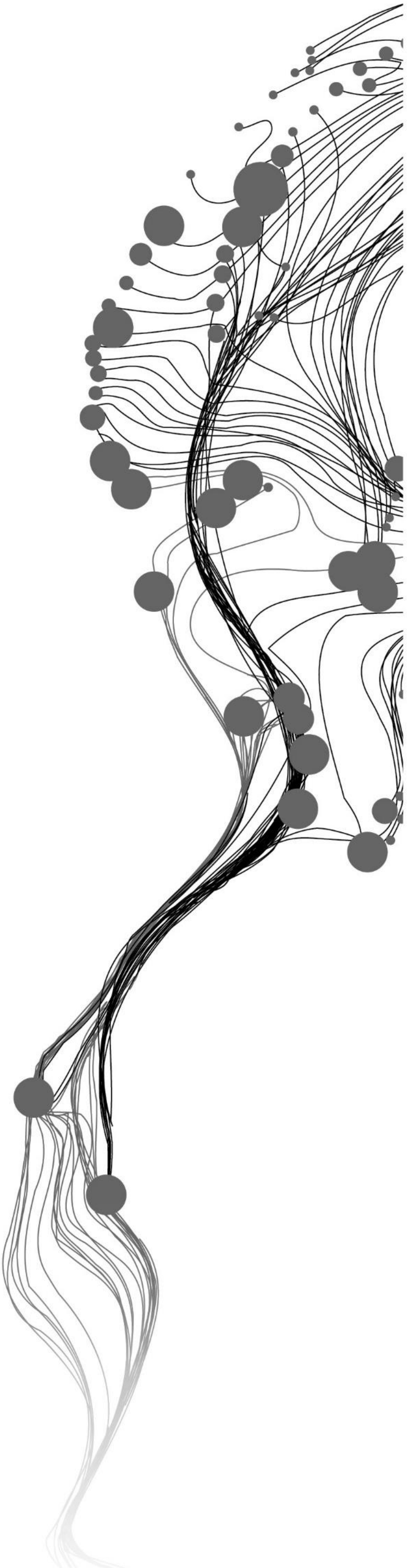
MODELLING AND MAPPING OF ULTRAFINE PARTICLES IN SPACE AND TIME IN THE CITY OF EINDHOVEN, THE NETHERLANDS

GERARDO MACEDO RODRIGUEZ
February, 2019

SUPERVISORS:
Dr. F.B. Osei
Dr. P. Truong

ADVISOR:
Msc. V.M. van Zoest



MODELLING AND MAPPING OF ULTRAFINE PARTICLES IN SPACE AND TIME IN THE CITY OF EINDHOVEN, THE NETHERLANDS

GERARDO MACEDO RODRIGUEZ
Enschede, The Netherlands, February, 2019

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

SUPERVISORS:
Dr. F.B. Osei
Dr. P. Truong

ADVISOR:
Msc. V.M. van Zoest

THESIS ASSESSMENT BOARD:
Prof. dr. ir. A. Stein (Chair)
Prof. dr. ir. G.B.M. Heuvelink (External Examiner, Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing)

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

Air pollution represents one of the most significant issues in the world. In this context, ultrafine particles (UFP) are one the most dangerous pollutants due to its small diameter. Hence, it is crucial to characterize the spatiotemporal distribution of air pollution which motivates to a civil initiative (AiREAS) to set up an innovative network (ILM) for measuring air quality in the city of Eindhoven. ILM network is conformed by several sensors, among which are UFP sensors that are usually in fixed locations. Nevertheless, during the research phase, they were rotated for five periods of three weeks among different locations in the city.

This research aims to model and map UFP in space and time in the city of Eindhoven using meteorological variables and the five rotation periods as fixed and random effect variables respectively. To do so, a multiple linear regression analysis was applied to 24 different hourly timestamps getting the mixed-effect models' parameters and the residuals.

Then, the covariates values were obtained in the entire study area using Thiessen polygons, and the residuals were predicted using ordinary kriging after estimating their variogram parameters, applying maximum likelihood estimation.

Afterwards, the models were run to get the 24 prediction UFP maps. Results showed a higher concentration of UFP during traffic peak hours and a lower concentration out of peak hours. Furthermore, the highest concentration of UFP is in the north-east of the city.

The applied methodology was able to show the spatiotemporal variability of UFP. Moreover, RMSE and ME were assessed.

Keywords: UFP, air quality, mixed-effect models, spatiotemporal

ACKNOWLEDGEMENTS

First, I would like to thank God for giving me enough strength to get over all the difficulties presented during the development of the research and for lighting up my path with his wisdom.

This MSc would not have been possible without the support of the Peruvian Navy, especially the Directorate of Hydrography and Navigation which I'm proud to belong to, and I will always be grateful with them for giving me this opportunity.

I owe my deepest gratitude to my supervisors who have guided me during this process. I want to show my gratitude to my advisor, MSc. Vera van Zoest for always be disposed to help me and to encourage me to overcome myself.

To my friends, I feel blessed to have met a fantastic bunch of people during my stay in Enschede. Definitely, you represented significant support during all the MSc program.

Finally, but not least, I dedicate this thesis to my family for giving me their unconditionally love, in particular to my dear aunt Dora who I saw for last time before starting the MSc, and now she is guiding me from heaven.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and problem statement

Nowadays it is known that air pollution represents one of the most significant health, environmental and social issues in the world. Additionally, the connection between air pollution and adverse health effects for humanity was already established and proved in the 1990s (Panis, 2010). Recent studies from World Health Organization (WHO) showed that nine out of ten people breathe air containing high levels of pollutants; the last estimations reveal a shocking death toll of seven million people every year caused by air pollution ("WHO," 2018).

Human exposure to air pollution is related to different types of illnesses. For instance, lung cancer can be attributed to the exposure to various PM components and different sources like cars combustion (Raaschou-Nielsen et al., 2016).

In this respect, it is essential to understand clearly the definition of air pollution which refers to a complex composition of both solid and gaseous pollutants. Accurate identification of the different contaminants of the air mixture contributing most to the health threat might have relevant implications for environmental, health and social policies, and for crucial decisions of local governments in taking steps to protect population health (Chen et al., 2012).

In this context, particulate matter (PM) is a pollutant of particular concern. Smaller-diameter particles referred to as PM2.5 (2.5 µm or smaller) are generally more hazardous for human health, as they can reach deeper into the small airways of the body ("WHO," 2018). Nevertheless, there is another particle even more dangerous than PM2.5, called ultrafine particles (UFP). UFP are particles with a diameter smaller than 0.1 µm. They can penetrate tissues and organs, posing an even greater risk of systemic health impacts ("WHO," 2018).

From the previous paragraphs can be inferred that there is the need to create tools and mechanisms to address this global issue. In that sense, it is required to have an accurate evaluation of air quality, its compliance with legal limits and its potential health impacts. Likewise, to know the pollution level and its spatiotemporal distribution. Therefore, it is required to have air quality sensor networks with proper spatial density making measurements continuously.

Despite the relevance of this worldwide issue which is affecting millions of people, there are not enough air quality sensors. As stated by van de Kassteele et al. (2009): "the increased distance between stations has caused a substantial loss of information and resulted in higher uncertainties in the maps." Furthermore, there is even a smaller number of UFP sensors since there is not a regulation in Europe that establishes UFP thresholds levels.

The rise of low-cost microsensors with the capacity to measure different components of air pollution has made more feasible the development of high-resolution mapping of air quality (Schneider et al., 2017). The recent emergence of low-cost microsensors triggers diverse projects to implement them in different cities. One of them is the city of Eindhoven where a civil initiative, AiREAS, has set up the sensor network (ILM—

Dutch: Innovatief Lucht Meetsysteem, English: Innovative Air Measurement System) and where the present study will be carried out.

The University of Twente through the Faculty of Geoinformation and Earth Observation (ITC) is involved in this civil project in the city of Eindhoven in cooperation with other scientific institutions. In this context, 35 low-cost air quality air boxes with various environmental sensors were installed in 2013 and are currently part of a network that measures Nitrogen Dioxide (NO_2), particulate matter (PM10, PM2.5, PM1), UFP and Ozone (O3) (AiREAS, 2014). This network makes it possible to obtain a better spatial distribution of data than traditional (high-cost) sensor networks.

Even though the air quality sensor network in Eindhoven has an acceptable spatial density, there are still spaces between the air stations without measurements and even more between UFP sensors since there are just six because they are more expensive than the other sensors. Nonetheless, during the research period, not all of them were working as will be explained in detail in Chapter 3.

Hence, it is necessary to model and map air pollution to have an approximation of it, in the entire city of Eindhoven and not just in the points where the stations are located. In that sense, the present study is focused on finding out a proper method for modelling and mapping UFP in the urban area of Eindhoven using the data of the air quality sensors network.

Finally, once a suitable method has been selected to modelling the UFP in the city of Eindhoven using data from the air quality sensors network, this work aims to find out whether five sensors are enough to study the spatial variability in the city. Likewise, to represent the result depicting a map of the air quality in space and time.

This result will create consciousness in Eindhoven citizens since they will be able to see the pollution in the city, which is not possible without a map that represents air quality. Likewise, the information will be beneficial for local governmental authorities to make important decisions to take care of citizen-health.

1.2. Research identification

The current research can be identified through the following objectives and questions.

1.2.1. Overall research objective

The primary objective of this study is modelling and mapping of hourly average ultrafine particles in the city of Eindhoven.

1.2.2. Specific research objectives

1. Find out whether five sensors rotated for periods of 21 days over eighteen different locations in total are enough to predict the spatiotemporal variability of UFP in the city of Eindhoven.
2. Identify a set of suitable covariates for the prediction of UFP in the study area.
3. Develop prediction maps of hourly average UFP in the city of Eindhoven.

1.2.3. Research questions

1. Are eighteen locations enough to predict the spatiotemporal variability of UFP concentrations in the city of Eindhoven?
2. Which covariates can be used to predict UFP concentrations in the study area?
3. Which method is suitable to predict UFP concentrations in space and time in the city of Eindhoven?

1.2.4. Innovation aimed at

UFPs are amongst the most dangerous pollutants for human health (“WHO,” 2018). However, previous studies have used data measured from mobile sensors or non-permanent sensor networks (Berghmans et al., 2009; Kwasny et al., 2010; Klompmaker, Montagne, Meliefste, Hoek, & Brunekreef, 2015; Farrell et al., 2016). To the best of my knowledge, there not study that has modellled and mapped this pollutant at an urban level using data from a permanent sensor network.

2. LITERATURE REVIEW

2.1. UFP and health effects

Air pollution is a composition of different noxious substances of various sizes and is believed that the small ones (UFP) are the most harmful for human health. UFP is a pollutant that has a diameter smaller than 0.1 µm and is related to road traffic emissions (Panis, 2010). Moreover, the primary source of UFP is the vehicle emissions, and highest concentrations of UFP are found nearby major roads (Farrell et al., 2016).

Another particularity of UFP is its dynamic nature which in contradistinction to other pollutants as PM_{10} or $PM_{2.5}$, can vary significantly in space and time at small scales. This variation depends on meteorological conditions as temperature, wind direction or wind speed (Kumar et al., 2014).

The small size of UFP makes it more likely to penetrate tissues and organs. Thus "UFP is deemed to be a major risk for human health" (Buonanno, Stabile, & Morawska, 2014). Different mortal diseases are directly related to exposure to air pollution. In the case of lung cancer, the risk is associated with various components of particulate matter (Raaschou-Nielsen et al., 2016).

Other studies have suggested that UFP concentrations are related to cardio-respiratory diseases (Hoek et al., 2010), and they are a possible causative agent for an increment of mortality with increases in UFP concentrations (Zhu, Hinds, Shen, & Sioutas, 2010).

2.2. Related work

Many studies have been carried out to model and map air pollution in different areas around the world. However, most of them are focused on pollutants like NO_2, PM10, and PM2.5. For instance, Hamm, Finley, Schaap, & Stein (2015) presented a study on modelling and mapping PM10 concentrations in Europe, whereas a research of NO_2 concentrations in Oslo, Norway was carried out by Schneider et al. (2017) in which they also used data of low-cost sensors.

There are also some studies based on modelling and mapping UFP such as the estimation of the exposure of cyclists to UFP in the town of Mol in Flanders, Belgium using mobile sensors installed on the bicycle (Berghmans et al., 2009). They fixed the sensors in a bicycle, and the sampling path was selected considering different parts of the city as major roads and residential areas. The results indicates a a high variability of UFP even within the same transport microenvironment. Furthermore, since it is a relative small city, the concentrations of UFP were smaller than in major cities around the world as Amsterdam (The Netherlands), Los Angeles (USA) or Zurich (Switzerland).

There is also a more recent study presented by Weichenthal et al. (2016) in which a land use regression model was developed for UFP in Montreal, Canada using mobile monitoring data during the summer and winter months between 2011 and 2012. This research applied multivariable linear regression model and kernel-based regularized least squares (KRLS). The results showed that KRLS approach explained better the UFP distribution.

Hoek et al., (2010) developed a research to characterize UFP in Amsterdam, the Netherlands applying a land use regression model. They measured UFP outside of 50 homes around the city for one week in different periods, and continuous measurements close to the city center. Covariates as land use, address density, traffic intensity, and others were used to model UFP. However, the product of traffic intensity and inverse distance to nearest major road was the most significant predictor.

Ragettli et al., (2014) built multivariate regression models to predict UFP in the city of Basel, Switzerland. They sampled UFP in 60 locations among the city for periods of 20 minutes. The results showed a higher averaged concentration of UFP in winter than in summer and spring. They demonstrated that short-term measurements surveys are effective to characterize short-term UFP concentrations. Nonetheless, correlations between short and long term UFP concentrations requires to be assessed in the future.

Zhu et al., (2010) developed a study to compare the UFP concentrations between winter and summer in Los Angeles. They found a significantly higher UFP concentration in winter than in summer which suggested a weaker atmospheric dilution in winter. This effect contributes to a major accumulation of particles in the air, increasing UFP concentrations.

2.3. Modelling and analysis

The main goal of this study is to model and map the ultrafine particles in the city of Eindhoven and to do so it is necessary to apply different techniques considering the distribution of the data based on descriptive statistics.

To this end, geostatistical methods will be used during the development of the thesis, being used for the data analysis and also for the modelling and mapping of the ultrafine particles. For instance, Diggle & Ribeiro Jr. (2007) presented basic concepts and more advanced techniques about geostatistical data like the covariance functions and the variogram, different approaches for model parameters estimations (e.g., maximum likelihood) and spatial prediction methods (e.g., kriging).

For spatial data quality of pollutants, there are different methodologies like a novel outlier detection presented by van Zoest, Stein, & Hoek (2018) which is focused on NO2 that also has high spatial variability in urban areas. On the other hand, Wang et al. (2017) focused their study on spatiotemporal variation of PM2.5 and carbon monoxide (CO2), and they dropped all the data that was out of $\pm 3\sigma$ of the mean.

It is also important to know which external factors influence the spatial variability of the UFP. Regarding this, we know that the meteorological parameters like temperature, relative humidity, wind direction, and wind speed might have a significant influence on the spatial variation of the pollutants in the air (Jayamurugan, Kumaravel, Palanivelraja, & Chockalingam, 2013). Likewise, other pollutants as NO_2 can be used to explain UFP mostly in winter (Kwasny et al., 2010). Other spatial covariates are also valid to predict UFP as the distance from the nearest highway or distance from the nearest restaurant as used by Farrell et al. (2016).

Since the observations were taken in different periods,they will be cosidered as a random effect and the other covariates as a fixed effect. In order to include both effects into the model, the mixed effect model method will be used. Regarding this, several studies based on mixed effect model has been reviewed like "Model selection in linear mixed effect models" in which a simple procedure was developed for estimation and selection of fixed and random effects for linear mixed models (Peng & Lu, 2012). Another research based on linear mixed effect models was carried out to mapping nighttime PM2.5 concentrations in Beijing (Fu et al., 2018). More related to the present research, a linear mixed model was developed to modelling and mapping UFP in space and time, using spatial and temporal covariates (Farrell et al., 2016).

To carry out the complete analysis, R software version 3.4.3 was used for data pre-processing and for developing the methodology to get the predicted values of UFP, including the resulting final maps (R core team, 2017). To do so, the main package that will be used is GeoR version 1.7-5.2 which contains a great variety of geostatistical functions (Ribeiro Jr. and Diggle, 2016). This package includes useful functions for reading and preparing the data, exploratory analysis applying descriptive statistics, inference on model parameters including variogram based and likelihood-based methods which will be used to get the coefficients of the model to predict UFP, and spatial interpolation (Jr., Christensen, & Diggle, 2003).

3. STUDY AREA AND DATA DESCRIPTION

3.1. Study area

The present study aims to predict the spatial and temporal variability of UFP in the city of Eindhoven, The Netherlands. The study area is delimited by the border of the municipality of Eindhoven. In Figure 1 we can observe the eighteen locations where the UFP sensors were rotated during the study period.

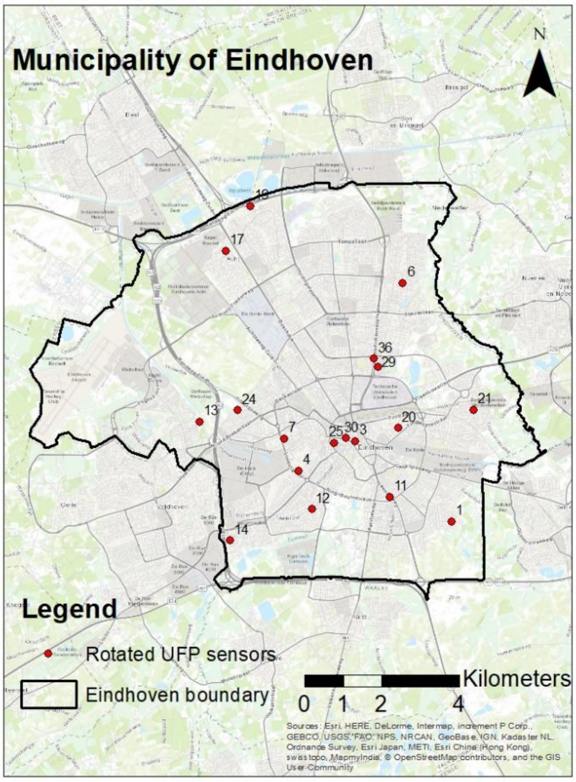

Municipality of Eindhoven
N
19
17
6
36
29
21
24
13
7
25
30
3
20
4
11
12
1
14
Legend
Rotated UFP sensors
Eindhoven boundary
Kilometers
0
1
2
4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 1. Map of the municipality of Eindhoven with the eighteen locations over the UFP sensors were rotated.

