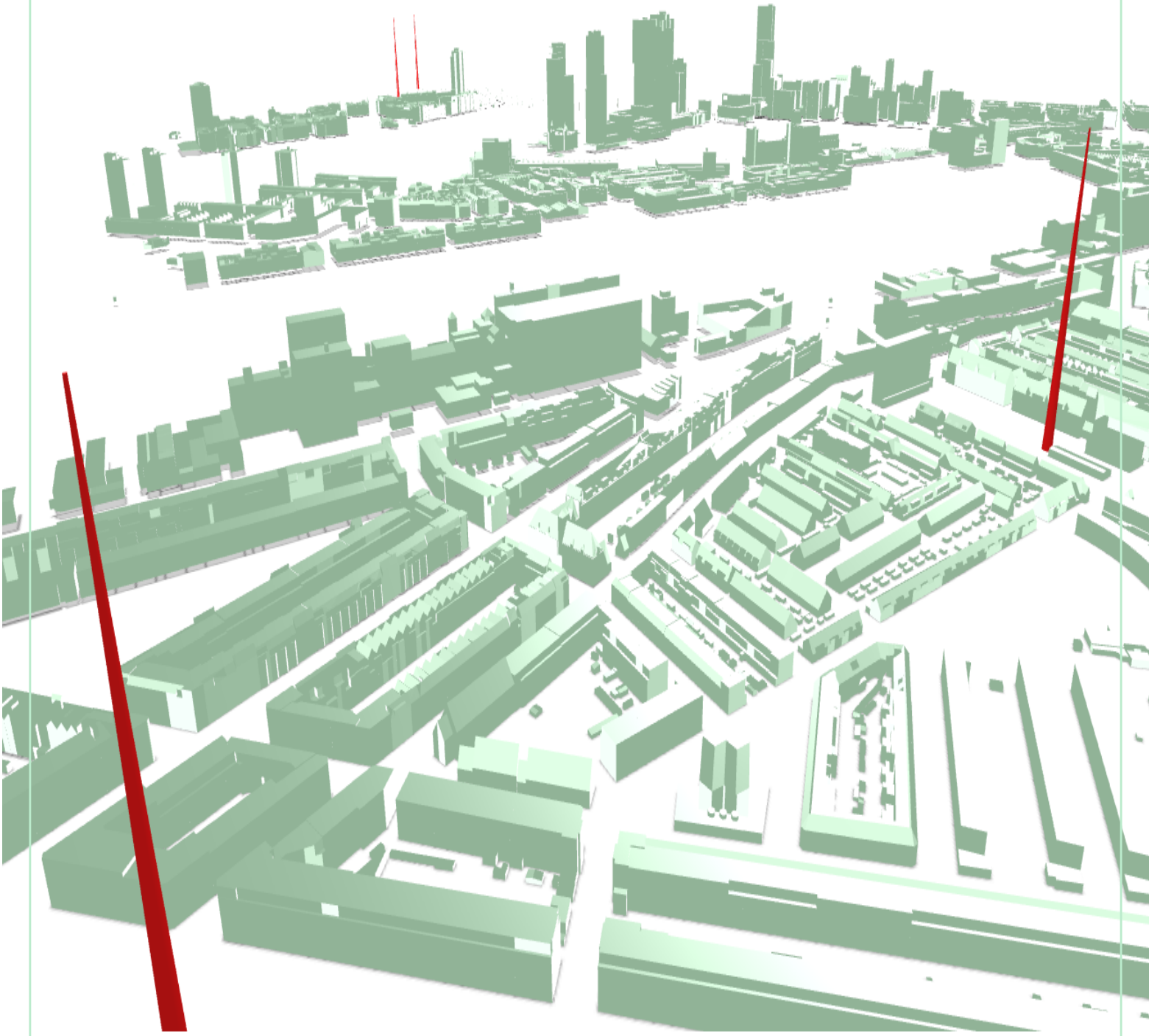
MSc thesis in Geomatics for the Built Environment

Correction Model for Particulate Matter Measurements with a Low-Cost Sensor Network in Rotterdam

Niek Bebelaar

April 2019



TU Delft

DCMR milieudienst Rijnmond

CORRECTION MODEL FOR PARTICULATE MATTER MEASUREMENTS WITH A LOW-COST SENSOR NETWORK IN ROTTERDAM

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Geomatics for the Built Environment

by

Niek Bebelaar

April 2019

Niek Bebelaar: *Correction Model for Particulate Matter Measurements with a Low-Cost Sensor Network in Rotterdam* (2019)

The work in this thesis was supported by:

OTB group
Department of OTB
Faculty of Architecture & the Built Environment
Delft University of Technology

Dienst Centraal Milieubeheer Rijnmond
Environmental agency for the province of South Holland
Schiedam

Supervisors:	Dr. M.J.P.M. Lemmens
	Dipl.Ing. A. Wandl
Co-reader:	Dr.Ir. S.C. van der Spek
Delegate of the Board of Examiners:	Drs. D.J. Dubbeling

Front cover: Screenshot of the Rotterdam 3D City Model visualized in QGIS3, with locations of DCMR air quality monitoring stations highlighted.

ABSTRACT

Low-cost air quality sensors can fill gaps between the sparse measurements done with high-quality national monitoring grids and might contribute to creating a more complete understanding of air pollution in an urban area. However, until there is no agreement on what degree of sensor accuracy is acceptable, the sensor data quality should be validated before governmental bodies use it as input for decision-making [Lewis and Edwards, 2016].

This research proposes a method to assess and improve the data quality of low-cost air quality sensors measuring *Particulate Matter* (PM). To answer the research question "How can accuracy and precision of Particulate Matter measurement results from a low-cost outdoor sensor network be improved by using a correction model, using data from reference sensors and additional sensors measuring inferencing phenomena?" an experiment setup with sensors operating under real-world conditions is applied.

Two *low-cost sensor nodes*, both containing a microcontroller, two low-cost PM sensors, and a temperature and humidity sensor, are placed at two locations in the city of Rotterdam. At those two locations, they are placed next to a high-quality air quality monitoring station from the environmental agency of Rotterdam. These *monitoring stations* provide benchmark data for the *low-cost sensor nodes*. A third data source provides data on air pressure and wind speed for the whole city of Rotterdam.

The data that originates from both sensor nodes and monitoring stations are matched and correlated with each other. Subsequently, the measurements from the low-cost sensor nodes are evaluated. Correlations and cross inferences of PM with other independent variables such as humidity, ambient temperature, wind speed and air pressure are investigated. Thereafter, utilizing the Stepwise Multiple Linear Regression method, various correction models are created that take various combinations of external variables into account. The correction models vary with respect to the amount of included external environmental variables and the polynomial degree. From all those possible correction models, the best correction model per location is selected by evaluating the Root Mean Square Error (RMSE) of the corrected dataset.

Consequently, the results of the chosen correction model are validated. It is found that the best performing correction models are those that include only the original PM data and the effect of adding more independent variables is limited. The best correction models for the four low-cost PM sensors are able to decrease the RMSE of the observations: the original normalized RMSE ranged from 0.0918 to 0.1249, while the corrected normalized RMSE range from 0.03110 to 0.03759. So, it is possible to improve the data quality of low-cost PM sensors with the stepwise MLR method and setup as shown in this research. However, including parameters for independent variables humidity, temperature, air pressure or wind speed does not improve the data quality significantly.

Besides, when an extra sensor node is placed in an air quality monitoring network as described in this research, it is necessary to create a correction model for that specific sensor. Like Castell et al. [2017] and Mukherjee et al. [2017] also found, it is necessary to calibrate each individual low-cost sensor before adding it to an air quality measuring network of the type as described in this research. Namely, it is found that for each low-cost PM sensor in the network different correction models are created.

ACKNOWLEDGEMENTS

In this section I would like to thank a few people who helped me during the graduation research.

First, I would like to thank my supervisors Mathias Lemmens and Alexander Wandl for the regular meetings which motivated me to keep going. Thank you for the guidance during the graduation process and the useful advice and valuable feedback. Furthermore, thanks to Ed van der Gaag and his colleagues from DCMR for the support. Thank you for the information regarding air pollution monitoring techniques and for the support with placing my own sensor nodes. Also thanks to Teun Verkerk and his colleagues from the Science Centre, where I was welcome to create the sensor nodes. Martijn Meijers, thank you for configuring the database and data collection software on the TU Delft server. Thank you Stefan van der Spek for giving useful comments and feedback as co-reader. Finally, I would like to thank my family, friends and fellow students for all the support they gave during the graduation process!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Algorithms

ACRONYMS

API Application Program Interface . . . 35
CO Carbon Monoxide . . . 2
CO_2 Carbon Dioxide . . . 9
CSV Comma Separated Values . . . 43
DCMR Dienst Centraal Milieubeheer Rijnmond . . . xi
EM Electromagnetic . . . 17
ESA European Space Agency . . . 4
HDMR High-Dimensional Model Representation . . . 19
JSON JavaScript Object Notation . . . 35
LML Landelijk Meetnet Luchtkwaliteit . . . 2
LSA Least Squares Adjustment . . . 24
MAE Mean Absolute Error . . . 19
MBE Mean Bias Error . . . 19
MLR Multiple Linear Regression . . . 23
NO Nitrogen Oxide . . . 2
NO_x Nitrogen Oxides . . . 9
NO_2 Nitrogen Dioxide . . . 4
O_3 Ozone . . . 2
PDF Probability Density Function . . . xii
PM Particulate Matter . . . xi
PM_1 Particulate Matter of 1μm . . . 9
$PM_{2.5}$ Particulate Matter of 2.5μm . . . xi
PM_{10} Particulate Matter of 10μm . . . xi
RIVM Rijksinstituut voor Volksgezondheid en Milieu . . . 2
RMSE Root Mean Square Error . . . xiii
SD Standard Deviation . . . 25
SO_2 Sulfur Dioxide . . . 4
VAR Vector Autoregression . . . 18
VECM Vector Error Correction Model . . . 18
VIF Variance Inflation Factor . . . 26
VOC Volatile Organic Compounds . . . 9
WSN Wireless Sensor Network . . . 19

1 INTRODUCTION

1.1 SMART CITIES AND THEIR TECHNOLOGIES

Dutch municipalities like Utrecht, Eindhoven and Enschede deploy sensor networks in their built environment and use those for monitoring various environmental conditions, for example noise levels, urban air quality, and movement of people, products and vehicles [Naafs, 2017]. The European Union provides subsidies to governments and knowledge institutions for developing so called *smart city* initiatives, such as with the Urban Innovative Actions program [Barrosso, 2014]. Besides being a business concept used by businesses worldwide the "smart city" has also technical aspects, for example sensor hardware and software [Marshall, 2017; Naafs, 2017]. Next to that, (open) standards are developed in order to improve communication between sensors, applications, and end users of data from the smart city [Liang et al., 2016].

