CITY LEVEL AIR POLLUTION MODELLING AND MAPPING

LINGYUE KONG
February, 2015

SUPERVISORS:
Dr. Nicholas Hamm
Prof.Dr.Ir. Afred Stein

CITY LEVEL AIR POLLUTION MODELLING AND MAPPING

LINGYUE KONG
Enschede, the Netherlands, February, 2015

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISOR:
Dr. Nicholas Hamm
Prof.Dr.Ir. Afred Stein

THESIS ASSESSMENT BOARD:
Prof.dr.ir. M.G. Vosselman (chair)
Dr.ir.G.Hoek (External examiner, University Utrecht)
Dr Zhang wei (Supervisor, Chang'an University)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Due to the air pollution problem is becoming more and more serious these years, how to measure the air quality is a new topic for research. In Eindhoven, an innovative network is established for real-time measuring air quality through the whole city. To evaluate the spatial data quality, spatial data analysis is needed in this research.

For this research, the data used comes from the ILM, the first week of October 2014, especially October 6th. R is the main software to storage, deal with and display the data. After pre-processing of the data, descriptive statistics is used as the overall introduction to this study. Afterward regression analysis leads to the relationship between different variables. Analysis the spatial and temporal variability then make the prediction map in order to visualize the air quality data.

Results of spatial and temporal variability of specific air pollutant matter provide the reliable of measurement network and develop a protocol for spatial data quality. Furthermore, the study identify the appropriate temporal scale for modelling and mapping.

ACKNOWLEDGEMENTS

I would like to take this chance to thank for all people who supported me during the 18-months longs MSc study period.

To my supervisors Dr. Nicholas Hamm and Prof.Dr.Ir. Afred Stein, I would like to express my sincere gratitude for the technique support and continues suggestion. It was my honour and pleasure to work with them.

To my friends, the girls live in 7th floor in ITC hotel, thanks for accompany and sharing during the study period.

To my family, thanks for the endless love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and problem statement

Air pollution refers to harmful material that enters the earth's atmosphere. It may cause respiratory diseases, cardiovascular disease, cancer or even lead to death to human beings. It also damages the living environment and thus affects both human well-being and the quality of natural resources.

The atmosphere of the earth is a surrounding layer that contains several gases. A tiny change of some specific component will affect human health. It may cause many problems. One of the most serious problems we facing today is the ozone depletion (CAMBRIDGE, 1998), but in this project, more attention is paid on air quality in city level. During this case, there is a higher need to consider fine particles.

Indoor and outdoor urban pollution are considered to be the two of the world's worst poisonous pollution problems. According to the report of 2014 WHO, the death of about 7 million people's linked to air pollution every year (WHO, 2014). The major hazard of air pollution is as follows:

1. There has been shown a strong relationship between the health problem with aerosol concentration in the environment (Křůmal, Mikuška, & Večeřa, 2013). If the concentration of the pollutant is high, it may cause acute illness. Even if the density is slightly less high, chronic diseases like bronchitis, asthma, and lung cancer are common if people inhale in such air during many years.

2. Air pollution damages plants. Specifically, chlorine dioxide and fluoride influence the plant seriously (Han, Xu, Wei, Shi, & Ma, 2013).

3. Third, air pollution affects the climate. This includes reducing the amount of solar radiation that reaches the ground, increasing precipitation in some areas and the occurrence of acid rain (Fernando, Klaić, & McCulley, 2012).

This research focuses on urban environmental pollution. It is the emission that contains various pollutants during the city's life. It is a pollution, if its level exceeds the self-purification capacity of natural environment. In that case, it will damage the ecological balance and may also cause health problems.

In my country, China, air pollution is a serious problem since 1970's. Following the economic development and the increase of industrial activity, the pollution caused by human activities became progressively worse. Especially in the east of China, the increase of coal and wood burning for heating system in winter cause haze weather over a wide area. Even worse, the east of China is the main collection of cities, containing one third of the total population. For these reasons, the main purpose for me to do this research is to find a helpful methodology to analyse this problem and to help solving it. .

In Eindhoven, a new monitoring system has been installed to provide information about air pollutant (Gurp, 2014)by AiREAS. This system is focused on finding the relationship between human health and fine particles as well as coarse fractioning urban environments. It provide direct display on their official webpage for citizens to get the real time data every ten minutes. Spatial analysis is an important method to do this. After the data are collected, there is still a long way for dealing and analysis the fine particles. Making use of this combined data will lead to the result of this topic.

Spatial data quality evaluation of measurement network is the principle motivation of the research. The final output will be a map that shows the spatial distribution of particulate matter air pollution. In order to do this, data quality aspects need to be addressed, e.g. to the fit of the regression-based model (Briggs et al., 2000). This would be an important methods after there-processing of the data.

1.2. Research identitication

1.2.1. Overall research objective

The main objective is to evaluate the spatial data quality of the Eindhoven innovative air quality measurement network and to provide recommendations on an SDQ protocol.

1.2.2. Specific research objective

1. Exploratory spatial data quality analysis and development of a SDQ protocol.
2. Estimate the number of measurements that are necessary to represent spatial-temporal variability
3. Identify the appropriate temporal scale for modelling and mapping.

1.2.3. Research question

Table 1 Research Question that link to specific research objective

1 Exploratory spatial data quality analysis and development of a SDQ protocol	1.1 What is the distribution of the data considering basic descriptive statistics, time and space? What is the relationship between the different variables (e.g., PM1 vs. PM2.5)? What are the ratio's between different variables (e.g., PM2.5 / PM10)?
	1.2 How does the temporal distribution of specific pollutants vary between sites? Are they systematically lower or higher between sites? Are they comparable?
	1.3 What is the spatial distribution of the data at different time epochs? Are the concentrations influenced by different types of locations?
	1.4 How do the above questions lead to a protocol for SDQ evaluation?
2 Estimate the number of measurements that are necessary to representing spatial-temporal variability	2.1 How well does a single site represent temporal variability in air quality?
	2.2 What is the optimal number of observations, balancing between the cost of sensor and the precision of the modelling?
3 Identify the appropriate temporal scale for modelling and mapping	3.1 What is the appropriate temporal scale for mapping
	3.2 Which area have higher concentration changes and why?

1.3. Innovation aimed at

The research provides an innovation example for city level air pollution observation. It defines spatial data quality criteria and an optimal sampling strategy for a low-cost air pollution sensor network.

1.4. Thesis Structure

This thesis is divided into seven chapters to achieve the objectives of the research.

Chapter 1 makes the overall introduction of the thesis including the motivation of the research, the objectives and the questions.
Chapter 2 provides the literature review that both theoretic background and the method that are relevant to the research.
Chapter 3 describes the study area and original dataset used in the research.
Chapter 4 explains the methodology in detail.
Chapter 5 provides the result and analysis.
Chapter 6 the discussion and limitation of the research has been carried out to answer the objective question.
Chapter 7, the last chapter, the conclusion and the recommendation of the research has been presented.

2. LITERATURE REVIEW

2.1. Data related work

Air pollutants which is harmful to human health when exposed to it (WHO, 2003),has become one of the most dangerous effect factors to human health from environment(WHO, 2014).European commission has established a standard for air quality (European Commission, 2013), including various pollutants though different exposure time. The annual limit for PM2.5 and PM10 is 25μg/m^3 and 40μg/m^3, respectively. Meanwhile, PM10 also has the daily limit for 50μg/m^3. However, some particles like PM1 (Křůmal et al., 2013) and UPF that are believed to be much effective to human health, are not involved in this standard.

To measure the air quality and find the relationship between air quality and human health, a new monitoring system has been installed to provide information in Eindhoven (AiREAS, 2014a). As a brand new system, it has seldom research has been depended on it. One of them provides fundamental spatial data quality of this system (Gurp, 2014). In addition, multiple elements are utilized for determine the quality of data, including lineage, positional accuracy, attribute accuracy, logical consistency, completeness, etc.(Oort, 2006)

ILM is established by AiREAS, which stands for a clean city to encourage its citizens, businesses, academic institutions and government all work together to reach this goal. The official website of AiREAS is serve as an important source (AiREAS, 2014a), which contains the study area and basic description of measurements. Besides, all the weather data tend to affect the air quality (KNMI, 2014). Thus, not only the temperature and humidity, but also plenty of weather data are provided.

Based on the study of previous literature, this research focuses on the PM1, PM2.5 and PM10, and closely connects with it to city level lives.

2.2. Modelling and analysis

The objective of this research is to do the analysis for spatial data quality. To reach this goal, spatial and temporal variability are taken into consideration (Heuvelink & Webster, 2001). Actually, previous studies have been working on the geostatistics model (Buttafuoco, Castrignanò, Busoni, & Dimase, 2005) mostly in soil or water resources. However, regarding to measurement and analysis of city level air quality, seldom research has focused on it (Johansson, Norman, & Gidhagen, 2007)

Webster & Oliver, (2007) described the basic principle of geostatistics, which is the fundamental of environmental scientists. Unlike the traditional statistics works on the similar sample space, the distribution of environment elements, such as weather, soil, plant and air pollution vary from place to place. Thus, geostatistics technology plays an important role in dealing with spatial variance and spatial structure. In this research, geostatistics is the most important tool.

Temporal variations of atmospheric aerosol have already been studied in some European areas (Lianou et al., 2011). Although only the concentrations of PM10 and PM2.5 have been measured for one and a half year in urban areas, this paper focus on the analysis to estimate the spatial and temporal variability of four main cities across Europe. Some methods which have been applied to the analysis is also suitable for this research. They provide quantitative and qualitative analyses for PM2.5 and PM10 according to the regional environment and policies. Meanwhile, the ratio between PM2.5 and PM10 is also involved.

The spatial variability about concentrations of PM2.5 and PM10 in the United States has been evaluated (Li, Wiedinmyer, & Hannigan, 2013). PM10 concentrations demonstrate strong spatial variability, and the ratio between PM2.5 and PM10 shows different spatial patterns depending on the locations. The season, weather, and the time of day affect the spatial patterns.

Once the spatial and temporal models have been developed, the proper software that is needed to deal with and post these models. R is the main software in this research. Packages *gstat* with *rgdal* is applied in the analysis and make prediction based on kriging method. A book has been developed to introduce the how to handle with R (Bivand, Roger S., Pebesma, Edzer J., Gómez-Rubio, n.d.), as well as visualize and analyse the spatial data.

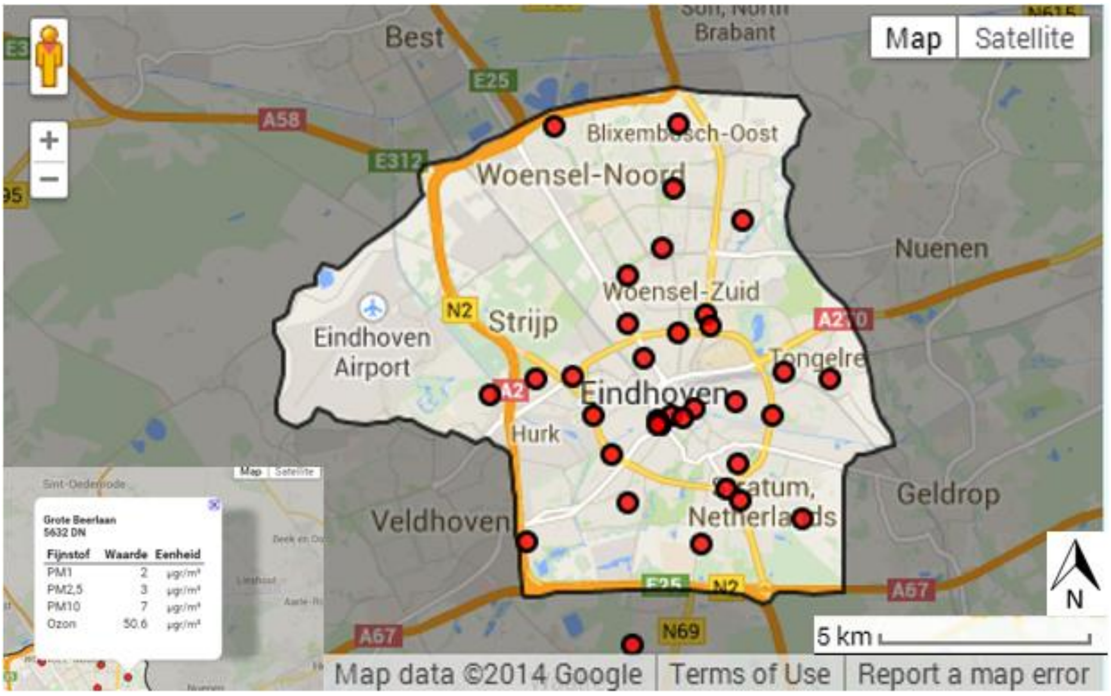
Spatial data analysis cannot be directly seen from the dataset. To make the hypothesis and generate the result of measurement data, spatial data analysis which uses R is considered as one of the best methods (Bivand, Roger S., Pebesma, Edzer J., Gómez-Rubio, n.d.). This book not only introduces the background about the relationship between R and Geoinformatics, but also includes the types of spatial data, the storage and display of data, as well as the usage of R for spatial data analysis. Besides, this book provides some exercises on R with code. It is a very beneficial to the beginners of R software.

A blogger, called R graph gallery, collects a series of examples with data and output in R for data analysis (RGraphGallery, 2013). Different authors work on it and share the function to people who is interested in analysis and display in R. It contains a great amount of articles with a clear title for search. One of the most important example for spatial data in R is about how to enhance the visualization of data (Bivand, 2011).

The spatial interpolation on the Netherlands (Hengl, 2010), introduces the basic geostatistical analysis on geographic data. This article shows the basic step for interpolation and prediction in different areas. It contains code and result to make readers have a better understanding during participation. Meanwhile, the projection coordinate transformation of Netherlands has been completed.

3. STUDY AREA AND DATA DESCRIPTION

The research has been working in Eindhoven, one of the main cities in the south of Netherlands. An innovative air measurement system (ILM) has been built in Eindhoven by AiREAS (AiREAS, 2014a) so as to provide information about air quality. The measurement system has dozens of Airboxes to measure pollutant matter since 2013, and presents the observation on the webpages to let citizen recognise the real time pollutants. Not only the common particle matters, such as PM1, PM2.5 and PM10, but also the ozone and ultrafine particles (UFPs) are measured. This harmful substance can be visualized through the Airboxes on the webpage.



Grote Beerlaan
5632 DN

Fijnstof	Waarde	Eenheid
PM1	2	µgr/m³
PM2,5	3	µgr/m³
PM10	7	µgr/m³
Ozon	50.6	µgr/m³

Figure 3-1 Study area that is given by AiREAS on their official webpage with locations and a small example to show how it works

Figure 3-1 shows the study area that is given by AiREAS on their official website. The area with higher brightness refers to the main city of Eindhoven. Besides, the red points indicate to the location of Airboxes. The small windows is the screenshots when clicking each red point on the map.

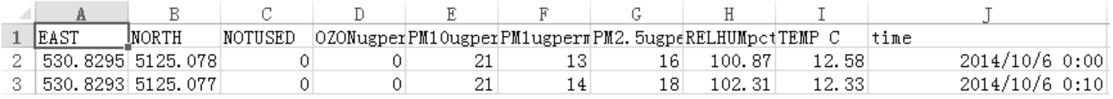
The base map is provided by Google Earth, with scale and compass on it.

3.1. Data description

The data are collected every ten minutes since 2013. However, they are not always collected at the same time during each day and each sensors.The web server (AiREAS, 2014b) provides two original formats of data, including HDF5 and CSV.

HDF5 format collects the data for each day, which means a single files everyday.

CSV format contains the data for each sensor, which means there are over 40 files depending on the different sensors.



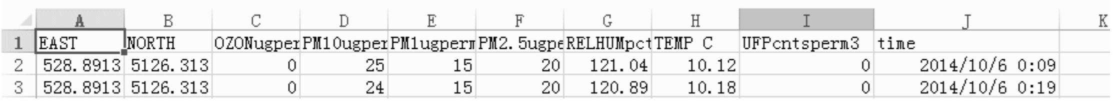
	A	B	C	D	E	F	G	H	I	J
1	EAST	NORTH	NOTUSED	OZONugper	PM10ugper	PM1ugperm	PM2.5ugpe	RELHUMpct	TEMP C	time
2	530.8295	5125.078	0	0	21	13	16	100.87	12.58	2014/10/6 0:00
3	530.8293	5125.077	0	0	21	14	18	102.31	12.33	2014/10/6 0:10

Figure3-2 Screenshot of CSV data format that is extracted from original dataset from sensor number 1 on October 6th, 2014

The means of each header are given as below:

EAST is the transformed longitude measured in degree and, decimal minutes,
NORTH is the transformed latitude measured in degree and, decimal minutes,
NOTUSED is the not used column that is blank in most tables of sensors,
OZONugperm3 is the ozone measured in µg/m^{-3}
PM10ugperm3 is the particulate matter that size 10 micrometres that measured in µg/m^{3}
PM1ugperm3 is the particulate matter that size 1 micrometres that measured in µg/m^{3}
PM2.5ugperm3 is the particulate matter that size 2.5 micrometres that measured in µg/m^{3}
RELHUMpct is the relative humidity measured in percentages
TEMP C is the temperature measured in degrees Celsius
Time is the time measured.

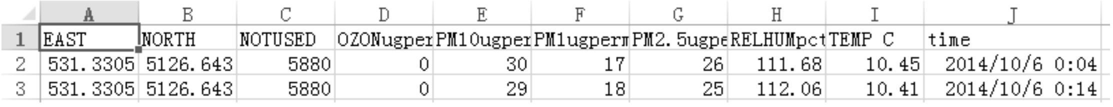
Two of the total sensors, which are Sensor number 26 and Sensor number 35, have a column called UPFcntsperm3. It seems to measure UFPs, but all of the values equal to zero. This is a blank column so that probably concerns to NOTUSED.



	A	B	C	D	E	F	G	H	I	J	K
1	EAST	NORTH	OZONugper	PM10ugper	PM1ugperm	PM2.5ugpe	RELHUMpct	TEMP C	UFPcntsperm3	time	
2	528.8913	5126.313	0	25	15	20	121.04	10.12	0	2014/10/6 0:09	
3	528.8913	5126.313	0	24	15	20	120.89	10.18	0	2014/10/6 0:19	

Figure 3-3 The screenshot of the header of Sensor 26 on October 6th

Five of the total sensors, which are sensor number 21, 22, 25, 30 and 36, are not blank in the column NOTUSED in Figure 3-4. Since there should be five sensors measured in UFPs, I suppose these values are the measurements of UFPs. Because on the official webpage of AiREAS, five of the measurements provide the values of UFPs which should be found from the original dataset. After checking the postcode and link the address to the location in the original dataset, these five sensors are truly the sensor which measure the UFPs. This can be seen from Appendix-1.