3.2. Data description

3.2.1. Sources

The data used for this research has been taken from two sources: ILM and the Royal Netherlands Meteorological Institute (KNMI) and consist in UFP as the response variable and PM_1, temperature, relative humidity, wind speed and wind direction as covariates.

ILM is a network composed of 35 airboxes spread in the city of Eindhoven. Each airbox has sensors of PM_{10}, $PM_{2.5}$, PM_1, temperature, relative humidity, ozone and just six of them have a sensor of UFP. The data used from this source is UFP, PM_1, temperature, and relative humidity. In Figure 2, we can observe one airbox of the ILM network in the city of Eindhoven, where the UFP sensor is installed inside the small white box.

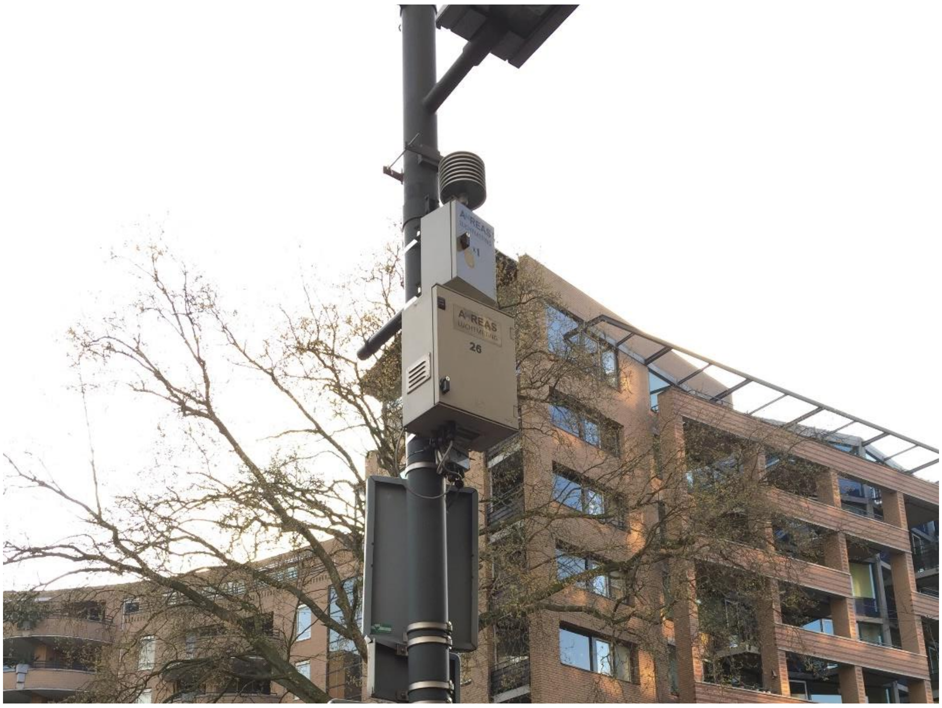


Figure 2. ILM airbox installed in the city of Eindhoven.

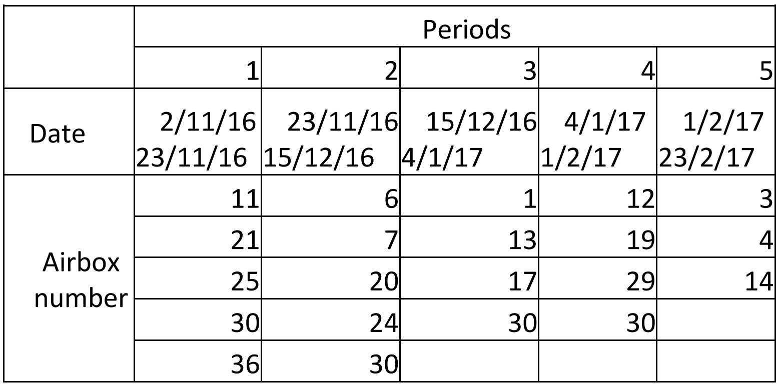
KNMI is the Dutch national weather service and the data obtained from that source is considered highly reliable. The data used from this source is wind speed and wind direction.

3.2.2. Variables

UFP: This is the main variable in this study since it is the one that will be predicted, also known as the response variable. UFP-unit is counts per cubic centimeter (Counts/cm^3).

Sensors of UFP are usually fixed in six locations in Eindhoven. However, they were rotated over eighteen different airboxes during 113 days in five different periods of rotation as we can see in Table 1. As it is shown in Figure 2, the UFP sensor is installed inside a separate box which makes its transport over different locations easy. When the rotation started one of the sensors was out of order, so just five sensors were available to get the data from.

Additionally, during the rotation periods, two more UFP sensors had problems, finally having available only three sensors as can be seen in Table 1.



	Periods				
	1	2	3	4	5
Date	2/11/16 23/11/16	23/11/16 15/12/16	15/12/16 4/1/17	4/1/17 1/2/17	1/2/17 23/2/17
Airbox number	11	6	1	12	3
	21	7	13	19	4
	25	20	17	29	14
	30	24	30	30	
	36	30			

Table 1. Periods of UFP sensors-rotation among the airboxes

In Appendix 1, we can observe maps of where the sensors were located during each of the five different periods.

Covariates:

PM_1: There are 35 sensors of PM_1 in the city of Eindhoven installed in the airboxes of the ILM network. This variable is measured in micrograms per cubic meter ($\mu g/m^3$).

Temperature: The temperature sensors are located in 35 airboxes in the city of Eindhoven that belong to the ILM network. This variable is measured in degrees Celsius (^{0}C).

Relative humidity: There are also thirty-five sensors and are located in the same airboxes where PM_1 and temperature sensors are located. This variable is measured in percentage (%).

Wind speed: There is just one sensor for the whole study area. This variable is measured in kilometers per hour (km/h).

Wind direction: There is just one sensor for the whole study area. This variable is measured in north azimuth degrees (0-359), but it was classified to wind rose directions and treated as a factor.

Periods: It refers to the five periods in which the UFP sensors were rotated. This covariate will be treated as a factor and as a random effect since the values of UFP, and the other covariates might have differences between periods. If it is true that every period differs from each other in time, they also have another particularity; every period is related to a different area in which the UFP sensors were located during the corresponding period as we can observe in Appendix 1. Hence, periods differ in time and space.

UFP, PM_1, relative humidity, and temperature were downloaded from the AiREAS website ("Download - ECN dustmonitoring Aireas," 2018). While wind speed and wind direction were downloaded from the KNMI website ("KNMI - Weather data from the Netherlands - Download," 2018).

All the variables mentioned before were downloaded considering an hourly average.

After introducing the problem statement, the research objectives, and data and variables description, in the next chapter, the methodology applied to develop the present study and to reach the research objectives is presented.

4. METHODOLOGY

The methodology used in the present research can be explained by the flowchart in Figure 3. The initial input, which is the response variable (UFP) and all the initial covariates (PM_1, relative humidity, temperature, wind direction, and wind speed) went through a pre-processing analysis to clean the data, removing possible errors and outliers. The Pearson correlation test was applied between the covariates and the response variable, leaving out the covariates that present a low correlation and significance.

The cleaned data and the selected covariates were split into 24 hourly timestamps before applying the multiple regression linear analysis with the aim of removing the temporal trend between hours. However, there was still a temporal trend between the different days, which will not be considered in the present study due to time limitation.

Hereafter, the term timestamp will be referred to as the hourly classification of the data considering each timestamp as a set of observations (Log-transformed UFP and the covariates) taken in different days but at the same time (e.g., 0700, 0800, etc.). It means that in the end, 24 timestamps were obtained in total.

The multiple regression linear analysis was applied to every timestamp with the aim of getting the model coefficients required for predicting UFP. This process was useful to verify the significance of each covariate and find out again which ones should be considered for further analysis.

When the covariates were defined, they were prepared to have their respective values in every cell raster into the study area. Furthermore, ordinary Kriging was applied to the residuals to get their values in the entire research area as well.

Knowing the models' coefficients, and the covariates and residuals values; the 24 different models corresponding to each hourly timestamp were executed to get the maps of predicted UFP in the study area.

Finally, the models were validated using the mean error (ME), and root mean square error (RMSE) approaches.

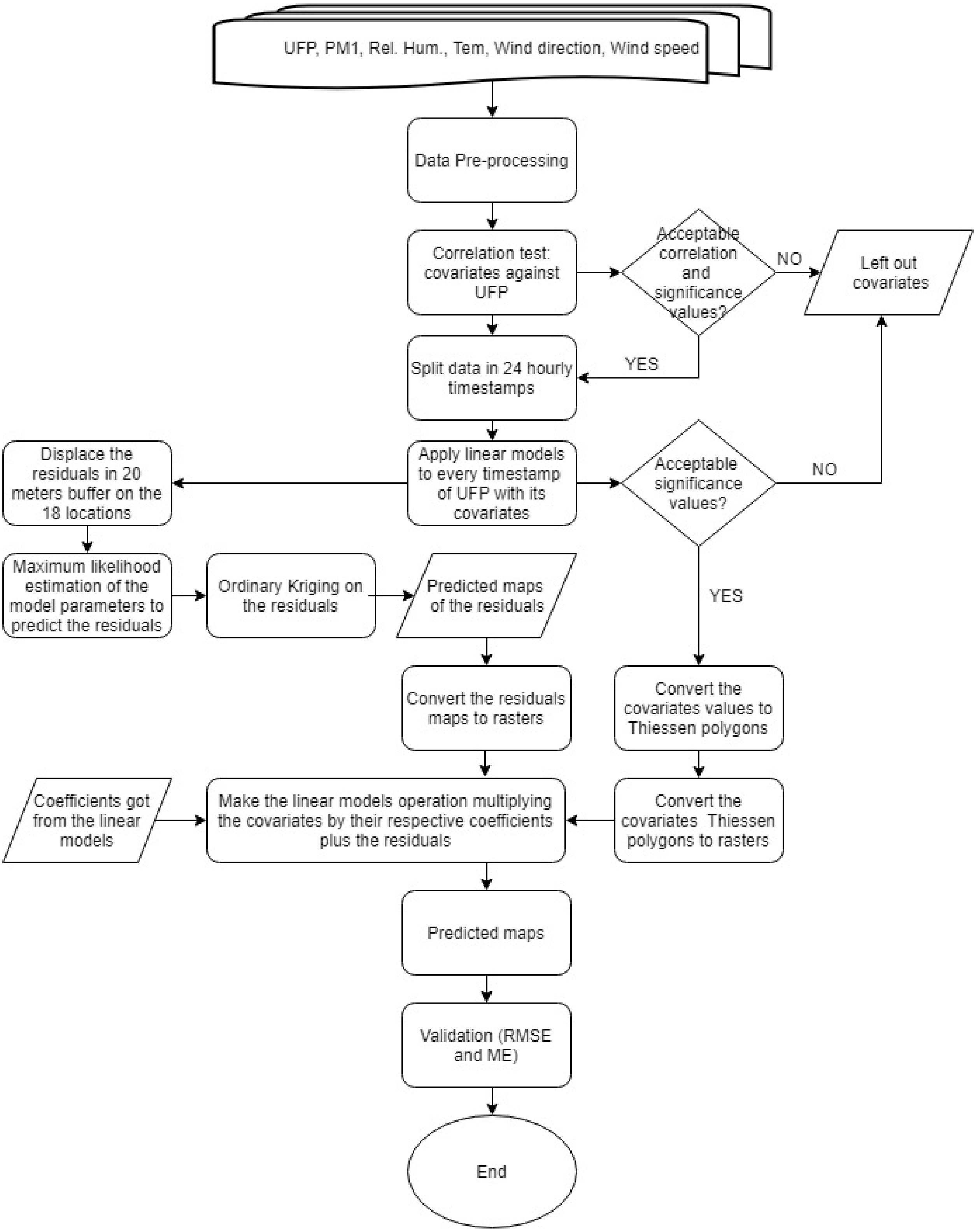

UFP, PM1, Rel. Hum., Tem, Wind direction, Wind speed
Data Pre-processing
Correlation test: covariates against UFP
Acceptable correlation and significance values?
NO
Left out covariates
YES
Split data in 24 hourly timestamps
Apply linear models to every timestamp of UFP with its covariates
Acceptable significance values?
NO
YES
Displace the residuals in 20 meters buffer on the 18 locations
Maximum likelihood estimation of the model parameters to predict the residuals
Ordinary Kriging on the residuals
Predicted maps of the residuals
Convert the residuals maps to rasters
Convert the covariates values to Thiessen polygons
Convert the covariates Thiessen polygons to rasters
Coefficients got from the linear models
Make the linear models operation multiplying the covariates by their respective coefficients plus the residuals
Predicted maps
Validation (RMSE and ME)
End

Figure 3. Flowchart of the methodology

4.1. Data pre-processing

Data quality is a concept that embraces more than data accuracy since it is already known that it also consists of completeness, consistency, and currency which are relevant characteristics of quality of data (Batini & Scannapieca, 2006).

The data used in the present study consist of different variables as is explained in Chapter 3. However, inspecting the data just by checking at it, is difficult to read and analyze due to its size. In that sense, descriptive statistics offer a great range of tools and procedures like graphics and mathematical calculations that make the data readable and interpretable (Miles & Banyard, 2008).

The following procedures and calculations of descriptive statistics were used to analyze the data:

- Histogram
- Box plot
- Q-Q plot
- Timeline
- Mean, quantiles, standard deviation, maximum and minimum value, etc.

As a result, we found outliers since devices that usually have mobility such as the UFP sensors, and are exposed to different environmental conditions, tend to present outliers (Cho & Choi, 2017).

Finally, the outliers were removed considering a suitable criterion according to the data distribution.

4.2. Correlation between covariates and response variable

The Pearson correlation coefficient was used to assess the correlation degree between the covariates and the response variable.

This coefficient can be defined as "the covariance of the two variables divided by the product of their standard deviations" (Zhou, Deng, Xia, & Fu, 2016).
The Pearson correlation coefficient measures the degree at which variables are linearly dependent. This coefficient has a value within the range of -1 and 1.

Since periods and wind direction are covariates treated as factors, it will be not possible to assess their correlation against the response variable. Instead, their significance were checked while the multiple linear regression analysis was applied.

4.3. Linear model

A multiple linear regression model analysis was applied to establish the linear relation between the response variable and the covariates to get the respective coefficients. Furthermore, the significance values of the covariates were checked to determine which of them explain better the response variable and which of them should be removed. The linear regression model analysis was applied after splitting the data into 24 hourly timestamps. Hence, 24 different set of coefficients were calculated, one for every timestamp.
The multiple linear regression model can be represented by the following equation:

$$y_{p,s,d,h} = \beta_{0\ p,h} + \sum_c \beta_{c,p,h} x_{c,p,s,d,h} + \varepsilon_{p,s,d,h} \qquad (1)$$

Where:

$y_{p,s,d,h}$ = Response variable (log-transformed UFP) at period p, location s, day d, and timestamp h
$\beta_{0\ p,h}$ = Intercept at period p, and timestamp h
$\beta_{c,p,h}$ = Linear parameter estimates for covariate c at period p, and timestamp h
$x_{c,p,s,d,h}$ = Covariate c at period p, location s, day d and timestamp h
$\varepsilon_{p,s,d,h}$ = Residuals at period p, location s, day d, and timestamp h

After applying the multiple linear regression model to every timestamp, we got the model coefficients and residuals considering the hourly temporal variation since the data were split in hourly timestamps. Likewise, the significance level of the obtained coefficients was tested. The model obtained was the following:

$$\hat{y}_{p,s,h} = \hat{\beta}_{0,p,h} + \hat{\beta}_{hum,p,h}x_{hum,p,s,h} + \hat{\beta}_{pm,p,h}x_{pm,p,s,h} + \hat{\beta}_{wd,p,h}\,x_{wd,p,h} + \hat{\beta}_{ws,p,h}x_{ws,p,h} + \varepsilon_{p,s,h} \quad (2)$$

Where:

$\hat{y}_{p,s,h}$ = Predicted log transformed UFP at period p, location s, and timestamp h
$\hat{\beta}_{0,p,h}$ = Interception at period p, and timestamp h
$\hat{\beta}_{hum,p,h}$ = Estimated coefficient for relative humidity at period p, and timestamp h
$\hat{\beta}_{pm,p,h}$ = Estimated coefficient for log-transformed PM$_1$ at period p, and timestamp h
$\hat{\beta}_{wd,p,h}$ = Estimated coefficient for wind direction at period p and timestamp h
$\hat{\beta}_{ws,p,h}$ = Estimated coefficient for wind speed at period p and timestamp h
$x_{hum,p,s,h}$ = Relative humidity value at period p, location s, and timestamp h
$x_{pm,p,s,h}$ = Log-transformed PM$_1$ value at period p, location s , and timestamp h
$x_{wd,p,h}$ = Wind direction at period p, and timestamp h
$x_{ws,p,h}$ = Speed direction value at period p, and timestamp h
$\varepsilon_{p,s,h}$ = Residuals at period p, location s, and timestamp h

As we can observe in Equation 2, the temperature variable was removed from the model because in most of the timestamps it was not significant

4.4. Ordinary Kriging of the residuals

Once the data were split in timestamps, it was necessary to select a proper method to estimate the variogram parameters (sill, range, and nugget) before applying Kriging to predict the residuals in the study area. To do so, it is essential to consider that UFP has a small range.

The empirical variogram estimator has a limitation since the selection of lag classes, or bins is very critical because if they are too narrow then the estimated variogram will be noisy, but if the classes are too broad then it will lead a loss of the spatial structure of the variable (Lark, 2000).

On the other hand, the maximum likelihood estimator (MLE) estimates better the variograms of fields with a weak spatial structure (short range and small spatial dependence) than the empirical variogram estimator (Lark, 2000). The MLE seems to be the most appropriate approach to estimate the variogram of the residuals. Consequently, this approach was chosen because the fitted model takes into consideration all the pairs of distances between points, which makes it possible to model spatial autocorrelation in a small range.