Proliferation of available low-cost micro-controllers, sensors and actuators has helped the development of new digital and electronic kits [Salim, 2012]. These sensors are now easier to configure and accessible to a wider group of users. They can measure phenomena such as air quality, temperature, humidity, noise, solar radiation, or the current location of people or assets. This data regarding the urban microclimate can be useful for professionals such as architects and urban planners [Pijpers-van Esch, 2015]. For example, the Delft University of Technology and other stakeholders have the ambition to use a new city park as laboratorium for – urban – climate research.[1]. At various locations in the new Van Leeuwenhoekpark (see figure 1.1) climate aspects such as windspeed and temperature will be sensed, giving insight in the effect of a "green" area in a mainly concrete environment.

When two or more sensors are connected to each other with communication protocols such as Wi-Fi, Bluetooth or LoRa, it becomes a *sensor network*. Advantages of a reliable but low-cost sensor network is that they can be deployed in high quantities in order to understand microclimates in cities better. Moreover, due to the existence of a sensor network platform that uses wireless communication such a platform can be extended with other types of sensors. Some of the projects provide the gathered information as open data. For example the Dutch government has the ambition to provide government data as open data [Ministerie I&M, 2015], thereby contributing to a more transparent government. Using data that originates from low-cost sensor networks would contribute to that ambition. However, a problem is that in most countries there is no legislation and regulation regarding testing and verification of the data from these type of sensor networks [Lewis and Edwards, 2016]. Thus, the *quality of the data* from these sensor networks is unknown.

1.2 MONITORING AIR POLLUTION

Epidemiological studies show positive associations between exposure to outdoor *air pollution* and human mortality [Bentayeb et al., 2015]. A literature review by Pope and Dockery [2006] concluded that "the exposure of humans to fine particulate air pollution has adverse effects on cardiopulmonary health." Human health and air pollution are closely linked [Mead et al., 2013]. Emissions from different sources such as industry, road traffic and intensive livestock farming all have an influence on local air quality [Van Alphen and Pot, 2014]. Humans who are exposed to *Particulate Matter* (PM) have a higher risk of developing cardiovascular and respiratory diseases and lung cancer [Pijpers-van Esch, 2015]. Air pollution can be in the form of gas-phase species or in the form of PM.

Since air pollution is omnipresent and affects human health limits for air pollution concentrations are defined. Monitoring the concentrations of air pollution is a legal obligation for

1 https://nieuwdelft.nl/portfolio/van-leeuwenhoekpark/

Figure 1.1: Impression of the new Van Leeuwenhoekpark in Delft by LODEWIJK BALJON landscape architects. When the construction of the park is ready the Delft University of Technology is going to deploy sensors sensing the microclimate and use it as a "living lab".

public health agencies in The Netherlands.[2] This monitoring is often performed with location bound *monitoring stations* equipped with certified instruments [Castell et al., 2017]. The instruments at the monitoring stations – *monitoring instruments* – measure regulatory pollutants such as Carbon Monoxide (CO), Nitrogen Oxide (NO), Ozone (O_3) and PM. An example of a monitoring instrument is the Beta Attenuation Monitor "BAM-1020" from the manufacturer MetOne Instruments. This instrument is capable of monitoring Particulate Matter of 1 mm (PM1), Particulate Matter of 2.5 mm (PM2.5), and Particulate Matter of 10 mm (PM10). These various sizes of PM particles affect visibility, human health, global climate and the urban microclimate. When installed, operated and calibrated according to established procedures the instrument is certified as PM monitoring method for environmental agencies in the United States and The Netherlands [Mukherjee et al., 2017]. One BAM-1020 monitoring instrument has a price tag of around €15000.[3]

In the Netherlands the Rijksinstituut voor Volksgezondheid en Milieu (RIVM) – the National Institute for Public Health and the Environment – is responsible for monitoring air quality. Broadly, RIVM uses three complementary approaches to get a clear insight in the air quality in the whole of the Netherlands: using a nationwide measurement network, the registration of emissions, and interpolated pollution models.

- One approach is using measurements on specific locations that together constitute a nationwide air quality monitoring network: the "Landelijk Meetnet Luchtkwaliteit (LML)" [Van Alphen and Pot, 2014]. Most of the monitoring stations in the LML do the analysis and calibration automatically. RIVM currently uses BAM-1020 monitoring instruments on these locations to monitor concentrations of PM. In The Netherlands the PM monitors are located countrywide, though most of them are in and nearby the Randstad metropolitan area, see figure 1.2.
- The Emissieregistratie – Emission Registration – method takes account of collecting, managing, editing and reporting of Dutch emission data. With this method are emissions calculated based on the activities per location, the expected air pollution due to these activities, and nationwide statistical information regarding the extent of these

2 https://wetten.overheid.nl//BWBR0003245//2019-01-01//#Hoofdstuk5
3 https://www.alibaba.com//product-detail//BAM-1020_118816030.html

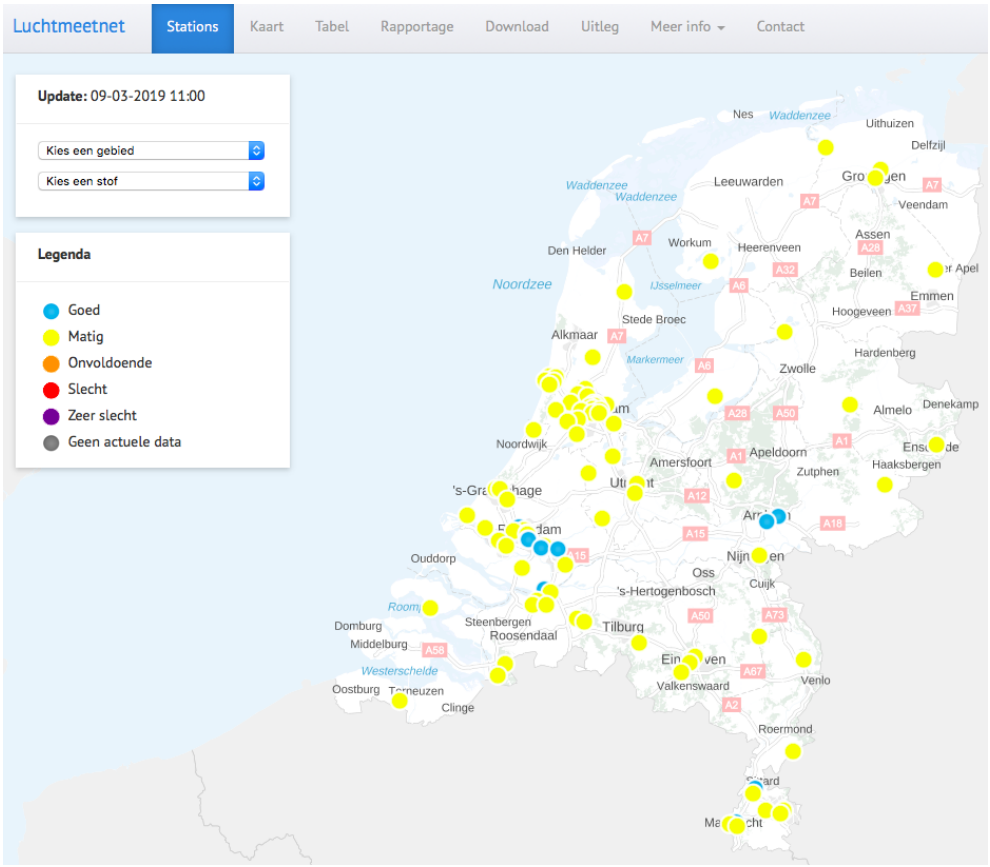


Figure 1.2: Locations of the monitoring stations in the Landelijk Meetnet Luchtkwaliteit (screenshot from `www.luchtmeetnet.nl`). The colors indicate the real-time air quality (blue = ”good”, yellow = ”moderate”).

kind of activities. The supply of input data for the Emission Registration is decentralized: it is done by a large amount of stakeholders [Van Alphen and Pot, 2014].

- The third approach is calculation of pollution models. The measurements from the monitoring instrument network and Emission Registration are input for these interpolation models.

Another initiative to monitor air quality is done by the European Space Agency (ESA). In October 2017, researchers from this European institution launched the Sentinel 5-p satellite. This satellite contains the TROPOMI atmosphere measuring instrument, capable of measuring O_3, CO, Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2) and aerosol properties [De Vries et al., 2016]. Data from this satellite has already been provided. However, this initiative require high investment costs and the resulting data is still on a small – coarse – spatial scale.

Next to that, in another study air quality is monitored with instruments mounted on vehicles. This research demonstrated a measurement approach that revealed local air pollution patterns in an urban region. The result was a spatial precision that is higher than monitoring with static monitoring instruments [Apte et al., 2017].

Since a broad range of sources influence air quality, the temporal and spatial distribution of PM concentrations may vary significantly in a region [Wang et al., 2015]. This asks for an approach that accounts for those spatial differences. Low-cost sensor networks could contribute to policy and decision making processes that might result in a healthier built environment. Especially sensors that monitor air quality and air pollution.

1.3 A LOW-COST SENSOR NETWORK AND DATA QUALITY

An earth-based low-cost sensor network can help densifying the measurement grid of the RIVM and can yield measurement results on a larger scale – less coarse – than ESA provides. Low-cost sensors can fill gaps between the sparse measurements that are done with the monitoring network of RIVM and might contribute to creating a more complete understanding of air pollution in an area, i.e. on the micro climate level. However, until there is no agreement on what degree of sensor accuracy is acceptable, those sensor results should be validated before governmental bodies use it as input for decision-making [Lewis and Edwards, 2016]. As Lewis and Edwards [2016] state it: "Even sensors that are designed for entertainment or awareness-raising need appropriate labelling to define their capabilities."

