha_paper.doc (accessed on: March 15th, 2012)

Chamida. (2006). Air Quality Improving - Experiences of Surabaya City Indonesia. The Sixteenth Asia Pasific Seminar on Climate Change *"Asia Pasific Regional Approach to Climate Friendly and Climate Change - Resilient Society"*, Jakarta, Indonesia, 5 - 8 September 2006

Cheung, Y., and Xu, L. (2001). Independent component ordering in ICA time series analysis. *Neurocomputing*, 41, 145–152.

Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185(2-4), 513–529.

Dai, W., Liu, B., Ding, X., and Huang, D. (2013). Modeling dam deformation using independent component regression method. Transactions of Nonferrous Metals Society of China, 23(7), 2194–2200.

Damon, J., and Guillas, S. (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. Environmetrics, 13(7), 759–774. doi:10.1002/env.527

Diebold FX, Lee J-H, Weinbach G (1994) Regime switching with time-varying transition probabilities. In C. Hargreaves (ed.), Nonstationary Time Series Analysis and Cointegration, 283-302. Oxford: Oxford University Press.

Dijkema, M. B., *et al.* (2011). Long-term exposure to traffic-related air pollution and type 2 diabetes prevalence in a cross-sectional screening-study in the Netherlands. *Environmental Health*, 10:76

Djuraidah, A. (2007). Model Aditif Spatio-Temporal untuk Pencemar Udara PM_{10} dan Ozon di Kota Surabaya dengan Pendekatan Model Linear Campuran. Disertasi di Sekolah Pascasarjana, Departemen Statistika, Institut Pertanian Bogor

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., and Kenski, D. (2009). $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications*, 36(5), 9046–9055.

Dou, Y., Le, N. D., and Zidek, J. V. (2012). Temporal Forecasting with a Bayesian Spatial Predictor: Application to Ozone. *Advances in Meteorology*, 1–13

Elbir, T., Mangir, N., Kara, M., Simsir, S., Eren, T., and Ozdemir, S. (2010). Development of a GIS-based decision support system for urban air quality management in the city of Istanbul. *Atmospheric Environment*, 44(4), 441–454.

Fassò, A., Cameletti, M., and Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. Environmetrics, 18, 245–264. doi:10.1002/env

Finazzi, F., Scott, E. M., and Fassò, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. Journal of the Royal Statistical Society. Series C, Applied Statistics, 62(2), 287–308. doi:10.1111/rssc.12001

Fuller, G.W., D.C. Carslaw and H.W. Lodge, 2002. An empirical approach for the prediction of daily mean PM concentrations. Atmos. Environ., 36(9): 1431-1441.

Gao, H. O. (2007). Day of week effects on diurnal ozone/NOx cycles and transportation emissions in Southern California. Transportation Research Part D: Transport and Environment, 12(4), 292–305.

Gardner M.W., and Dorling S.R. (1998). Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. *Atmospheric Environment* (32), 2627-2636

Gardner, M. W., and Dorling, S. R. (1999). Neural network modelling and prediction of hourly NO and NO concentrations in urban air in London. *Atmospheric Environment* 33, 709-719

Geladi P., Hadjiiski L., and Hopke P. (1999). Multiple regression for environmental data: Nonlinearitas and prediction bias. *Chemometrics and Intelligent Laboratory Systems*, (47), 165-173

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, America

Ghosh, S. K., Bhave, P. V., Davis, J. M., and Lee, H. (2010). Spatio-Temporal Analysis of Total Nitrate Concentrations Using Dynamic Statistical Models. Journal of the American Statistical Association, 105(490), 538–551.

Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., and Boyadzhiev, D. T. (2013). Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment*, 28(4), 1045–1060

Goldfeld S, Quantd R (2005) A Markov model for switching regression. Journal of Econometrics 135: 349–376.

Hampton, S. E., Holmes, E. E., Scheef, L. P., Scheuerell, M. D., Katz, S. L., Pendleton, D. E., and Ward, E. J. (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (MAR) models. Ecology, 94(12), 2663–2669.

Han, S. (2011). Analysis of the Relationship between O3, NO and NO_2 in Tianjin, China. *Aerosol and Air Quality Research*, (2), 128–139.

Han, S., Bian H., Feng Y., Liu. A., Li, X., Zeng, F., and Zhang, X. (2011). Analysis of the Relationship between O3, NO and NO2 in Tianjin, China. Aerosol and Air Quality Research, (2), 128–139. doi:10.4209/aaqr.2010.07.0055

Henderson, S. B., Beckerman, B., Jerrett, M., and Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environmental Science and Technology, 41(7), 2422–8.