Maximum likelihood estimation approach determines values for the parameters of the variogram such that they maximize the likelihood that the process described by the model produced the observed data (Brooks-Barlett, 2019). The MLE can be simplified to its natural logarithm, which is the log-likelihood function to be maximized to get the parameters estimates.

$$L(\theta_h) = -\frac{1}{2}\{n_h \log(2\pi) + \log\{|\boldsymbol{C}(\boldsymbol{\theta_h})|\} + \boldsymbol{\varepsilon}_{\boldsymbol{h}}^T(\boldsymbol{C}(\boldsymbol{\theta_h}))^{-1}\boldsymbol{\varepsilon_h}\} \qquad (3)$$

Where:

$L(\theta_h)$ = Log-likelihood function to estimate set of parameters θ at timestamp h
θ_h = Parameters (sill, nugget, and range) to be estimated at timestamp h
$\boldsymbol{C}(\boldsymbol{\theta_h})$ = Covariance matrix of θ at timestamp h
$\boldsymbol{\varepsilon_h}$ = n length vector of residuals at timestamp h, where $\boldsymbol{\varepsilon_h} \sim N(0, \boldsymbol{C}(\boldsymbol{\theta_h}))$
n_h = Number of observations at timestamp h defined by the number of locations multiplied by the number of days

Likewise, the covariance of θ at timestamp h can be expressed in terms of the nugget, sill, and range:

$$\boldsymbol{C}(\boldsymbol{\theta_h}) = \sigma_h^2 R(\phi_h) + \tau_h^2 \boldsymbol{I} \qquad (4)$$

Where:

$\boldsymbol{C}(\boldsymbol{\theta_h})$ = Covariance matrix of θ at timestamp h
σ_h^2 = Sill of the variogram at timestamp h
ϕ_h = Range of the variogram at timestamp h
τ_h^2 = Nugget of the variogram at timestamp h
R = Exponential function which determines the shape of the variogram
$\boldsymbol{I}$ = Identity matrix

After estimating the variogram parameters of the residuals, ordinary Kriging was applied to predict them in the entire study area. Ordinary Kriging is an interpolation method for spatial data (Diggle, Tawn, & Moyeed, 1998) with constant spatial mean and can be written as follows:

$$\hat{\varepsilon}_{s_0,h} = \sum w_{s_0,h}\, \varepsilon_{s,h} \qquad (5)$$

Where:

$\hat{\varepsilon}_{s_0,h}$ = Predicted residual at any unobserved locations s_0, and timestamp h
$w_{s_0,h}$ = Kriging weight for residual at location s, and timestamp h derived from the estimated spatial mean and covariance of the given data
$\varepsilon_{s,h}$ = Residual at timestamp h and location s

4.5. Preparation of the covariates

Once the covariates have been selected, they must be prepared before being considered as inputs of the model. The first step was average all the values at every location and timestamp, obtaining 24 values (one for every hourly timestamp) in every position. However, it was possible just for the covariates that have numeric values but not being treated as factors.

Later, Thiessen polygons were generated to obtain the covariates values in the entire study area for every timestamp. Finally, they were converted to rasters, considering the same cell size as the one selected for the predicted maps. In the end, 24 rasters (one for every timestamp) were created for all the covariates. The covariates treated as factors were also converted to rasters with the same cell size of the predicted maps, and that process will be explained in detail in Chapter 6.

4.6. Predicted maps

To assess the hourly temporal variation during the day, 24 maps were generated, one for every timestamp using the obtained model after applying the multiple linear regression analysis.

A suitable cell size was chosen for the predicted maps according to the following criterion:

- Cell size smaller than the UFP range, to be able to observe these variations.
- A small enough cell size that allows observing variations of UFP in the urban area of Eindhoven.
- A cell size that allows running the model with the computational sources available for this study.

Once the model has been set, we can split it into two parts:

$\beta_{0\,p,h} + \sum_c \beta_{c,p,h} x_{c,p,s,d,h}$ = Non-constant mean at timestamp h
$\varepsilon_{p,s,d,h}$ = Residuals at timestamp h

The mean was calculated multiplying the covariates by their respective coefficients. However, it is necessary to have the covariates values in every single cell of the raster of the study area. To do so, Thiessen polygons were generated for all the covariates, and afterwards, they were converted into rasters with the cell size selected for the predicted maps.

The residuals from the multiple linear regression analysis were predicted in the whole study area using ordinary Kriging. Since there are 24 different models, there are also 24 sets of residuals, one for every timestamp.

Finally, the resultant raster of the non-constant mean was summed with the residual raster for every timestamp, getting the 24 prediction maps of log-transformed UFP, then a back-transformation was applied to obtain the UFP values for predicted maps.

4.7. Validation

Once the 24 UFP predicted maps were obtained, the mean error (ME) and root mean square error (RMSE) were used to validate the models for every timestamp.

The validation was applied using two vectors of the same size, the first one with all the observations values and the second one with the predicted values in the corresponding locations where the observations were taken at every timestamp.

The ME refers to the average of the sum of all differences between observations and predicted values. It can be represented with the following equation:

$$ME_h = 1/n_h \sum_{i=1}^{n} y_{h,s,d} - \hat{y}_{h,s} \qquad (6)$$

Where:

ME_h = Mean error at timestamp h
n_h = Number of observations at timestamp h, it is equal to the number of locations multiplied by the number of days
$y_{h,s,d}$ = Observation of log-transformed UFP at timestamp h, location s, and day d
$\hat{y}_{h,s}$ = Predicted value of log-transformed UFP at timestamp h, and location s

The ME was used to detect if there is a bias in the model.

The RMSE can be expressed as follows:

$$RMSE_h = \sqrt{1/n_h \sum_{i=1}^{n} (y_{h,s,d} - \hat{y}_{h,s}})^2 \qquad (7)$$

Where:

$RMSE_h$ = Root mean square error at timestamp h
n_h = Number of observations at timestamp h, it is equal to the number of locations multiplied by the number of days
$y_{h,s,d}$ = Observation at timestamp h, location s, and day d
$\hat{y}_{h,s}$ = Predicted value at timestamp h, and location s

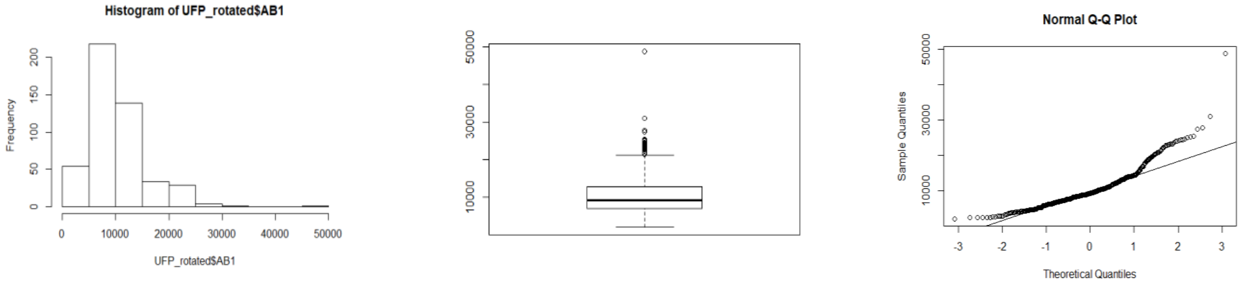
5. DATA PRE-PROCESSING

5.1. UFP

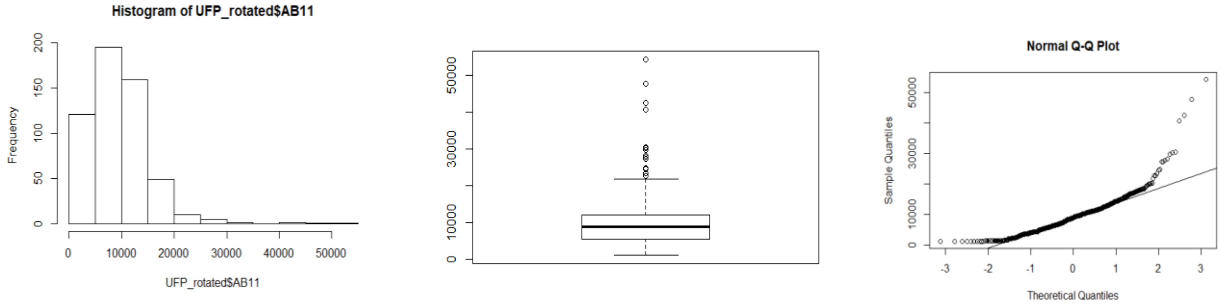
First, the summary of the data was observed checking at the minimum and maximum value, quantiles, median and mean for the 18 airboxes, finding a high variability and outliers. This data summary can be observed in Appendix 2.

The next step was to visualize the data using graphics like histograms, boxplots, and Q-Q plots, so three sets of data (3 airboxes) were selected to confirm what was already observed in the data summary as it is noted in Figure 4.

Airbox 1:



Airbox 11:



Airbox 30:

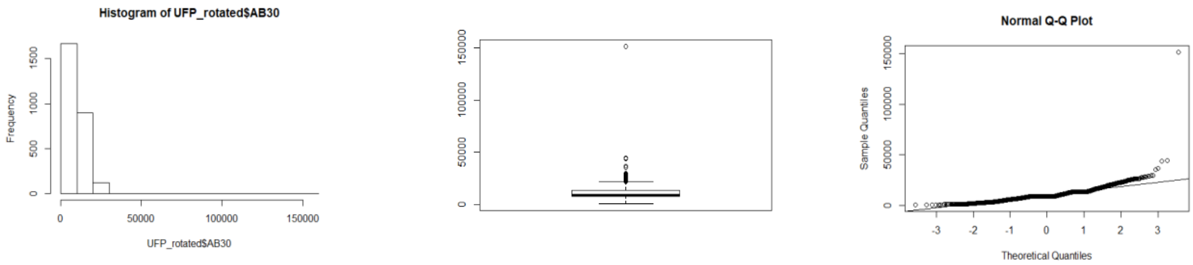


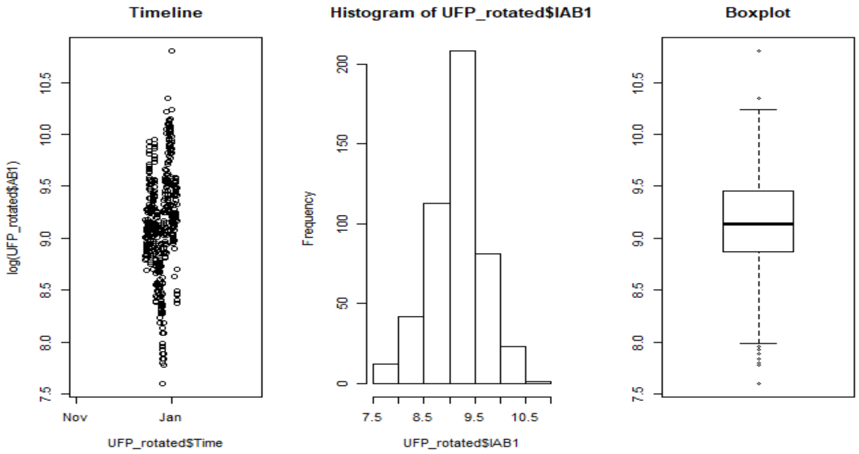
Figure 4. Descriptive statistics of UFP (Histogram, boxplot, and Q-Q plot)

The graphics confirmed that the UFP data is not normally distributed and very skewed. Hence, a log transformation was applied. Likewise, we can see outliers checking at the boxplots.

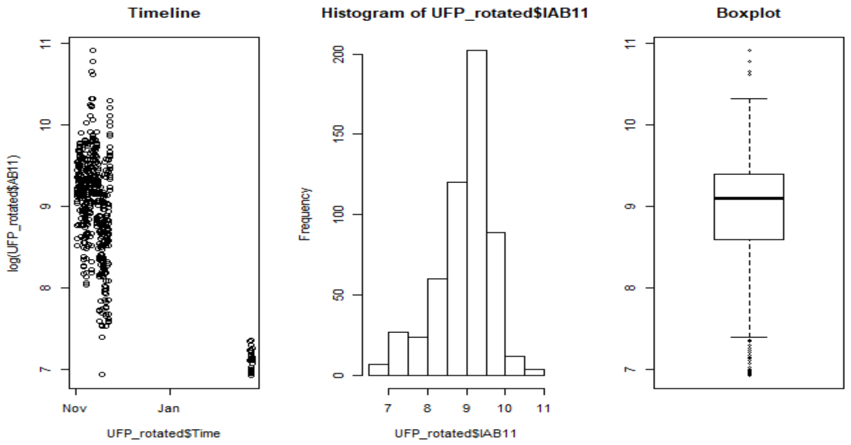
After the log transformation was applied and continuing with the descriptive statistics, timelines, histograms, and boxplots were checked to find out if the data has a normal distribution or at least presents fewer outliers and look less skewed.

Figure 5 shows that the data now is more similar to a normal distribution and there are fewer outliers than in the non-transformed data. Hence, the log-transformed data will be used to get the predicted maps. The complete list of mentioned graphics can be found in Appendix 3.

Airbox 1:



Airbox 11:



Airbox 30:

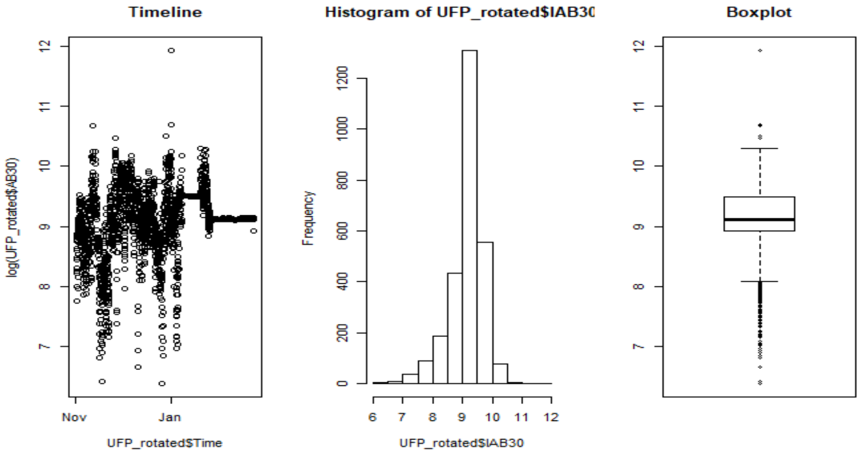


Figure 5. Descriptive statistics of log-transformed UFP (Timeline, histogram, and boxplot)

Nevertheless, we can still observe outliers and also some data that apparently does not present any variability during some periods, but the dynamic nature of UFP makes it highly temporal and spatial variable (Kumar et al., 2014). Hence, it is inferred that those periods in which UFP seems to present a lack of variability, are errors during the measurement. This sort of error will be called flat data in the present study for purposes of practicality, and they were deleted from the dataset.

In order to remove the outliers, I considered removing all the observations above (μ+3σ) and below (μ-3σ), but in that case, the number of removed observations would be more than expected, and probably most of them are valid observations, hence, to avoid eliminating correct observations I decided to use (±4σ) criterion.

Considering that criterion, the inferior and superior limits established for the observations are shown in Figure 6 (just the airboxes that have outliers are represented by a timeline graphic):

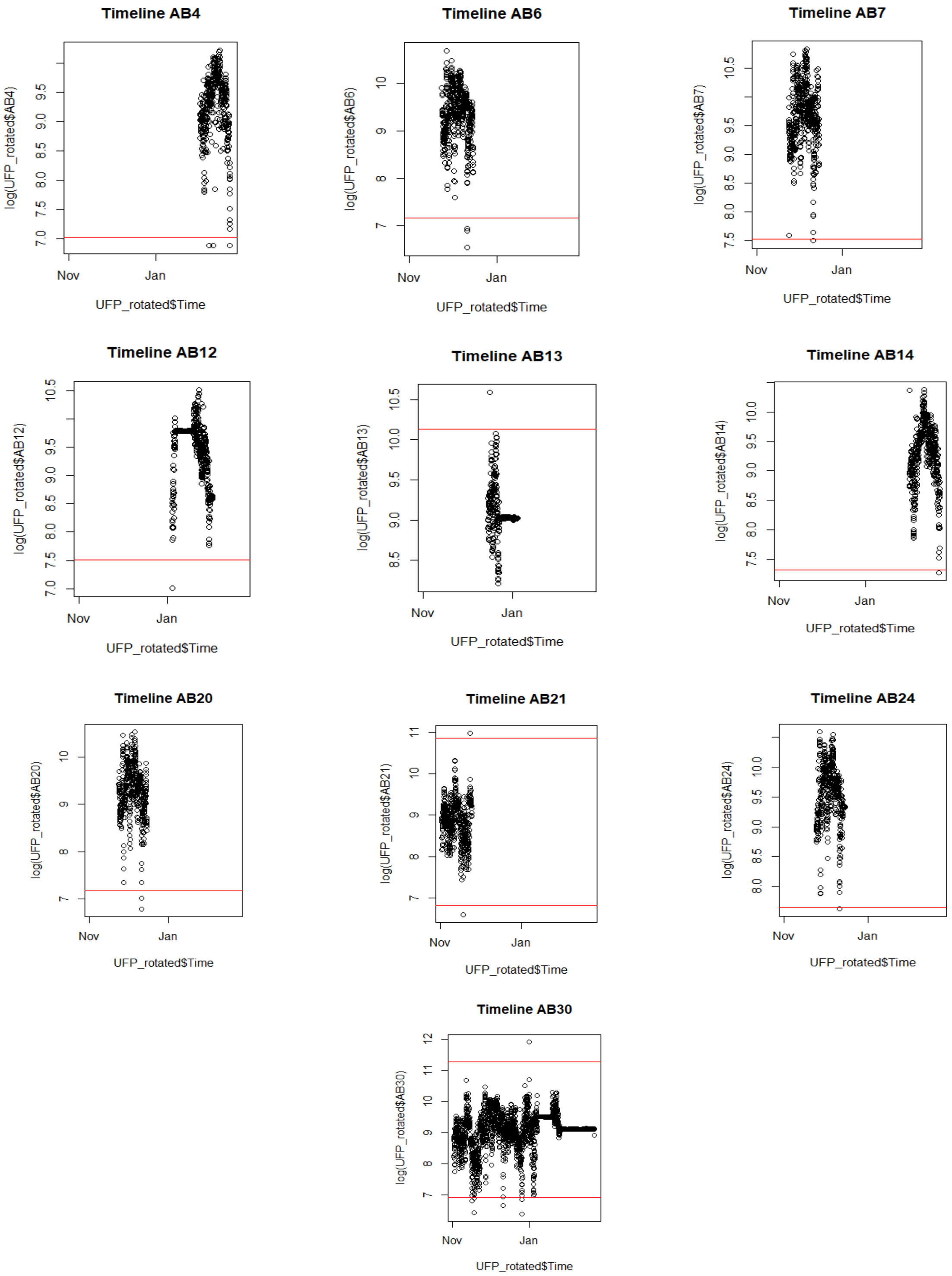


Figure 6. Outliers detection of log-transformed UFP

After removing the outliers, there was still remaining flat data in airboxes 3, 12, 13 and 30. So, checking at the timelines, periods that correspond to flat data were identified. The next step was to remove the flat data of those identified periods.