Low-cost sensors – having prices ranging from €10 to €100 – measuring air pollution can be categorized in two groups: the ones measuring gasses (e.g. CO, NO, O_3) and sensors measuring PM [Castell et al., 2017]. The microcontroller that is part of these sensors transforms the output signal from the sensor into a digital value. Often, such an air quality sensor is placed on a *sensor node*. A sensor node can contain multiple sensors measuring gasses and/or PM, and a microcontroller that integrates the electronics, stores data, and eventually transmits the data to a central server via a communication protocol. Kumar et al. [2017] concluded that for designing a low-cost, energy efficient, portable real-time system for monitoring indoor air quality a multidisciplinary approach is needed. One needs to have knowledge about accuracy of measurements, effects of other gases on air quality (*interference*), wireless communication networks, data storage and power consumption.

Li and Biswas [2017] performed a laboratory experiment and proposed a method to improve the results from low-cost sensors measuring air quality. Doing outdoor air quality improvements with more complex aerosols is not investigated yet [Li and Biswas, 2017].

Moreover, the influences of cross-sensitivities, temperature, and humidity on air quality measurements implies a need for data processing on the sensor node [Postolache et al., 2009]. They suggest four types of data processing on the sensor node: data smoothing, continuous data calibration to overcome the problem of cross sensitivity, correction for temperature and humidity dependency, and an aggregation algorithm that sends only when the values change significantly.

Data quality

Data quality is recognized as a relevant performance issue of decision making activities, inter-organizational cooperation requirements and operating processes [Batini et al., 2009; Lewis and Edwards, 2016]. Batini et al. [2009] focus on two main steps in the research on data quality methodologies: assessment and improvement. *Assessment* is comparing a value to a

reference value in order to diagnose the data quality. *Improvement* relates to the selection of steps, strategies and techniques in order to reach a new level of data quality [Batini et al., 2009].

Further, regarding the improvement of data quality Batini et al. [2009] distinguish two general strategies: data-driven and process-driven. Data-driven strategies improves data quality by directly modifying the value of a data record, while process-driven strategies improve quality of data by redesigning the process that create or modify the data [Batini et al., 2009]. This research utilizes the data-driven strategy for quality improvement. More precise, this research proposes a method that localizes the error and corrects the error in the datasets, i.e. *identifying and eliminating data quality errors by detecting the records that do not satisfy a given set of quality rules"* [Batini et al., 2009]. Therewith, the proposed method improves the *accuracy* – the closeness of a value v to the elements of the corresponding definition domain D [Batini et al., 2009] – of the data. Other strategies to improve data quality, such as acquiring completely new datasets when the quality is poor, selecting data sources based on their trustworthiness, or using the minimization of costs as definition for quality improvement are not included in this research. Also, other dimensions wherein data quality can be expressed, such as completeness and consistency, are no part of this research.

1.4 LINK WITH GEOMATICS FOR THE BUILT ENVIRONMENT

This research is about tools for measuring spatial-temporal phenomena, which is the focus of Geomatics for the Built Environment. Namely, the measurements take place in the built environment and thus have a geo-component. Further, phenomena that are sensed for this inquiry – i.e. air pollution, relative humidity, temperature, wind speed and wind direction – change on a regular basis and can be affected by how the built environment is designed and/or used. Therefore, quality of the urban environment can be increased by doing targeted interventions that eventually change the values of the phenomena. In order to do so, tools should be able to deliver data on a large – fine – spatial scale, and these tools should deliver reliable data from which information and knowledge can be derived.

1.5 RESEARCH OBJECTIVES

The problem statement is that the quality of the data from low-cost sensors measuring air quality in the built environment is unclear, i.e. the accuracy of the data is unknown. On the other hand, these low-cost sensor systems are deployed and will be deployed in the future in several cities [Naafs, 2017; Ministerie I&M, 2015]. It is then possible that citizens, professionals, and/or decision makers attach meaning to information derived from data of those sensors, which may result in problems when the data is not validated since the data may be incorrect in the first place [Lewis and Edwards, 2016]. Therefore, the first aspect of the research objective is to *assess* the data quality of the original PM data from these low-cost sensors.

In this research it is assumed that the accuracy and precision of the low-cost sensors is lower than those of the PM monitors from the Dutch RIVM. Therefore, the second aspect of the research objective is to *improve* the data quality from the low-cost sensors. How can the data quality from the low-cost sensor observations be improved? The proposed approach in this research is to use an *error correction model*, which utilizes a formula to manipulate the accuracy of an observation value. Sensors measuring inferencing phenomena might be needed for acquiring data to correct the PM measurement results. The performance of the correction model should be expressed in an unambiguous evaluation metric.

Therefore, the research objective is to investigate how a low-cost sensor network can support the established high-cost and high-quality PM monitoring network. This makes the air monitoring network denser, gives more insight in local air quality variances as a result of different designs of the built environment, and it reduces operation and maintenance costs. The rationale is that the low-cost sensors are less accurate, precise and reliable compared with the high-quality monitoring instruments – the ones used by RIVM – although that can be improved by applying a correction model that corrects for atmospheric and environmental

phenomena such as temperature, humidity, air pressure, wind speed and others. Investigating how this correction model should look like is the goal of this research.

1.6 RESEARCH QUESTIONS

Based on the research objectives is the following main question defined:

How can accuracy and precision of Particulate Matter measurement results from a low-cost outdoor sensor network be improved by using a correction model, using data from reference sensors and additional sensors measuring inferencing phenomena?

Quality of the low-cost sensor measurements is expressed in accuracy, i.e. how close are the observed values to the true value of the quantity being measured? The measurement results of high quality reference air quality monitors is in this research considered as the 'true' value. The precision of a PM measuring instrument is also considered: the same type of instrument should yield the same values under similar conditions (time and location). Moreover, the main research question implies that the sensors are relatively low-cost. This implies that the proposed methodology can be used in order to scale up sensor projects and in order to make the sensor grid denser. Another key concept in the main research question is the validation of the correction model. Because a correction model is created with data from one location, but should also be able to improve data quality on another location. The following sub-questions are formulated:

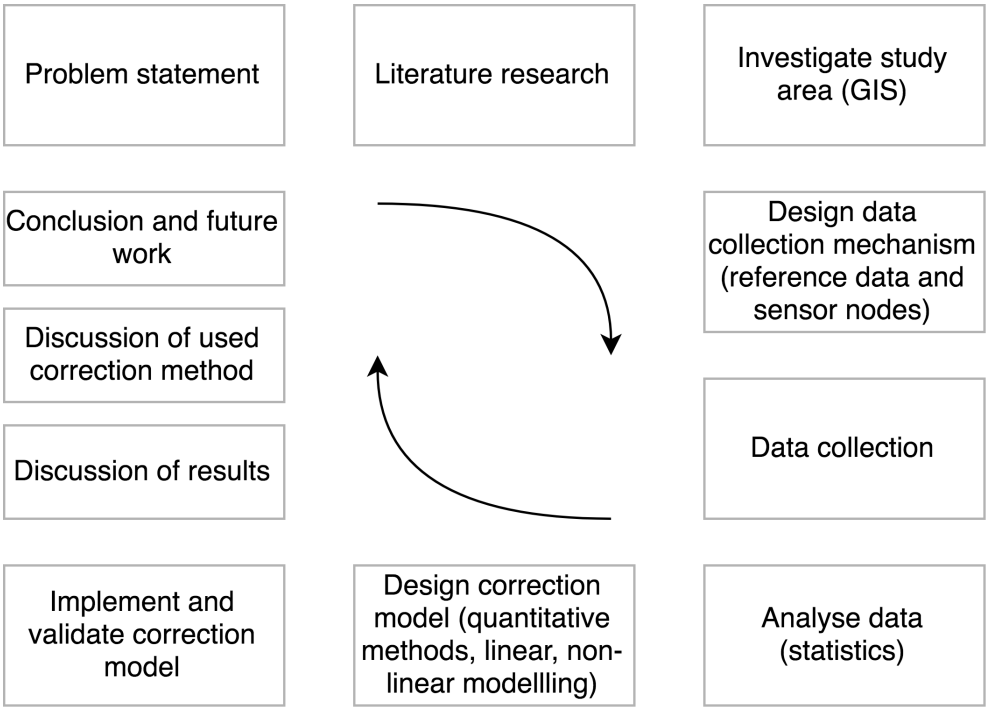


Figure 1.3: Research design

1. *How do temperature, humidity, air pressure, and wind speed affect Particulate Matter measurements?* The first sub-question focuses on which environmental phenomena influence the air quality measurements, depicted from a literature review. This first sub question clarifies which additional sensors are to be included on the sensor nodes and which external datasets need to be included in the research. Those additional sensors sense phenomena and this data is input for the correction model.
2. *What is a good experimental setup for calibrating air quality measurements and how to develop this sensor setup?* This sub-question focuses on designing, developing and implementing the sensor system: the specific hardware, choices regarding storing and transmit-

ting the data, sampling interval of the measuring system, and the synchronization with high quality reference data.

3. *What can be acquired from existing correction models in the field of air quality monitoring and related fields of research?* This question implies a literature review on correction models that are currently used within the field of air quality and related research fields. And the aspects of these correction models that can be used for the newly developed model as proposed in this research.
4. *How to create the new correction model?* Sub-question 4 is about making the correction model. Results from the air quality sensors and the additional sensors are read and plotted in graphs in order to find relations between the different phenomena. Insights from sub question 3 can be used to create the new correction model. Statistics from the datasets, such as the analysis of variance, can provide insight into the results.
5. *How to validate the correction model?* The time and domain when the correction model is valid is investigated in the final sub-question. If there is no typical domain when the correction model is valid, a generic correction model can be applied. That model corrects the data with one and the same formula for all domains. However, if there are specific domains for which different formulas need to be created, a specific correction model is created. These specific domains are for example the time of the day (peak hours, off-peak hours, night), specific seasons, or when the temperature is within a certain range. Such a specific correction model consists of more formulas, which one to take depends on the domain.

With the sub-questions are all steps in the research design of figure 1.3 covered.