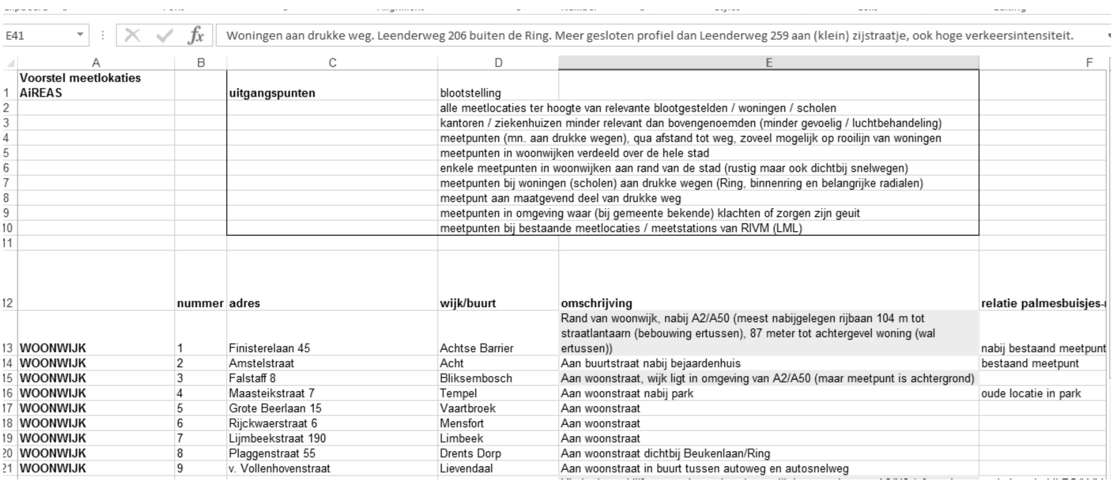


	A	B	C	D	E	F	G	H	I	J
1	EAST	NORTH	NOTUSED	OZONugper	PM10ugper	PM1ugperm	PM2.5ugpe	RELHUMpct	TEMP C	time
2	531.3305	5126.643	5880	0	30	17	26	111.68	10.45	2014/10/6 0:04
3	531.3305	5126.643	5880	0	29	18	25	112.06	10.41	2014/10/6 0:14

Figure 3-4 The screenshot of the header of Sensor 21 header on October 6th

3.2. Address description

There are three main documents to describe the spatial distribution of the sensors, including the description of location when established the sensors, the coordinate extract from the dataset, as well as the address and postcode given by AiREAS on the webpage.



E41 | Woningen aan drukke weg. Leenderweg 206 buiten de Ring. Meer gesloten profiel dan Leenderweg 259 aan (klein) zijstraatje, ook hoge verkeersintensiteit.

	A	B	C	D	E	F
1	**Voorstel meetlokaties AiREAS**		**uitgangspunten**	blootstelling		
2				alle meetlocaties ter hoogte van relevante blootgestelden / woningen / scholen		
3				kantoren / ziekenhuizen minder relevant dan bovengenoemden (minder gevoelig / luchtbehandeling)		
4				meetpunten (mn. aan drukke wegen), qua afstand tot weg, zoveel mogelijk op rooilijn van woningen		
5				meetpunten in woonwijken verdeeld over de hele stad		
6				enkele meetpunten in woonwijken aan rand van de stad (rustig maar ook dichtbij snelwegen)		
7				meetpunten bij woningen (scholen) aan drukke wegen (Ring, binnenring en belangrijke radialen)		
8				meetpunt aan maatgevend deel van drukke weg		
9				meetpunten in omgeving waar (bij gemeente bekende) klachten of zorgen zijn geuit		
10				meetpunten bij bestaande meetlocaties / meetstations van RIVM (LML)		
11						
12		**nummer**	**adres**	**wijk/buurt**	**omschrijving**	**relatie palmesbuisjes-**
13	**WOONWIJK**	1	Finisterelaan 45	Achtse Barrier	Rand van woonwijk, nabij A2/A50 (meest nabijgelegen rijbaan 104 m tot straatlantaarn (bebouwing ertussen), 87 meter tot achtergevel woning (wal ertussen))	nabij bestaand meetpunt
14	**WOONWIJK**	2	Amstelstraat	Acht	Aan buurtstraat nabij bejaardenhuis	bestaand meetpunt
15	**WOONWIJK**	3	Falstaff 8	Bliksembosch	Aan woonstraat, wijk ligt in omgeving van A2/A50 (maar meetpunt is achtergrond)	
16	**WOONWIJK**	4	Maasteikstraat 7	Tempel	Aan woonstraat nabij park	oude locatie in park
17	**WOONWIJK**	5	Grote Beerlaan 15	Vaartbroek	Aan woonstraat	
18	**WOONWIJK**	6	Rijckwaerstraat 6	Mensfort	Aan woonstraat	
19	**WOONWIJK**	7	Lijmbeekstraat 190	Limbeek	Aan woonstraat	
20	**WOONWIJK**	8	Plaggenstraat 55	Drents Dorp	Aan woonstraat dichtbij Beukenlaan/Ring	
21	**WOONWIJK**	9	v. Vollenhovenstraat	Lievendaal	Aan woonstraat in buurt tussen autoweg en autosnelweg	

Figure 3-3-5 Describe the definition of the location

Figure 3-5 shows the definition of each location of the sensors, which can be linked to the current location by address. This definition divides the total sensors into four types, three of them are shown in Figure 3-5, including the busy road, the residential area and the others. However, not all the sensors are concluded in this definition and some of the location have changed during the time. Besides, there is one sensor located away from the downtown city, and can be realized as the background area.

The address and postcode linked to the sensor number and coordinate can be seen from Appendix-1.

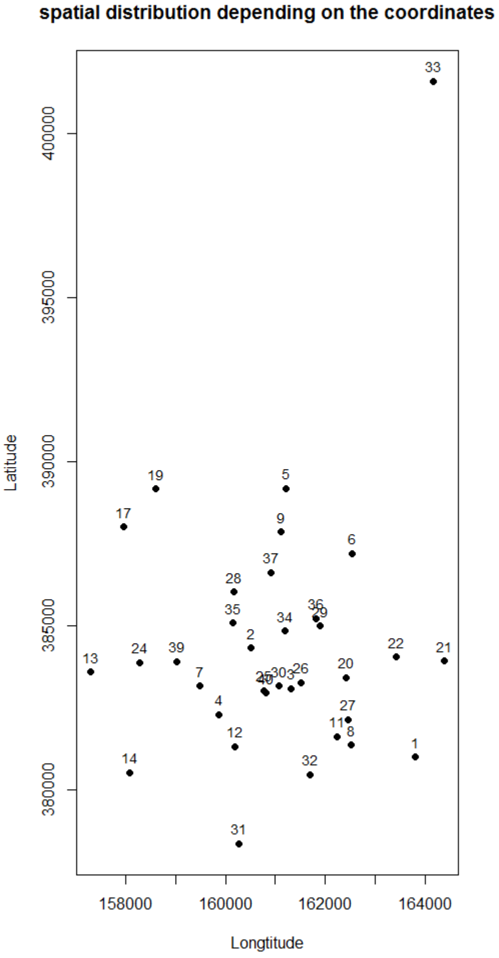


Figure 3-6 Spatial distribution of the observation sites with sensor number on it.

Figure 3-6 is the spatial distribution of the observation sites which depends on the location from each sensor on October 6th, 2014. The coordinate has been transformed into the Dutch Coordinate Reference System (Amersfoort /RD New).

According to the description, the total sensors can be divided into four parts which are $\mathrm{S} = \{s_1, s_{2,}s_3, s_4\}$.

$s_1 = \{1, 2, 5, 6, 12, 13, 14, 17, 20, 21, 24, 27, 30,32\}$
$s_2 = \{3, 4, 7, 8, 11, 24, 25, 26, 34, 35, 36,39\}$
$s_3 = \{22, 29, 32, 37,40\}$
$s_4 = \{33\}$

S_1 is considered the residential area, while S_2 refers to busy road. S_3 contains the definition "other" and the point without any description. S_4, which is a little away from the city downtown is consider as the background.

4. METHODOLOGY

4.1. Framework of methodology

The framework of the methodology is provided in Figure 4-1 as below. There are five main phases in the research, including data pre-processing, spatial data quality analysis 1, represent spatial and temporal variability, appropriate scale for mapping and spatial data quality analysis 2.

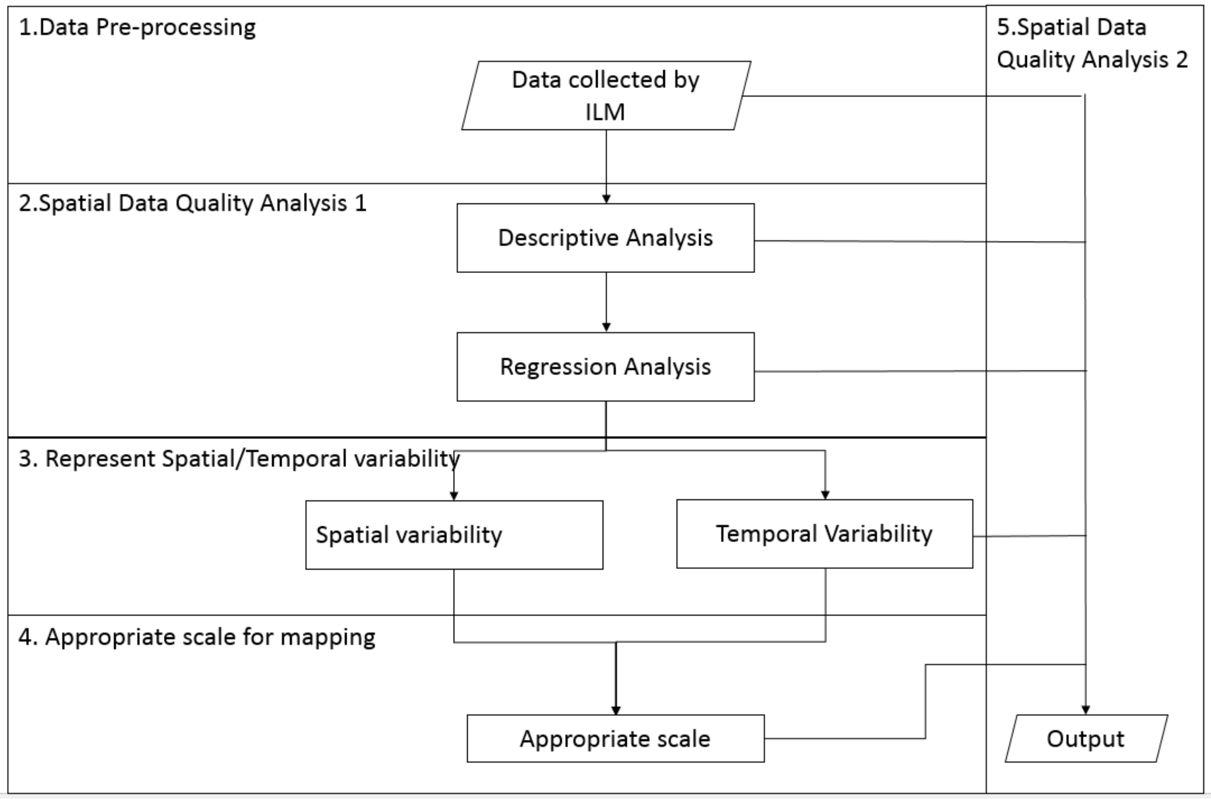


Figure 4-1 Framework of Methodology

4.2. Data pre-processing

Before the data analysis, all the datasets need to be prepared in a convenient and complete way. For further analysis, the original data need to be processed in a good structure, while the missing data should be filled and the outlier should be removed.

The outlier is an observation point that has much distance from the other observations (Hannes, Reuter, & Gerboles, 2013). It may be caused by the variability or error during the measurements. Outlier detection is important for real-time data collection (Zhang et al., 2012). In this research, outliers can be detected using boxplots. The normal default outer fences are $\mathrm{Q1} - 3\mathrm{IQR}$ and $\mathrm{Q1} + 3\mathrm{IQR}$ (Dawson, 2010), which refers to the values above $\mathrm{Q1} + 3\mathrm{IQR}$ or below $\mathrm{Q1} - 3\mathrm{IQR}$. Here, $\mathrm{Q1}$ is the first quartile, and IQR is the third quartile minuses the first quartile.

4.3. Descriptive statistics

Descriptive statistics (Maria, 2000) is used to describe the overall information about the data quantitatively, which provide simple summaries about the sample. Descriptive statistics is distinguished from inferential statistics that make conclusions beyond the data or with any hypothesis. This means descriptive statistics cannot be developed on the basic of probability theory.

Descriptive statistics mainly focuses on univariate analysis that involves the description of single variable over a period of time. The major characteristics of the single variable which is mostly used are the distribution, the measure of central tendency and the measure of variability.

The distribution is a summary of the frequency of individual values for a variable. Histogram is the main method. The measures of central tendency include the mean, median and the mode. Besides, the variability is the spread of values around the central tendency, including the standard deviation, the minimum and maximum values of variables, as well as the skewness.

4.4. Regression analysis

Regression analysis (Rawlings, Pantula, & Dickey, 1981) is a statistical method to analyse the data in order to estimate the relationship between two or more variables. It includes the modelling and analysis the variables for observations purpose, especially focusing on the relationship between a dependent variable and one or more independent variables.

Linear regression is the method for modelling this relationship using least square approach, while simple linear regression only focuses on the relationship between a dependent variable and one explanatory variable. The formula is shown as follows:

$$\mathrm{Z} = \mathrm{X}\beta + \varepsilon$$

$$e_i = z_i - \widehat{z_i}$$

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2$$

Where,

Z - response variable,

X - design matrix consist of explanatory variables.

β - corresponding element of intercept which is also called effects or regression coefficients.

ε - error term, which is also called the noise or the residual.

e_i - difference between the value that is predicted by the model and the true value.

z_i - true value of response variable,

$\widehat{z_i}$ - value of response variable which is predicted by the model

SSE - the sum of squared errors.

Regression analysis is also utilised to test and estimate the concentration of pollutants (Briggs et al., 2000). The p-value with regression coefficients (intercept and slope) and standard errors of the estimates (SSE) are all taken into consideration.

4.5. Spatial and temporal variability

4.5.1. Spatial variability

In terms of spatial variability, when measure the variable in different spatial locations at the same time (Heuvelink & Webster, 2001), the exhibit values are different across the locations. The spatial process has a mean $\mathrm{E}[\mathrm{Z}(\mathrm{u})] = \mu(\mathrm{u})$ and the variance of $\mathrm{Z}(\mathrm{u})$ exists.

The bubble map is a simple way to visualize the spatial variability. Bubble map is a function in R that uses the circle to exhibit both the location and the values of the variable. The size of circle shows the range of the variable. The bigger the circles, the larger the values.

4.5.2. Temporal variability

Temporal variability measures the difference during a period of time under the same location (Lianou et al., 2011). For better visualization, the flow line of the whole day and the boxplot for hourly data are used in this research. Connecting the values according to the time period can provide the trend of values, such as increasing and decreasing of the trend. Besides, the boxplot for hourly data tends to reduce the noise during measuring.

4.6. Spatial variogram and fiting the variogram

The main method to model the spatial air quality data is to use the variogram (Hannes et al., 2013). Variogram is a tool to describe the degree of spatial dependence in a spatial random field and then model it. It can be defined as the variance of the difference between field values in two dimensions(x and y) across the fields. The value of a property is called Z, at any place called u. An intrinsic random variable can be donated as$\mathrm{Z(u)}$.

$$\mathrm{Z(u)} = \mu + \mathrm{S(u)} + \varepsilon$$

$\mathrm{S(u)}$ is a location related error term while ε is a random error term, which is also called the noise, so the

$$\mathrm{E}\big(\mathrm{Z(u)}\big) = \mu$$

Besides, the variogram does not increase without the boundary, and the intrinsic hypothesis assumes that there should be stationary of the increment. Hence,

$$\mathrm{E}[\mathrm{Z(u)} - \mathrm{Z(u+h)}] = 0$$

$$2\gamma(\mathrm{u}) = \mathrm{E}[\mathrm{Z(u)} - \mathrm{Z(u+h)}]^2 = \mathrm{var}[\mathrm{Z(u)} - \mathrm{Z(u+h)}]$$

Where,

$\gamma(\mathrm{u})$ - Variogram

var - Variance

E - Expectation

μ - mean

u - location of all possible points,

h - lag distance between two measured points

Here, the variogram for h is defined as average squared difference of value that is separated by h

$$2\gamma(\mathrm{h}) = \frac{1}{N(h)}\sum_{N(h)}[Z(u) - Z(u+h)]^2$$

Where $\mathrm{N(h)}$refers to the number of pairs of measurements.

4.7. Prediction map

After the variogram has been set, ordinary kriging is used for interpolation. The kriging (BLUP) is the best linear unbiased estimator, which means it has the lowest variance amongst all linear unbiased predictors.

The kriging prediction and kriging variance map should be made.

The constant and unknown spatial mean need to be estimated in ordinary kriging. In ordinary kriging, the predictor function is shown as follows:

$$\hat{T} = \hat{\mu} + c_0^T C^{-1}(z - \hat{\mu}1)$$

Where,

T is a linear predictor, and to take the best predictor $\mathrm{var}(\hat{T} - T)$

$\hat{\mu}$ is the estimated spatial mean

c_0 is a vector giving the covariance between all observations and across the locations

C is the n*n covariance matrix

z is the vector of observations of length n.

1 is a vector of 1's of length n

The formulate can be written using the kriging weights

$$\hat{T} = \sum \lambda_i Z(u_i)$$

4.8. Cross Validation

To evaluate the accuracy of model that fits the variogram, cross validation is the main method (Kohavi, 1995). As the estimated model cannot be better than the real observed model, cross validation can help find the best estimated model.

Cross validation leaves one of the data-point out, and re-calculates the model to make prediction at that point. The mean error (ME) and the root mean square error (RMSE) are mentioned in this research.

4.9. Software

R refers to a language and software environment that for statistical computing and graphics. It is commonly used in analysis to solve a series of statistical problem, including linear modelling, time-series analysis, graphical techniques etc. R is an open source software project that allows inspection and modification so that everyone that can see how the methods and algorithms work.

Furthermore, R provides the platform for users to submit packages for specific application. In this research,*sp, gstat* and *rgdal* are three main important packages.

sp provide sthe classes and methods for spatial data. It is a basic packages for spatial data analysis.

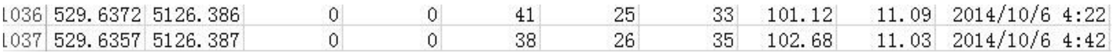
gstat is a packages focusing on modelling , prediction and simulation of Geostatistical Spatial and Spatio-Temporal variability.

rgdal is the binding for Geospatial Data Abstraction Library.

5. RESULT AND ANALYSIS

5.1. Data pre-processing

In the original dataset, some data are missing that make the amount of data less than expected. This can be seen from Figure 5-1.



L036	529.6372	5126.386	0	0	41	25	33	101.12	11.09	2014/10/6 4:22
L037	529.6357	5126.387	0	0	38	26	35	102.68	11.03	2014/10/6 4:42

Figure 5-1 Screenshot of original dataset from Sensor number 20 on October 6th, 2014

In Figure 5-1, the data set from 4:32 is missing. This missing data will cause error when calculating the mean and other statistical values for each hour. After reviewing the values from other sensor of the same time period, this outlier is an individual event but will have great influence in later analysis. However, since the values of other sensors are stable during this period of time, the simplest way to solve this data missing issue is to utilise the value from last time to fill in the blank.

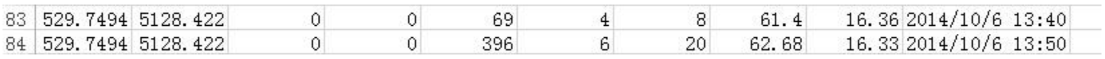
On October 6th, there are total six sets of ten minutes data series missing.

Besides, there are extreme values higher than expected, which are called outlier. The outliers may be caused by the error on sensors when collecting the data. The boxplot of each PM values in total sensors can lead to outlier.

The outliers are the values that stand far away from the mean. In other word, if make a plot of temporal distribution using line chart, the sensor with outlier will lose detail at low values. There is no difference for prediction, that one top value will reduce the changes of fluctuation in smooth area. In the example of Figure 5-2, if the PM10 value 396 is kept, there will be one highlight in the prediction map, but the values from other sensors will not be presented well.

To reduce these bad effect, it is better to replace the outliers by the value of last 20 minutes before the error happen. Only the outliers (PM10 and PM2.5) are changed.