5.2. Covariates

The procedure applied to UFP was also used to check the data quality of the covariates, removing errors in the data. However, outliers were not found but flat data was removed.

In the case of relative humidity of airbox 11, the data presented illogical values during the whole period, for that reason the entire dataset was removed.

Regarding wind direction, 0 degrees is very close to 359 degrees, but the software would interpret 0 degrees very far from 359 degrees. Hence, to avoid that mistake, the data were categorized as it is shown in Table 2.

Wind direction categories
North
North-east
East
South-east
South
South-west
West
North-west
Calm/variable

Table 2. Wind direction categories

6. RESULTS AND ANALYSIS

6.1. Correlation between covariates and response variable

The results of Pearson correlation between candidate covariates and log-transformed UFP are shown in Table 3.

Correlation between covariates and log-transformed UFP				
Air boxes	log-transformed PM1	Temperature	Rel. Humidity	Wind speed
AB1	0.63	-0.56	-0.11	-0.49
AB3	0.55	-0.22	-0.31	-0.51
AB4	0.68	-0.53	-0.17	-0.28
AB6	0.63	-0.35	0.06	-0.29
AB7	0.48	-0.23	-0.24	-0.48
AB11	0.70			-0.57
AB12	0.47	-0.52	-0.36	-0.50
AB13	0.46	-0.14	-0.34	-0.25
AB14	0.79	-0.54	-0.12	-0.18
AB17	0.58	-0.51	-0.09	-0.41
AB19	0.49	-0.30	-0.27	-0.49
AB20	0.56	-0.40	-0.07	-0.41
AB21	0.70	-0.33	0.10	-0.41
AB24	0.54	-0.41	-0.18	-0.37
AB25	0.70	-0.36	0.08	-0.54
AB29	0.58	-0.33	-0.30	-0.54
AB30	0.65	-0.42	-0.12	-0.41
AB36	0.61	-0.20	-0.02	-0.55

Table 3. Correlation values between candidate covariates and log-transformed UFP

Temperature and relative humidity values in airbox 11 presented illogical values during the whole period of interest, and that data was removed, which will not have a significant influence on the final predicted maps since there are 34 more sensors of those variables in the study area.

In most cases, we can observe an acceptable correlation. However, the correlation values of relative humidity are very low in comparison with the other covariates. Nevertheless, its significance value is acceptable; consequently, relative humidity was regarded as a covariate.

Since the wind direction was categorized as a factor, it was not possible to assess its correlation with log-transformed UFP. However, it was considered as a covariate because in further steps it was verified that it had an acceptable significance level when the linear regression models were applied.

6.2. Residuals of 24-hourly timestamps

The R-Squared value of the 24 models is between 0.4 and 0.7 as can be observed in Figure 7, which I consider an acceptable value.

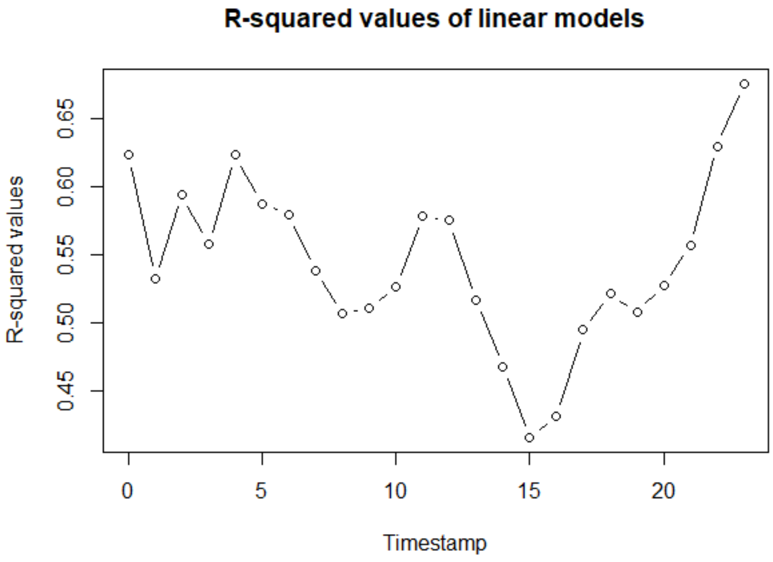


Figure 7. R-squared values of every timestamp

After applying the linear regression model to get the residuals and the model coefficients, residuals were included to the 24 different data frames (split by hourly timestamps) and with the aim of avoid computational problems and to have locations close enough to observe spatial autocorrelation on the residuals, the observations within the same airbox were artificially displaced in a buffer of 20 meters.

6.3. Ordinary Kriging to the residuals

Before applying ordinary Kriging to predict the residuals in the study area, maximum likelihood estimator was used to estimating the variogram parameters of the residuals for every timestamp as we can see in Figure 8 and Figure 9.

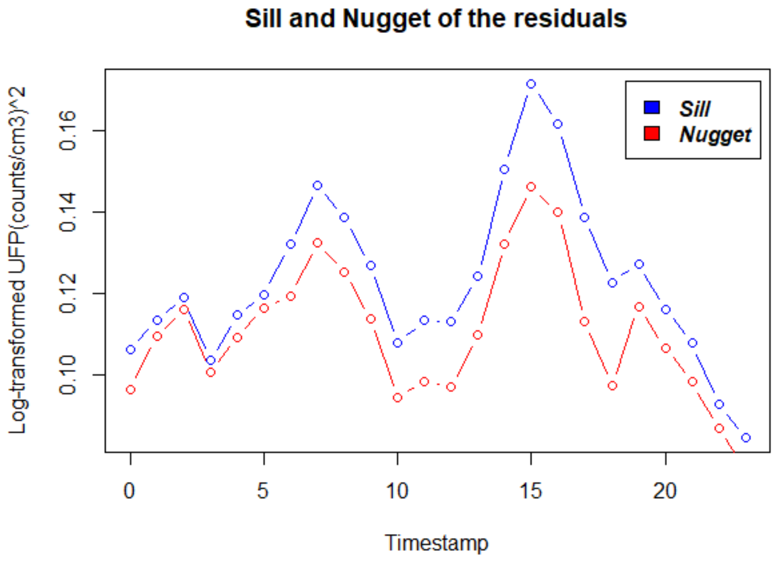


Figure 8. Nugget and sill of the residuals for every timestamp

Figure 8 shows that the nugget and sill have the same pattern for every timestamp, and they present very similar values, being the partial sill significantly smaller than the nugget. Hence, the variability is mainly explained by the nugget.

On the other hand, Figure 9 does not show a pattern in the variation of the range among the timestamps. However, the lowest range is generated at 04:00.

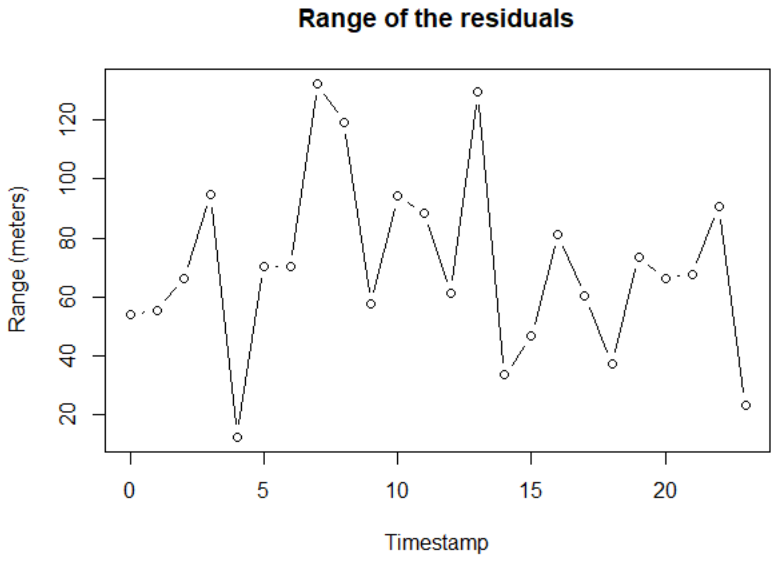


Figure 9. Range of the residuals for every timestamp

After getting the model parameters with maximum likelihood estimation, ordinary kriging was applied to the residuals. Since in most of the timestamps the range of the model presented a value between 50 and 100 meters, the selected cell size for the predicted maps was 50 meters, regarding also the available computational capacity.

In Figure 10, we can observe the raster map of the predicted residuals of the timestamp 07:00.

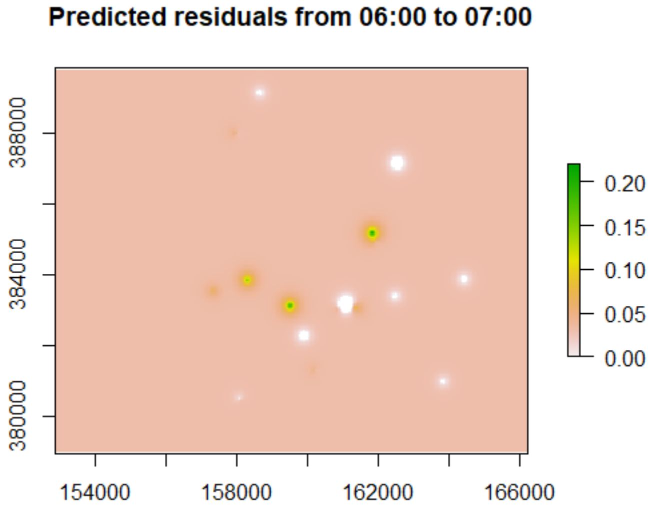


Figure 10. Raster map of the predicted residuals at timestamp 07:00

6.4. Resultant covariates as inputs of the model

A group of five covariates has already been selected with the linear model analysis, leaving out temperature due to its non-significant value. Log-transformed PM_1, relative humidity, wind direction, and wind speed are considered as fixed effect covariates while, while periods are regarded as random effect. The reason why periods are viewed as random effect is that UFP were taken in five different periods according to Table 1, so different conditions between periods might influence the UFP values. The next step was to convert the covariates values in rasters for every timestamp regarding the same cell size selected for the residuals (50 meters). To do so, Thiessen polygons were generated to get the values of the covariates in the whole study area as we can see in Figure 11.

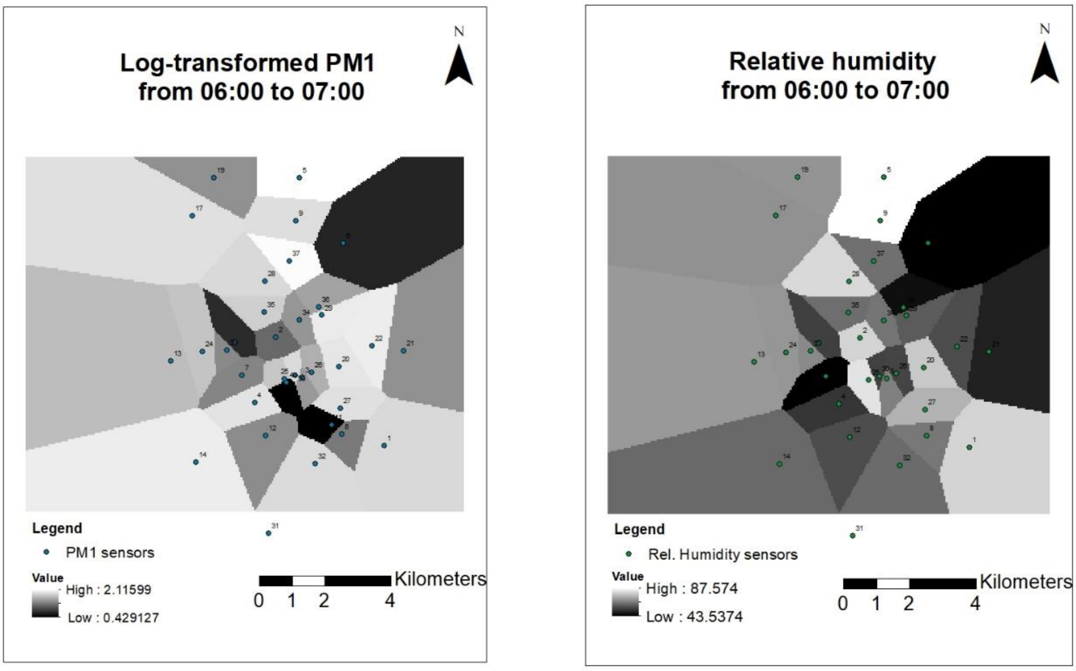


Figure 11. Thiessen polygons of log-transformed PM1 and relative humidity at timestamp 07:00

For wind speed, there is just one sensor available in the study area, so one raster with one single value was created for every timestamp. Those values were obtained averaging all the observations at each timestamp.

However, since the wind direction is treated as a factor, the most repetitive category was selected for every timestamp. In Table 4, we can observe the chosen categories and the averaged wind speed.

Timestamp	Wind direction	Wind speed
2300 to 0000	South-west	31.59
0000 to 0100	South	30.80
0100 to 0200	South-west	31.50
0200 to 0300	South-west	31.95
0300 to 0400	South	31.68
0400 to 0500	South	31.42
0500 to 0600	South	31.77
0600 to 0700	South-west	32.39
0700 to 0800	South-west	32.74
0800 to 0900	South-west	34.51

0900 to 1000	South-west	36.55
1000 to 1100	South-west	38.94
1100 to 1200	South-west	40.62
1200 to 1300	South-west	40.97
1300 to 1400	South-west	39.64
1400 to 1500	South	38.23
1500 to 1600	South	35.13
1600 to 1700	South-west	32.57
1700 to 1800	South	32.04
1800 to 1900	South	33.10
1900 to 2000	South	32.21
2000 to 2100	South-west	32.21
2100 to 2200	South-west	32.92
2200 to 2300	South-west	32.30

Table 4. Selected wind direction category and averaged speed for every timestamp

As it is shown in Table 4, the predominant wind directions in Eindhoven are South and South-west.

Finally, the different five periods are also treated as factors, so one different raster was generated for every single period. This is possible because UFP sensors were rotated during five periods among different locations in Eindhoven. Consequently, each period is related to a set of locations or an area.

First, a Thiessen polygon was created to have the periods in the entire study area and later the period rasters were generated considering the value of 1 in the area related with the period and value of 0 in the rest of the raster as we can see in Figure 12 and Figure 13.

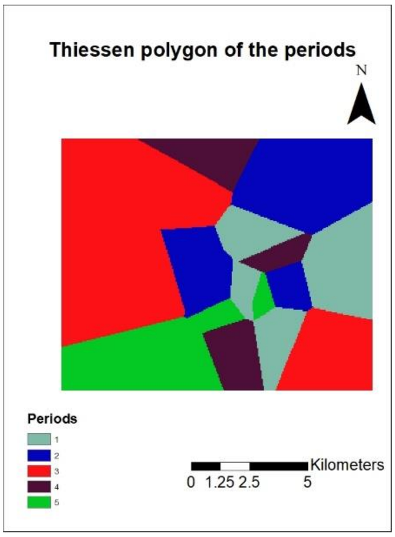


Figure 12. Thiessen polygons of the 5 rotations periods in the study area

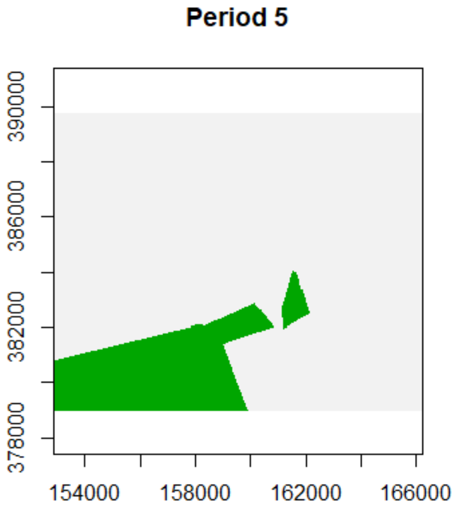


Figure 13. Raster of period 5

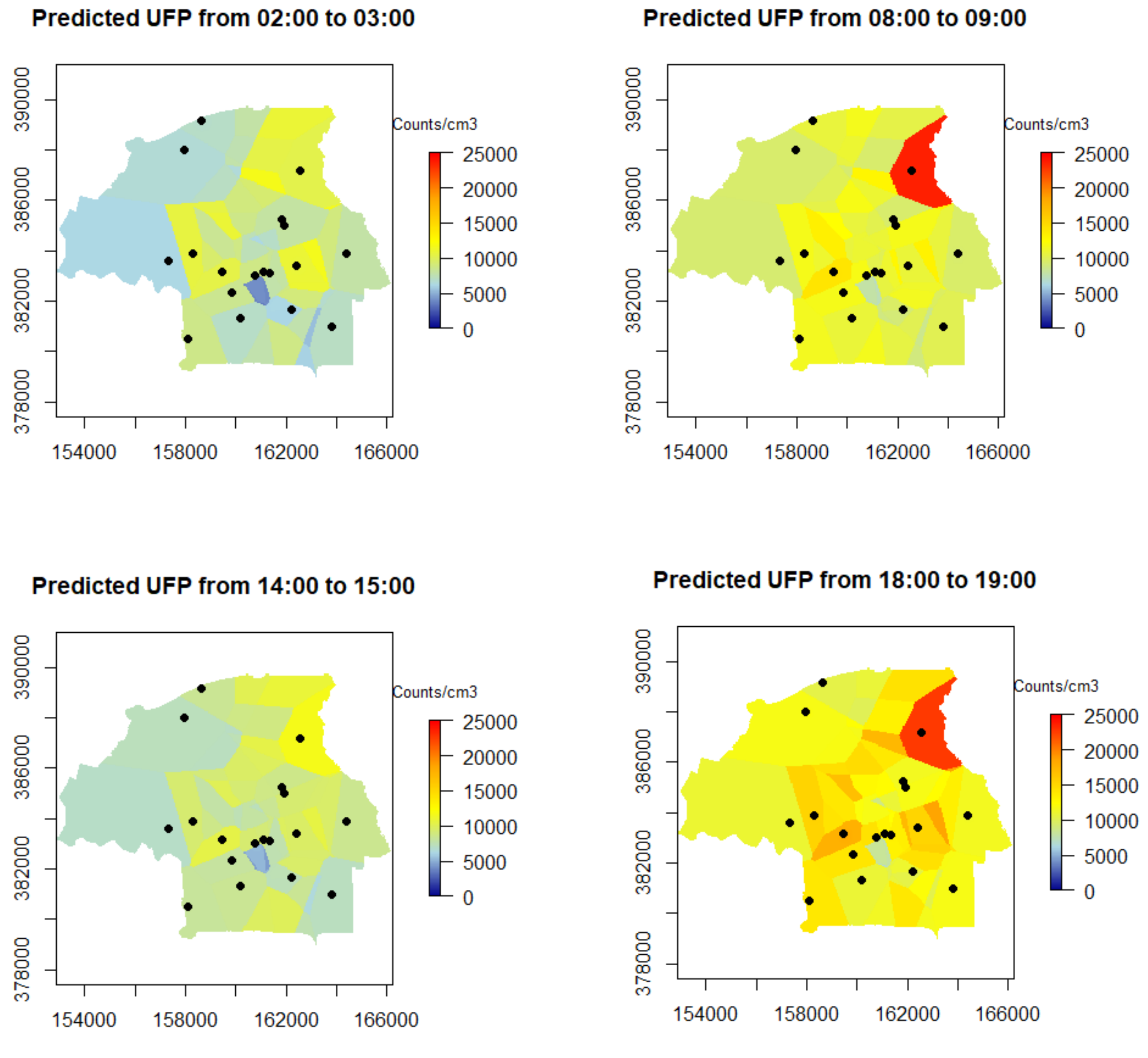
As we can observe in Figure 12, a Thiessen polygon was generated considering all the periods in the entire study area, and then rasters were made independently for each period, in Figure 13 is shown the raster of period 5.