1.7 READING GUIDE

In the next Chapter 2 are the theoretical framework and related work in the fields of air quality monitoring and (error) correction models discussed. A number of relevant air quality sensor network projects are reviewed and implications for this research are taken into account. In Chapter 3 is the methodological framework of the research described based on findings from theory. Chapter 4 elaborates on the implementation of the proposed methodology. The relationships between the independent and dependent variables and the *baseline measurement* are also discussed here. In Chapter 5 are the results of the implemented methodology shown and discussed. Finally, Chapter 6 concludes this report by answering the research questions and with recommendations for future work.

2 THEORETICAL FRAMEWORK AND RELATED WORK

This chapter elaborates on the theoretical background of this thesis. The focus is on research in the fields of air quality monitoring, low-cost sensor networks in an outside environment, and error correction models. Theoretical concepts that are related to the research question are air quality and air pollution, PM and the monitoring or measurement of PM concantrations, low-cost and low-energy sensor networks, accuracy and precision. After explaining these topics an overview of related research in this field is given.

2.1 AIR QUALITY AND AIR POLLUTION

One branch of the study of urban metabolism is studying a city's flows of water, materials and nutrients in terms of fluxes of mass [Kennedy et al., 2011]. Example applications of urban metabolism studies are urban sustainability indicators, urban greenhouse gas emission calculations, urban metabolism expressed in mathematical models for policy analysis, and sustainable urban design. Air pollution and air quality monitoring and analysis can be part of the *sustainable urban design* branch.

Definitions of air pollution and Particulate Matter

Air pollution is defined as *"when gases or aerosol particles that are emitted anthropogenically, are build up in concentrations sufficiently high to cause direct or indirect damage to plants, animals, other life forms, ecosystems, structures, or works of art"* [Monks et al., 2009]. Amounts of several chemical compositions such as Nitrogen Oxides (NOx), CO, and Carbon Dioxide (CO_2), can affect air quality. Those chemical compositions are not the focus of this research: it only focuses on PM. The definition of PM is: *"air suspended mixture of solid and liquid particles that vary in number, size, shape, surface area, chemical composition, solubility, and origin"* [Postolache et al., 2009]. The chemical mixture and mass concentrations of PM can differ per region [Monks et al., 2009]. PM is subdivided in the following classes [Environmental Protection Agency, 2016]:

- PM_{10} are course particles with mass concentrations with sizes of 10μm to 2.5μm and smaller in aerodynamic diameter. Those particles originate from suspension or resuspension of soil, dust and sea salt [Hendriks et al., 2012]. Or from other crustal materials or events such as roads, mining, farming, volcanic eruptions, and wind storms. However, the origin for most of the PM_{10} dust particles observed the Netherlands is undetermined, see figure 2.1.
- $PM_{2.5}$ are fine particles with a diameter of 2.5μm to 1μm. Coal burning, wood burning, use of vehicles with gasoline and diesel engines and industrial proceses are examples of combustion processes from which fine particles are directly emitted. Volatile Organic Compounds (VOC) also contribute to the amount of ambient fine particles. Those are chemical, gaseous products originating from the transformation of organic aerosols to other products. See figure 2.2.
- Particulate Matter of 1μm (PM_1) are called ultra-fine particles or submicron particles and have a diameter of less than 1μm. Ultra-fine particles also originate from vehicle exhausts and atmospheric photochemical reactions. These particles can move from the human lungs into the veins and via there to other parts of the human body [Monks et al., 2009; Mukherjee et al., 2017].

Figure 2.3 gives insight in the relative sizes of these PM classes. Although PM_1 is the most dangerous class regarding human health, it is also hardest to monitor or sense with technologies currently available. Therefore RIVM does not measure PM_1 but rather $PM_{2.5}$. Therefore, since there is no reliable PM_1 data available this research focuses only on $PM_{2.5}$.

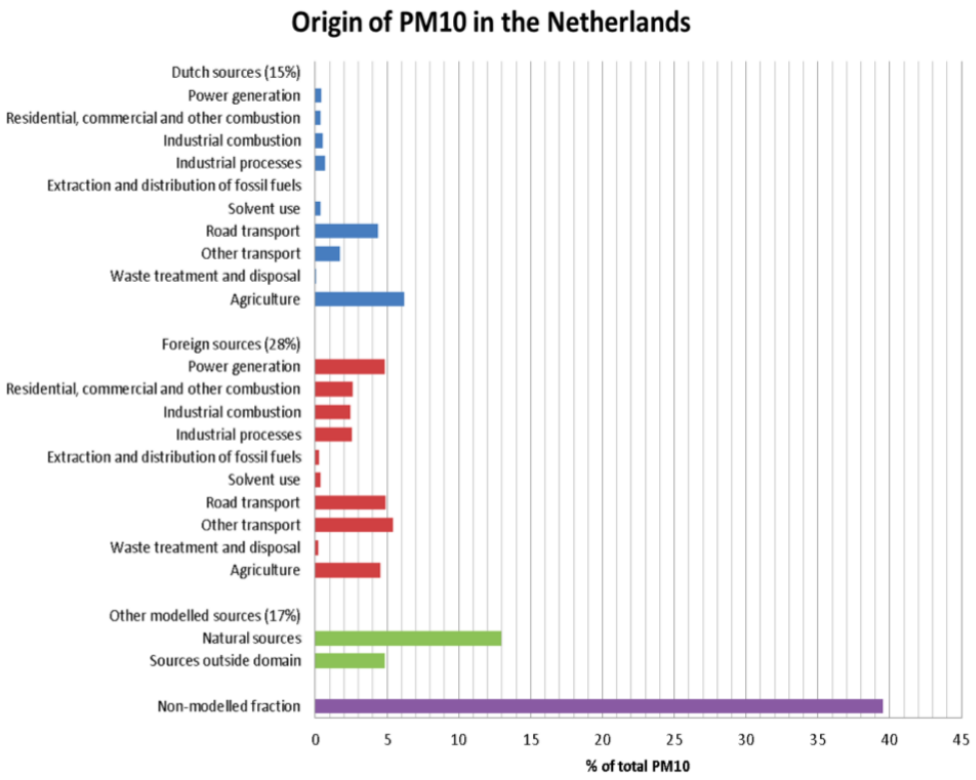


Figure 2.1: Origin of PM10 in the Netherlands. 70 to 80% of PM2.5 in the Netherlands is anthropogenic [Hendriks et al., 2012]. Notice the high percentage of non-modelled fraction, which shows the uncertainty of the sources of PM.

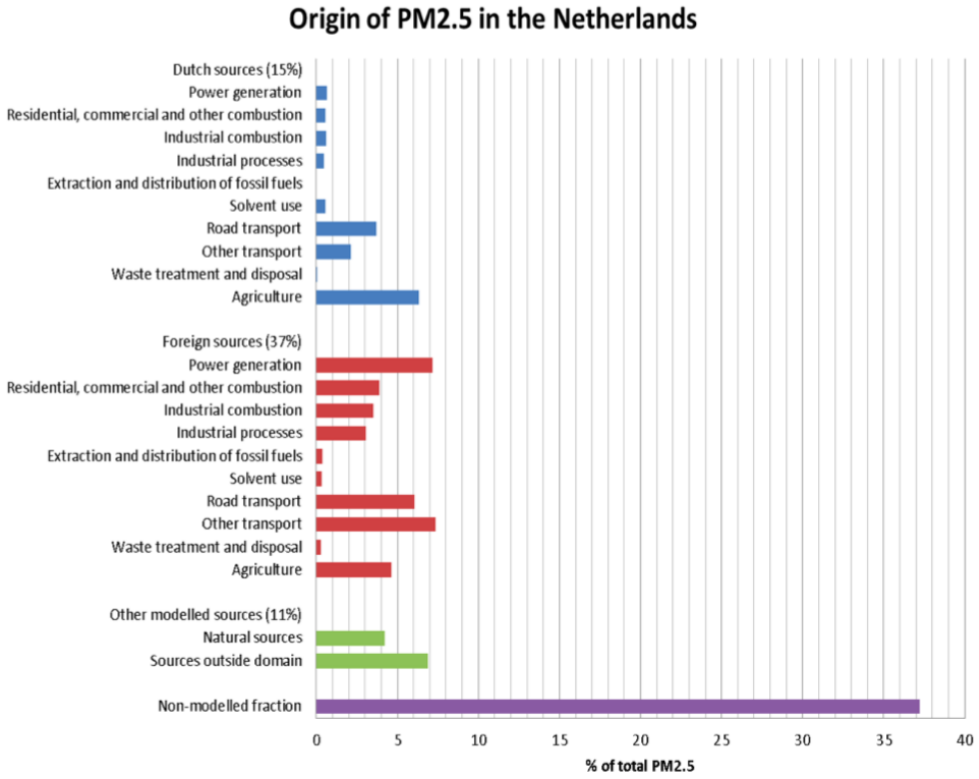


Figure 2.2: Origin of PM2.5 in the Netherlands. 80 to 95% of PM2.5 in the Netherlands is anthropogenic [Hendriks et al., 2012].

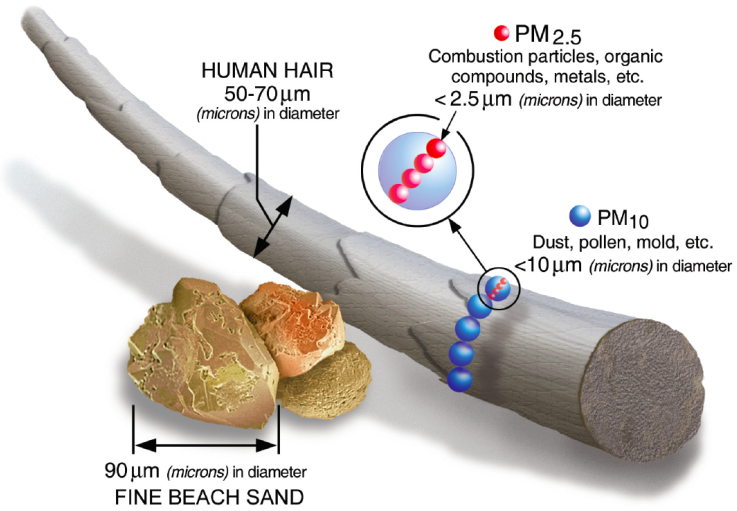


Figure 2.3: PM classes visualized [Environmental Protection Agency, 2016]

One common way to express the amount of PM is in units of mass per volume of ambient air: it is conventional to use micrograms per cubic meter ($\mu m/m^3$)[1]. Another way is to express it in Parts Per Million (PPM) or Parts Per Billion (PPB), i.e. the volume of pollutant per million or billion volumes of ambient air. In the Netherlands RIVM expresses PM in ($\mu m/m^3$) [Hendriks et al., 2012], the concentration unit which is also used in this research.