Holmes E.E., Ward, E.J., and Scheuerell, M.D. (2014) Analysis of multivariate time-series using the MARSS package. Available on: *cran.r-project.org/web/**packages/MARSS**/vignettes/**User**Guide.pdf*f

Holmes, E. E., Ward, E. J., and Wills, K. (2012). MARSS : Multivariate Autoregressive State-space Models for Analyzing Time-series Data. *Contributed Research Article in The R Journal* 4(1), 11–19.

Honaker J., King G., and Blackwell M (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7), 1-47

Hyndman R.J., Athanasopoulos G., Razbash S., Schmidt D., Zhou Z., and Khan Y. (2014) Forecast: Forecasting functions for time series and linear models. R package version 5.1; http://CRAN.R-project.org/package=forecast

Hyvärinen, A, and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13(4-5), 411–30.

Ibrahim, M. Z., Zailan, R., Ismail, M., and Lola, M. S. (2009). Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia. *American Journal of Environmental Sciences*, 5(5), 625–632.

Irsyada, W.S. (2011). *Study of People Behavior on the Implementing of Air Pollution Mitigation Program in Surabaya*, Undergraduate Theses of Dept of Environmental Engineering, Institut Teknologi Sepuluh Nopember, Surabaya.

Ji J, Shen YT, Tan MS, Wu LN, Zhang JH, Liu ZQ (2013) Short-term wind speed forecasting based on time series-state transition method. Applied Mechanics and Materials 291-294: 2298–2301

Jorquera, H. (2002). Air quality at Santiago, Chile: a box modeling approach—I. Carbon monoxide, nitrogen oxides and sulfur dioxide. *Atmospheric Environment*, 36(2), 315–330.

Khoder, M. I. (2009). Diurnal, seasonal and weekdays-weekends variations of ground level ozone concentrations in an urban area in greater Cairo. *Environmental monitoring and assessment*, 149(1-4), 349–62.

Kim AC, Nelson CR, Kim C (2013) Has the US Economy become more stable ? A bayesian approach based on a markov-switching model of the business cycle. *The Review of Economics and Statistics* 81(4): 608–616.

Kim, Y., Kim, M., Lim, J., Kim, J. T., and Yoo, C. (2010). Predictive monitoring and diagnosis of periodic air pollution in a subway station. *Journal of Hazardous Materials*, 183(1-3), 448–59.

Kukkonen, J. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, 37(32), 4539–4550.

Kumar, U., and Jain, V. K. (2009). ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). *Stochastic Environmental Research and Risk Assessment*, 24(5), 751–760.

Kuo, Y.-M., Wang, S.-W., Jang, C.-S., Yeh, N., and Yu, H.-L. (2011). Identifying the factors influencing PM2.5 in southern Taiwan using dynamic factor analysis. *Atmospheric Environment*, 45(39), 7276–7285.

Lee MH, Haizum N, Rahman A, Latif MT, Nor ME, Arina N, Kamisan B (2012) Seasonal ARIMA for forecasting air pollution index : a case study. *American Journal of Applied Sciences* 9(4): 570–578.

Liu, Y., Guo, H., Mao, G., and Yang, P. (2008). A Bayesian hierarchical model for urban air quality prediction under uncertainty. Atmospheric Environment, 42(36), 8464–8469.

Long, F., Wu, T., Movellan, J. R., Bartlett, M. S., and Littlewort, G. (2012). Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing*, 93, 126–132. doi:10.1016/j.neucom.2012.04.017

Lu, C.-J., Lee, T.-S., and Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2), 115–125. doi:10.1016/j.dss.2009.02.001

Lui, G. C. S., Li, W. K., Leung, K. M. Y., Lee, J. H. W., and Jayawardena, a. W. (2007). Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter. Ecological Modelling, 200(1-2), 130–138.

Madhavi Latha, K., and Highwood, E. J. (2006). Studies on particulate matter (PM_{10}) and its precursors over urban environment of Reading, UK. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 101(2), 367–379.

Malaroiu S, Kiviluoto K., Oja E. (2000) Time series prediction with independent component analysis, in: E. Tonkes, C. Tan, S. Sugden, K. Kumar (Eds.), Proceedings of the International Conference on Advanced Investment Technology, Gold Coast, Australia.

Mamtimin, B., and Meixner, F. X. (2011). Air pollution and meteorological processes in the growing dryland city of Urumqi (Xinjiang, China). *The Science of the total environment*, 409(7), 1277–90.

Martínez-Beneito MA, Conesa D, López-Quílez A, López-Maside A (2008) Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine* 27(22): 4455–68.

Minguillón, M. C., Monfort, E., Escrig, a., Celades, I., Guerra, L., Busani, G. Sterni, A., and Querol, X. (2013). Air quality comparison between two European ceramic tile clusters. *Atmospheric Environment*, 74, 311–319.