For example, in Figure 5-2, PM10 and PM2.5 are relatively higher.



83	529.7494	5128.422	0	0	69	4	8	61.4	16.36	2014/10/6 13:40
84	529.7494	5128.422	0	0	396	6	20	62.68	16.33	2014/10/6 13:50

Figure 5-2 Screenshot of outliers of Sensor 6 on October 6th, 2014

Besides, the attribute of location type and sensor number are added as the column in the dataset, which will bring more convenience to the further analysis in R.

All available data has been transformed into the Dutch Coordinate System (RD-NEW, Rijksdriehoekstel, EPSG: 28992).

Appendix-2 includes the prepared dataset which has been used in the research.

5.2. Descriptive statistics

This research focuses on the second week of October, 2014. October 6th is the first day for analysis.

5.2.1. Location types analysis

The table of address and coordinate of each sensor are put in Appendix-1.

In Appendix-1, five sensors with UFPs refer to the sensor with values in column NOTUSED. The red word means that sensors are not currently located within Eindhoven.

The description of location provides the definition and type of each sensor in the process of setting the network. It has totally 33 sensors, 4 of them are currently no longer shown on the web of AiREAS.

5.2.2. Descriptive statistics of October 6th,

The first thing needs to do after data collection is to gain an overall understanding of it. Summary statistics is an effective tool to achieve this. Because summary statistics (Table 2) are used to make a summary of all the observations and simplify the information in order to get some basic results.

Table 2 Summary statistics table of all sensors for PM values on October 6th, 2014

2014/10/6	PM1	PM2.5	PM10
minimum	1	2	4
1st quarter	3	3	7
median	5	6	11
mean	10.85	15.65	20.16
3nd quarter	20	30	35
maximum	27	46	57
variance	77.74	196.69	222.87
standard deviation	8.82	14.02	14.92

The means and medians are different, illustrating that PM1, PM2.5 and PM10 values of this day are skew.

In all situations, the values of PM 10 are bigger than PM2.5, and those of PM2.5 is bigger than PM1. As PM2.5 is part of PM10, while PM1 is part of PM2.5.

5.3. Temporal variobility analysis

5.3.1. Temporal variability through the day

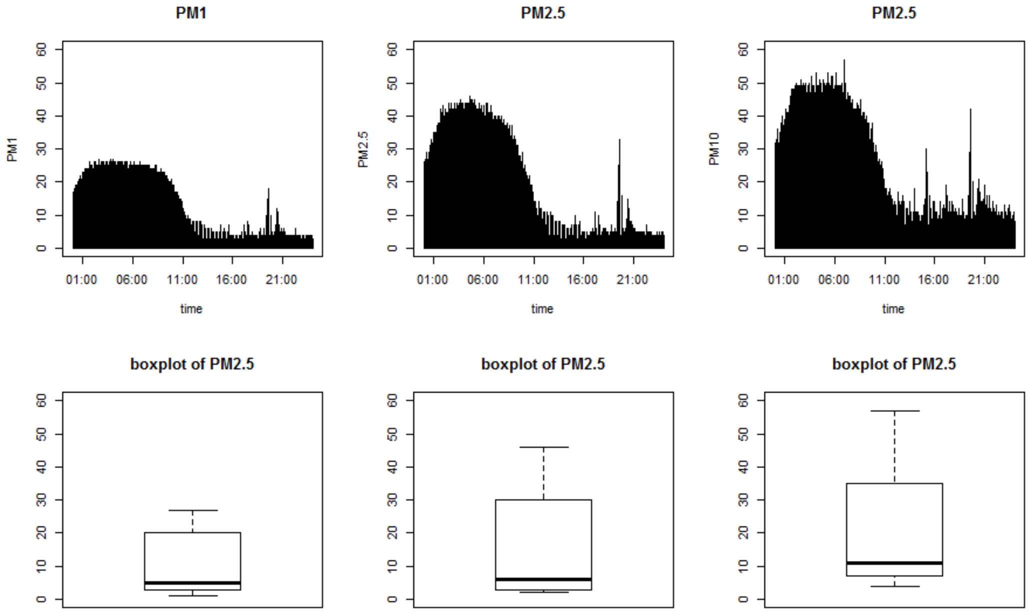


Figure 5-3 Combine the whole data that each plot has 34*6*24=4896 sets of values on October 6th, 2014.

In Figure 5-3, the time plots of concentration against the time of shows the temporal distribution of particulates matter values. It provides an overall trend of PM, PM2.5 and PM10. The fine particles arise from the midnight of the day, and reach the peak at around 6:00. Subsequently, they decrease rapidly until 11:00 am, then maintaining at a stable level. At around 20:00, some specific sensors (2, 3, 8, 11, 22 and 33) show the increases of values, and these sensors will be discussed later in the thesis.

The measurements are collected every 10 minute, but are not always exact in the same time period. However, it may collected on 00:03 or 00:45. Since it combines all sensors, the histograms of each PM value look more than 144 times of collection during the whole day.

The boxplot is another simple way to visualize the basic attribute of the dataset. In comparison with the image as discussed above, it can be clearly shown that the distribution of PM values is skew. .

5.3.2. Temporal variability on hourly data

In order to make the thesis more readable, only the example of each location type for different particles can be shown here and, the other images of single sensor will be put in Appendix-2.

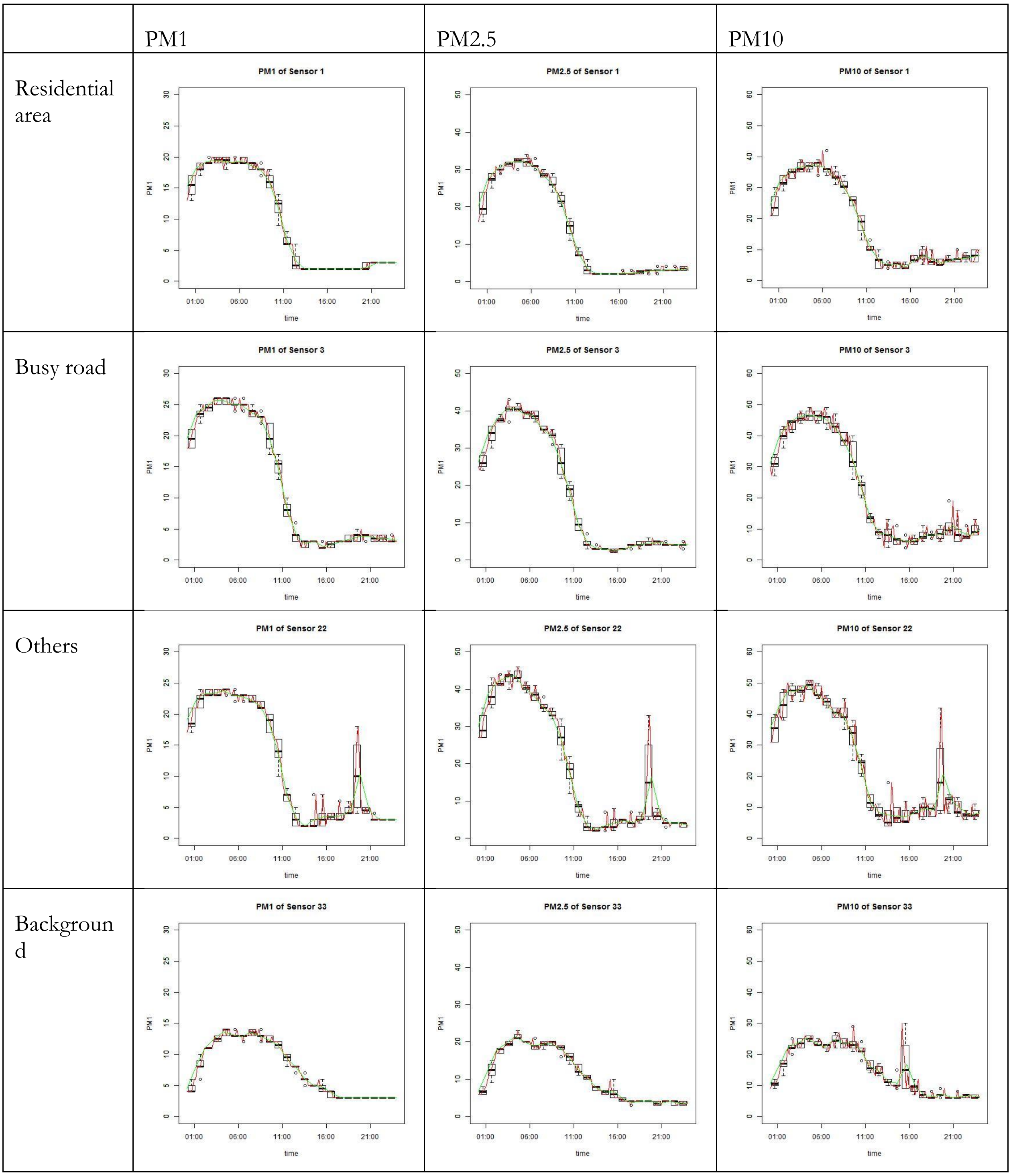


Figure 5-4 Temporal variability of different location types on October 6th, 2014.

Each single image has its own title to show the sensor number and the name of fine particles. The x-axes is the time zone, the y-axes is the concentration of fine particles, and the unit is μg/m^3. The red line is the ten minutes data while the blue line is the hourly data. Thus, the red line has more extreme value to generate the fluctuant.

The green line using the mean that calculated from hourly data and link the mean on each sharp time.

Since the boxplot also depends on the hourly data, the peak of green lines always lies on the means of boxplots.

5.3.3. Relationship and ratio between fine or coarse particles

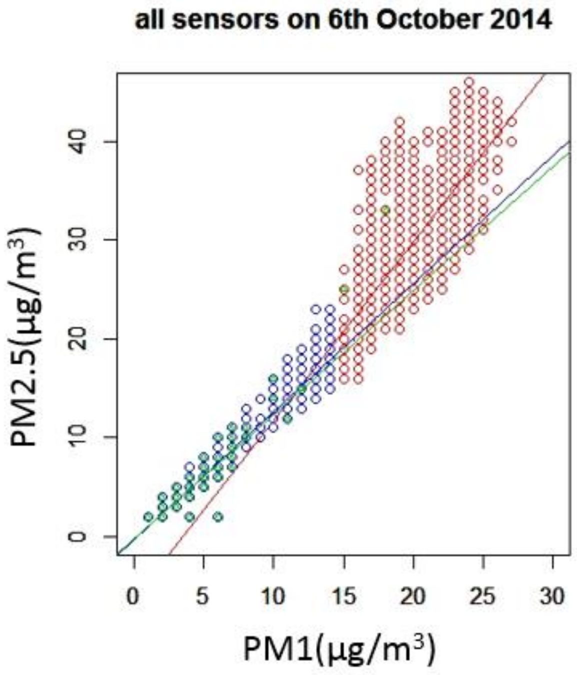


Figure 5-5 Relationship between PM1 & PM2.5

Figure 5-5 shows the regression line of PM1 and PM2.5 in the whole day on October 6th, 2014.

Table 3 Estimate and t-values of intercepts PM1

	Estimate	Standard Error	t-values	p-value	Adjust-R2	Shown colour
PM1<15	1.8	0.7	-9.01	$<2.2*10^{-6}$	0.60	blue
PM1>15	1.29	0.03	-12.61	$<2.2*10^{-6}$	0.96	red
Time before 11:00 am	1.25	0.04	-4.53	$<2.2*10^{-6}$	0.79	green

The blue line is the regression line while PM1 values are smaller than 15μg/m3. The blue points are the measurements showing that the concentration of PM1 less than 15μg/m3.

The red line is the regression line while PM1 values are equal or bigger than 15. The red point shows the measurements.

The green line is the regression line of PM1 values collected before 11:00 am. This data are chosen since both PM1 and PM2.5 show a stable and low values after 11:00 am.

In Figure 5-5 the green line and blue line have only tiny difference, which also can be seen from the similar estimate values in Table 3. Some of the blue points have been overlapped by small green triangle. Then the conclusion can be made, that PM1 values after 11:00 am are nearly totally smaller than 15μg/m3. The two green triangles encircled by red refer to the increase of concentration in rush hour.

Table 3 shows different patterns of relationship in the data collecting after 11:00 am. According to Figure 5-5, it has clearly distinguish between the blue line and red line, which refers to different patterns in the relationship between PM1&PM2.5 values. Besides, the estimate in Table 3 shows this difference as well. Otherwise, the green point in Figure 5-5 locates inside the blue point in most of the time, so the green point collected before 11 am must have different a pattern within the rest of time.

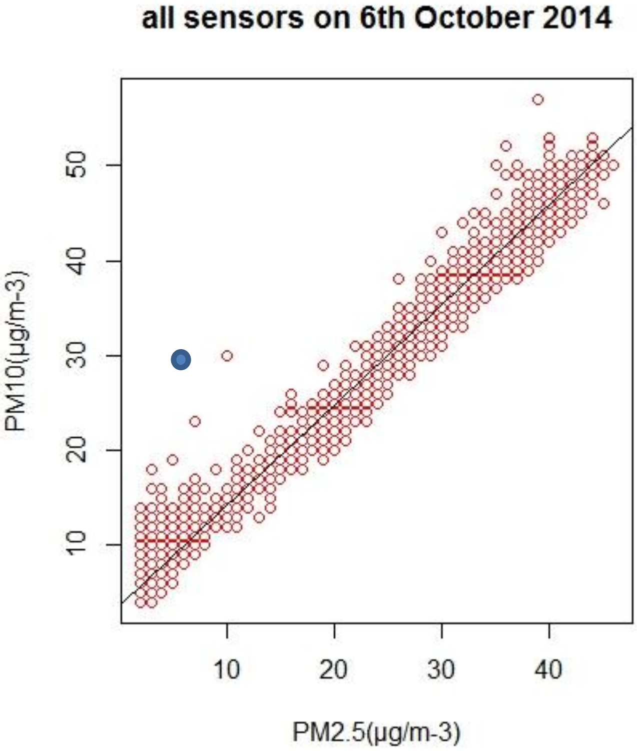


Figure 5-6 Relationship between PM2.5&PM10

Figure 5-6 is the regression line between PM2.5&PM10. The points are located in the narrow area around the line, which means PM2.5&PM10 on October 6th, 2014 has a clear relationship.

There is one outlier shown in the blue point. It is one measurement from sensor number 33. Since it is far away from the downtown city, it may have different air quality pattern.

The ratio between PM2.5&PM10 is important as it can estimate for each one if it already with this ratio, especially when the fine particles are difficult to measure in comparison with coarse particles. In most cities in China and India, only PM10 can be measured. Meanwhile, PM2.5 is tends to be more harmful to people because smaller particle has more probability to go deeper into the lungs (WHO, 2003). Special attention needs to be paid to such particles.

Table 4 Ratio between PM1&PM2.5, PM2.5&PM10, PM1&PM10

6-Oct	PM1/PM2.5	PM1/PM10	PM2.5/PM10
sensor1	0.69	0.53	0.77
sensor2	0.63	0.49	0.77
sensor3	0.7	0.54	0.77
sensor4	0.67	0.52	0.78
sensor5	0.76	0.57	0.75
sensor6	0.58	0.4	0.68
sensor7	0.72	0.58	0.81
sensor8	0.7	0.54	0.77

sensor9	0.8	0.6	0.75
sensor11	0.75	0.57	0.76
sensor12	0.77	0.59	0.77
sensor13	0.72	0.55	0.76
sensor14	0.67	0.52	0.77
sensor17	0.67	0.52	0.78
sensor19	0.63	0.5	0.79
sensor20	0.81	0.61	0.76
sensor21	0.66	0.51	0.78
sensor22	0.63	0.5	0.8
sensor24	0.58	0.46	0.79
sensor25	0.77	0.58	0.76
sensor26	0.7	0.54	0.77
sensor27	0.72	0.55	0.77
sensor28	0.58	0.46	0.79
sensor29	0.65	0.51	0.78
sensor30	0.73	0.56	0.77
sensor31	0.83	0.62	0.74
sensor32	0.69	0.54	0.79
sensor33	0.69	0.51	0.73
sensor34	0.72	0.54	0.76
sensor35	0.76	0.59	0.77
sensor36	0.69	0.55	0.8
sensor37	0.76	0.58	0.76
sensor39	0.67	0.54	0.81
sensor40	0.65	0.54	0.82

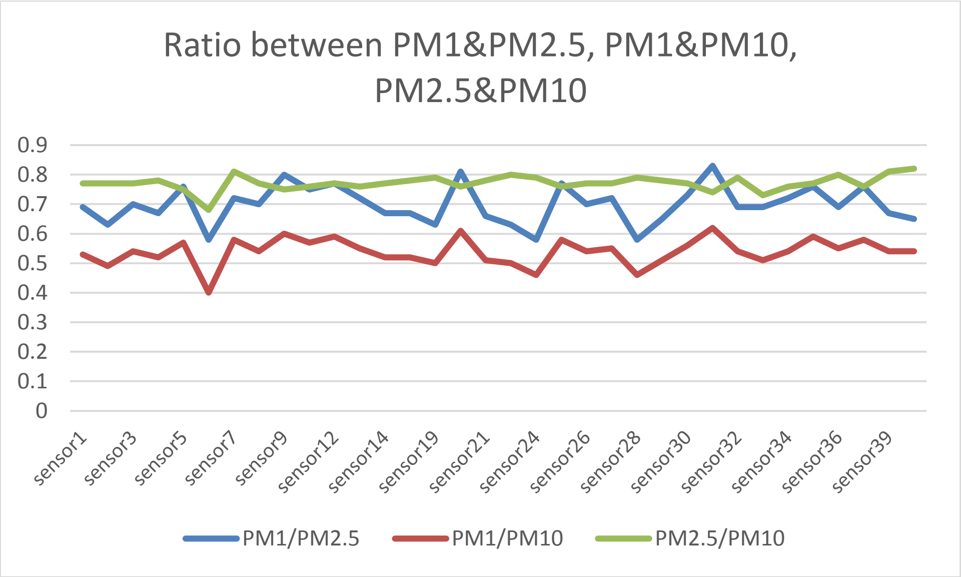


Figure 5-7 Ratio between PM1&PM2.5, PM1&PM10, PM2.5&PM10

Table 4 is the ratio table about the mean of every sensors in the whole day of October 6th, 2014. The ratio ranges from 0.4 to 0.9 in Table 4.

From Table 4 and Figure 5-7, Sensor 6 has all three lowest ratios. According to the aggregation figures from Appendix-2 , the PM1 values collected by this sensors are lower than the other sensors almost at anytime of the day.

The ratio between PM1&PM2.5, PM1&PM10, PM2.5&PM10 may be caused by different types of location in each sensor. For example, in PM1&PM2.5, the mean of residential area is 0.62, and the mean of busy road is 0.71. Because different pollutant source can lead to different ratio between coarse and fine particles(Li et al., 2013).

Sensor numer 33 has the biggest value in the ratio between PM1/PM2.5. This can be seen from Appendix-2 that if PM2.5 values are relatively lower on October 6th 2014 in Eindhoven , the estimated PM1 values will be higher in ratio. The ratio between PM1&PM2.5 tends to increase on sensor number 33.