Legal norms for PM concentrations

The national legal norm for PM10 is a year average of $40\mu m/m^3$. For PM is the year average norm a concentration of $25\mu m/m^3$ [Van Breugel and Van den Elshout, 2018; Juliana et al., 2009]. The norm for PM2.5 is stricter because it is more dangerous than PM10 since these particles can penetrate deeper into human lungs. Next to the year average norms are also peak norms defined, which focus on day peak averages. Per year, the PM10 concentration may exceed a limit of $50\mu m/m^3$ for 35 times. For PM2.5 there is no such norm for day peak concentration.

In 2017 are the norms for PM2.5 and PM10 in the Rijnmond region in the Netherlands – the study area – not exceeded [Van Breugel and Van den Elshout, 2018]: see the "*Gem*" column of figure 2.4.

Further, the PM concentrations are close to the World Health Organization (WHO) guideline of 20 and $10\mu m/m^3$ for PM2.5 and PM10, respectively [WHO, 2006]. Moreover, figure 2.4 shows more key figures for the distribution of the particle concentrations. The "P50" and "P98" are the year averages of concentrations in the 50th and 98th percentile, respectively. The relatively high differences between those averages indicate days with extreme values. In the "Max" column are maximum extreme values indicated per street. For PM10 there is the extra norm that $50\mu m/m^3$ may not be exceeded for more than 35 days per year. The information in the "D50" column shows how many days that happened: the norm is not exceeded. Finally, the "Aantal" – "Amount" – column indicates the number of days in 2017 that the BAM-1020 monitor on each specified location was working correctly.

2.2 MEASUREMENT OF PM

The concentration of PM is monitored for various application scenarios. For example in waste incinerators of industrial plants, indoor air pollution monitoring, or outdoor urban air quality monitoring. The latter is the application scenario of this study. Ambient air pollution is monitored with several types of detection systems. Those can be "static" – non-moving, location bound – monitoring instruments, mobile units mounted on vehicles, or

1 https://www2.dmu.dk/AtmosphericEnvironment/Expost/database/docs/PPM_conversion.pdf

Tabel III. Kentallen van de concentratieverdeling van fijnstof ($PM_{2.5}$) in 2017 (in µg/m^3)

$PM_{2.5}$	Kalenderjaar 2017					
	Middelingstijd in uren	**24**	**24**	**24**	**24**	**24**
	Kental	**Gem**	**P50**	**P98**	**Max**	**Aantal**
	EU Grenswaarde	25				
V	Rotterdam-Pleinweg	14,7	11,5	47,0	79,3	357
S	Rotterdam-Zwartewaalstraat	11,6	8,9	42,3	75,8	351
V	Ridderkerk-Hogeweg	13,0	10,1	40,5	77,4	357
V	Rotterdam-Oostsidelinge	12,7	10,6	41,2	72,1	353
S	Schiedam-Alphons Ariënstraat	12,6	9,6	39,8	76,7	347
S	Maassluis-Kwartellaan	11,9	9,2	36,5	70,5	354
S	Hoek van Holland-Berghaven	11,3	8,3	41,5	76,8	344

Tabel II. Kentallen van de concentratieverdeling van fijnstof (PM_{10}) in 2017 (in µg/m^3)

PM_{10}	Kalenderjaar 2017						
	Middelingstijd in uren	**24**	**24**	**24**	**24**	**24**	**24**
	Kental	**Gem**	**P50**	**P98**	**Max**	**D50**	**Aantal**
	EU Grenswaarde	40				35^{15}	
S	Hoogvliet-Leemkuil	19,6	17,0	52,3	87,2	9	358
V	Rotterdam-Pleinweg	22,6	19,9	54,3	79,7	10	346
S	Rotterdam-Zwartewaalstraat	18,1	15,5	48,8	79,7	7	350
V	Ridderkerk-Hogeweg	21,2	19,1	50,2	80,4	8	355
V	Rotterdam-Oostsidelinge	21,7	19,5	48,4	81,8	7	352
V	Rotterdam-Statenweg	21,1	18,0	50,1	81,3	8	351
S	Schiedam-Alphons Ariënsstraat	20,6	18,1	51,5	81,8	7	319
S	Maassluis- Kwartellaan	19,7	17,4	47,5	75,1	4	355
S	Hoek van Holland-Berghaven	23,8	21,3	56,8	99,1	13	358

Figure 2.4: Key figures for concentration distribution of PM2.5 and PM10 in 2017 [Van Breugel and Van den Elshout, 2018].

hand-held instruments [Apte et al., 2017]. Ambient air pollution monitoring instruments differ in terms of physical portability, sensitivity, and level of automation [Carminati et al., 2011], and therewith in market price. Though the most important difference between those types of detection systems is the used PM measurement technique. These techniques are the gravimetric weight method, opacimeter light scattering method, triboelectric method, beta absorption and laser scattering. Since the beta absorption and laser scattering methods are applied in the current research they are discussed in-depth, while the other three are now only discussed shortly.

Gravimetric weight method

This is the traditional technique whereby polluted ambient air is forced to flow through a clean filter of which the mass is known [Yu et al., 2017]. The concentration of PM is measured by weighting the mass of the filter after a given time. The difference in weight of the sample filter before sampling and after sampling results in a value for the PM concentration [Carminati et al., 2011]. The weighting is performed in a laboratory setting and this method is most accurate.

Opacimeter light scattering

This is an optical technique, which uses the weakening-ratio of a beam of light as basis to estimate the PM concentration [Carminati et al., 2011]. The particulate matter in the air sample crosses visible light which is absorbed, reflected or scattered. This light scattering method uses the Mie scattering theory of particles [Yu et al., 2017].

Triboelectric sensors

Triboelectric sensors use a technique that measures the electrical current of an electrical charge which is generated by electrostatic mechanical friction. The mechanical friction is the result of flowing particles that impact with the sensing electrode in the sensor [Carminati et al., 2011].

2.2.1 Beta absorption method

The beta absorption technique uses beta radiation of a ribbon filter that has been exposed to the ambient air, which will include a concentration of PM. A Geiger counter measures the variation of the intensity of beta radiation before and after exposing to the ambient air. The microcontroller in the sensor calculates a concentration value based on the beta radiation intensity before and after exposure [Mukherjee et al., 2017; Met One Instruments, 2016].

Official methods for acquiring PM concentrations

In the United States and the European Union – including The Netherlands – there are various standardized methods to measure PM concentrations with Beta Attenuation Monitors. In the United States, the official methods are "EQPM-098-122" and "EQPM-0308-170" for PM2.5 and PM10, respectively [US EPA, 2016]. In The Netherlands is the proprietary "NEN EN 12341" method the official standardized method, which is in accordance with the "2008/50/EG" guideline from the European Union [EU, 2008; NEN, 2014]. The NEN method is the Dutch implementation for this guideline and describes one possible standard reference method for PM10 and PM2.5. Regional environmental agencies such as DCMR – responsible for monitoring in the study area Rotterdam – use BAM-1020 monitoring stations, which in turn use the Beta Absorption technique (see figure 2.5). When properly installed, operated and calibrated according to the national standard, this monitoring instrument is the official certified method for environmental agencies in The Netherlands [Mukherjee et al., 2017], and can therefore be regarded as high-quality monitoring instruments.

The USA EQPM and NEN standards prescribe how the BAM-1020 monitors should be installed and configured. For example which type of air inlet to use, the type of cyclone and fiber tape, which external temperature and barometric pressure sensors to use, valid firmware versions for the software, measurement times, and sample intervals [US EPA, 2016]. See figure 2.6 for a detailed description of the standard one-hour cycle timeline for the PM measurements, as done by DCMR. In short, 8 minutes after the start of the hour the particles

Figure 2.5: The BAM-1020 monitoring station from Met One Instruments.

in the ambient air is collected, for a total of 42 minutes. The BAM-1020 pulls ambient air through the machine and the result is a concentration of PM in microgram per m/m^3. How does this process from dust on a filter tape to data on particulate matter concentrations in ambient air go?

Minutes	Activity	Description
0 - 8	Baseline measurement	The BAM-1020 immediately advances the filter tape forward one "window" to the next fresh, unused spot on the tape. This takes a few seconds. The new spot is positioned between the beta source and the detector, and the BAM begins counting beta particles through this clean spot for exactly eight minutes (I_0).
8 - 50	PM collection	The BAM-1020 stops counting beta particles through the clean filter. This baseline measurement yields a value for (I_0). The tape is moved four windows forward, away from the beta detector and now positioned directly under a nozzle through which the ambient air is going to flow. This takes a few seconds. The unit then lowers the nozzle onto the filter tape and turns the vacuum pump on, pulling particulate-laden air through the filter tape on which I_0 was just measured, for 42 minutes at 16.7 liters per minute.
50 - 58	Beta ray detection	The BAM-1020 turns the vacuum pump off, raises the nozzle, and moves the filter tape backwards exactly four windows. This takes a few seconds, and puts the spot that was just loaded with particulate back between the beta source and the detector. The BAM begins counting beta particles through the now dirty spot of tape for exactly eight minutes (I_3).
58 - 60	Performing calculations, then idle	The BAM-1020 stops counting beta particles through the dirty spot (I_3). The unit uses the I_0 and I_3 counts to calculate the mass of the deposited particulate on the spot, and uses the total volume of air sampled to calculate the concentration of the particulate in milligrams or micrograms per cubic meter of air. The BAM then sits idle and waits a few moments for any remaining time in the hour to expire.

Figure 2.6: Overview of the one-hour timecycle for BAM-1020. Adapted and recreated from the manual [Met One Instruments, 2010].