6.5. Predicted UFP maps

To get the predicted maps, the models obtained from the linear regression analysis were used for each timestamp respectively.

The first step was to solve the mean of the model, multiplying the covariates by their respective coefficients.

The second step was to sum the resultant raster of the previous action with the resultant raster of the residuals for every timestamp.
Since the applied model predicts log transformed UFP values, those results were back-transformed before plotting the maps.



Predicted UFP from 02:00 to 03:00

Predicted UFP from 08:00 to 09:00

Predicted UFP from 14:00 to 15:00

Predicted UFP from 18:00 to 19:00

Figure 14. Predicted UFP maps at four different timestamps

As we can see in Figure 14, the concentration of UFP is higher in traffic peak hours like 09:00 and 19:00 and lower in timestamps when there is not a high density of traffic jam.

Another particularity that can be observed is that the highest concentration of UFP is not in the center of the city as expected but in the north-east of it probably because the predominant wind in the city comes from the south-west and brings the UFP and other solids and gases in the air to that area.
In appendix 4, all predicted maps are presented.

6.6. Validation of the model

The model was validating calculation the ME and the RMSE with the results shown in Table 5:

Timestamp	ME	RMSE
2300 to 0000	300	6848
0000 to 0100	-280	6223
0100 to 0200	-832	5444
0200 to 0300	-1836	5289
0300 to 0400	264	4956
0400 to 0500	-798	4868
0500 to 0600	-1723	4895
0600 to 0700	-836	5043
0700 to 0800	-590	6007
0800 to 0900	-491	6278
0900 to 1000	-481	5890
1000 to 1100	-455	5879
1100 to 1200	-576	6116
1200 to 1300	338	5434
1300 to 1400	1401	5775
1400 to 1500	-1841	6066
1500 to 1600	-1893	6337
1600 to 1700	120	6211
1700 to 1800	-693	6451
1800 to 1900	-375	7270
1900 to 2000	-794	7018
2000 to 2100	-1213	7027
2100 to 2200	-556	7326
2200 to 2300	-259	7554

Table 5. ME and RMSE comparison

The predicted maps presented an acceptable ME value, in most of the cases below of 10% of the UFP values, and we can infer that the model does not introduce bias since the resultant values are negatives and positives. Nevertheless, the calculated RMSE has high values between 25% and 35% of UFP values.

7. DISCUSSION

This chapter refers to an analysis of the entire research, reviewing some points of the methodology used to achieve the research objectives, the results obtained, beneficiaries of this study and limitations during the development of the thesis.

7.1. Methodology used

The first step that has been taken in this research was the pre-processing of the data to remove outliers and errors in UFP data. The criterion applied to remove outliers was $\pm 4\sigma$ out of the mean of the normal distribution after applying a log-transformation. This method was used to every location independently which allowed to keep the spatial variation. Nonetheless, it did not consider the temporal variation.

Other outlier-detection methods keep the spatial but also the temporal variation, for instance, van Zoest et al., (2018) made a spatiotemporal classification of the data before applying an outlier-detection using the mean and standard deviation of the normal distribution underlying the truncated normal distribution of the NO_2 observations. This method would be suitable for the present research, but due to the time limitation to achieve the research objectives, $\pm 4\sigma$ out of the mean of the normal distribution was applied.

The method used is probably leaving some outliers during the hours when there is not a high traffic jam but removing more values than expected during peak hours since it is not considering the temporal variation. Thus, this fact might have influenced the estimation of the model coefficients and consequently in the results, decreasing the accuracy of the predicted UFP maps.

After the pre-processing, the data was split into 24 hourly timestamps to partially keep the temporal variability, and just partial because there are also differences between days, weekdays and weekends, and between months. Those temporal variations are not considered in this research due to time limitation, and they will negatively influence the accuracy of the results.

Another critical process to be reviewed is the artificial spatial displacement of the observations. This process was made due to the limited number of locations where the UFP data have been taken and the long distances between them. As it was presented in previous chapters, UFP have a high dynamic spatial nature which means that vary significantly in short ranges, making hard to establish a spatial autocorrelation of UFP.

Since it is known that UFP can vary significantly in ranges of 10 meters, the artificial displacement of measurements was made in a maximum buffer of 20 meters; even though this may have led some errors in the results, it was necessary to establish the spatial autocorrelation of the residuals before estimating its variogram and fitted model to apply ordinary Kriging.

Likewise, the selection of covariates is a critical process for the accuracy of the model. In this case, the covariates selection was based on the literature review (Jayamurugan et al., 2013) and data availability, pre-selecting just atmospheric variables. In the end, log-transformed PM_1, relative humidity, wind direction, and wind speed were selected as fixed effect variables and the periods were selected as a random effect variable. The lack of sensors of wind direction and wind speed (one sensor in the study area) might affect the accuracy of the results. However, spatial covariates as population or distance to main roads are not considered, and they would significantly increase the accuracy in the results.

About the validation of the model, the RMSE presented high values because the data has a high variability as was shown in chapter 4, so all the peaks that initially the data presented are not represented in the predicted maps.

In spite of the results can be improved, the methodology applied is suitable to predict UFP in the city of Eindhoven, and regarding the lack of data, the few sensors available, and the time limitation, the research objectives were achieved, and the maps give an idea about the spatiotemporal distribution of UFP with an acceptable spatial resolution.

7.2. Spatiotemporal variability of UFP in Eindhoven

The results show the spatiotemporal variability of UFP in the city of Eindhoven. Hence, the main goal of the present research was achieved. The maps of predicted UFP show in general a small concentration between 00:00 and 06:00 but in peak hours like 08:00 or 19:00 when there is more circulation of vehicles in the city, the concentration of UFP increases, showing that the UFP quantities in urban areas are directly proportional to vehicles combustion.

Another particularity of the results is that in all the 24 maps, the major concentration of UFP is in Eindhoven north-east as we can see in Appendix 4, which is not expected since that area is relatively far from the center, and with a small vehicles circulation. A possible reason may be that the predominant wind in Eindhoven comes from the south-west and carries the particles in the wind towards that area.

Furthermore, in Figure 15 we can observe that the highest UFP concentration was predicted at 20:00 and the lowest one at 03:00, which makes sense due they are opposite in terms of vehicles circulation.

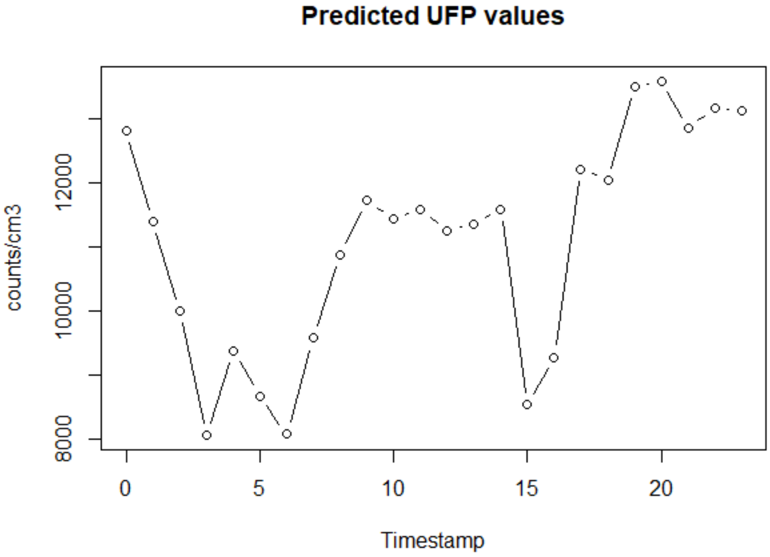


Figure 15. Variation of averaged predicted UFP among timestamps

7.3. Beneficiaries of the research

This study will benefit to Eindhoven authorities to have an idea about how UFP concentration is spatiotemporally distributed in the city. Therefore, they can make important decisions to mitigate the fatal health consequences that people who are exposed to air pollution can suffer.

It will also benefit to Eindhoven citizens since they will know which areas are more dangerous to stay in and it will create consciousness to improve habits in order to decrease air pollutants emissions in the city.

7.4. Research limitations

The first limitation was the lack of UFP data since there are just six sensors. However, when they started to be rotated, there were five of them working, and during the rotation two of them had problems and stopped working. Likewise, the availability of data in the maximum amount of locations is vital because the primary goal of this study is to predict UFP values in the city of Eindhoven, so every location with available data makes the results more accurate.

Furthermore, all the observation were taken between the end of November 2016 and the end of February 2017, which means that this research was able to predict UFP just in the winter season. Likewise, it can be inferred from literature review that UFP concentrations are higher in winter session (Zhu et al., 2010; Ragettli et al., 2014).

Finally, the computational capacity available was not enough to try out different methods. In a first moment, maximum likelihood estimation was applied including the trend to get the coefficients of the models with the end of utilizing regression Kriging. However, this led to a memory overload even on a workstation with 128GB of RAM.

8. CONCLUSIONS AND RECOMMENDATIONS

8.1. Conclusions

The main conclusion is that the methodology applied is suitable to predict UFP in the city of Eindhoven, despite the limitations described in the previous chapter. Nevertheless, the RMSE showed a high degree of uncertainty, whereby the method can be improved to reduce the uncertainty and consequently increase the accuracy.

Even though it was possible to predict UFP in the study area using five sensors rotated over eighteen locations, the spatial resolution of the maps seems to be large for an urban area since we cannot observe smooth variations of the values but defined differences between the regions that compound the generated Thiessen polygons of the covariates. So, to get a better spatial resolution in the prediction maps, it would be necessary to have more than eighteen locations available.

Likewise, the distances between observations are not short enough to assess spatial autocorrelation of UFP, because this pollutant presents a high spatial dynamic, making it challenging to apply various geostatistical approaches to predict UFP in the city of Eindhoven.

The covariates used to estimate the model, explain well the response variable. Nonetheless, the inclusion of other spatial covariates (e.g., population, distance to main roads) would increase the accuracy of the predictions.

8.2. Recommendations

For future researches, it would be suggested to look for different methods (e.g., including Bayesian approach, regression kriging, etc.) to improve the accuracy of the results. The Bayesian approach has the advantage of accounting for a prior belief which will increase the accuracy of the obtained model, if the prior belief is informative but also will allow calculating the uncertainty for each coefficient of the model, giving a better idea of which covariates should be included.

Also, one could include more explanatory variables (e.g., population, distance to main roads) to estimate the model. These variables will explain better the spatial distribution of the response variable, which will help to obtain more accurate results.

For people involved in AiREAS project and developers of air quality sensor networks in general, it would be adviced to purchase more UFP sensors or to rotate those which are currently available for more extended periods and more locations to have a better spatial and temporal resolution of predicted UFP in Eindhoven.

LIST OF REFERENCES

AiREAS. (2014). Retrieved from http://www.aireas.com

Batini, C., & Scannapieca, M. (2006). Introduction to Data Quality. In *Data Quality: Concepts, Methodologies and Techniques* (pp. 1–18). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5_1

Berghmans, P., Bleux, N., Panis, L. I., Mishra, V. K., Torfs, R., & Van Poppel, M. (2009). Exposure assessment of a cyclist to PM10 and ultrafine particles. *Science of The Total Environment*, *407*(4), 1286–1298. https://doi.org/10.1016/J.SCITOTENV.2008.10.041

Brooks-Barlett, J. (2019). Probability concepts explained: Maximum likelihood estimation. Retrieved February 19, 2019, from https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Buonanno, G., Stabile, L., & Morawska, L. (2014). Personal exposure to ultrafine particles: The influence of time-activity patterns. *Science of The Total Environment*, *468–469*, 903–907. https://doi.org/10.1016/J.SCITOTENV.2013.09.016

Chen, R., Samoli, E., Wong, C.-M., Huang, W., Wang, Z., Chen, B., & Kan, H. (2012). Associations between short-term exposure to nitrogen dioxide and mortality in 17 Chinese cities: The China Air Pollution and Health Effects Study (CAPES). *Environment International*, *45*, 32–38. https://doi.org/10.1016/J.ENVINT.2012.04.008

Cho, W., & Choi, E. (2017). Big data pre-processing methods with vehicle driving data using MapReduce techniques. *The Journal of Supercomputing*, *73*(7), 3179–3195. https://doi.org/10.1007/s11227-017-2014-x

Diggle, P. J., & Ribeiro Jr., P. J. (2007). *Model-based Geostatistics*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-48536-2

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). *Model-Based Geostatistics*. *Source: Journal of the Royal Statistical Society. Series C (Applied Statistics)* (Vol. 47). Retrieved from https://blackboard.utwente.nl/bbcswebdav/pid-1116202-dt-content-rid-2892986_2/courses/M18-EOS-100/Diggle1998.pdf

Download - ECN dustmonitoring Aireas. (n.d.). Retrieved January 31, 2019, from https://aireas.site.dustmonitoring.nl/download?lang=en-US

Farrell, W., Weichenthal, S., Goldberg, M., Valois, M.-F., Shekarrizfard, M., & Hatzopoulou, M. (2016). Near roadway air pollution across a spatially extensive road and cycling network. *Environmental Pollution*, *212*, 498–507. https://doi.org/10.1016/J.ENVPOL.2016.02.041

Fu, D., Xia, X., Duan, M., Zhang, X., Li, X., Wang, J., & Liu, J. (2018). Mapping nighttime PM2.5 from VIIRS DNB using a linear mixed-effect model. *Atmospheric Environment*, *178*, 214–222. https://doi.org/10.1016/J.ATMOSENV.2018.02.001

Hamm, N. A. S., Finley, A. O., Schaap, M., & Stein, A. (2015). A spatially varying coefficient model for mapping PM10 air quality at the European scale. *Atmospheric Environment*, *102*, 393–405. https://doi.org/10.1016/j.atmosenv.2014.11.043

Hoek, G., Beelen, R., Kos, G., Dijkema, M., van der Zee, S. C., Fischer, P., & Brunekreef, B. (2010). Land Use Regression Model for Ultrafine Particles in Amsterdam. https://doi.org/10.1021/es1023042

Jayamurugan, R., Kumaravel, B., Palanivelraja, S., & Chockalingam, M. P. (2013). Influence of

Temperature, Relative Humidity and Seasonal Variability on Ambient Air Quality in a Coastal Urban Area. *International Journal of Atmospheric Sciences*, *2013*, 1–7. https://doi.org/10.1155/2013/264046

Jr., P. J. R., Christensen, O. F., & Diggle, P. J. (2003). Geostatistical software-geoR and geoRglm. *Proceedings of DSC*, 15. https://doi.org/10.1134/S1063783410020319

Klompmaker, J. O., Montagne, D. R., Meliefste, K., Hoek, G., & Brunekreef, B. (2015). Spatial variation of ultrafine particles and black carbon in two cities: Results from a short-term measurement campaign. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2014.11.088

KNMI - Weather data from the Netherlands - Download. (n.d.). Retrieved January 31, 2019, from http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi

Kumar, P., Morawska, L., Birmili, W., Paasonen, P., Hu, M., Kulmala, M., … Britter, R. (2014). Ultrafine particles in cities. *Environment International*, *66*, 1–10. https://doi.org/10.1016/J.ENVINT.2014.01.013

Kwasny, F., Madl, P., Hofmann, W., Kwasny, F., Madl, P., & Hofmann, W. (2010). Correlation of Air Quality Data to Ultrafine Particles (UFP) Concentration and Size Distribution in Ambient Air. *Atmosphere*, *1*(1), 3–14. https://doi.org/10.3390/atmos1010003

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, *51*(4), 717–728. https://doi.org/10.1046/j.1365-2389.2000.00345.x

Miles, J., & Banyard, P. (2008). Descriptive Statistics. In *Understanding and Using Statistics in Psychology: A Practical Introduction: Or, How I Came to Know and Love the Standard Error* (pp. 11–51). New York, NY: Springer New York. https://doi.org/10.4135/9781446215722.n2

Panis, L. I. (2010). New Directions: Air pollution epidemiology can benefit from activity-based models. *Atmospheric Environment*, *44*(7), 1003–1004. https://doi.org/10.1016/J.ATMOSENV.2009.10.047

Peng, H., & Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, *109*, 109–129. https://doi.org/10.1016/J.JMVA.2012.02.005

Raaschou-Nielsen, O., Beelen, R., Wang, M., Hoek, G., Andersen, Z. J., Hoffmann, B., … Vineis, P. (2016). Particulate matter air pollution components and risk for lung cancer. *Environment International*, *87*, 66–73. https://doi.org/10.1016/J.ENVINT.2015.11.007

Ragettli, M. S., Ducret-Stich, R. E., Foraster, M., Morelli, X., Aguilera, I., Basagaña, X., … Phuleria, H. C. (2014). Spatio-temporal variation of urban ultrafine particle number concentrations. *Atmospheric Environment*, *96*, 275–283. https://doi.org/10.1016/J.ATMOSENV.2014.07.049

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., & Bartonova, A. (2017). Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International*. https://doi.org/10.1016/j.envint.2017.05.005

van Zoest, V. M., Stein, A., & Hoek, G. (2018). Outlier Detection in Urban Air Quality Sensor Networks. *Water, Air, & Soil Pollution*, *229*(4), 111. https://doi.org/10.1007/s11270-018-3756-7

Wang, Z., Lu, Q.-C., He, H.-D., Wang, D., Gao, Y., & Peng, Z.-R. (2017). Investigation of the spatiotemporal variation and influencing factors on fine particulate matter and carbon monoxide concentrations near a road intersection. *Frontiers of Earth Science*, *11*(1), 63–75. https://doi.org/10.1007/s11707-016-0564-5