5.4. Spatial variability analysis

5.4.1. Spatial variability of the whole city

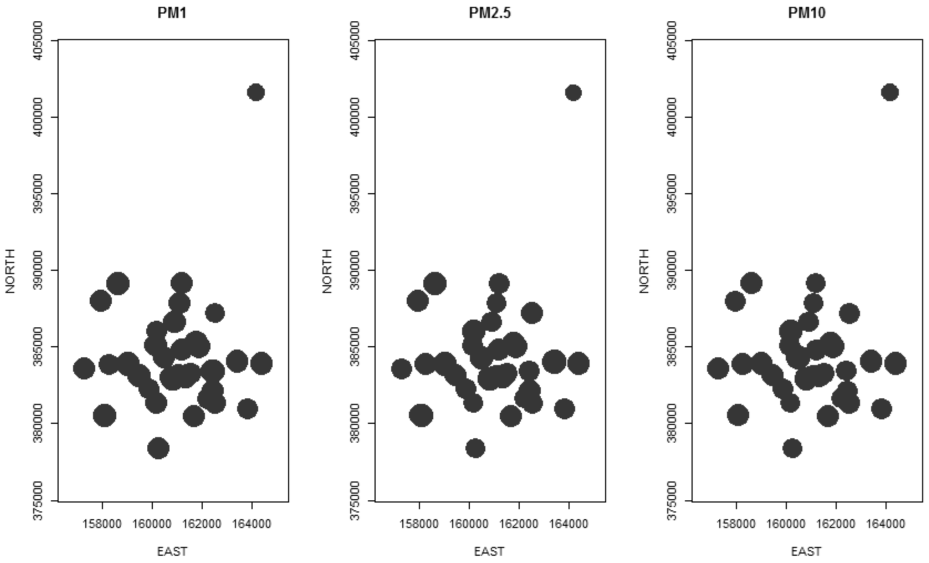


Figure 5-8 Bubble plot to show the location of measurements

Figure 5-8 shows the spatial distribution of all three PM values in the whole day of October 6th, 2014. The sizes of the points show the concentration of PM values.

The point on the right top of the Figure 5-8 is sensor number 33.

5.4.2. Location type analysis

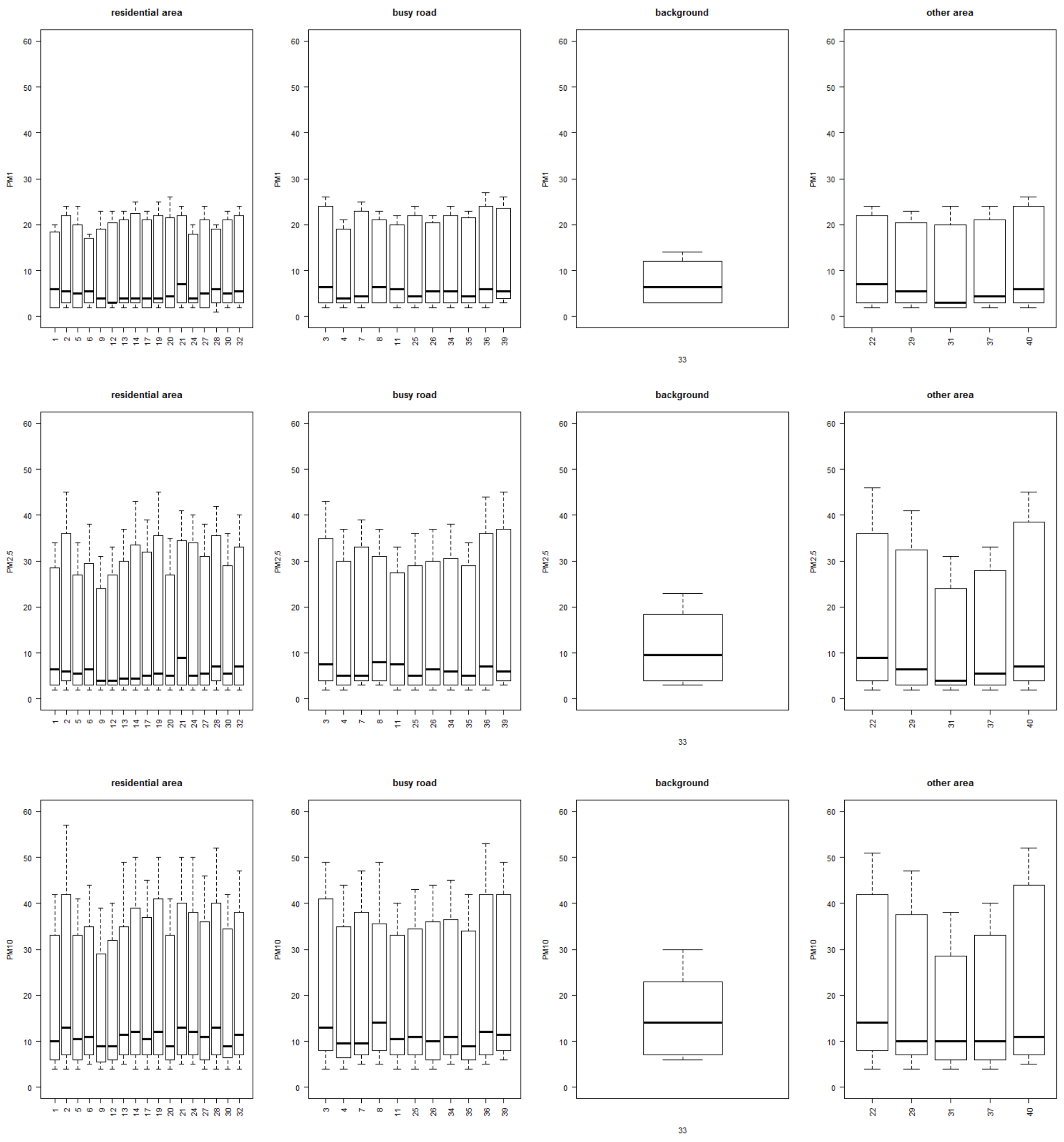


Figure 5-9 boxplot of different types of location on October 6th, 2014, with sensors number

From Figure 5-9, it shows that the specific sensors with high mean do not always have high top values. In this case, Sensor number 33 has the smallest range of values, although it has nearly the highest mean in PM2.5 and PM10. However, Sensor number 2 in residential areas has the top values in PM2.5 and PM10, which is unusual because Sensor number 2 located in a park in residential area.

The boxplots of the other days during that week (October 8th, October 10th and October 11th) will be put in Appendix-4.

Under the same scale of boxplots (from 0 to 60μg/m^{-3}), the high value of other days (from October 7th to October 12th, 2014)is lower than that on Monday, October 6th, 2014. The data change in a small range on the other days rather than on Monday. Furthermore, at the beginning of Monday, the values of PM are much higher than that on the other days.

Each particulate values tends to show higher values on Saturday than on Wednesday. And the minimum of each is also lower on weekdays.

The boxes in Appendix-4 are narrow than those in Figure 5-9. Besides, that the range between the first quarter and third quarter is relatively small. The fine particles cluster around the means, which implies that they have not changed in a big event that cover the whole city. Some specific time periods or extreme values on sensors would not affect this result.

5.5. Prediction map

To make the prediction map, the histograms of the mean value and the variogram should be done before the prediction map for every hour on October 6th, 2014.

Since the histograms of the hourly data are skew, the log-transform is supposed to be used after getting the mean of each measurement, and the back transform is made before plotting the prediction map. However, at this time, after comparing both the variogram and prediction map of no-transform and log-transform, they show seldom difference. The objectivity to use log-transform is to transform the original data into normal distribution. However, most data are so skewed that even with log-transform, so it is better to use the no-transform data to obtain a clearly visualized data.

In every hour, PM1, PM2.5&PM10 are all taken into consideration.

The histogram tests all of three particles, including PM1, PM2.5&PM10. It contains only the data that will be used in the variogram.

The showing coordinate is x ranging from 157000 to 165000 and y ranging from 377000 to 390000 to contain the whole city inside the predicted area. As there is not the base map in grid in city level to hold the kriging inside the counter of city exactly. This working area is approximately close to the cover area of Eindhoven.

All of model used is Spherical model after making the prediction map for hourly data, it turns out that Spherical model can fit most of the situations.

The legend of the prediction map uses grey scale to show the concentration of fine particles. Light area is the low concentration whereas dark area is the high concentration. There is not a same scale for all the prediction map.

The images are showed as follows:

1. One prediction map covers the whole day October 6th, 2014(Figure5-10).
2. Divide the day into four parts(7:00- 9:00 morning rush hour, 9:00-16:00 daytime 16:00-19:00 afternoon rush hour, 19:00-tomorrow midnight Figure5-11 to Figure5-14)
3. Hourly prediction map uses the mean of every hour from that day (7:00-8:00 and 17:00-18:00 as example Figure5-15 to Figure 5-16)
4. Ten minute prediction map during one hour (17:30-17:40 Figure 5-17).

	PM1	PM2.5	PM10
Minimum	7.653	11.01	15.06
First quarter	10.36	14.14	18.69
Median	10.84	15.53	20.06
Mean	10.85	15.65	20.16
Third quarter	11.34	17.34	22.17
maximum	12.54	19.15	24.02

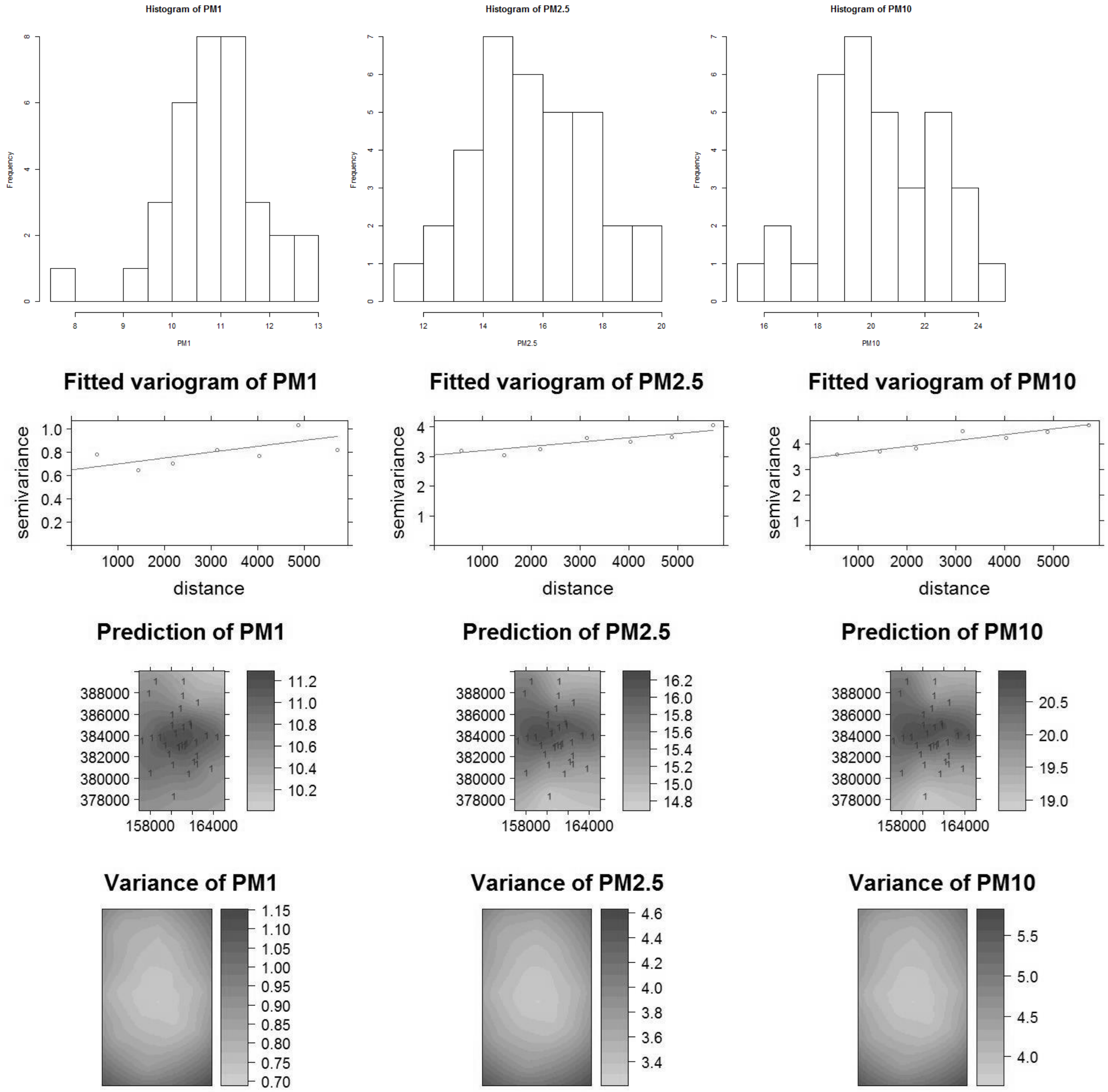


Figure 5-10 Summary statistics, histogram, variogram and prediction map of the whole day, October 6th, 2014.

Table 5 Pairs of measurements

vaPM10			
	np	dist	gamma
1	20	556.94	3.6
2	60	1445.58	3.71
3	90	2185.01	3.81
4	101	3135.96	4.51
5	78	4034.59	4.25
6	68	4885.93	4.47
7	30	5721.71	4.93

	PM1	PM2.5	PM10
Minimum	13.00	19.50	23.00
First quarter	21.10	31.67	37.27
Median	22.50	34.83	41.08
Mean	21.89	34.40	40.74
Third quarter	23.21	38.19	44.77
maximum	25.00	42.50	49.67

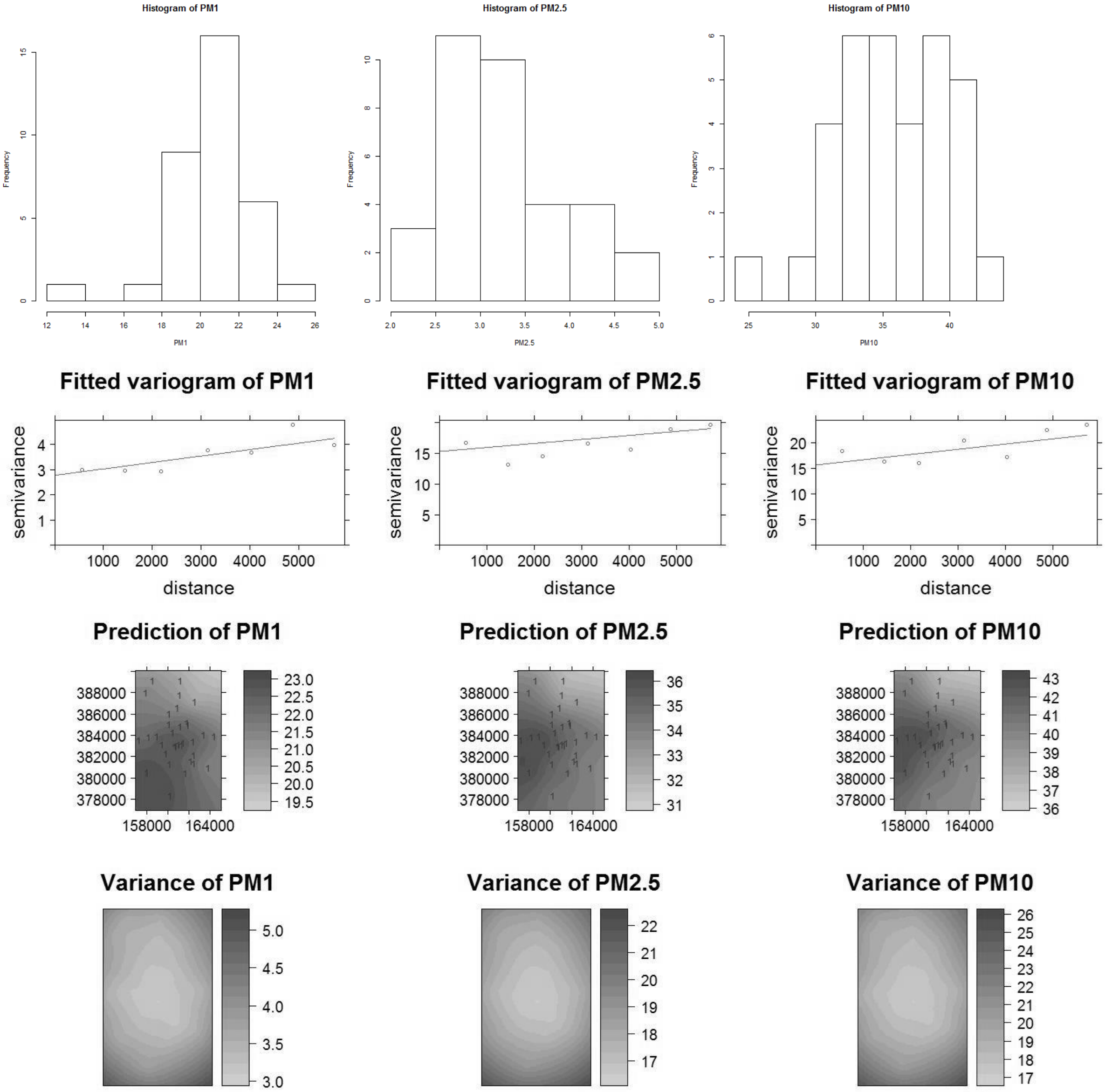


Figure 5-11 Summary statistics, histogram, variogram and prediction map for morning rush hour (7:00-9:00)

	PM1	PM2.5	PM10
Minimum	5.67	6.48	9.74
First quarter	6.46	7.72	11.51
Median	6.75	8.66	12.61
Mean	6.88	8.57	12.59
Third quarter	7.18	9.29	13.73
maximum	8.17	11.10	16.31

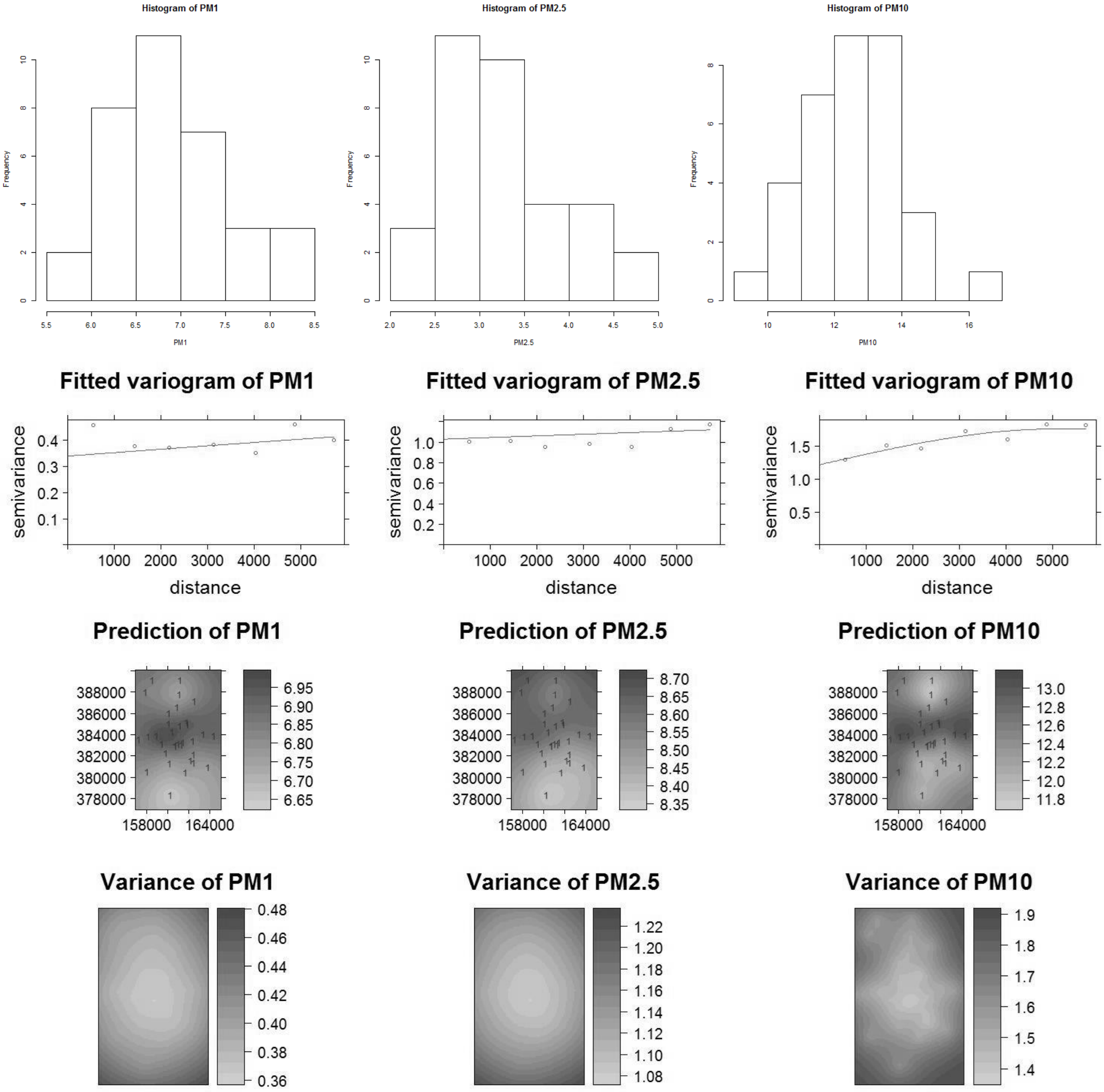


Figure 5-9 Summary statistics, histogram, variogram and prediction map for daytime (9:00-16:00)