Retrieving a value for PM from beta attenuation

The dust being collected on the filter tape in the Beta Attenuation Monitor contains carbon-14, a naturally occurring carbon isotope [Met One Instruments, 2010]. The carbon-14 undergoes a beta decay process whereby high-energy electrons are emitted through radioactive decay. These high-energy electrons are also called beta rays, and the process of the decay of those beta rays is called beta ray attenuation. Hence the name "Beta Attenuation Monitor" for this type of particulate matter monitors. Due to the radioactive decay of carbon-14 the number of particles reduces over time. Before a clean filter tape in the monitor is exposed to ambient dust, its beta ray attenuation is determined by a beta ray detection unit, resulting in I_0 of clean filter tape. After collecting dust on this clean filter tape the tape is moved to a unit that detects beta rays from the – in the meantime dirty – filter tape. Thus, the magnitude of reduction of counted beta particles is a function of mass of absorbing matter between the carbon-14 beta source – the filter tape – and the detector [Met One Instruments, 2016]. Formula 2.1 shows this relationship.

$$I = I_0 e(-\frac{\mu M}{S}) \tag{2.1}$$

Where I is the measured beta ray intensity from the dusted filter tape in counts per unit time, I_0 is the measured beta ray intensity of the clean filter tap, μ the beta ray absorption cross section of the material on which the beta rays are absorbed (cm^2/mg), M is the aerosol mass deposited on the filter tape (mg) and S the spot area on the filter tape (cm^2) [Met One Instruments, 2016].

Absorption cross sections μ for species in ambient particulate matter such as iron oxide, silica, salt or soot are all approximately the same: μ depends only on mass of the absorbing

species and not on its chemical composition [Met One Instruments, 2016]. Therefore, one does not need to know ahead of time the chemical compositions of aerosols that are sampled in order to perform accurate mass measurements with a BAM. The constant value for μ is determined in the calibration process of the BAM. In this calibration process is a membrane used of which the mass density (M/S) is known. Repeated measurements of I and I_0 are used to acquire μ, according to formula 2.2. Each BAM1020 monitor will give small variations in the measured μ due to small differences in the types of aerosols in the air and manufacturing tolerances of the instrument itself. With the membrane calibration process of formula 2.2 is the response of each BAM1020 for monitoring projects of environmental agencies – such as DCMR – standardized.

$$\mu = \frac{S}{M} ln(\frac{I_0}{I}) \tag{2.2}$$

To further improve μ for more BAM1020 monitors, another calibration process is performed whereby a reference BAM1020 monitor is used. They both measure the same aerosol for a longer period, i.e. 48 or 72 hours. A linear regression of hourly outputs of one BAM1020 that is tested versus the other BAM1020 provides a slope k. This slope k is used for final calibration of μ. Data from a time period is corrected with the k parameter. In the cases when μ deviates too much the data from that time period can be labelled as invalid [Met One Instruments, 2016].

While μ is determined in the calibration process and I and I_0 are measured with the beta ray detector, densities of dusted air x can now be acquired. From Met One Instruments [2010] is the following formula 2.3 adapted.

$$x = \frac{1}{\mu} ln[\frac{I_0}{I}] \tag{2.3}$$

In the formula, x is mass density in mg/cm^2. It is not a concentration over a specified time period. In the BAM-1020 monitor however the ambient air is sampled at a constant flow rate Q for a specified time Δt. That sampled air is passed through a filter with surface A. The ambient concentration of particulate matter can be determined when x is determined. The following formula 2.4 then yields the quantities of ambient particulate matter, in $\mu g/m^3$

$$c = \frac{10^9 A}{Q \Delta t \mu} ln[\frac{I_0}{I}] \tag{2.4}$$

Where c is the concentration of ambient particulate matter in $\mu g/m^3$, A the cross sectional area in cm^2, Q the rate at which the particulate matter is being collected on that filter tape $(liters/minute)$, and Δt the sampling time $(minutes)$.

Thus the BAM1020 monitor uses beta radiation of the dust aerosols in order to acquire PM concentrations. With a calibration process is the μ parameter for the filter membrane in the BAM achieved. The beta attenuation of the clean filter, compared with the attenuation of the dirty filter and constant factors as shown in equation 2.4 result in a PM concentration in ambient air. Since Δt is in minutes, the particles collection time is 42 minutes, and the whole timecycle is 60 minutes, the resulting PM concentration c relates to an average concentration for one hour.

2.2.2 Laser scattering

Like with the light scattering technique, laser scattering is a technique whereby particle sizes can be obtained based on collision of light/laser rays with particles [Yu et al., 2017]. Carminati et al. [2011] consider laser scattering as the current state-of-the-art technique for real-time air quality monitoring because of its ability to analyse single particles and their sizes. With this technique, particles are focused into a single stream of fast flowing air which is generated by a ventilator. A laser beam is placed orthogonally relative to this air flow. Particles interacting with the laser beam reflect a scattered light. When the incident light of the laser passes through the dust particles, as shown in figure 2.7, the light scattering occurs. The amount of scattered light is counted with the photodetector. A mirror ensures that particle light scatter in the opposite direction is also collected by the photodetector. The transmitted

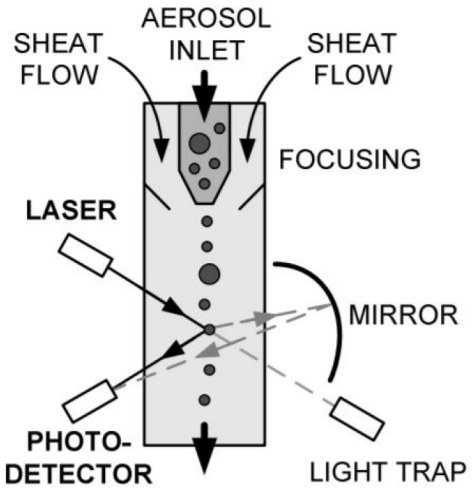


Figure 2.7: The working principle of laser scattering. Particles are focused in a single stream by flows of clean air. They are detected when a light pulse, originating from the laser, is scattered by the particle into the photodetector [Carminati et al., 2011]. The low-cost sensor used in the current research contains an internal microcontroller that outputs a PM concentration in digital format.

laser beams should not be counted with the photodetector: those laser beams are absorbed in a light trap.

Scattering is the reflection of Electromagnetic (EM) energy by particles suspended in the atmosphere. This scattering in the EM spectrum can be subdivided in three types: Rayleigh, Mie, and non-selective scattering [Lemmens, 2016a]. Rayleigh scattering occurs when the size of particles are small compared to the wavelength of the EM radiation. The shorter the wavelength, the more scatter. Mie scattering occurs when the size of the particles are similar to the wavelength of the EM radiation. For example dust or water vapor in the lower part of the atmosphere. Finally, non-selective scatter occurs when the size of particles are large compared to the wavelength of the EM radiation.

The laser scattering method for obtaining the number of PM particles in the lower atmosphere is Mie scattering. The basic assumption of Mie scattering theory is the following: light scattering occurs when a light beam illuminates an inhomogeneous medium, when there are particles with sizes that are approximately equal to the wavelengths of the light beam. Then, the relative scattering intensity varies as a function of the angle of the light beam [Yu et al., 2017].

The theoretical determination of which variables to use in order to calculate a PM sensor output, empirical verification of these variables which are sensor-dependent, and determining the ideal wavelengths and incidence angles of the laser beam fall outside the scope of this research, therefore I refer to Yu et al. [2017]. In this research the low-cost sensor is perceived as black-box: the digital sensor output is used for the correction model.

2.3 CORRECTION MODELS

To create a correction model for low-cost ambient air quality sensors empirical data is needed. Empirical data is data or evidence that originates in or is based on observation or experience. Data for environmental phenomena such as temperature, relative humidity, air pressure, wind speed and wind direction, and particulate matter are also *time series* datasets. Time series datasets show long or short-term behavior of multiple phenomena, plotted against a time unit [Liao and Phillips, 2014].

2.3.1 Vector Autoregression models and cyclostationary processes

For analysis of time series is the Vector Autoregression (VAR) model one of the most commonly used, successful, flexible and easy to use [Zivot and Wang, 2006]. An autoregressive regression model uses previous values of a time series in order to do forecasting [Xiong and Connor, 2002]. Besides, VAR models are useful to describe the dynamic behavior of time series. For VAR models it does not matter if time series are stationary or non-stationary, i.e. if the data fluctuates around a deterministic trend or if the data does not fluctuate around a deterministic path [Zivot and Wang, 2006; Nelson and Plosser, 1982].

Testing if a time series variable is stationary or non-stationary can be conducted with the augmented Dickey-Fuller test [Tseng et al., 2017]. The goal of the test if to check if the dataset has a unit root characteristic. In the test, the null hypothesis is that the dataset has a unit root characteristic, also known as random walk, and is therefor non-stationary. The alternative hypothesis is that the dataset has no unit root and is therefore stationary [Dickey and Fuller, 1979].

Environmental phenomena such as temperature and humidity follow a day-night and seasonal rhythm, and amounts of PM have over days a rush hour - non rush hour rhythm. Therefore, like many processes encountered in nature, they have a periodic rhythm. This type of data has than statistical characteristics that vary periodically with time and they are called cyclostationary processes [Gardner et al., 2006].

In the field of real-time river flow forecasting, Xiong and Connor [2002] analyzed four types of autoregressive models: the linear autoregressive model, a "piecewise" linear model, a "piecewise" linear model with fuzzy thresholds, and a neural network. Xiong and Connor [2002] compared those four error correction models, to improve the model simulated river flows with daily time resolution [Pianosi et al., 2014], i.e. to improve one dataset with help of another reference dataset of "ground truth". They concluded that the simpler linear models provide equivalent performance compared to the more sophisticated models [Pianosi et al., 2014].

2.3.2 Vector Auto Correction Models

Like economic variables, variables regarding environmental phenomena may exhibit upward and downward movement through time. When in a stationary time series dataset a set of integrated variables change jointly through time, they are called co-integrated [Engle and Granger, 1987]. Then, linear combinations of those integrated variables are stationary. The linear combinations that link those variables of different datasets to a common trend path are the co-integration relationships [Kilian and Lütkepohl, 2017].

Using a Vector Error Correction Model (VECM) is a convenient parametrization of a VAR model [Kilian and Lütkepohl, 2017]. VECMs models are mostly used for predicting interdependent time series systems and for analyzing dynamic impulses from random interferences in a system [Tseng et al., 2017]. A VECM can be used for empirical research, to observe dynamic relationships among factors affecting a phenomenon.