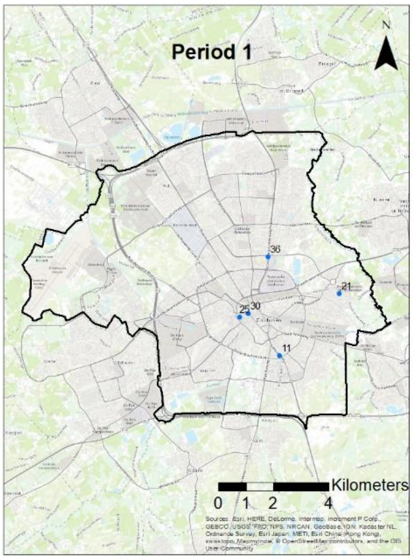
Weichenthal, S., Ryswyk, K. Van, Goldstein, A., Bagg, S., Shekkarizfard, M., & Hatzopoulou, M. (2016). A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. *Environmental Research*. https://doi.org/10.1016/j.envres.2015.12.016

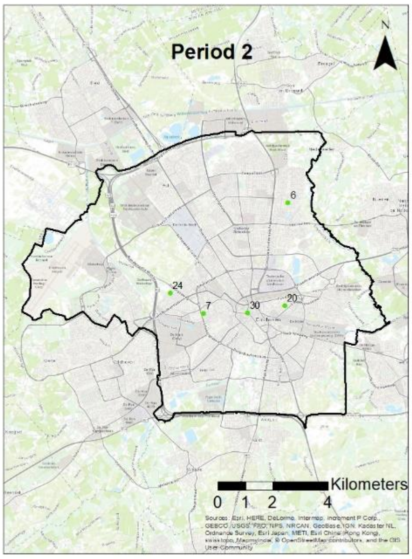
WHO. (2018). Retrieved from http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

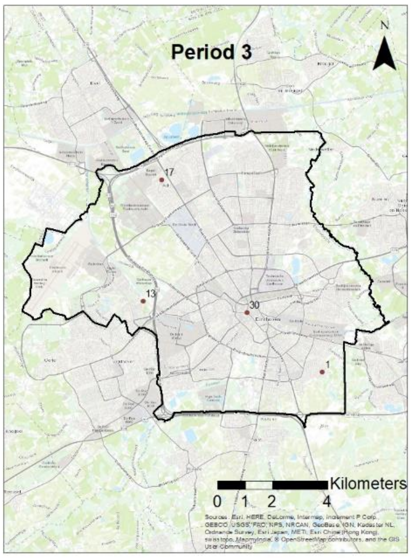
Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*, *216*, 208–215. https://doi.org/10.1016/J.NEUCOM.2016.07.036

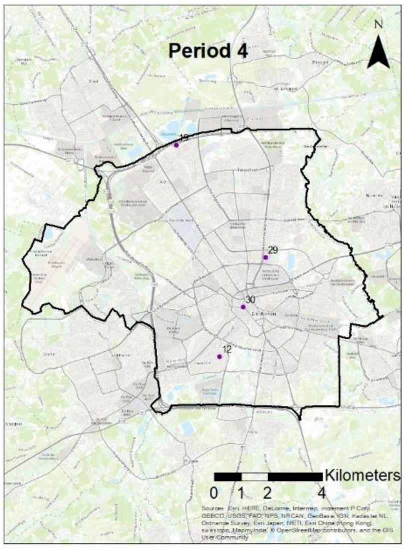
Zhu, Y., Hinds, W. C., Shen, S., & Sioutas, C. (2010). Aerosol Science and Technology Seasonal Trends of Concentration and Size Distribution of Ultrafine Particles Near Major Highways in Los Angeles Special Issue of Aerosol Science and Technology on Findings from the Fine Particulate Matter Supersites Program. https://doi.org/10.1080/02786820390229156

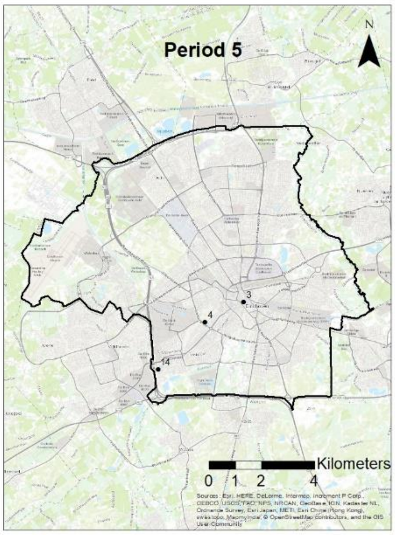
Appendix-1 UFP sensors-rotation among the airboxes


Period 1
N
36
21
25
30
11
Kilometers
0 1 2 4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

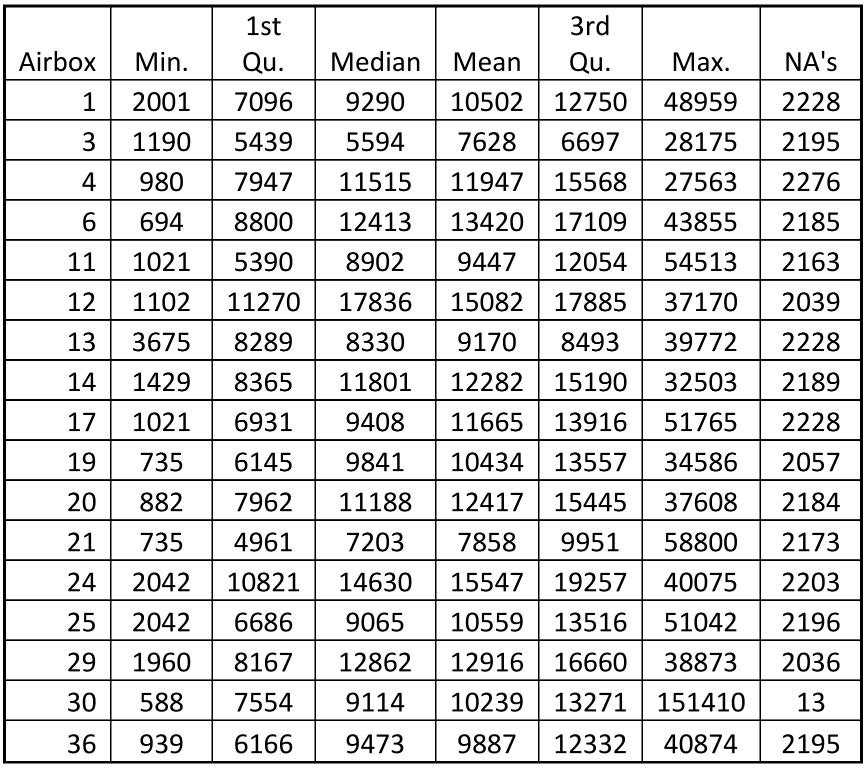

Period 2
N
6
24
7
30
20
Kilometers
0 1 2 4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community


Period 3
N
17
13
30
1
Kilometers
0 1 2 4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community


Period 4
N
19
29
30
12
Kilometers
0 1 2 4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

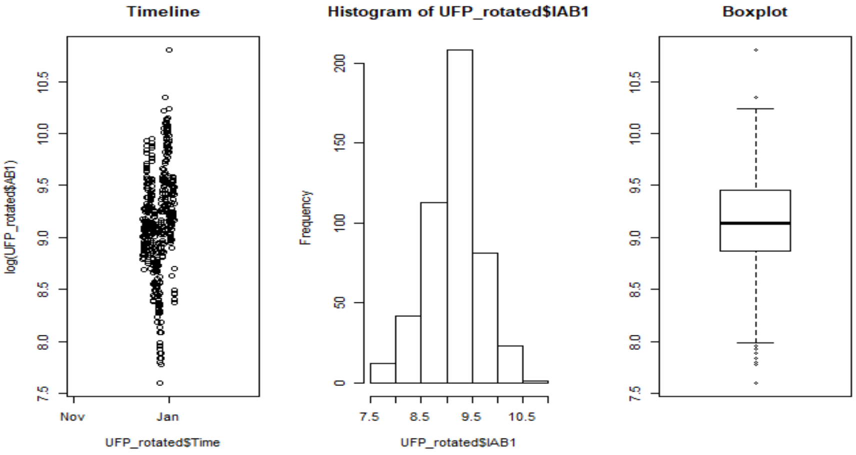

Period 5
N
3
4
14
Kilometers
0 1 2 4
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

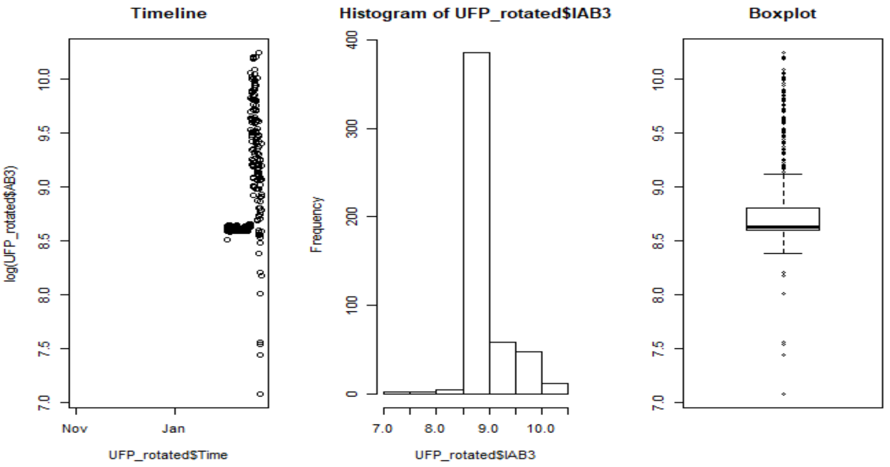
Appendix-2 Summary of UFP data in the 18 airboxes

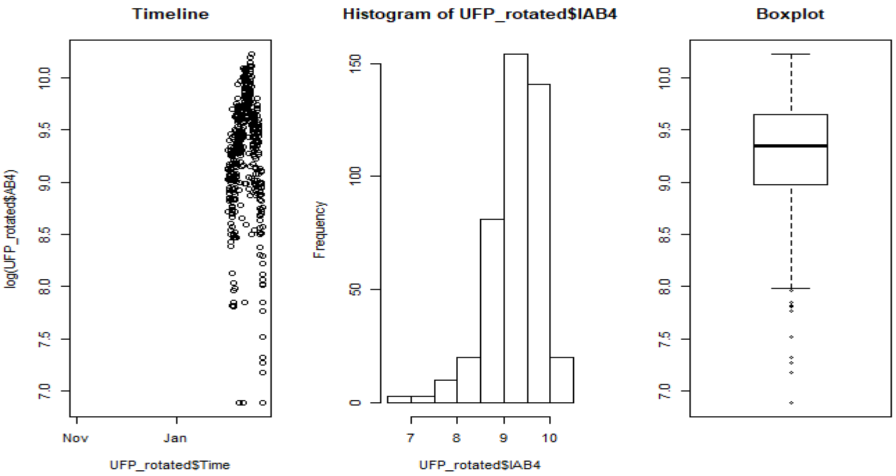


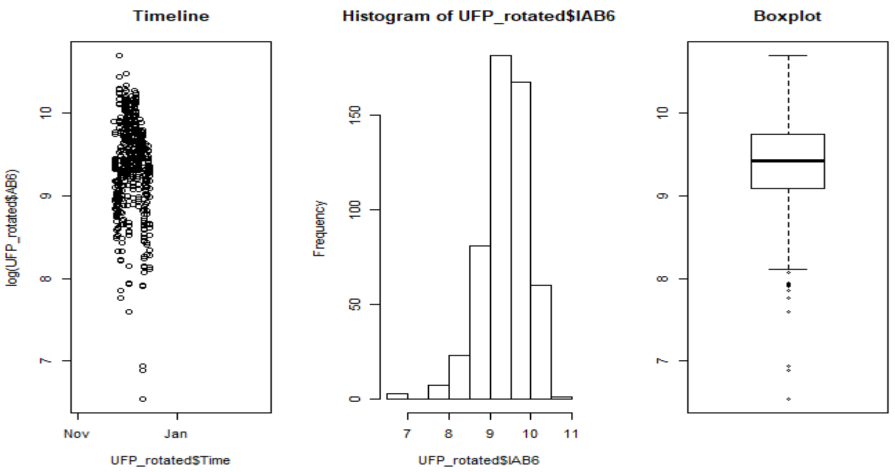
Airbox	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	2001	7096	9290	10502	12750	48959	2228
3	1190	5439	5594	7628	6697	28175	2195
4	980	7947	11515	11947	15568	27563	2276
6	694	8800	12413	13420	17109	43855	2185
11	1021	5390	8902	9447	12054	54513	2163
12	1102	11270	17836	15082	17885	37170	2039
13	3675	8289	8330	9170	8493	39772	2228
14	1429	8365	11801	12282	15190	32503	2189
17	1021	6931	9408	11665	13916	51765	2228
19	735	6145	9841	10434	13557	34586	2057
20	882	7962	11188	12417	15445	37608	2184
21	735	4961	7203	7858	9951	58800	2173
24	2042	10821	14630	15547	19257	40075	2203
25	2042	6686	9065	10559	13516	51042	2196
29	1960	8167	12862	12916	16660	38873	2036
30	588	7554	9114	10239	13271	151410	13
36	939	6166	9473	9887	12332	40874	2195

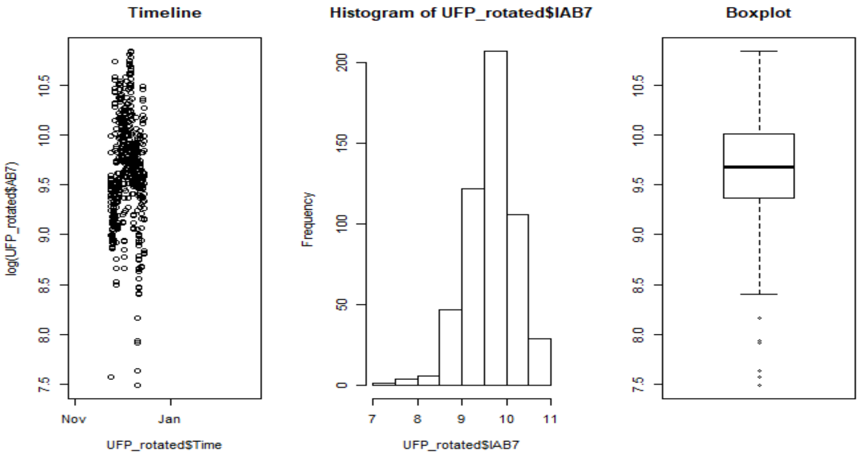
Appendix-3 Timelines, histograms and boxplots of log-transformed UFP data for the 18 Airboxes

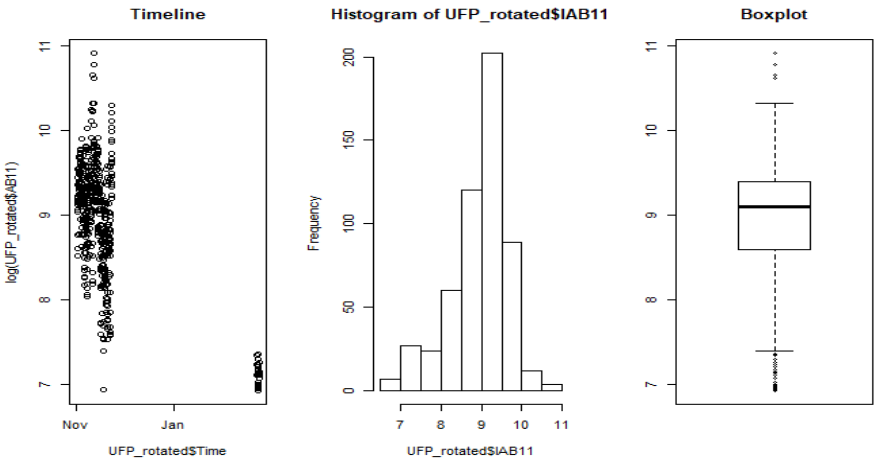


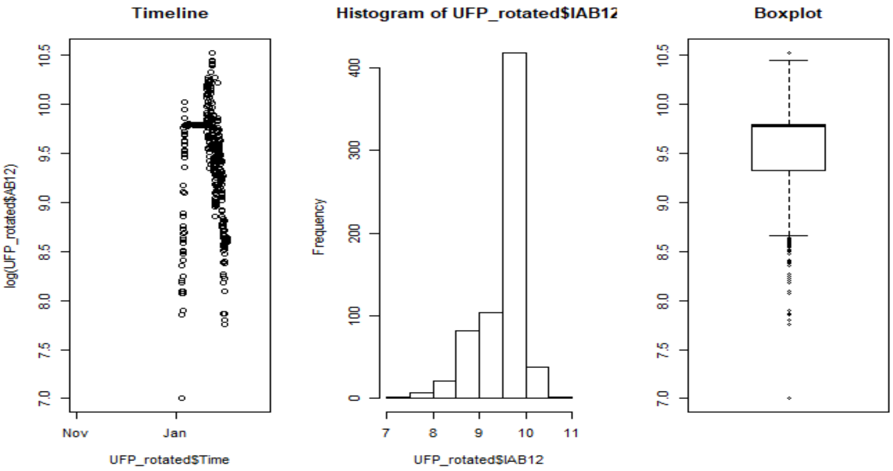


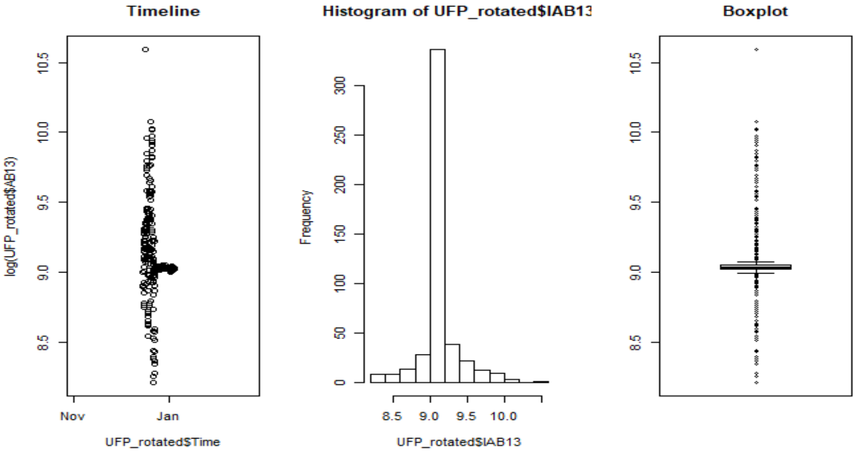


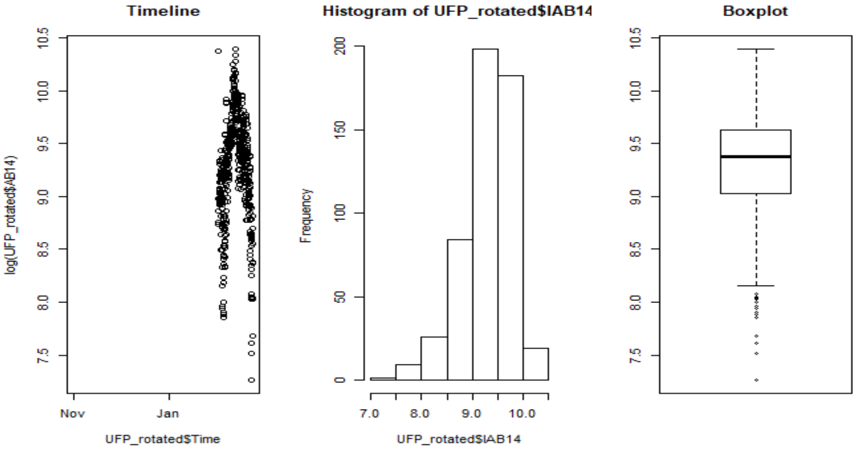

Timeline
log(UFP_rotated$IAB6)
7
8
9
10
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB6
Frequency
0
50
100
150
7
8
9
10
11
UFP_rotated$IAB6
Boxplot
7
8
9
10