	PM1	PM2.5	PM10
Minimum	2.00	2.27	5.22
First quarter	2.40	2.89	6.57
Median	2.67	3.14	6.83
Mean	2.78	3.30	7.02
Third quarter	3.06	3.65	7.47
maximum	4.11	4.83	9.11

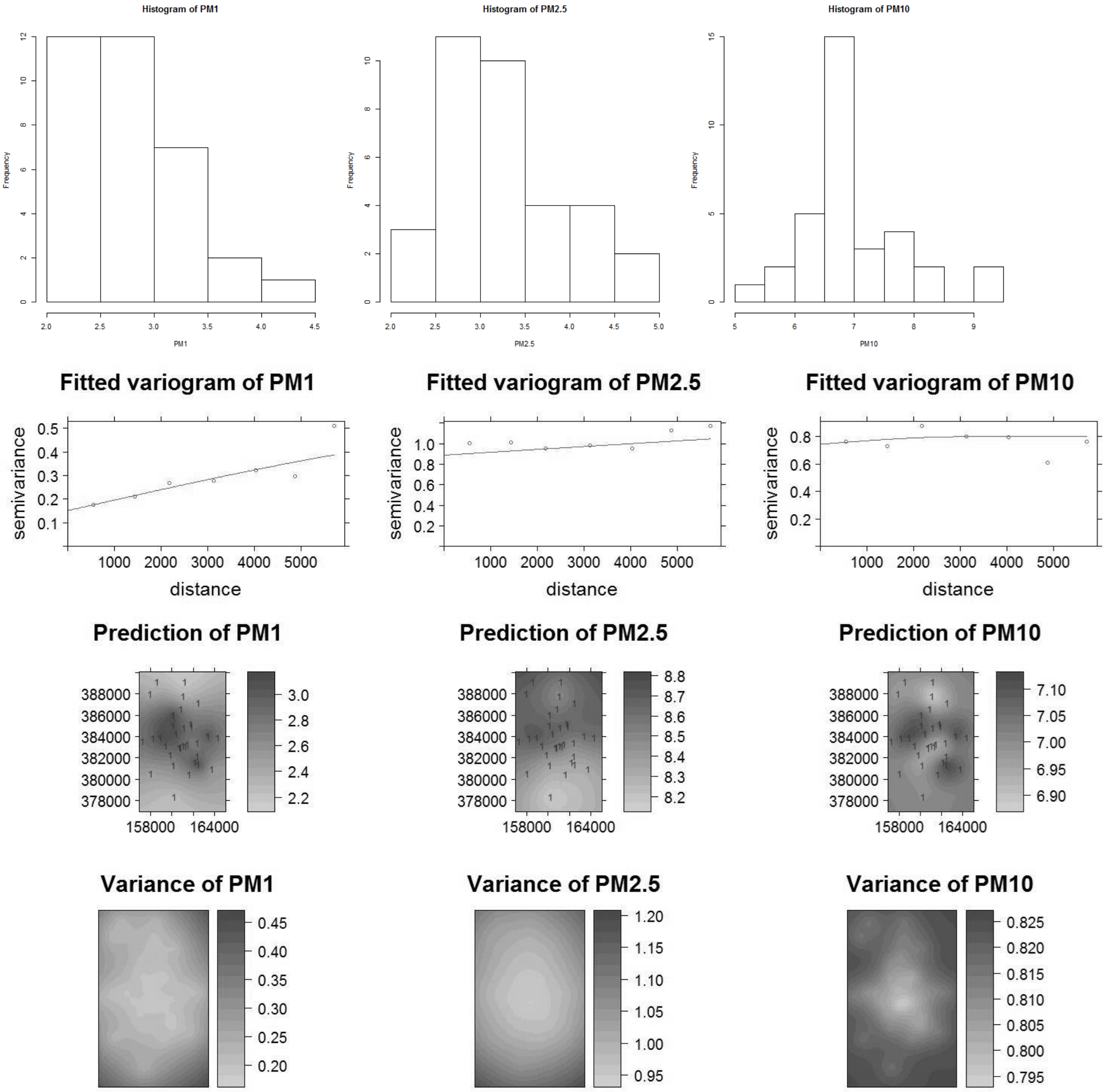


Figure 5-10 Summary statistics, histogram, variogram and prediction map for afternoon rush hour (16:00-19:00)

	PM1	PM2.5	PM10
Minimum	2.40	2.87	6.50
First quarter	2.97	3.40	7.10
Median	3.03	3.65	7.63
Mean	3.11	3.74	7.75
Third quarter	3.16	3.90	8.21
maximum	4.80	6.87	11.50

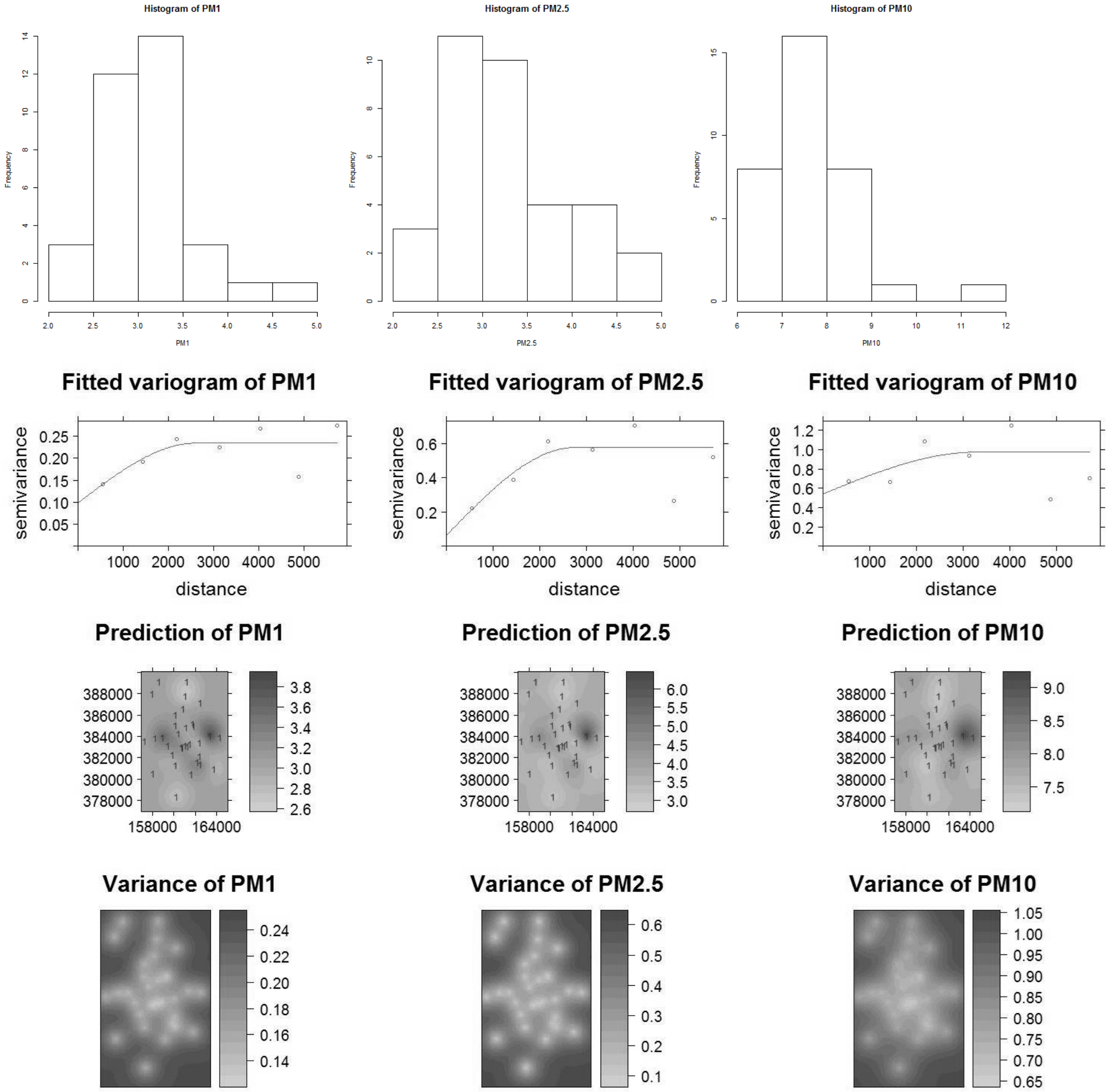


Figure 5-11 Summary statistics, histogram, variogram and prediction map for midnight (19:00-24:00)

	PM1	PM2.5	PM10
Minimum	2.00	2.00	5.17
First quarter	2.00	2.37	6.00
Median	2.17	2.67	6.33
Mean	2.10	2.84	6.53
Third quarter	2.67	3.00	7.21
maximum	3.67	4.67	9.33

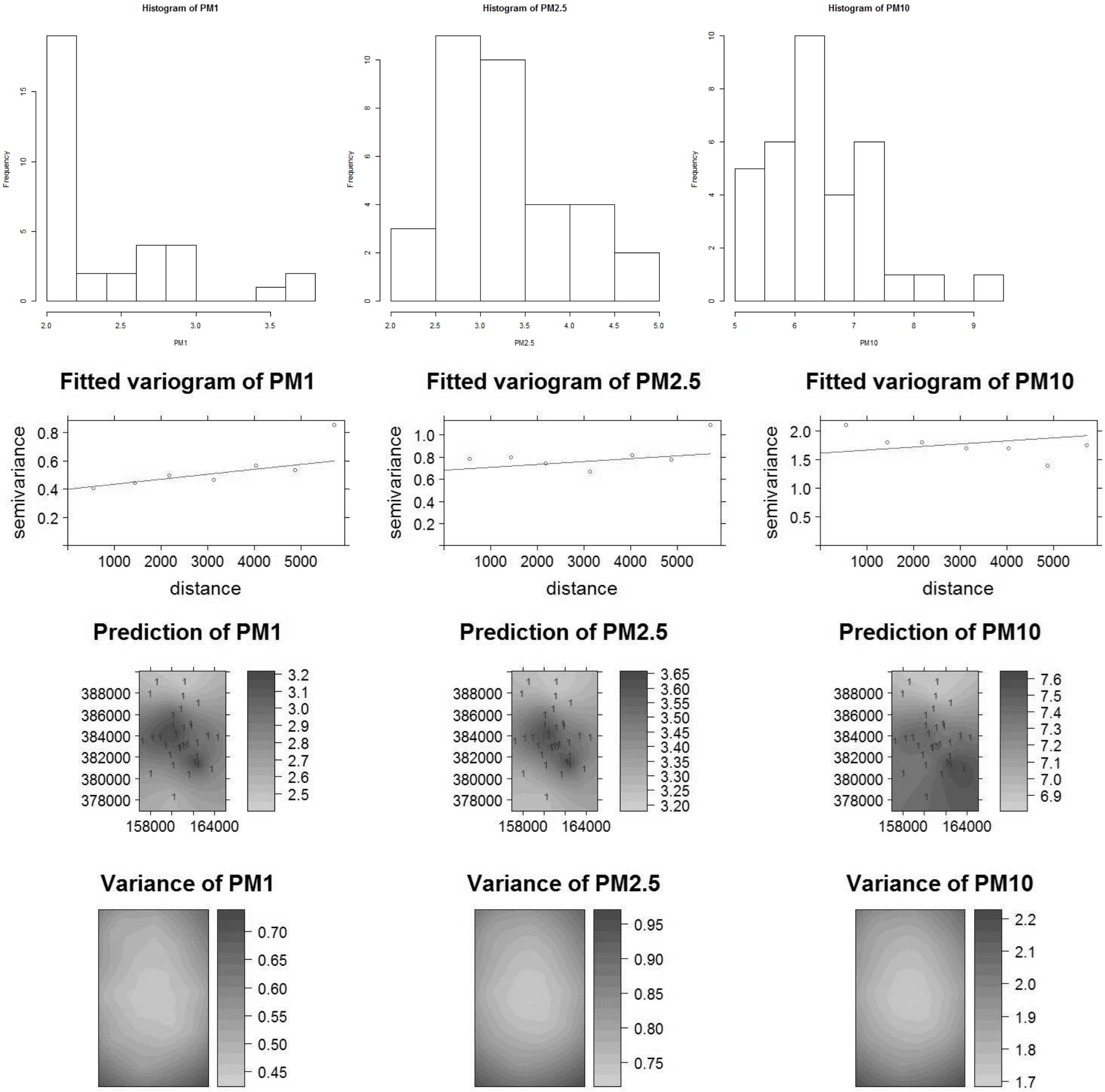


Figure 5-15 Summary statistics, histogram, variogram and prediction map for 17:00-18:00

	PM1	PM2.5	PM10
Minimum	12.83	19.00	22.50
First quarter	20.88	31.04	36.83
Median	22.33	33.83	40.25
Mean	21.79	33.86	40.00
Third quarter	22.96	37.62	43.75
maximum	25.17	42.17	50.17

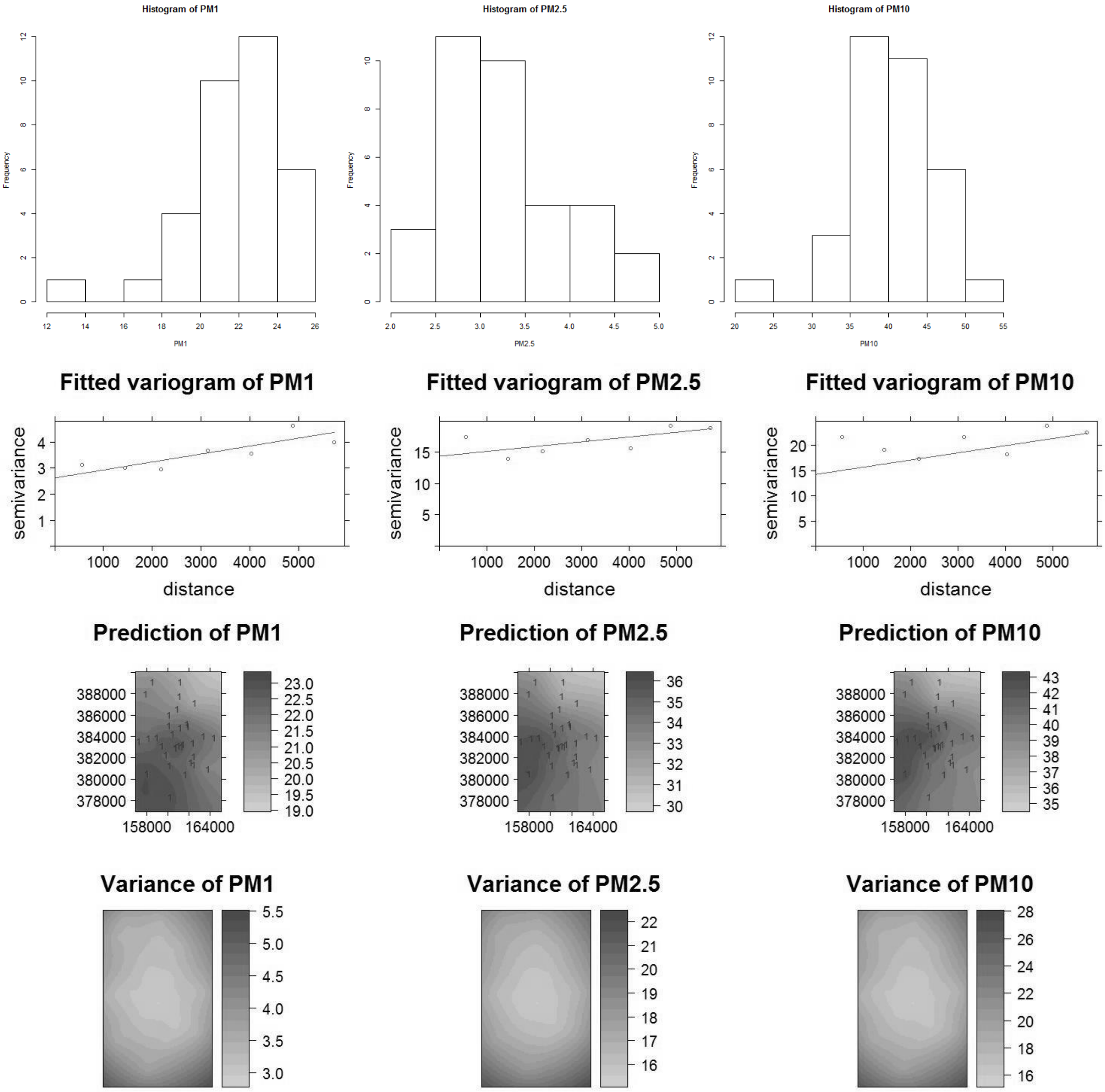


Figure 5-16 Summary statistics, histogram, variogram and prediction map for 7:00-8:00

	PM1	PM2.5	PM10
Minimum	2.00	2.28	4.00
First quarter	2.00	2.89	5.00
Median	2.00	3.14	6.00
Mean	2.59	3.30	6.67
Third quarter	3.00	3.65	7.75
maximum	5.00	4.83	13.00