2.3.3 Correction model as part of a correction system

Pianosi et al. [2014] their research was on the improvement of forecasts of water flows in river catchment areas by applying a correction model. In their research one correction model would be applicable if particular meteorological or hydrological conditions were met, while another correction model should be used if for other meteo-hydro conditions. Thus, depending on these different modes of the system, different correction models should be used. Therefore, Pianosi et al. [2014] proposed the use of a classification system that identifies the current mode of the river system, before applying a correction model that would improve the forecast of the water flows in the river area.

In the research of Pianosi et al. [2014] three system modes were identified: I) low rainfall forecast (r) and low-flow conditions (q), II) low rainfall forecast and high-flow conditions, and III) high rainfall forecast. In their work, the classification is reproduced with if-then rules:

- if r $<$ R and q $<$ Q, then use "Mode I"
- if r $<$ R and q $\geq$ Q, then use "Mode II"
- if r $\geq$ R, then use "Mode III"

Where R and Q are thresholds for rainfall forecast and the height of the river flow, respectively. The "*Mode*" specifies which correction model should be applied under those conditions [Pianosi et al., 2014].

In the current research, on improving data from low-cost PM sensors with help of a correction model from high-quality reference monitors, also different "modes" of the system could be identified. For example the wind direction, the time of the day, or the season could be the system classifiers in the research.

2.4 RELATED WORK IN THE FIELD OF LOW-COST AIR QUALITY MONITORING

This section elaborates on four studies regarding the utilization of a low-cost air quality sensor networks. For some of these research projects is data acquisition – and in some cases also data processing – executed using a distributed Wireless Sensor Network (WSN).

2.4.1 Multivariate correction model HDMR

Cross et al. [2017] researched the performance of low-cost gas sensors measuring NO in an ambient setting. Their tests with sensors from the manufacturer Alphasense showed that drift resulting from temperature changes can exceed a bias of 600 parts-per-billion, which is equivalent to $480 \mu m/m^3$ of NO, if temperature changes are unaccounted for in calibration of the sensor. Thus, temperature has a significant effect on the sensed value of NO concentration. Although the Alphasense organization provided instructions for correcting for temperature, accompanied with a correction table, that approach gave only stable results when temperature was below 20 degrees Celsius.

Therefore, [Cross et al., 2017] used a multivariate model: they created a High-Dimensional Model Representation (HDMR). That is a numerical method for capturing input-output system behavior without reliance on a physics-based model or an empirical correction procedure which would be provided by the sensor manufacturer. It consists of a general set of quantitative model assessments and analyses of this input-output behavior. A HDMR can produce a model that captures interdependencies of the input variables and can provide a mathematical description of the system. When applied for air quality data, this method can identify and quantify the sensor response to interfering gas species and multiple environmental variables simultaneously. The modeling procedure of HDMR involves the following steps:

1. Specify the maximum amount of variables that are used in the algorithm;
2. Do a statistical test – F-test – to identify the input variables and combinations of input variables that contribute significantly to the variation in the output of the phenomenon of interest;
3. Calculate coefficients for the correction formulas using Least Squares Analysis, minimizing the deviation between the HDMR model prediction and training data.

In the work of Cross et al. [2017] were various metrics used in order to evaluate the model. Those were the slope and intercept of a linear least squares regression of the model output with reference measurements, the coefficient of determination of the linear fit (R^2), RMSE, Mean Absolute Error (MAE), and Mean Bias Error (MBE). Outputs of the HDMR model are plotted against raw sensor data, training data, and test data. See table 2.1 for more of their results and conclusions.

2.4.2 Evaluate performance of each individual sensor node

Castell et al. [2017] researched performance of low-cost AQMesh air quality sensors measuring CO, NO, NO_2, O_3 and PM2.5. The research objective was performance evaluation of the individual sensors: not creating or applying a correction method on their output data. Their research was performed in both a laboratory and ambient setting, of which only the latter is relevant for this current research. Castell et al. [2017] their conclusion was that each of the AQMesh nodes yield slightly different data outputs for the same location and time. Therefore, examining the quality of output data from each node before deploying it in the sensor network is important.

	Cross et al, 2017	*Castell et al, 2017*	*Mukherjee et al, 2017*	*Mead et al, 2013*
Motivation paper	Demonstration of a field-based calibration technique that utilizes co-located measurements and a model representation of interference effects.	Performance evaluation of low-cost air quality sensors. Are they able to monitor air pollution for applications requiring high accuracy or only lower accuracy?	Quantification of performance (accuracy, precision, reliability) of two PM sensors under ambient conditions.	Providing evidence for performance of electrochemical air quality sensors. Showing results from a sensor network deployed in the built environment.
City	Boston, USA	Oslo, Norway	Cuyama Valley, USA	Cambridge, UK
Data collection period	7 July 2016 - 23 November 2016	April 2015 - September 2015	14 April 2016 - 6 July 2016	12 March 2010 - 26 May 2010
Pollutants sensed	CO, NO, NO2, O3	CO, NO, NO2, O3, PM10, PM2.5	PM2.5, PM10	CO, NO, NO2
Sensor (and amount of sensors)	ARISense system with Alphasense sensors (2x)	AQMesh v3.5 (24x)	AirBeam (3x); Alphasense (3x); OPC-N2 (3x)	Alphasense (46x)
PM monitoring technique	n/a	Laser scattering	Laser scattering	n/a
Reference monitor	MassDEP with Teledyne sensors; Met One BAM-1020	AQM stations; Thermo TEOM; Grimm 180	Grimm 11-R optical particle counter; MetOne BAM-1020	Thermo Environmental Model 42C
Interference variables	Temperature; relative humidity; air pressure; wind direction; wind speed; solar intensity; PM2.5; PM10; black carbon	Temperature; relative humidity; air pressure	Temperature; relative humidity; wind direction; wind speed	Temperature; relative humidity; ozone
Correction method	High-dimensional model representation (HDMR)	No correction method applied	No correction method applied	Linear regression
Sampling interval for data collection	Simultaneous and real-time	15 minutes	1 minute	Approx. 3 minutes
Coefficient of determination (R2) against low-cost sensor	n/a	n/a	0.80 - 0.99	n/a
Coefficient of determination (R2) against reference	CO: 0.88; NO: 0.84; NO2: 0.69; O3: 0.39	PM2.5 dense traffic: 0.40; PM2.5 calm traffic: 0.84	0.6 - 0.76	CO: n/a NO: 0.80 - 0.95 NO2: 0.89 - 0.92
RMSE	CO: 39.2 ppb; NO: 4.52 ppb; NO2: 4.56 ppb; Ox: 9.71 ppb	PM2.5: 7 ppb PM10: 64 ppb	n/a	n/a
Conclusion paper	Distributed air pollution measurements can be enabled with those sensors.	Performance varies from unit to unit thus it is necessary to examine the data quality of each node before it is used. The data quality is good enough for applications that require lower accuracy.	Low-cost sensors show a moderate correlation with reference monitors. Results from the sensor measurements are influenced by the meteorological conditions and the size distribution of the aerosols.	Low-cost sensor are feasible for widespread use for monitoring outdoor concentrations and can complement other measurement methodologies.

Table 2.1: Overview of related work in the field of low-cost air quality monitoring networks

2.4.3 Quantify performance of sensors under real-world conditions

Mukherjee et al. [2017] quantified the performance of two types of PM sensors over a period of 12 weeks in an ambient setting. Performance was expressed in accuracy, precision and reliability. One of the objectives of their study was whether the PM sensors could be used as part of an "early warning system" for supporting decision making, in order to reduce human exposure to excessively high concentrations. The OPC-N2 and AirBeam PM sensors used in the study cost respectively around $450 and $250, and are therewith in a higher price class than the sensor nodes used for the current study. Outputs from these PM sensors are not corrected for interference effects: the results are compared with reference monitor results without applying a correction.

The authors found that sampling orientation has a major major effect on the correlation coefficient and the linear regression coefficients, when compared to the reference monitors. Further, Mukherjee et al. [2017] concluded that measurements were influenced by aerosol size distribution and the meteorological environment, and that "quantification of performance of sensors under real-world conditions is a requisite step to ensure that sensors will be used in ways commensurate – proportional – with their data quality" [Mukherjee et al., 2017].

Mead et al. [2013] acknowledged the importance of PM in air quality but focused in their research on the capability of electrochemical sensor which do gas-phase measurements of NO, NO_2 and CO concentrations. In the study were mobile and static sensor nodes used, where only the latter is relevant for this study. For an overview of the discussed studies in the field of low-cost air quality monitoring see table 2.1.

2.5 CONCLUSION RELATED WORK

To conclude, when a sensor system is deployed in ambient conditions are calibration protocols of low-cost air quality sensors needed to overcome potential measurement error [Lewis and Edwards, 2016; Cross et al., 2017]. The lifetime of an air quality sensor can also imply a need for a time-dependent sensitivity that should be corrected for: contaminants can be trapped in the sensor, wide variations in temperature and relative humidity exposure, or evaporation of the electric components can influence the sensitivity of the sensor [Cross et al., 2017]. However, new sensors will be used during the data collection period of one month in the current study: drift caused by lifetime issues can be neglected.

Further, in related work is stated clearly that air pollution concentrations – among with PM – can be affected by interferences from other environmental conditions such as temperature, relative humidity, wind speed, wind direction, and air pressure [Postolache et al., 2009; Cross et al., 2017]. Therefore, these conditions should be measured, correlated with PM and eventually corrected for in the research.

3 METHODOLOGY

In this chapter the methodology for this research is discussed. The chapter is subdivided into sections that cover an important topic or step in the proposed methodology. These steps are: designing and engineering the sensor nodes, data collection, data preprocessing, reliability of the sensor node, baseline measurement, relations between independent and dependent variables, creating correction models with the stepwise Multiple Linear Regression (MLR) method, and finally the validation of the correction models. Figure 3.1 shows an overview of this proposed methodology. This chapter focuses on the theoretical background and requirements regarding the methodology. The next chapter – *Implementation of the methodology* – elaborates on the practical implementation and intermediate results.