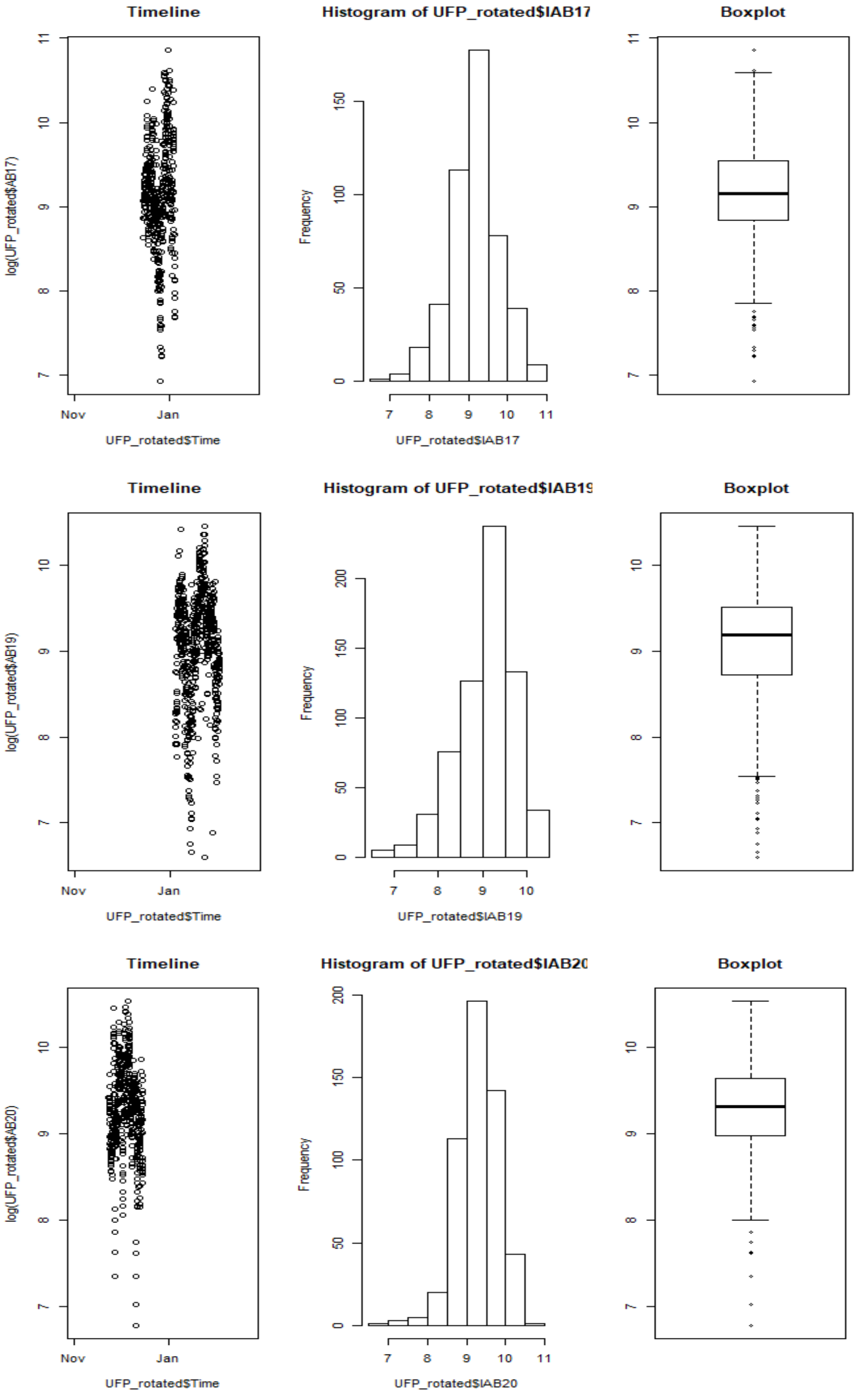

Timeline
log(UFP_rotated$IAB7)
7.5
8.0
8.5
9.0
9.5
10.0
10.5
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB7
Frequency
0
50
100
150
200
7
8
9
10
11
UFP_rotated$IAB7
Boxplot
7.5
8.0
8.5
9.0
9.5
10.0
10.5

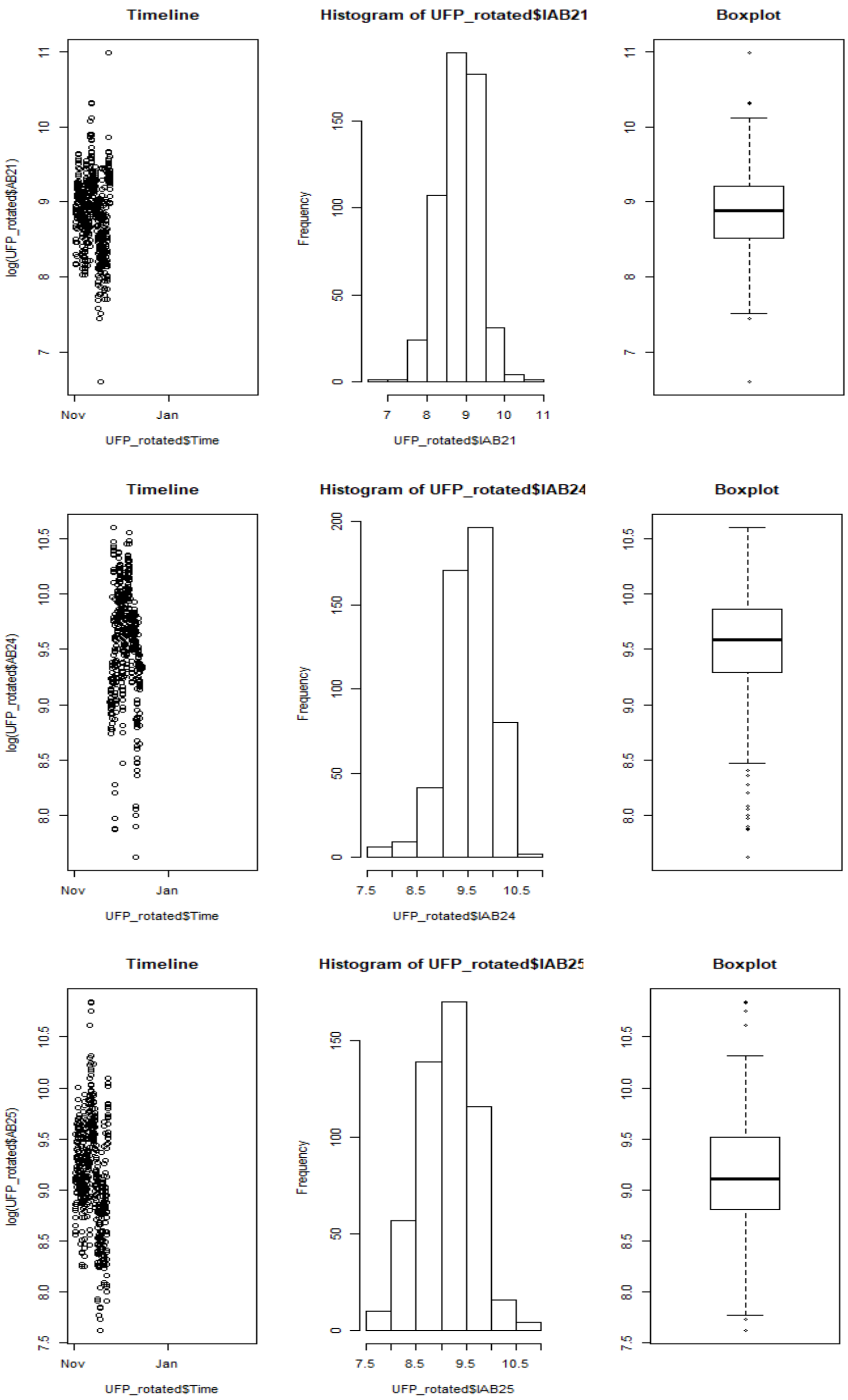

Timeline
log(UFP_rotated$IAB11)
7
8
9
10
11
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB11
Frequency
0
50
100
150
200
7
8
9
10
11
UFP_rotated$IAB11
Boxplot
7
8
9
10
11

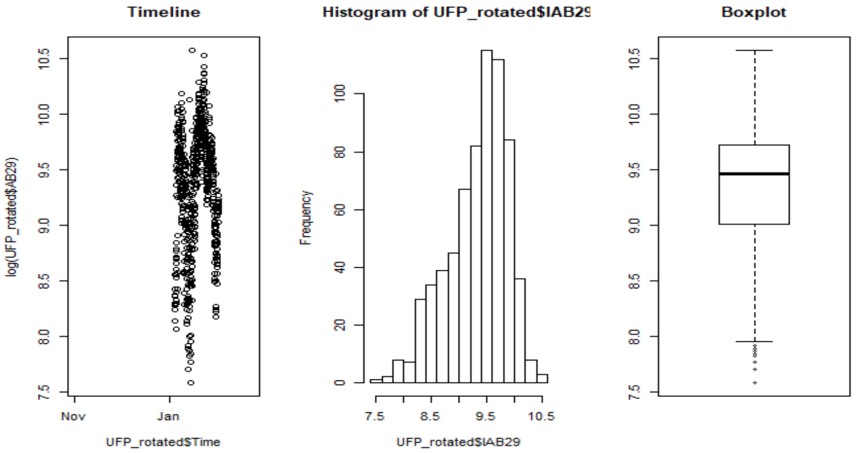

Timeline
log(UFP_rotated$IAB12)
7.0 7.5 8.0 8.5 9.0 9.5 10.0 10.5
Nov Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB12
Frequency
0 100 200 300 400
7 8 9 10 11
UFP_rotated$IAB12
Boxplot
7.0 7.5 8.0 8.5 9.0 9.5 10.0 10.5

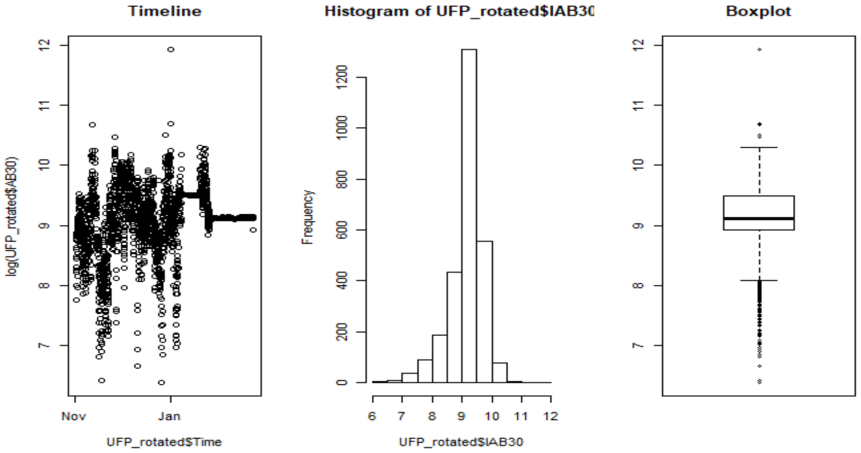

Timeline
log(UFP_rotated$IAB13)
8.5 9.0 9.5 10.0 10.5
Nov Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB13
Frequency
0 50 100 150 200 250 300
8.5 9.0 9.5 10.0
UFP_rotated$IAB13
Boxplot
8.5 9.0 9.5 10.0 10.5

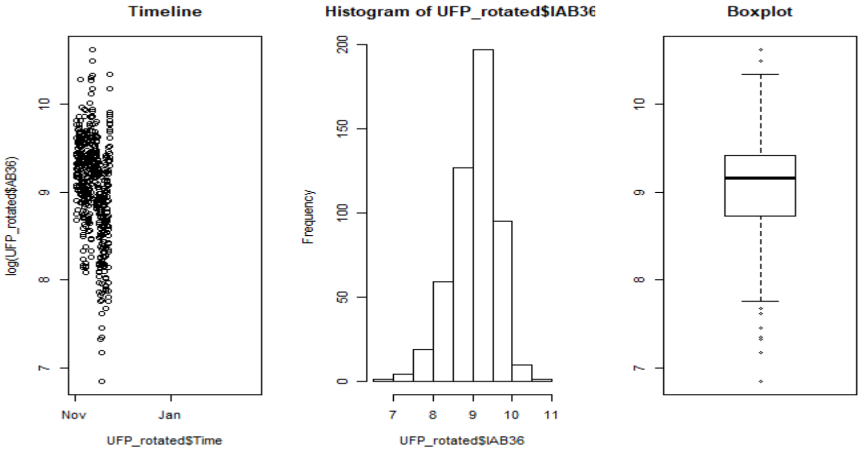

Timeline
log(UFP_rotated$IAB14)
7.5 8.0 8.5 9.0 9.5 10.0 10.5
Nov Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB14
Frequency
0 50 100 150 200
7.0 8.0 9.0 10.0
UFP_rotated$IAB14
Boxplot
7.5 8.0 8.5 9.0 9.5 10.0 10.5


Timeline
log(UFP_rotated$IAB17)
7
8
9
10
11
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB17
Frequency
0
50
100
150
7
8
9
10
11
UFP_rotated$IAB17
Boxplot
7
8
9
10
11
Timeline
log(UFP_rotated$IAB19)
7
8
9
10
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB19
Frequency
0
50
100
150
200
7
8
9
10
UFP_rotated$IAB19
Boxplot
7
8
9
10
Timeline
log(UFP_rotated$IAB20)
7
8
9
10
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB20
Frequency
0
50
100
150
200
7
8
9
10
11
UFP_rotated$IAB20
Boxplot
7
8
9
10


Timeline
log(UFP_rotated$IAB21)
7
8
9
10
11
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB21
Frequency
0
50
100
150
7
8
9
10
11
UFP_rotated$IAB21
Boxplot
7
8
9
10
11
Timeline
log(UFP_rotated$IAB24)
8.0
8.5
9.0
9.5
10.0
10.5
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB24
Frequency
0
50
100
150
200
7.5
8.5
9.5
10.5
UFP_rotated$IAB24
Boxplot
8.0
8.5
9.0
9.5
10.0
10.5
Timeline
log(UFP_rotated$IAB25)
7.5
8.0
8.5
9.0
9.5
10.0
10.5
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB25
Frequency
0
50
100
150
7.5
8.5
9.5
10.5
UFP_rotated$IAB25
Boxplot
7.5
8.0
8.5
9.0
9.5
10.0
10.5

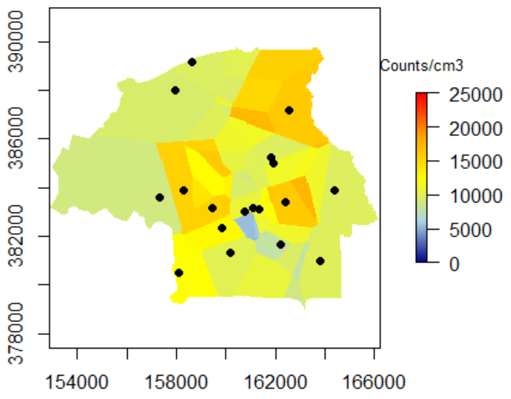

Timeline
log(UFP_rotated$IAB29)
7.5
8.0
8.5
9.0
9.5
10.0
10.5
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB29
Frequency
0
20
40
60
80
100
7.5
8.5
9.5
10.5
UFP_rotated$IAB29
Boxplot
7.5
8.0
8.5
9.0
9.5
10.0
10.5


Timeline
log(UFP_rotated$IAB30)
7
8
9
10
11
12
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB30
Frequency
0
200
400
600
800
1000
1200
6
7
8
9
10
11
12
UFP_rotated$IAB30
Boxplot
7
8
9
10
11
12

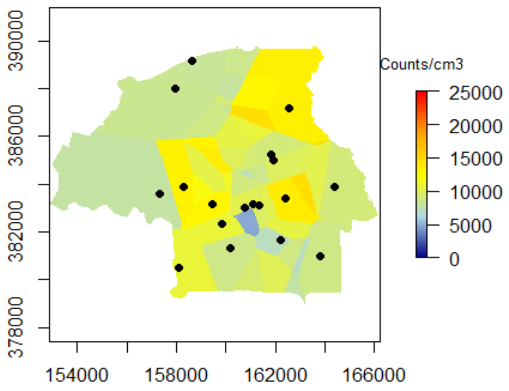

Timeline
log(UFP_rotated$IAB36)
7
8
9
10
Nov
Jan
UFP_rotated$Time
Histogram of UFP_rotated$IAB36
Frequency
0
50
100
150
200
7
8
9
10
11
UFP_rotated$IAB36
Boxplot
7
8
9
10