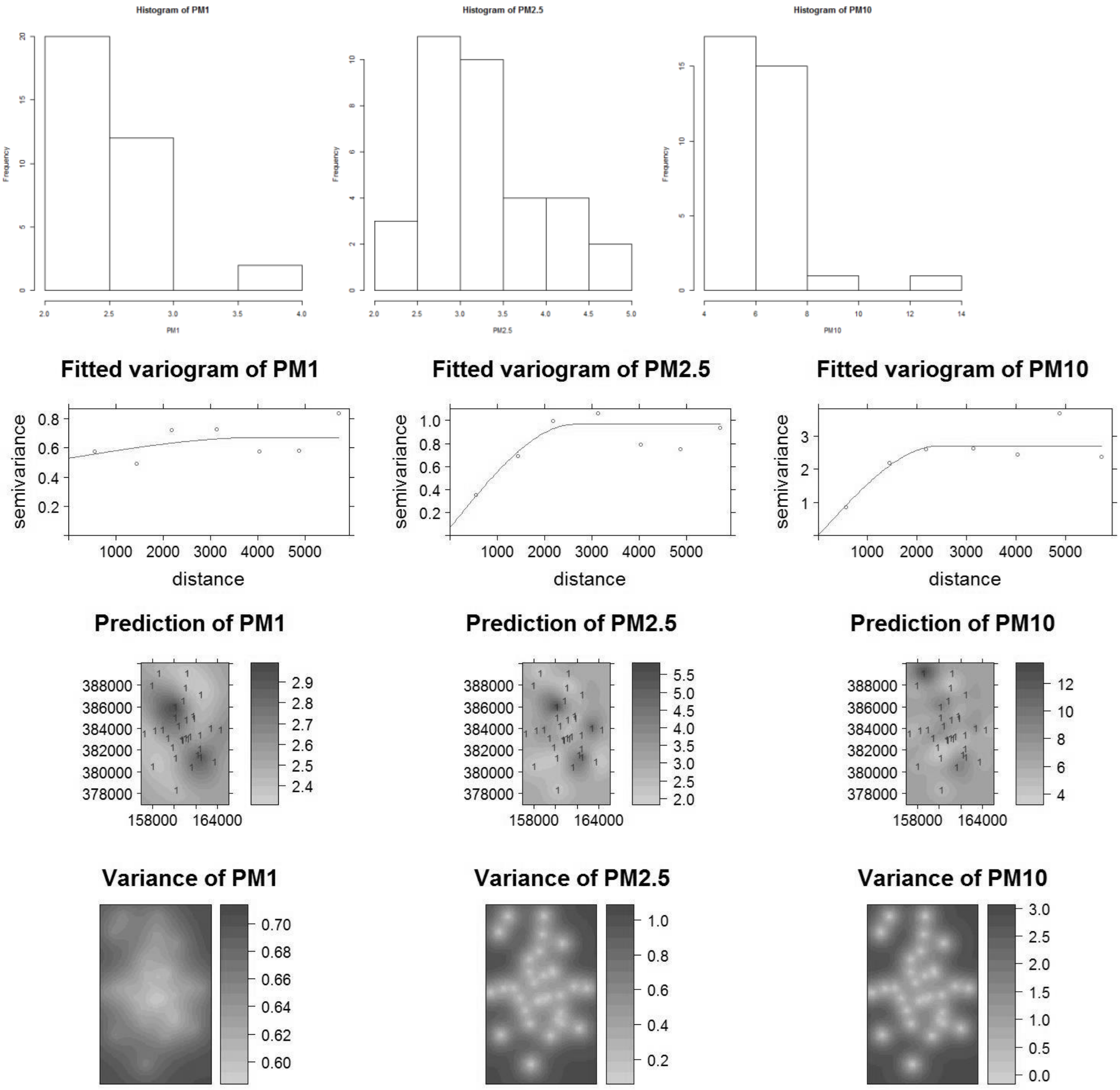


Figure 5-12 Summary statistics, histogram, variogram and prediction map for 17:30-17:40

Table 6 Cross validation of the interpolation above

	PM1	PM2.5	PM10
Whole day			
ME	-0.00025	-0.00023	-0.00035
RMSE	1.08	4.23	5.12
Morning rush hour(7:00-9:00)			
ME	-0.00036	-0.00017	-0.00064
RMSE	4.7	17.15	17.96
Daytime(9:00-16:00)			
ME	0.0000024	0.0000025	0.00034
RMSE	0.43	1.22	2.07
Afternoon rush hour(16:00-19:00)			
ME	-0.0000049	0.0000033	5.8E-07
RMSE	0.31	0.42	0.75
Midnight(19:00-24:00)			
ME	-0.00028	-0.0011	-0.00042
RMSE	0.22	0.5161648	0.93
17:00-18:00			
ME	0.0000051	0.000044	0.00029
RMSE	0.27	0.4	0.89
7:00-8:00			
ME	-0.00093	-0.00072	-0.0016
RMSE	6.05	23.76	29.79
17:30-17:40			
ME	0.00000012	0.0019	0.05
RMSE	0.63	0.91	4.54

From the 00:00 to 11:00 am on October 6th 2014, the highest concentration of all PM values occurs in the centre of the city. However, the darkest point changes from the east to the west of the city.

Between 11:00 and 12:00, the concentration drops rapidly, and the centre pollution moves back to the right of the city.

6. DISCUSSION

6.1. Discussion

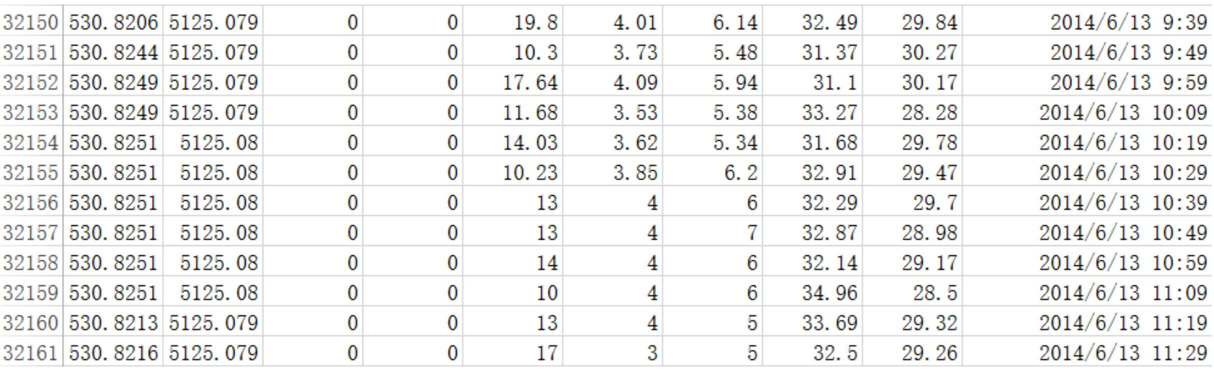
The purpose of this research is to do the analysis of spatial data quality of the measurements. The discussion starts from two sides, which is the spatial variability and the temporal variability.

6.1.1. Discussion about data quality

Since the measurement system has just been developed from 2013, it is not a well-established system. As a matter of fact, the data from August 27th, 2014 to September 17th, 2014 are all missing. Thus, it is difficult to analyse the air quality for the full year. Besides, the location of the sensors has recorded changes in time. There are sometimes clear error in locations.

The main difference between HDF5 format and CSV format is that HDF5 uses Unix Epoch time, which collects the time in second from January 1st, 1970. It is not a direct time that can be read from the original dataset. Besides, HDF5 are reported the data day by day, but it start from 22:30 the last night. This make the analysis for one single day difficult. That is why the CSV format is chosen as the suitable dataset for this research.

One of the biggest problem about the dataset is that the concentrations of fine particles were originally in float type, with two decimals, whereas, after June 13th, 2014 only the integer parts were reported, reducing the precision of measurement. However, t precision does not make much sense for analysis. So it is still an appropriate measurements for air quality.



32150	530. 8206	5125. 079	0	0	19. 8	4. 01	6. 14	32. 49	29. 84	2014/6/13 9:39
32151	530. 8244	5125. 079	0	0	10. 3	3. 73	5. 48	31. 37	30. 27	2014/6/13 9:49
32152	530. 8249	5125. 079	0	0	17. 64	4. 09	5. 94	31. 1	30. 17	2014/6/13 9:59
32153	530. 8249	5125. 079	0	0	11. 68	3. 53	5. 38	33. 27	28. 28	2014/6/13 10:09
32154	530. 8251	5125. 08	0	0	14. 03	3. 62	5. 34	31. 68	29. 78	2014/6/13 10:19
32155	530. 8251	5125. 08	0	0	10. 23	3. 85	6. 2	32. 91	29. 47	2014/6/13 10:29
32156	530. 8251	5125. 08	0	0	13	4	6	32. 29	29. 7	2014/6/13 10:39
32157	530. 8251	5125. 08	0	0	13	4	7	32. 87	28. 98	2014/6/13 10:49
32158	530. 8251	5125. 08	0	0	14	4	6	32. 14	29. 17	2014/6/13 10:59
32159	530. 8251	5125. 08	0	0	10	4	6	34. 96	28. 5	2014/6/13 11:09
32160	530. 8213	5125. 079	0	0	13	4	5	33. 69	29. 32	2014/6/13 11:19
32161	530. 8216	5125. 079	0	0	17	3	5	32. 5	29. 26	2014/6/13 11:29

Figure 6-1 Screenshot of original dataset sensor number 1 on June 13th, 2014

There are better methods rather than just removed of outliers and fill it the blanks with values passed by. In the literature review, the outlier detection in wireless sensor network (Zhang et al., 2012) has studied different methods for spatial outliers, temporal outliers and spatial-temporal outliers. As gathered from this literature, temporal and spatial real-time-based outlier detection is the final preferred technique. If this method can be used in this research that it may lead to better results. However, this is not the main objective of the thesis research and can be used for further research.

The concentration of every ten minutes is far more fluctuant than that of the hourly data. In Figure 5-13 and 5-14, based on the comparison of the hourly data and ten minutes data at the same time, it was shown that the hourly data can fit a smooth variogram while the variogram of ten minutes data is noisy. The other data variograms within that hour also has this problem. After combining the ten-minute data into hourly, however, that model has a better fit.

6.1.2. Discussion of temporal variability

Boxplots of every sensor hourly are not very useful as only 6 data collected by one sensor per hour is not sufficient to check both outliers and mean. Boxplots provide a direct method to make a graphic through quartiles. It contains the smallest non-outlier, the first quartile, median, the third quartile, and the biggest non-outlier. As there are only 6 data per boxplot, the outer fences can be only decided by the 2nd and 4th biggest data in the measurements. If most of the data are close to each other on values (it happens in PM1 rather than PM10), an outlier may have a large effect on the boxplot.

PM10 values are much more fluctuant than PM1 values at the same time, thus is it caused by the precision of data during collection. Since only the integer parts of values are collected, and most PM1 values are near zero, the difference among the PM1 values cannot be shown.

Most sensors have the values fluctuant from17:00 to 21:00, especially in PM10 values. However, the main hourly curve shows a stable level. Some specific sensors show obvious changes during this period of time as mentioned before. It can be seen clearly from the single image in Appendi-2, that these sensors are sensor number 2, 3, 8, 11, 22 and 33. Amongst these sensors, sensor number 3, 8, and 11 are from the types of busy road.

Sensor number 2 is a park in the residential area and holds the highest values of all three fine particles. Nevertheless, it is similar to other sensors for the rest of the week. On October 6th, 2014, the particle concentration (especially in PM10) is higher than that of other sensors, which could not be caused by some extreme values but intersection nearby.

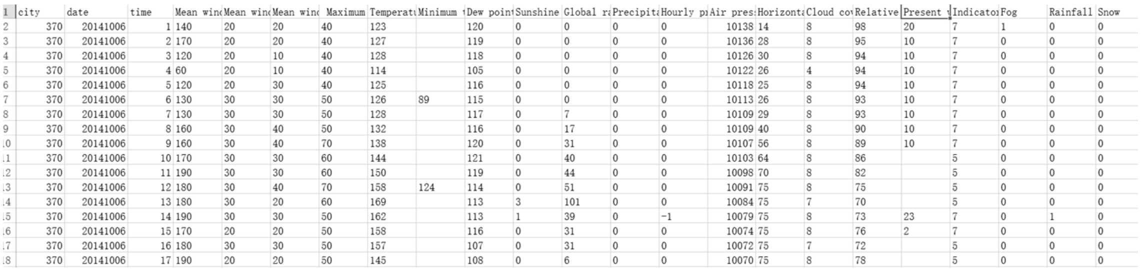
Sensor number 22 may have a road nearby. There is no description when established the sensor in the record, but according to the coordinate in Google Earth, it may still at this case.

Sensor number 33 is located at a road of Veghel according to the coordinate on Google Earth. That is a small town on the north of Eindhoven with a road close to it.

Based on the location types of all sensors and the time period in Appendix-2, the rise of fine particles may be caused by the rush hour after work.

All the sensors show a trend that the PM values increase from the midnight until 6:00 am, and drop to a stable level. This the phenomenon happens through the whole city. Eindhoven is not a heavy industry city which uses coal or other fossil fuels that make great pollutants. Therefore, it must be a great change in the environment rather than the influence of people's behaviours.

The only possible reason for this environment change is the weather (Demuzere, Trigo, Arellano, & Lipzig, 2009). In cloudy days, the airflow is stable to aggregate the fine particulates, and then the rain washes the atmosphere and the ground. Fine particulates existing in aerosol could be taken by water vapour and the PM values must fall down after the rain.



1	city	date	time	Mean wind	Mean wind	Mean wind	Maximum	Temperatu	Minimum t	Dew point	Sunshine	Global ra	Precipita	Hourly pr	Air pres	Horizonta	Cloud cov	Relative	Present w	Indicator	Fog	Rainfall	Snow
2	370	20141006	1	140	20	20	40	123		120	0	0	0	0	10138	14	8	98	20	7	1	0	0
3	370	20141006	2	170	20	20	40	127		119	0	0	0	0	10136	28	8	95	10	7	0	0	0
4	370	20141006	3	120	20	10	40	128		118	0	0	0	0	10126	30	8	94	10	7	0	0	0
5	370	20141006	4	60	20	10	40	114		105	0	0	0	0	10122	26	4	94	10	7	0	0	0
6	370	20141006	5	120	20	30	40	125		116	0	0	0	0	10118	25	8	94	10	7	0	0	0
7	370	20141006	6	130	30	30	50	126	89	115	0	0	0	0	10113	26	8	93	10	7	0	0	0
8	370	20141006	7	130	30	30	50	128		117	0	7	0	0	10109	29	8	93	10	7	0	0	0
9	370	20141006	8	160	30	40	50	132		116	0	17	0	0	10109	40	8	90	10	7	0	0	0
10	370	20141006	9	160	30	40	70	138		120	0	31	0	0	10107	56	8	89	10	7	0	0	0
11	370	20141006	10	170	30	30	60	144		121	0	40	0	0	10103	64	8	86		5	0	0	0
12	370	20141006	11	190	30	30	60	150		119	0	44	0	0	10098	70	8	82		5	0	0	0
13	370	20141006	12	180	30	40	70	158	124	114	0	51	0	0	10091	75	8	75		5	0	0	0
14	370	20141006	13	180	30	20	60	169		113	3	101	0	0	10084	75	7	70		5	0	0	0
15	370	20141006	14	190	30	30	50	162		113	1	39	0	-1	10079	75	8	73	23	7	0	1	0
16	370	20141006	15	170	20	20	50	158		116	0	31	0	0	10074	75	8	76	2	7	0	0	0
17	370	20141006	16	180	30	30	50	157		107	0	31	0	0	10072	75	7	72		5	0	0	0
18	370	20141006	17	190	20	20	50	145		108	0	6	0	0	10070	75	8	78		5	0	0	0

Figure 6-2 weather data that collected by KNMI （KNMI, 2014)

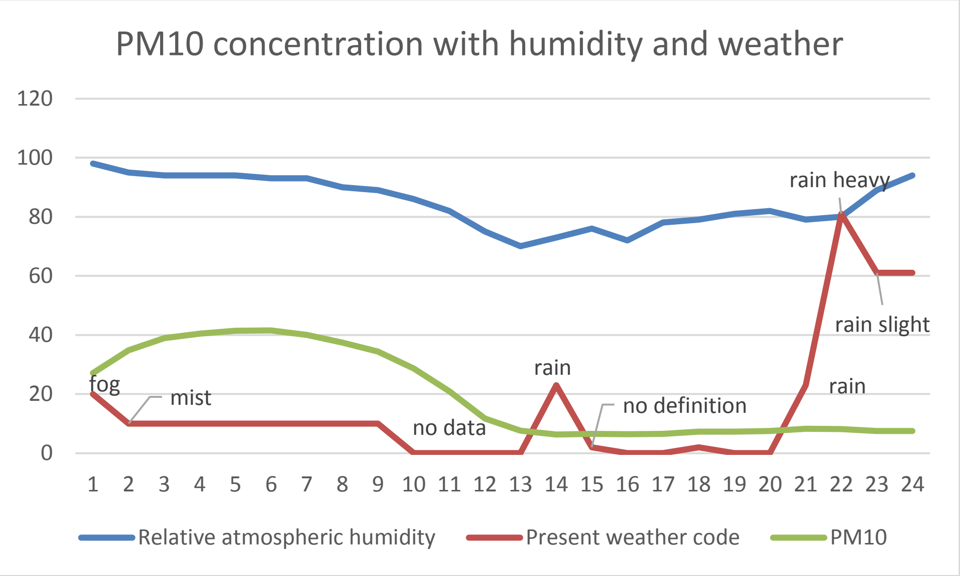


Figure 6-3 PM10 concentration comparing with humidity and weather on October 6th, 2014

The weather code represents the definition of weather. In Figure 6-3 Before 9 am in the morning on October 6th, 2014, it has fog inside the city. From 10 am to 1pm, there is no record for weather changes, and at that time the concentration of particulates changes strongly. After that, there are some heavy and light rain. Since the weather is collected across the whole city, the local changes are not taken into consideration.

In Figure 6-3, shows that before 13pm, PM10 and relative atmospheric humidity have a similar trend in their changes. Humidity, wind and rainfall are three main factors that affect the concentration of coarse and find particles inside the city (Chan, 2001). Besides, seasons also affect the concentration of pollutants (Röösli et al., 2001). Since the time period that PM10 changes rapidly is without any record in weather, it is difficult to say what the strongest influence to this concentration changes is. Fog, however, can carry the pollutants inside the fog layer (Cruz, 1999). This may increase the concentration of pollutants.

Reviewing the weather of the whole week, fog is considered as the most reasonable effect for the high values of PM10 on Monday morning, and the rain may in turn decrease the concentration of PM10.

6.1.3. Discussion of spatial variability

Figure 5-8 provides a general visualization of the whole day. It can be seen that in the whole day, the concentration of particulates is can well in modelled with a normal distribution. The fourth part of the day leads to general changes during the day that the highest value changes from the west to the east of the city. According to the hourly data, the prediction and changes can be provided in great detail. Thus, ten minutes data are so fluctuant that they nearly make no sense.

The variogram model in this thesis is taken as the spherical model. Since the choice of a model is not the main objective of this research, the evaluation between different models has not been completed due to time limit. After auto modelling for the twenty-four hours of the whole day, however, spherical model can fit most of the time periods. Therefore, we consider in spherical model as a suitable model to conduct this research.

As concentrations are only reported as integral values, most variables have a small range. The interpolation is also in the same small range. This makes the legend of prediction map narrow. And makes the predictions less meaningful, since the total values in the image are within a range of 2µg/m^3.

Traffic can bring mass of pollutant through both cars and roads (Querol, 2001), although the changes between sites are tiny, from 17:00 -24:00 in October 6th, 2014. Other effect on spatial distribution could be pressure of industry, the airport, the road structure, and the background pollutants.

6.2. Limitation of this research

Limited to the given data, they already has collected the data of humidity and temperature. These should be related to air quality.

Limited to the weather data, it only contain the whole city hourly.

7. CONCLUSION AND RECOMMENDATION

7.1. Conclusion

The conclusion are made by answering the question of research objective:

1.1 *What is the distribution of the data considering basic descriptive statistics, time and space? What is the relationship between the different variables (e.g., PM1 vs. PM2.5)? What are the ratio's between different variables (e.g., PM2.5 / PM10)?*

The measurement system is collecting air quality data, including basic information about the sensor. This research focused on the first week of October, 2014, especially 6th. Regression analysis has provide the relationship between different variables in Figure 5-5 and Table3. The ratio between different variables has been shown in Table 4. The study concluded that there are strong relationship between different variables, and the ratio's between different variables changes according to locations and times.

1.2 *How does the temporal distribution of specific pollutants vary between sites? Are they systematically lower or higher between sites? Are they comparable?*

The temporal pattern is similar between sensors and not much difference in observed between types of locations in Appendix-2. There is no strong evidence of variability between types of sites. In total, the centre of the city holds the highest concentrations of pollutants cf. Figure 5-10.