Monoura H. (1999). Some characteristics of surface ozone concentration observed in an urban atmosphere. *Atmospheric Research* (51), 153-169.

Murray, C. J., and Nelson, C. R. (2000). State-Space Modeling of the Relationship between Air Quality and Mortality. Journal of the Air and Waste Management Association, 50(7), 1075–1080. doi:10.1080/10473289.2000.10464158

Nugroho, S. B., Fujiwara, A., and Zhang, J. (2012). Spatial and Temporal Analysis of Surface Ozone in Urban Area: A Multilevel and Structural Equation Model Approach. *Air Pollution - A Comprehensive Perspective*, Dr. Budi Haryanto (Ed.), ISBN: 978-953-51-0705-7, InTech, DOI: 10.5772/50417. Available from: http://www.intechopen.com/books/air-pollution-a-comprehensive-perspective/spatial-and-temporal-analysisof-surface-ozone-in-urban-area-a-multilevel-and-structural-equation-mo. Accessed: 15th March 2013

O'Donoghue, R. T., Broderick, B. M., and Delaney, K. (2007). Assessing the impacts of infrastructural road changes on air quality: A case study. Transportation Research Part D: Transport and Environment, 12(8), 529–536.

Parkhurst, G. (2004). Air quality and the environmental transport policy discourse in Oxford. Transportation Research Part D: Transport and Environment, 9(6), 419–436.

Perez, P. and J. Reyes, 2002. Prediction of maximum of 24-h average of PM concentrations 30 h in advance in Santiago, Chile. Atmos. Environ., 36 (28): 4555-4561.

Perlin M (2010), MS Regress - The MATLAB Package for Markov Regime Switching Models. Available at SSRN: http://ssrn.com/abstract=1714016

Pires, J. C. M., Martins, F. G., Sousa, S. I. V, Alvim-Ferraz, M. C. M., and Pereira, M. C. (2008c). Prediction of the Daily Mean PM_{10} Concentrations Using Linear Models. *American Journal of Environmental Sciences*, 4(5), 445–453.

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M., and Martins, F. G. (2008a). Management of air quality monitoring using principal component and cluster analysis—Part I: SO_2 and PM_{10}. *Atmospheric Environment*, 42(6), 1249–1260.

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M., and Martins, F. G. (2008b). Management of air quality monitoring using principal component and cluster analysis—Part II: CO, NO2 and O3. *Atmospheric Environment*, 42(6), 1261–1274.

Pollice, A., and Lasinio, G. J. (2010). A multivariate approach to the analysis of air quality in a high environmental risk area z, (May), 741–754. doi:10.1002/env

Purwanto, D.S. (2007). Pola Hubungan Antara Volume Kendaraan Bermotor dengan Kadar Partikulat, Pb Udara dan Prediksi Kualitas Udara di Jalan Raya Kota Surabaya. Master Theses of Postgraduate of Environmental Management Engineering, Institut Teknologi Sepuluh Nopember, Surabaya. In Indonesian Language

Rachmani, N. (2007). *Dampak Pencemaran Udara terhadap Tumbuhan di Kebun Bibit Bratang Surabaya.* Master Theses of Postgraduate of Environmental Management Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, In Indonesian languange.

Rahmawati, L. D. (2008). *Numerical Solution on The Dispersion Of Carbon Monoxide Pollutant in the Air*. Undergraduate Thesis from Department of Mathematics, Institut Teknologi Sepuluh Nopember (ITS) Surabaya

Ramanathan, V., and Feng, Y. (2009). Air pollution, greenhouse gases and climate change: Global and regional perspectives. *Atmospheric Environment*, 43(1), 37–50.

Rubin, E.S. (2001). *Introduction to Engineering and the Environment*. McGraw-Hill International Edition, Singapore.

Sahu SK, Yip S, Holland DM (2009) A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics* 18(1): 185–207

Sahu, S. K., Yip, S., and Holland, D. M. (2009). A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics*, 18(1), 185–207

Samino. (2010). *Analysis Levels Pollution Ambient Air PM_{10}, SO_2, CO, O_3 and NOx Surabaya City.* Master Thesis of Faculty of Public Health, Airlangga University, in Indonesian Language

Sanchez-ccoyllo, O.R., Ynoue, R.Y., Martins, L.D., Andrade, M.F. (2006). Impacts of O_3 precursor limitation and meteorological variables on ozone concentration in Sao Paulo, Brazil. *Atmospheric Environment* (40), 552-562.