Appendix-4 Predicted UFP maps

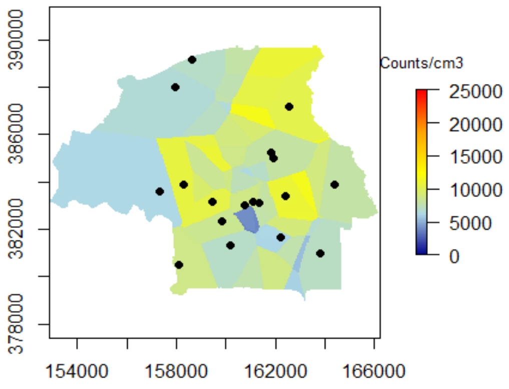
Predicted UFP from 00:00 to 01:00



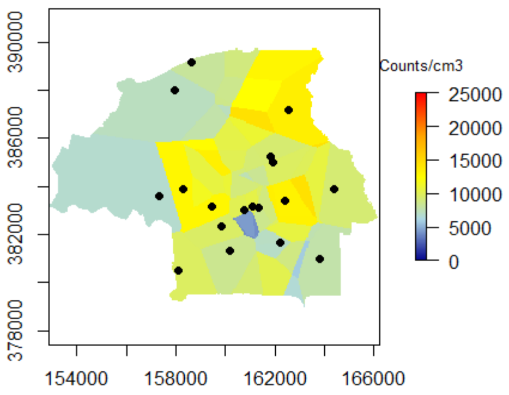
Predicted UFP from 01:00 to 02:00



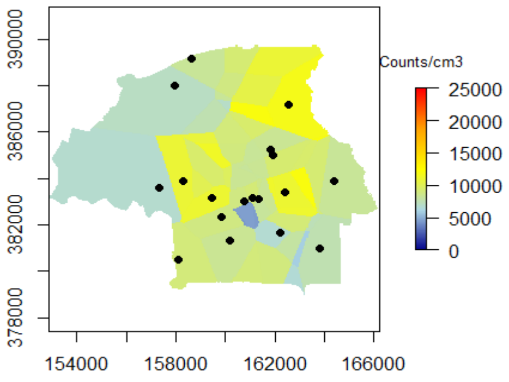
Predicted UFP from 02:00 to 03:00



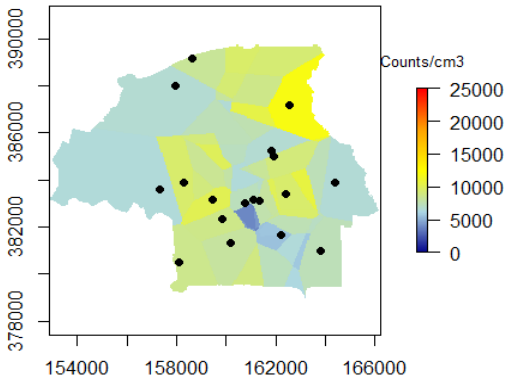
Predicted UFP from 03:00 to 04:00



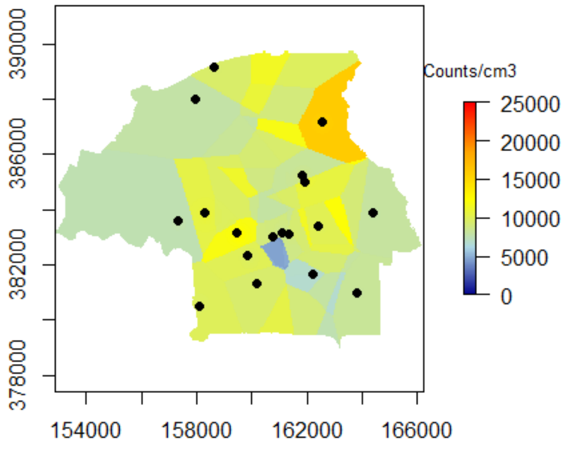
Predicted UFP from 04:00 to 05:00



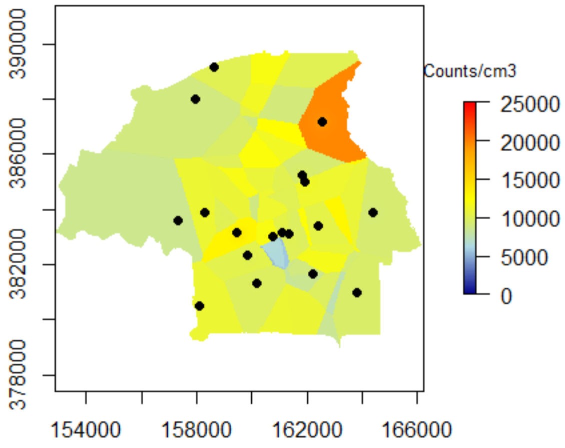
Predicted UFP from 05:00 to 06:00



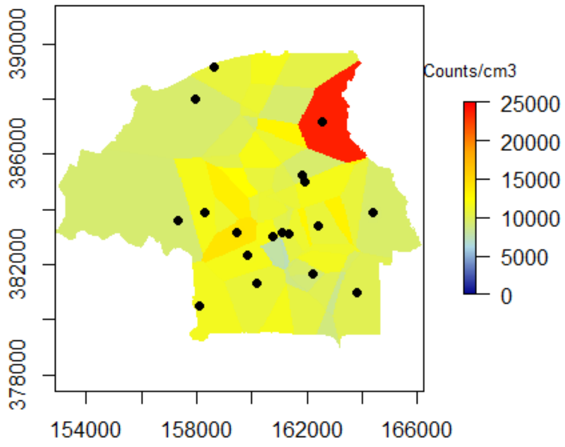
Predicted UFP from 06:00 to 07:00



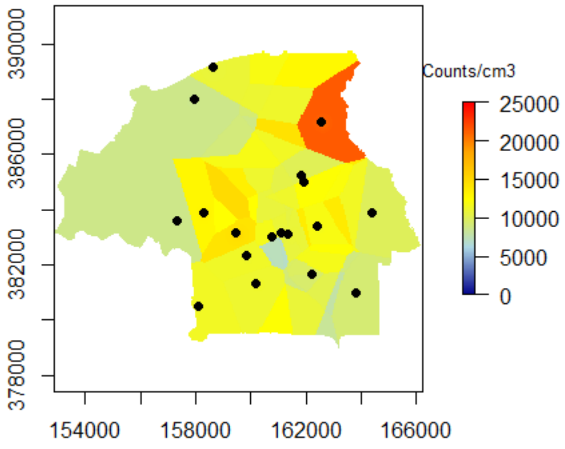
Predicted UFP from 07:00 to 08:00



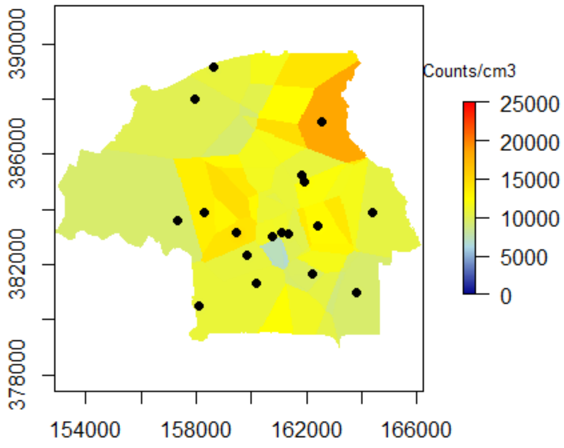
Predicted UFP from 08:00 to 09:00



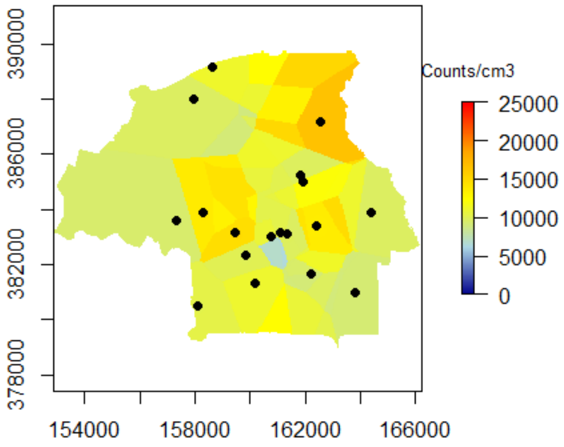
Predicted UFP from 09:00 to 10:00



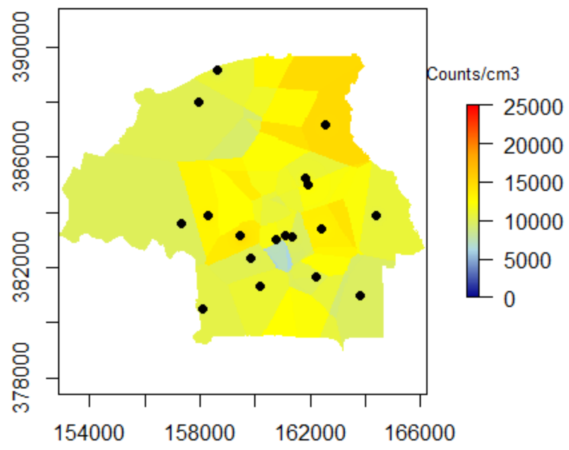
Predicted UFP from 10:00 to 11:00



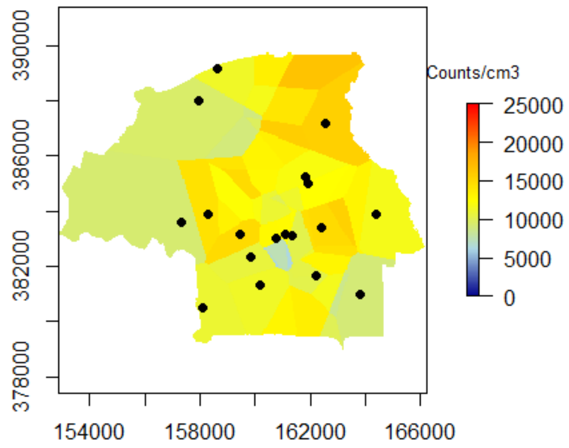
Predicted UFP from 11:00 to 12:00



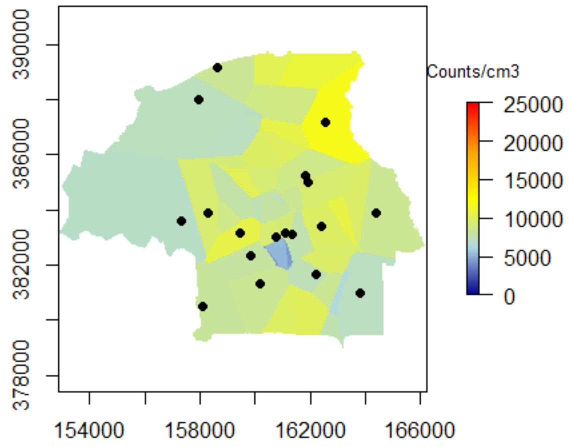
Predicted UFP from 12:00 to 13:00



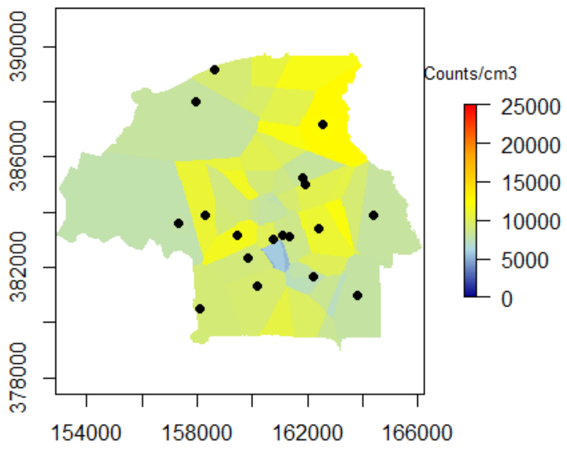
Predicted UFP from 13:00 to 14:00



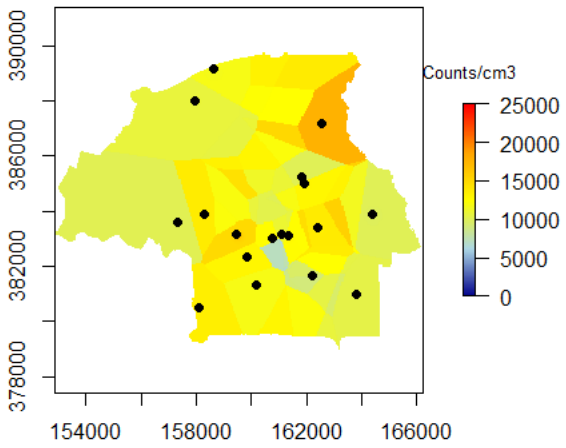
Predicted UFP from 14:00 to 15:00



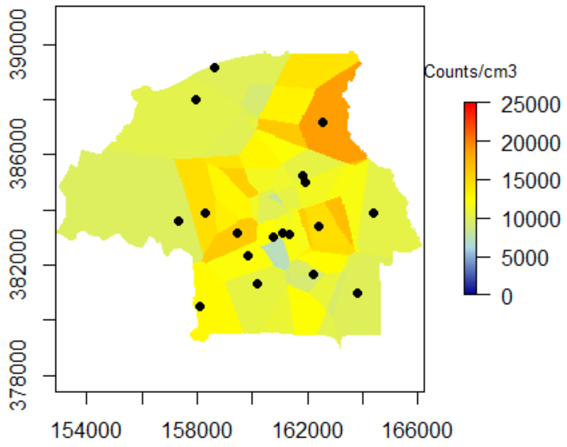
Predicted UFP from 15:00 to 16:00



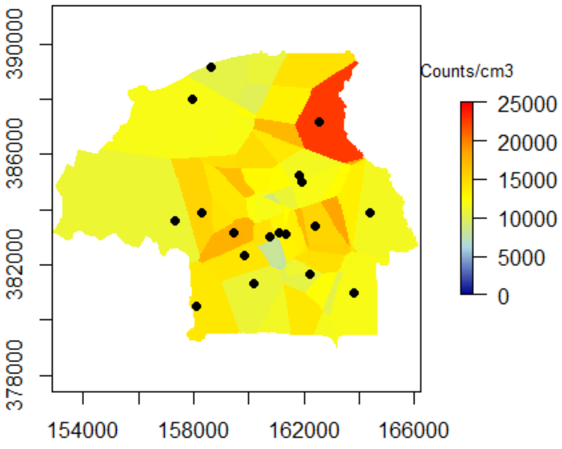
Predicted UFP from 16:00 to 17:00



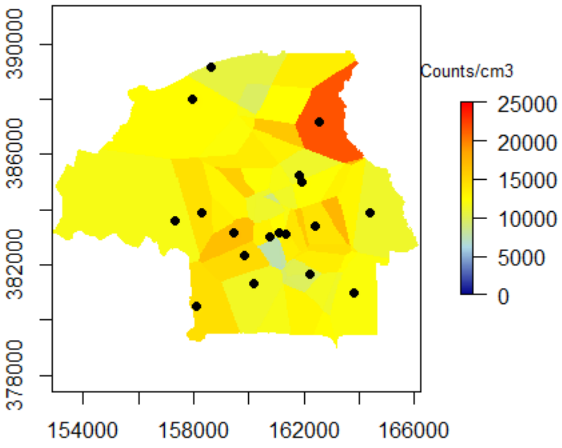
Predicted UFP from 17:00 to 18:00



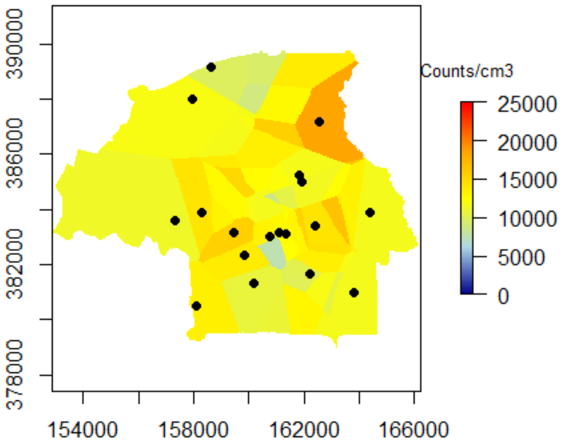
Predicted UFP from 18:00 to 19:00



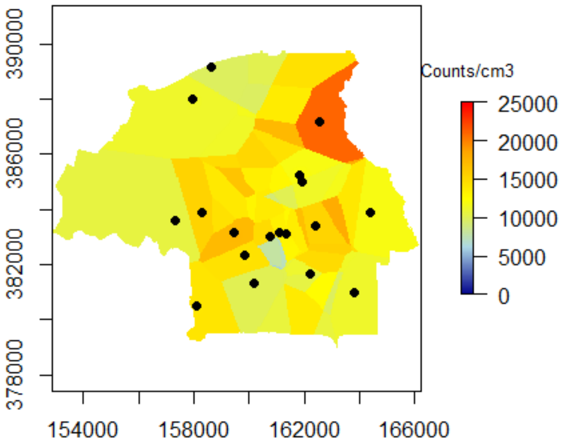
Predicted UFP from 19:00 to 20:00



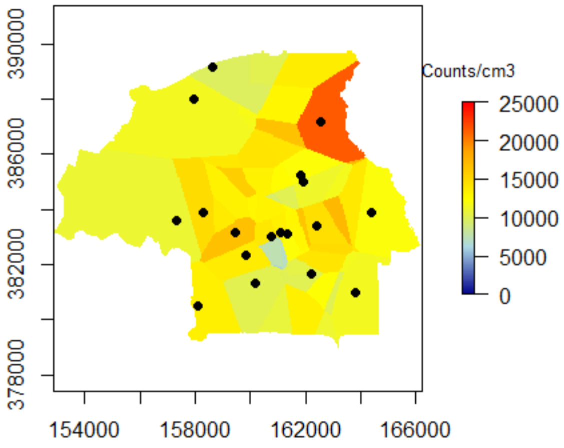
Predicted UFP from 20:00 to 21:00



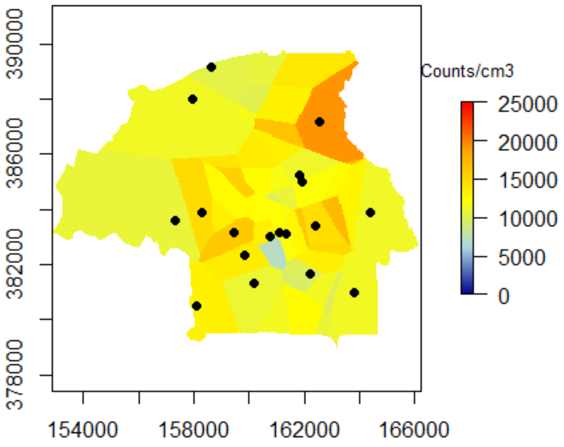
Predicted UFP from 21:00 to 22:00



Predicted UFP from 22:00 to 23:00



Predicted UFP from 23:00 to 00:00


390000
386000
382000
378000
154000
158000
162000
166000
Counts/cm3
25000
20000
15000
10000
5000
0