1.3 *What is the spatial distribution of the data at different time epochs? Are the concentrations influenced by different types of locations?*

On October 6th, 2014, the spatial distribution of pollutants changes as time goes, cf. Figure 5-11 to Figure 5-14. In the afternoon rush hour, some specific sensors (Sensor 2 8 11 22 25 and 33) have fluctuant values, as can be seen from Appendix-2. It has no significant evidence that the concentrations were influenced by location types cf. Figure 5-9 and Appendix 4.

1.4 *How do the above questions lead to a protocol for SDQ evaluation?*

Because the SDQ has been determined by spatial data accuracy, the above question has tested the accuracy of the spatial data. The spatial structure is an evidence that will be effective for outlier detection. Other SDQ types are either not relevant or are referred to further research.

2.1 *How well does a single site represent temporal variability in air quality?*

I have not managed to answer this question. It is still need to improve in further research.

2.2 *What is the optimal number of observations, balancing between the cost of sensor and the precision of the modelling?*

There isn't any exist evidence to show that the number of measurements can be reduced.

3.1*What is the appropriate temporal scale for mapping?*

In the discussion of spatial variability it was been reported that the ten minutes data are too fluctuant to use, and that the hourly scale is the minimum temporal scale for mapping, balancing the smooth changes of the data in space and time.

3.2 *Which area have higher concentration changes and why?*

The concentration changes rapidly during rush hour in the crossroads of the main street. At the start of the day, the highest concentration is in the west of city. During rush hour, however, the crossroads have shown increase of values as it on the prediction map.

7.2. Recommendation

1. Estimate the number of necessary measurements depending on accuracy and cost of sensors.
2. Further mining the original data, if humidity and temperature has relationship with particulates matter.
3. Redefine the location and location types of sensors since they have changed from the established of measurement network.
4. Try other transform method and model for kriging.

LIST OF APPENDIX

8. REFERENCE

AiREAS. (2014a). AiREAS. Retrieved December 09, 2014, from http://www.aireas.com/

AiREAS. (2014b). Index of data. Retrieved December 09, 2014, from http://193.172.204.137:8080/

Bivand. (2011). R sp graphics example figures. Retrieved February 02, 2015, from http://rspatial.r-forge.r-project.org/gallery/

Bivand, Roger S., Pebesma, Edzer J., Gómez-Rubio, V. (n.d.). Applied Spatial Data Analysis with R | Springer. Retrieved February 15, 2015, from http://www.springer.com/gp/book/9780387781716

Briggs, D. J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., & Smallbone, K. (2000). A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of The Total Environment*, *253*(1-3), 151–167.

Buttafuoco, G., Castrignanò, A., Busoni, E., & Dimase, A. C. (2005). Studying the spatial structure evolution of soil water content using multivariate geostatistics. *Journal of Hydrology*, *311*(1-4), 202–218.

CAMBRIDGE, T. U. O. (1998). The Ozone Hole Tour : Part III. The Science of the Ozone Hole. Retrieved August 13, 2014, from http://www.atm.ch.cam.ac.uk/tour/part3.html

Chan, L. (2001). Roadside suspended particulates at heavily trafficked urban sites of Hong Kong â " Seasonal variation and dependence on meteorological conditions. *Atmospheric Environment*, *35*(18), 3177–3182. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231000005045

Cruz, D. (1999). Production and removal of aerosol in a polluted fog layer: model evaluation and fog effect on PM. *Atmospheric Environment*, *33*(29), 4797–4816. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231099002642

Dawson, R. (2010). How Significant Is a Boxplot Outlier? *Journal of Statistics Education*, *19*(2).

Demuzere, M., Trigo, R. M., Arellano, J. V. De, & Lipzig, N. P. M. Van. (2009). The impact of weather and atmospheric circulation on O 3 and PM 10 levels at a rural mid-latitude site, 2695–2714.

EuropeanCommission. (2013). Air Quality Standards - Environment - European Commission. Retrieved January 23, 2015, from http://ec.europa.eu/environment/air/quality/standards.htm

Fernando, H. J. S., Klaić, Z., & McCulley, J. L. (Eds.). (2012). *National Security and Human Health Implications of Climate Change*. Dordrecht: Springer Netherlands.

Gurp, H. van. (2014). Spatial data quality of air quality data collected at the city level. *BSC Project, of Twente*.

Han, W., Xu, J., Wei, K., Shi, Y., & Ma, L. (2013). Estimation of N2O emission from tea garden soils, their adjacent vegetable garden and forest soils in eastern China. *Environmental Earth Sciences*, *70*(6), 2495–2500.

Hannes, O. K., Reuter, I., & Gerboles, M. (2013). A Tool for the Spatio-Temporal Screeninng of AirBase Datasets for Abnormal Values. *European Commission.*

Hengl, T. (2010). Spatial interpolation exercises (NL) - spatial-analyst.net. Retrieved December 07, 2014, from http://spatial-analyst.net/wiki/index.php?title=Spatial_interpolation_exercises_(NL)#Spatial_prediction_of_daily_precipitation_in_the_Netherlands

Heuvelink, G. B. ., & Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma*, *100*(3-4), 269–301.

Johansson, C., Norman, M., & Gidhagen, L. (2007). Spatial & temporal variations of PM10 and particle number concentrations in urban air. *Environmental Monitoring and Assessment*, *127*(1-3), 477–87.

KNMI. (2014). KNMI - Uurgegevens van het weer in Nederland - Download. Retrieved January 23, 2015, from http://www.knmi.nl/klimatologie/uurgegevens/#no

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, 1137–1143. Retrieved from http://dl.acm.org/citation.cfm?id=1643047

Křůmal, K., Mikuš ka, P., & Večeřa, Z. (2013). Polycyclic aromatic hydrocarbons and hopanes in PM1 aerosols in urban areas. *Atmospheric Environment*, *67*, 27–37.

Li, R., Wiedinmyer, C., & Hannigan, M. P. (2013). Contrast and correlations between coarse and fine particulate matter in the United States. *The Science of the Total Environment*, *456-457*, 346–58.

Lianou, M., Chalbot, M.-C., Kavouras, I. G., Kotronarou, A., Karakatsani, A., Analytis, A., … Hoek, G. (2011). Temporal variations of atmospheric aerosol in four European urban areas. *Environmental Science and Pollution Research International*, *18*(7), 1202–12.

Maria, C.-P. (2000). Descriptive statistics. Retrieved January 12, 2015, from http://www.pitt.edu/~super1/lecture/lec0421/index.htm

Oort, van P. A. J. (2006). Spatial data quality: from description to application. Retrieved from http://www.narcis.nl/publication/RecordID/oai%3Alibrary.wur.nl%3Awurpubs%2F346112

Querol, X. (2001). PM10 and PM2.5 source apportionment in the Barcelona Metropolitan area, Catalonia, Spain. *Atmospheric Environment*, *35*(36), 6407–6419. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231001003612

Querol, X., Alastuey, A., Ruiz, C. R., Artiñano, B., Hansson, H. C., Harrison, R. M., … Schneider, J. (2004). Speciation and origin of PM10 and PM2.5 in selected European cities. *Atmospheric Environment*, *38*(38), 6547–6555. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231004008143

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1981). Applied regression analysis 2nd ed. Retrieved from http://www.popline.org/node/423881

RGraphGallery. (2013). R graph gallery. Retrieved February 02, 2015, from http://rgraphgallery.blogspot.nl/#uds-search-results

Röösli, M., Theis, G., Künzli, N., Staehelin, J., Mathys, P., Oglesby, L., … Braun-Fahrländer, C. (2001). Temporal and spatial variation of the chemical composition of PM10 at urban and rural sites in the

Basel area, Switzerland. *Atmospheric Environment*, *35*(21), 3701–3713. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231000005112

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley & Sons, Ltd. Retrieved from https://ezproxy.utwente.nl:2174/abstract/9780470517260

WHO. (2003). Health Aspects of Air Pollution with Particulate Matter , Ozone and Nitrogen Dioxide, (January). Retrieved from http://apps.who.int/iris/handle/10665/107478

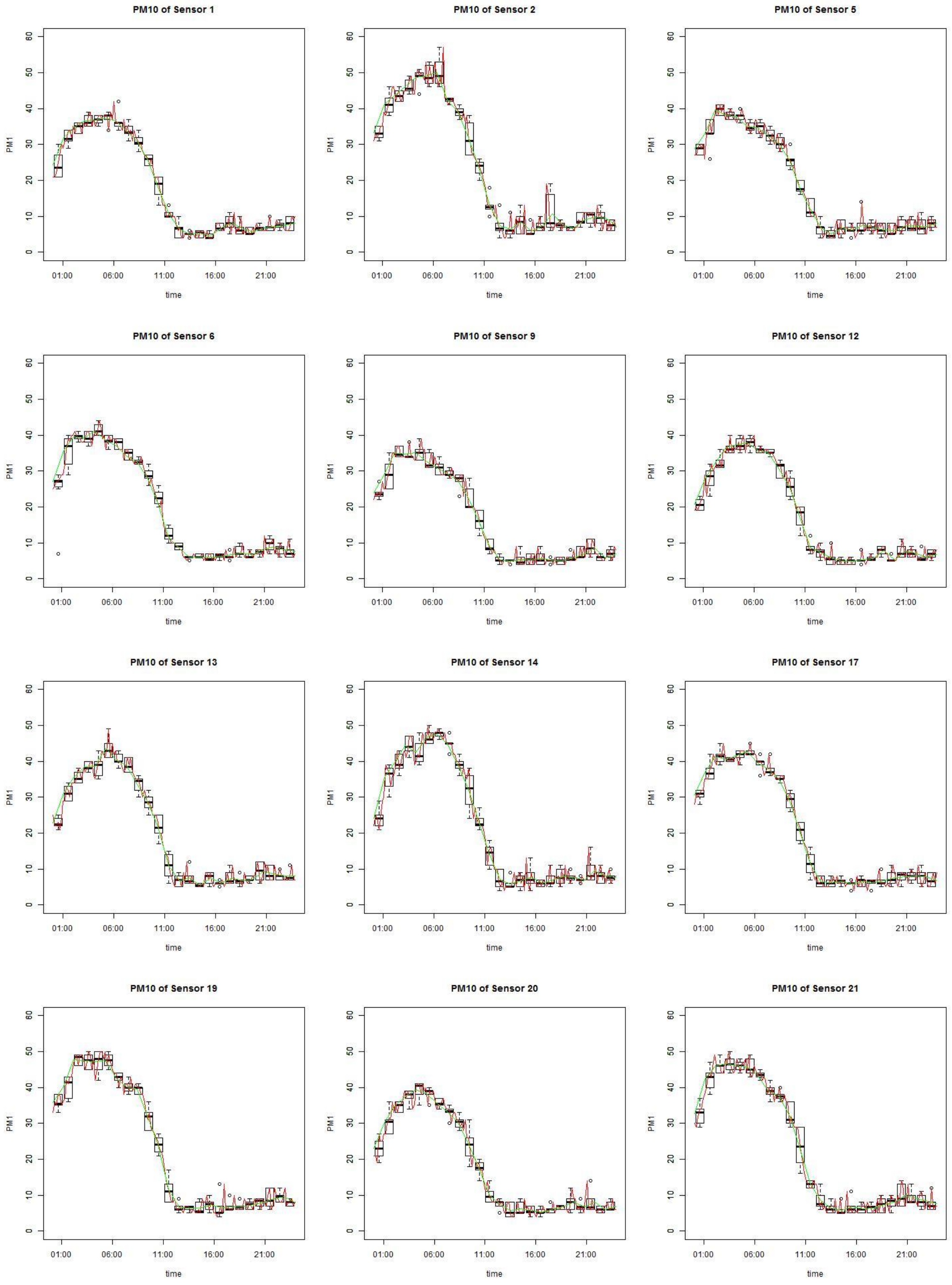
WHO. (2014). WHO | 7 million premature deaths annually linked to air pollution. Retrieved August 13, 2014, from http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

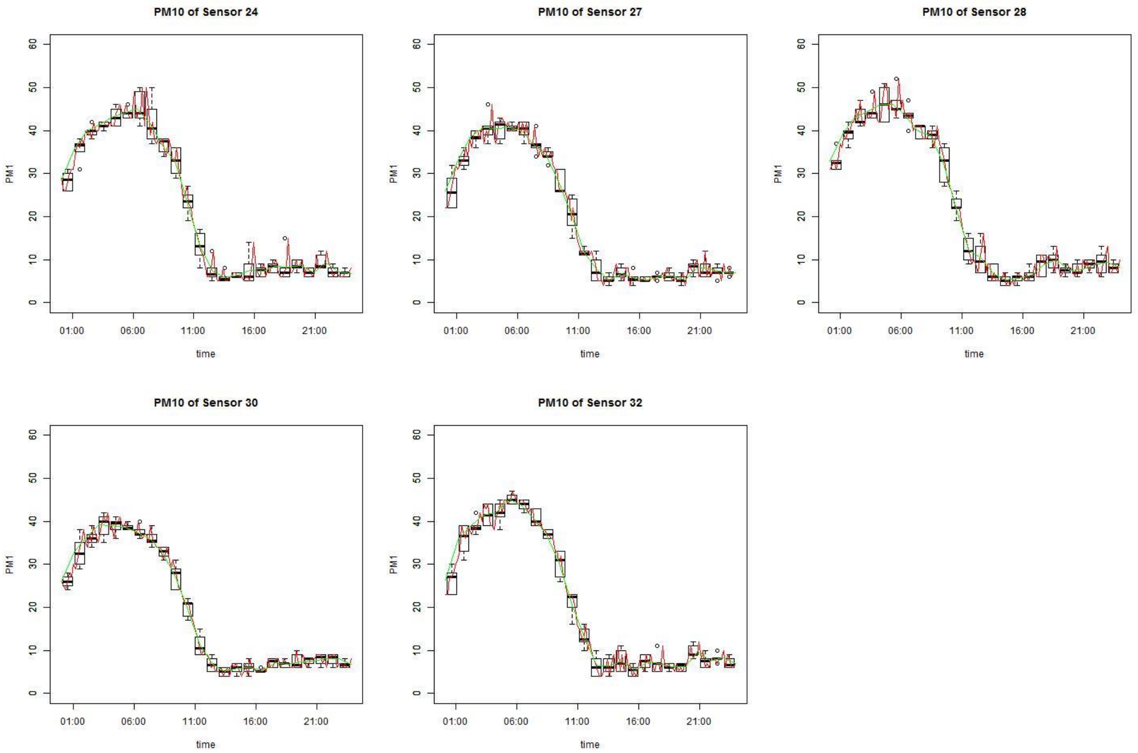
Zhang, Y., Hamm, N. A. S., Meratnia, N., Stein, A., van de Voort, M., & Havinga, P. J. M. (2012). Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science*, *26*(8), 1373–1392.

Appendix-1 Address description

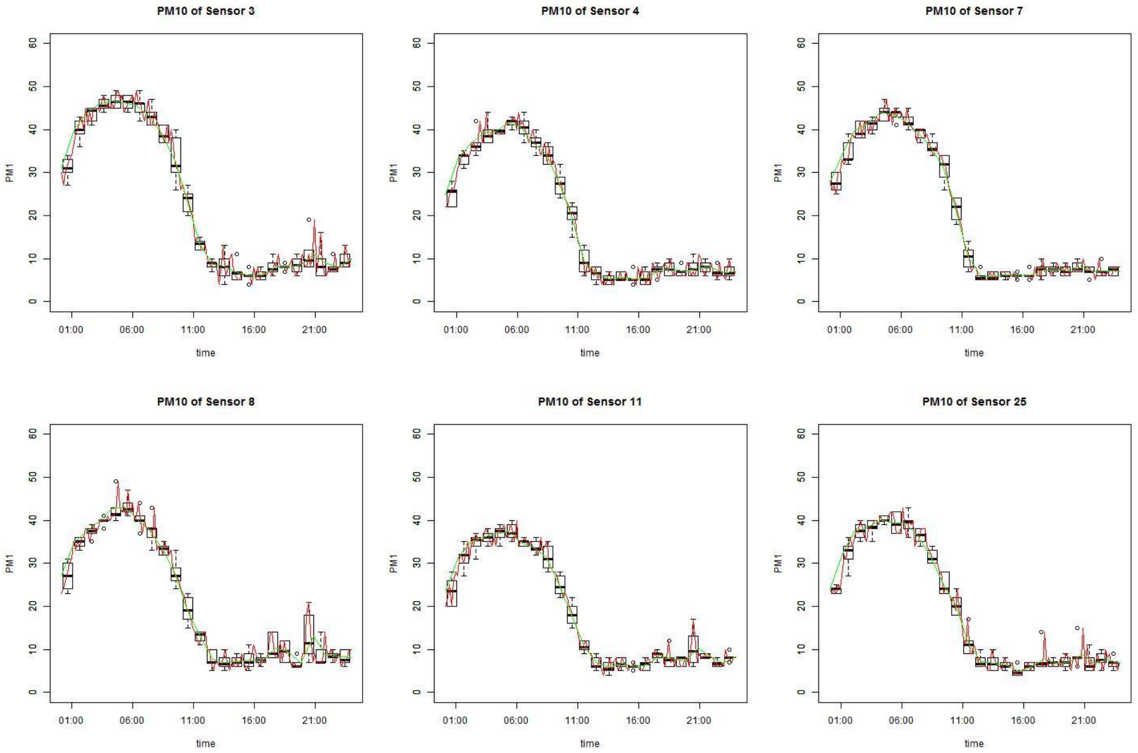
Sensor number	EAST	NORTH	Address	postcode	
1	530.829481	5125.07753	Eij-erven	5646 JM	
2	527.9814	5126.86782	Lijmbeekstraat	5612 NJ	
3	528.684006	5126.21157	Keizersgracht	5611 GD	
4	527.431915	5125.78627	X v. Weberstraat-Limburglaan	5653 AD	
5	528.5992	5129.49057	Falstaff	5629 NK	
6	529.7496	5128.41876	Grote Beerlaan	5632 DN	
7	527.1008	5126.24862	Botenlaan	5616 JG	
8	529.718	5125.27445	Leenderweg	5643 AJ	
9	528.512	5128.77574	Maasteikstraat	5628 PZ	
10	839.0714	4038.32945			
11	529.4744	5125.41376	Leostraat	5644 PA	
12	527.7038	5125.2516	Jan Hollanderstraat	5654 DT	
13	525.2223	5126.48024	Sliffertsestraat	5657 AR	
14	525.8829	5124.82356	Twickel	5655 JJ	
15	453.6295	5306.32682			
16	440.4792	5247.08644			
17	525.7942	5128.86846	Amstelstraat	5626 BN	
18	440.4949	5247.10688			
19	526.349	5129.48922	Finisterelaan	5627 TE	
20	529.6336	5126.3816	Sperwerlaan	5613 EE	
21	531.338	5126.65082	Donk	5641 PX	UFPs
22	530.5062	5126.72008	Hofstraat	5504 GD	UFPs
23	440.4764	5247.06988			
24	526.0618	5126.64603	v. Vollenhovenstraat	5652 SN	
25	528.2139	5126.18893	(tegenover) Mauritsstraat	5616 AB	UFPs
26	528.9411	5126.33113	Gedempte gracht / Vestdijk	5611 DM	
27	529.6661	5125.69714	St. Adrianusstraat	5614 EP	
28	527.7022	5127.79597	Rijckwaertstraat	5622 HV	
29	529.1943	5127.23678	Ds. Fliednerstraat	5631 BN	
30	528.4824	5126.24295	Spijndhof	5611 HV	UFPs
31	527.7846	5123.66015	Vincent Cleerdinlaan Waalre	5582 EJ	
32	529.0184	5124.78893	Vesaliuslaan	5644 HL	
33	531.1644	5136.18312	Moving airbox		
34	528.6015	5127.1574	Pastoriestraat	5612 EJ	
35	527.68	5127.28489	Boschdijk	5621 JC	
36	529.0904	5127.36137	Hudsonlaan / Kennedylaan	5623 NR	UFPs
37	528.3335	5128.10492	Genovevalaan	5625 EA	
38	440.806	5247.15116			
39	526.6951	5126.65722	Noord-Brabantlaan	5651 LZ	
40	528.2544	5126.12929	Zwijgerstraat	5616 AC	