Schwarz, C.J. (2013) Sampling, Regression, Experimental Design and Analysis for Environmental Scientists, Biologists, and Resource Managers. Department of Statistics and Actuarial Science, Simon Fraser University. Url: http://people.stat.sfu.ca/~cschwarz/Stat-650/Notes/PDFbigbook-JMP/. Accessed December 2013.

Seinfeld, J.H., and Pandis, S.N. (1998). *Atmospheric Chemistry and Physics*. John Wiley and sons, Inc. New York.

Sfetsos A (2002) A novel approach for the forecasting of hourly mean wind speed time series. Renewable Energy, 27(2): 163–174.

Shao, X., Wang, W., Hou, Z., and Cai, W. (2006). A new regression method based on independent component analysis. *Talanta*, 69(3), 676–80.

Shumway RH, Stoffer DS (2006) Time series analysis and its applications—with R examples. Springer Science + Business Media, LLC

Syafei A.D., Fujiwara A., Zhang Z. (2013) A comparative study on NO concentration interpolation in Surabaya City. *Proceedings of the Eastern Asia Society for Transportation Studies*, Vol. 9, 2013

Touloumi, G., Pocock, S. J., Katsouyanni, K., and Trichopoulos, D. (1994). Short-term effects of air pollution on daily mortality in Athens: a time-series analysis. *International Journal of Epidemiology*, 23(5), 957–67.

Vallero D.A. (2008) Fundamentals of Air Pollution. 4th Edition, Elsevier, USA

Verbeke, J.S., Hartog, J.C.D., Dekker, W.H., Coomans, D., Buydens, L., Massart, D.L., 1984. The use of principal components analysis for the investigation of an organic air pollutants data set. *Atmospheric Environment* 18 (11),2471–2478.

Wang, G., Bai, S., and Ogden, J. M. (2009). Identifying contributions of on-road motor vehicles to urban air pollution using travel demand model data. Transportation Research Part D: Transport and Environment, 14(3), 168–179.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G., and Brunekreef, B. (2012). Systematic evaluation of land use regression models for NO_2. *Environmental Science and Technology*, 46(8), 4481–9. doi:10.1021/es204183v

Wang, X., and Oliver Gao, H. (2011). Exposure to fine particle mass and number concentrations in urban transportation environments of New York City. *Transportation Research Part D: Transport and Environment*, 16(5), 384–391.

Ward, E. J., Chirakkal, H., González-Suárez, M., Aurioles-Gamboa, D., Holmes, E. E., and Gerber, L. (2010). Inferring spatial structure from time-series data: using multivariate state-space models to detect metapopulation structure of California sea lions in the Gulf of California, Mexico. *Journal of Applied Ecology*, 47(1), 47–56.

Westra, S., Brown, C., Lall, U., Koch, I., and Sharma, A. (2009). Interpreting variability in global SST data using independent component analysis and principal component analysis. *International Journal of Climatology*, 30, pp. 333-346.

Westra, S., Brown, C., Lall, U., Koch, I., and Sharma, A. (2010). Interpreting variability in global SST data using independent component analysis and principal component analysis. International Journal of Climatology, 30:333-346

Wu E.H.C., and Yu P.L.H. (2005). *Independent Component Analysis for Clustering Multivariate Time Series Data*. as Book Chapter of Advanced Data Mining and Applications, First International Conference, ADMA 2005, Wuhan, China, July 22-24, pp 474-482.

Zhai, S., Li, Z., and Zhao, B. (2014). State-space analysis of influencing factors on airborne particle concentration in aircraft cabins. *Building and Environment*, 74, 13–21.

Zhang, B.N., and Oanh, N.T.K. (2002). Photochemical smog pollution in the Bangkok Metropolitan Region of Thailand in relation to O_3 precursor concentrations and meteorological conditions. *Atmospheric Environment*, (36), 4211-4222.

Publications

Syafei AD., Fujiwara A., and Zhang, J. (2013) A comparative study on NO Concentration interpolation in Surabaya City. Proceeding of the Eastern Asia Society for Transportation Studies, Vol. 9

Syafei AD, Fujiwara A., and Zhang J. (2014) Spatial and Temporal Factors of Air Quality in Surabaya City: An Analysis based on a Multilevel Model. *Procedia – Social and Behavioral Sciences*, Vol. 138, pp 612-622

Syafei AD., Fujiwara A., and Zhang, J. (2014) Prediction Model of Air Pollutant Levels using Linear Model with Component Analysis. *International Journal of Environmental Sciences and Development. Accepted*

Syafei AD., Fujiwara A., and Zhang, J. (2014) Temporal and Spatial Analysis of Peak-concentration Times for NO and NO_2 in Morning and Evening Events: Case Study on Surabaya. *Procedia of Environmental Sciences* (*under review*). (*Extended abstract accepted*)