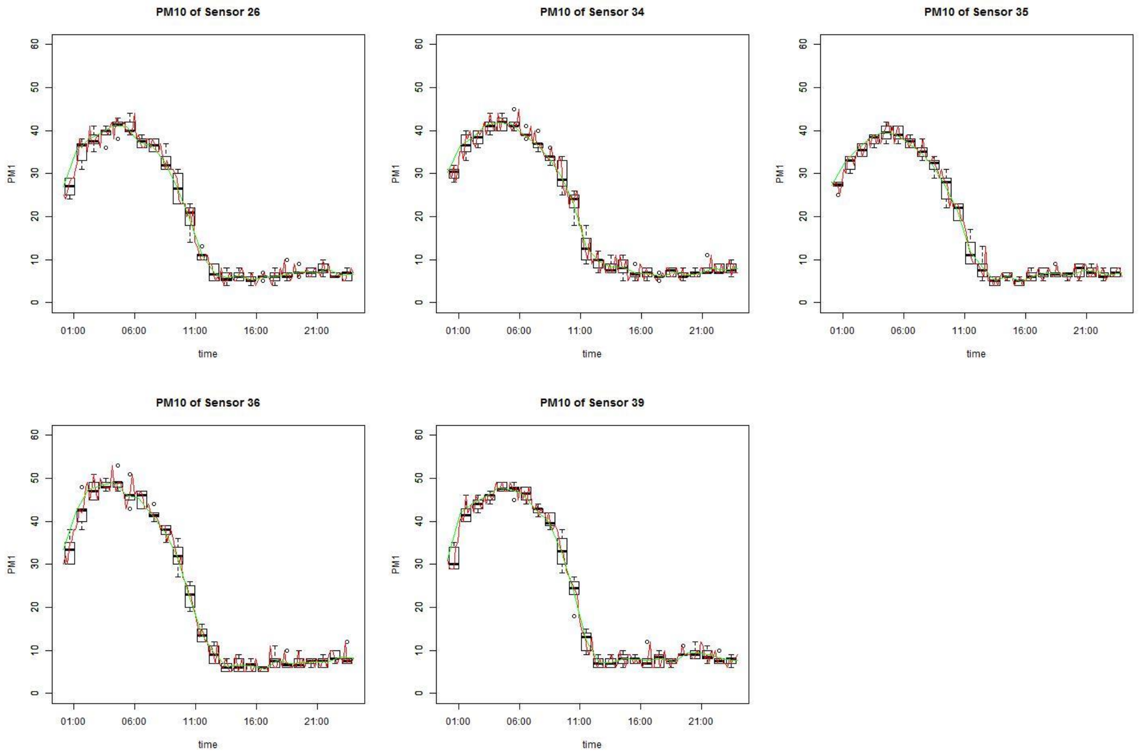
Appendix-2 Temporal distribtuion of each sensor of PM10 on October 6th, 2014



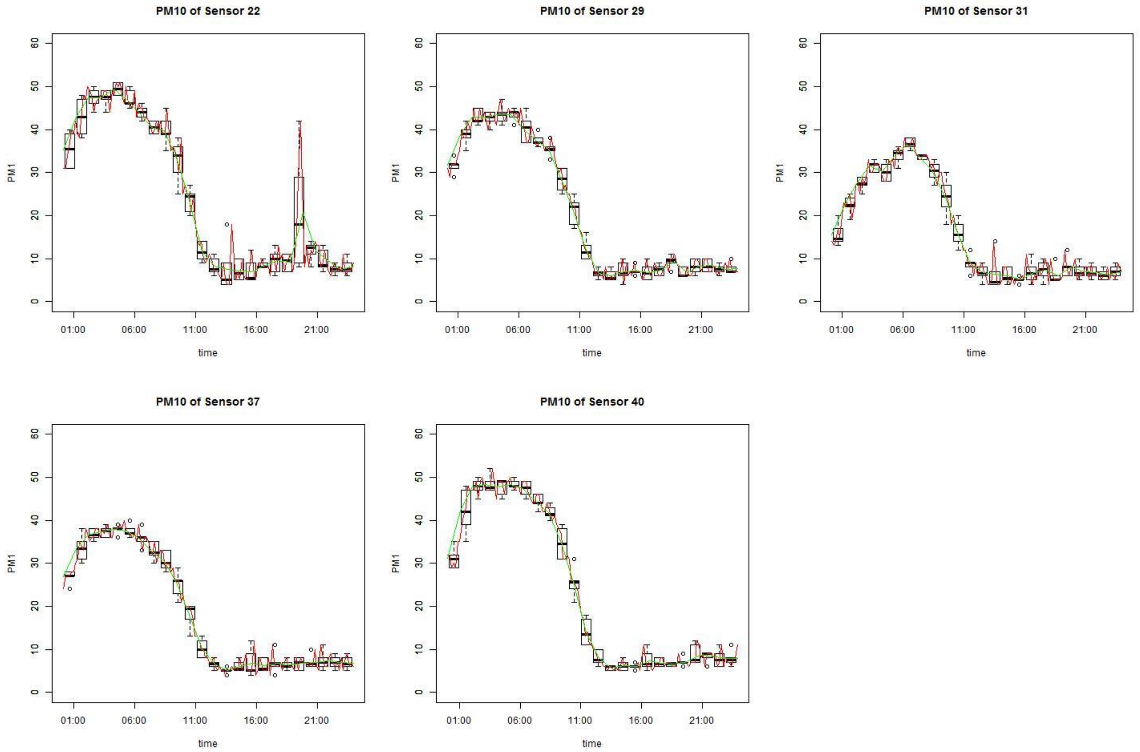


PM10 of sensors from residential area on October 6th

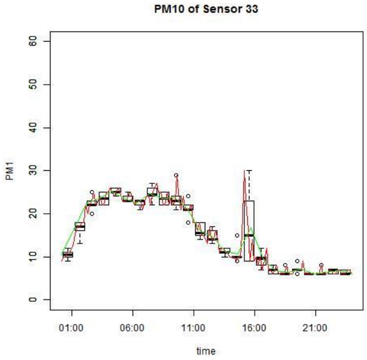




PM10 of sensors from busy road on October 6th

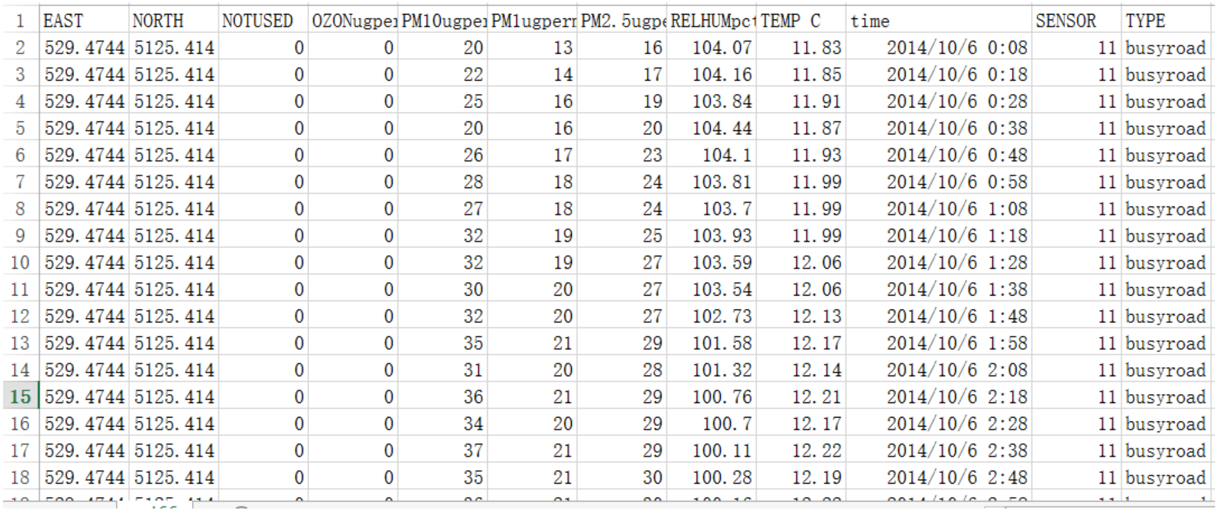


PM10 of sensors from the others on October 6th



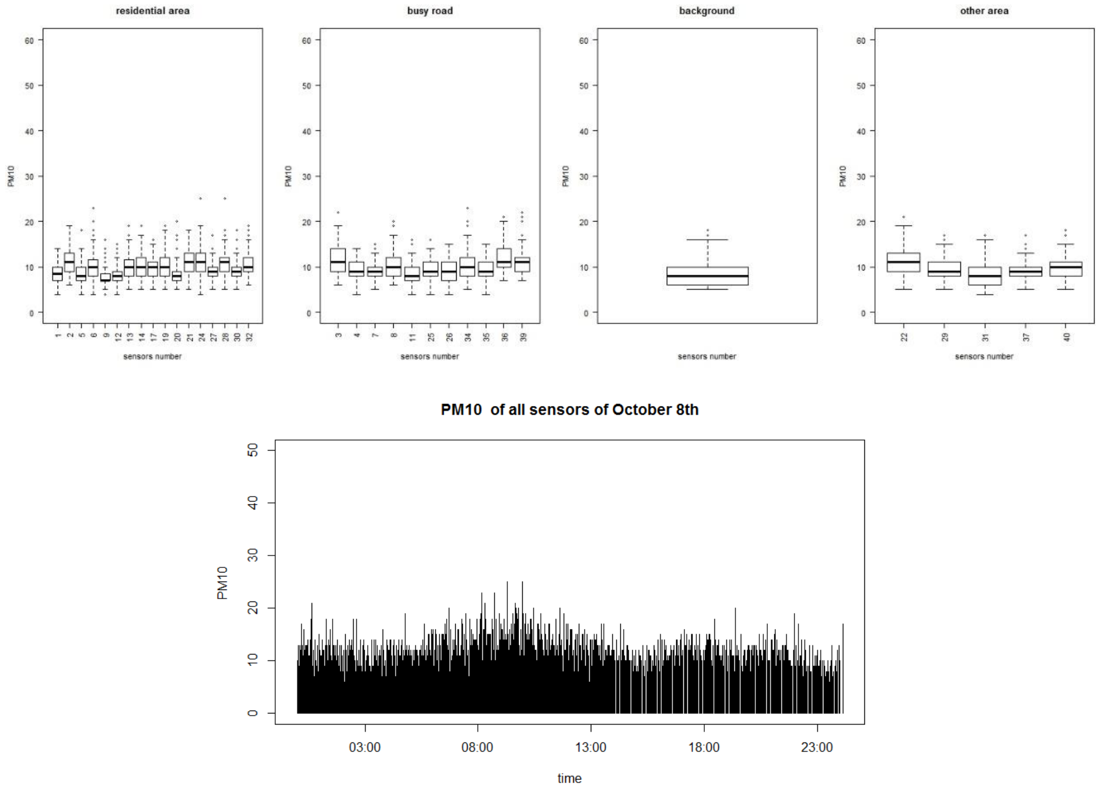
PM10 of sensors from background on October 6th

Appendix-3 Complete pre-processing data

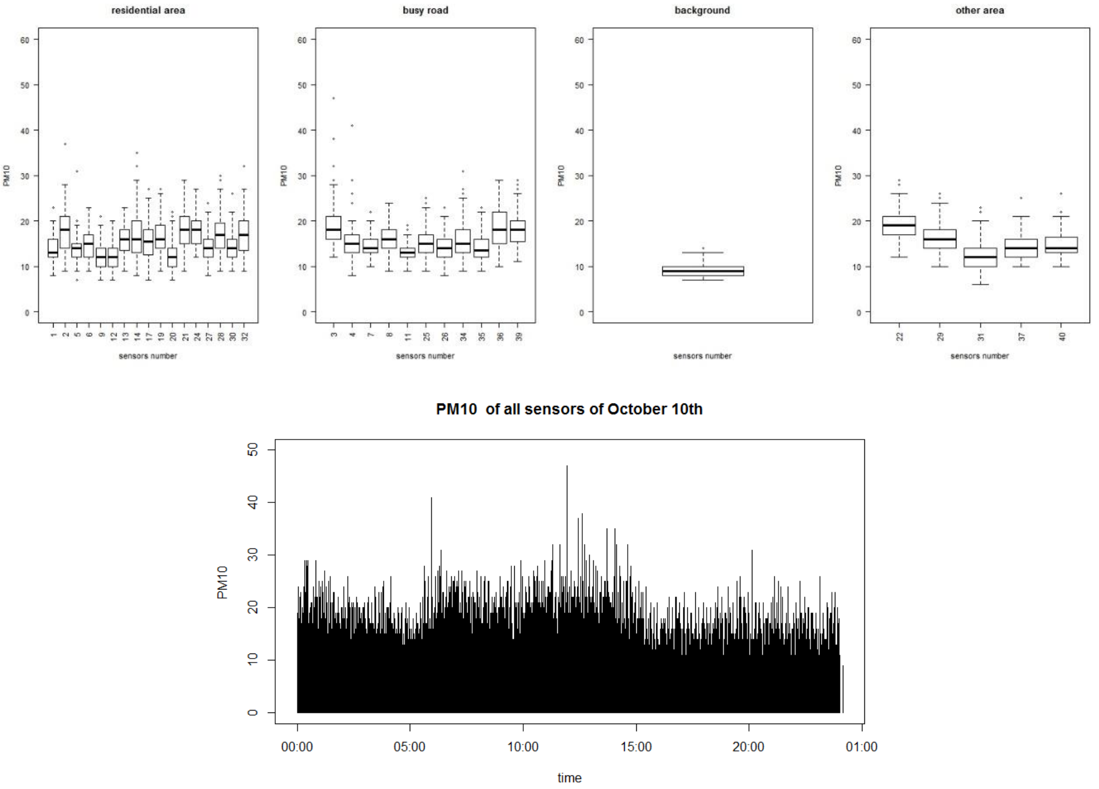


1	EAST	NORTH	NOTUSED	OZONugpe	PM10ugpe	PM1ugperr	PM2.5ugpe	RELHUMpct	TEMP C	time	SENSOR	TYPE
2	529.4744	5125.414	0	0	20	13	16	104.07	11.83	2014/10/6 0:08	11	busyroad
3	529.4744	5125.414	0	0	22	14	17	104.16	11.85	2014/10/6 0:18	11	busyroad
4	529.4744	5125.414	0	0	25	16	19	103.84	11.91	2014/10/6 0:28	11	busyroad
5	529.4744	5125.414	0	0	20	16	20	104.44	11.87	2014/10/6 0:38	11	busyroad
6	529.4744	5125.414	0	0	26	17	23	104.1	11.93	2014/10/6 0:48	11	busyroad
7	529.4744	5125.414	0	0	28	18	24	103.81	11.99	2014/10/6 0:58	11	busyroad
8	529.4744	5125.414	0	0	27	18	24	103.7	11.99	2014/10/6 1:08	11	busyroad
9	529.4744	5125.414	0	0	32	19	25	103.93	11.99	2014/10/6 1:18	11	busyroad
10	529.4744	5125.414	0	0	32	19	27	103.59	12.06	2014/10/6 1:28	11	busyroad
11	529.4744	5125.414	0	0	30	20	27	103.54	12.06	2014/10/6 1:38	11	busyroad
12	529.4744	5125.414	0	0	32	20	27	102.73	12.13	2014/10/6 1:48	11	busyroad
13	529.4744	5125.414	0	0	35	21	29	101.58	12.17	2014/10/6 1:58	11	busyroad
14	529.4744	5125.414	0	0	31	20	28	101.32	12.14	2014/10/6 2:08	11	busyroad
15	529.4744	5125.414	0	0	36	21	29	100.76	12.21	2014/10/6 2:18	11	busyroad
16	529.4744	5125.414	0	0	34	20	29	100.7	12.17	2014/10/6 2:28	11	busyroad
17	529.4744	5125.414	0	0	37	21	29	100.11	12.22	2014/10/6 2:38	11	busyroad
18	529.4744	5125.414	0	0	35	21	30	100.28	12.19	2014/10/6 2:48	11	busyroad

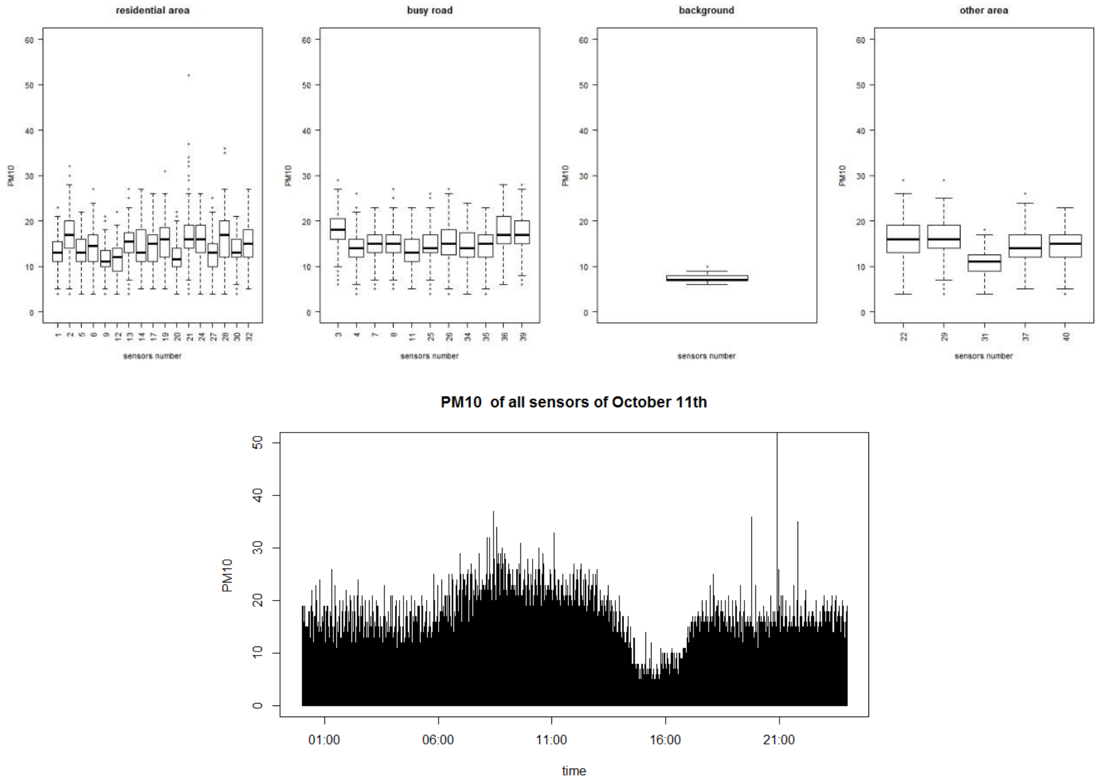
Appendix-4 Boxplots of different location types on October 8th, 10th,11th ,2014



Boxplots of different types of location on October 8th, 2014 with temporal distribution of PM10



Boxplots of different types of location on October 10th, 2014 with temporal distribution of PM10



Boxplots of different types of location on October 11th, 2014 with temporal distribution of PM10