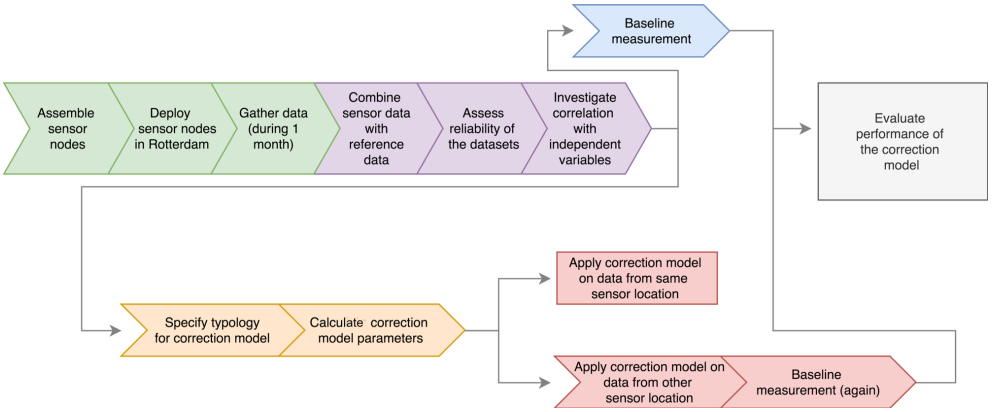


Figure 3.1: Systematic overview of the methodology

3.1 CREATE SENSOR NODES

Regarding hardware, the minimal requirements to conduct this research are two low-cost sensor nodes, placed at two locations. Those sensor nodes contain two low-cost particulate matter sensors, two temperature and humidity sensors, and a microcontroller which reads these sensors and stores the data. Besides, the hardware should be installed in a water-tight enclosure while at the same time allowing ambient air to flow along the sensors.

3.2 STUDY AREA AND DATA COLLECTION

For this research three different data sources are used from where data is acquired. The datasets contain data on PM2.5, PM10, relative humidity, air temperature, barometric air pressure, wind speed, and wind direction. For a period of at least 3 weeks the data will be collected.

3.3 SENSOR NODE RELIABILITY

The next steps in the methodology are combining the sensor data with reference data, based on time and location, and assessing the reliability of the sensor data. First, the sensor node

reliability is discussed, containing of the identification of gross errors, identification of systematic instrument error and creation of a noise model for visualizing the random errors. This data cleaning process starts with removing the gross errors from the PM data acquired with the low-cost sensors. Thereafter, the systematic errors are removed. The high-quality reference dataset is not adapted.

Removing gross errors

Gross errors can be a result of user error or equipment failure. The amount of times this type of error shows up is infrequent [Fisher and Tate, 2006]. When the sensing instrument malfunctions the dataset will contain such a gross error or outlier. Using the data snooping method the outliers are removed from the PM datasets. Data snooping utilizes the Least Squares Adjustment (LSA) approach in order to acquire a mathematical model fit of the data [Jazaeri and Amiri-Simkooei, 2013]. Based on the chosen polynomial degree, this model fit yields an amount of coefficients. Consequently, those coefficients are used to calculate new values, using the original observation. The difference of the new value with the original observation is the residual. If the maximum residual exceeds a threshold, then the original observation to which the residual belongs is removed from the dataset. This process is repeated until there are no more residuals exceeding the threshold. The threshold is often a value that is based on the dataset itself: for example one, two or three standard deviations of the error. How many standard deviations there will be used in this research is determined heuristically, i.e. by trial-and-error.

After the removal of outliers from the dataset, the systematic error is removed from the PM2.5 time series.

Removing systematic error

Systematic errors are the result of a deterministic system which, if known, may be represented by a relationship [Fisher and Tate, 2006]. Systematic error is also called "Mean Error" and indicates the systematic under- or over estimation of measured values – bias – in the dataset [Fisher and Tate, 2006]. It can be either positive or negative. If the systematic error is relatively big, there is a greater difference between RMSE and standard deviation. If the systematic error is small or zero, RMSE and standard deviation are the same. Systematic error is calculated with the following formula from Lemmens [2017]:

$$SE = \frac{1}{n}\sum_{i=1}^{n} z_i^M - \frac{1}{n}\sum_{i=1}^{n} z_i^R \tag{3.1}$$

Where z_i^M and z_i^R are the observations from the low-cost sensor and reference monitors, respectively.

Systematic error is removed from the datasets before evaluating the performance of the correction model with the RMSE metric. This implies that data from a reference monitor is necessary. However, in a real-world scenario not all low-cost sensor nodes will be colocated with reference monitors during the whole data collection period. Therefore, in order to calculate the systematic error of an individual sensor node it should then be colocated with a reference monitor instrument for a short period (e.g. one week) to collect enough data to assess the systematic error, and should then be relocated to the initial place in the network. This is in line with Castell et al. [2017] who stressed the importance of examining output data from each sensor node before deploying it in a sensor network.

Finally, each dataset contains some random error, which are errors with zero mean. Random errors are represented by random variations around the true reference value. The amount of random errors may increase if the amount of measurements increases [Fisher and Tate, 2006]. The objective for the correction model in this research is to decrease the random error of the low-cost sensor.

3.4 COMBINE VARIOUS DATASETS

Based on date, time and the location in the study area the datasets from different sources are combined into two separate datasets: one for each of the two sensor node locations. An

algorithm [1] performs this combination. The first steps for preprocessing in the proposed methodology are resampling and normalization. Resampling and normalization are necessary to make the time series comparable which is needed before the correction model can be created.

Resampling with interpolation

When the temporal sampling frequencies are unequal the values from the datasets may be incomparable. Therefore, in the reference datasets and sensor node datasets the temporal sample frequencies should be resampled. Sensor data that is not recorded on an hourly basis is averaged to one-hour values, therewith using the same approach as in the work of Borrego et al. [2016].

Normalization

With feature scaling is each value for the various variables normalized to a 0 to 1 scale, so the time series have a consistent distribution. The following equation 3.2 is used.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{3.2}$$

In equation 3.2 is each $min(x)$ and $max(x)$ the minimum respectively maximum value of a specific variable. For example, the minimum and maximum value for PM2.5 low-cost sensor 1, low-cost sensor 2 and the reference monitor. Further, x_i is the current data value and z_i is the normalized data value. The feature scaling procedure is applied after removing gross errors and systematic errors.

3.5 CALCULATE BASELINE MEASUREMENT STATISTICS

After removing the gross errors from the data collection of the PM sensors is the *baseline measurement* conducted. The baseline measurement is the "before" measurement: before any corrections are done on the datasets except for preprocessing. The data quality is expressed in accuracy and precision. Accuracy is how close a measurement is to the true value, expressed in RMSE, revealing the random error of the measurements. Precision is how close an estimate is to the mean estimate, expressed in Standard Deviation (SD), revealing the systematic measurement error. Accuracy and precision are equal if the mean error is zero. The error of the given set of low-cost sensor measurements is determined by comparison with another set of known and more accurate measurements. That data is the "reference" data and it is assumed that it is error free [Fisher and Tate, 2006].

Although Cross et al. [2017] investigated a gaseous air pollution indicator – Nitrogen Oxide – they used relevant criteria for evaluating the correction model. Those include Root Mean Square Error, coefficient of determination R^2, and the slope and intercept of the linear least squares regression of the model output with reference measurements. Next to those evaluation criteria are also the Standard Deviation and Systematic Error calculated in order to evaluate the instrument precision during the baseline measurement. Besides, a noise model is created: that shows the error distribution model graphically.

Evaluation metric: RMSE

In the current research is the RMSE used as evaluation metric for the assessment of the correction model performance, and not for example the MAE According to Chai et al. [2014], the requirements for choosing RMSE as evaluation metric are that the error is expected to follow a Gaussian distribution, the sample size is large (> 100), and the error should be unbiased, i.e. distributed randomly. The RMSE is chosen since the error distribution for the low-cost sensor is expected to be a Gaussian distribution, and there are over 600 samples. Moreover,

1 https://github.com/NiekB4/aqs

after removing the systematic measurement error the only error left in the dataset should be random error. The formula for RMSE is shown in equation 3.3:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(z_i^M - z_i^R)^2}{n}} \tag{3.3}$$

Where z_i^M is the observation from the low-cost sensor, z_i^R the observation from the high-quality reference monitor from the same location and time interval, and n the number of observations. In this study, RMSE is used as evaluation criterium, i.e. to assess the performance of various correction models and consequently select the best one.

Coefficient of correlation and coefficient of determination

The coefficient of correlation – R – is calculated according to equation 3.4. This coefficient is relevant for the baseline measurement and for investigating the correlations between the candidate independent variables. To calculate R^2 square R.

$$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.4}$$

Where x is the vector of observations for the low-cost sensor and y the vector of observations for the reference monitor.

Noise model

With the noise model the observation error of the low-cost sensors is indicated graphically. The graph can then reveal the type statistical distribution (e.g. normal (Gaussian), La Place, binomial, skewed distributions) [Forbes et al., 2011]. As indicated by Chai et al. [2014], a requirement for using the RMSE metric as evaluation criterium is that the unbiased observation error has a normal distribution. Thus, the error of the observations in this research should also be unbiased: the error distribution should have zero mean. Figure 3.2 shows the normal (Gaussian) distribution of the error of two randomly generated vectors (n=1000), with and without bias. The error distribution of the low-cost sensor dataset should look like the green graph in 3.2, before the baseline RMSE is calculated.

3.6 RELATIONSHIPS BETWEEN THE VARIABLES

Which independent variables should be included in the correction model? Datasets for the following candidate environmental phenomena are available at the study area: humidity, temperature, air pressure, wind speed, and wind direction. Each variable is investigated separately, whereby the correlation of the variables with PM are calculated.

Moreover, the multicollinearity between the candidate should be minimal [Ausati and Amanollahi, 2016]. The multicollinearity is the collinearity between each independent (candidate) variable. With the Variance Inflation Factor (VIF) metric multicollinearity between the independent variables can be detected. If the VIF between two independent variables is above 5 those independent variables are multi-collinear and should not be included in the model [Ausati and Amanollahi, 2016; Berninger et al., 2018]. VIF is calculated as:

$$VIF = \frac{1}{(1 - R^2)} \tag{3.5}$$

Where R^2 is the coefficient of determination, i.e. the square of the coefficient of correlation (equation 3.4).

Correlations with related environmental phenomena

Where the VIF is calculated to check for multicollinearity of two independent variables, the scatterplots and correlation coefficients are created and calculated to check for correlation between the independent and dependent variables. Using scatterplots and correlation coefficients – calculated using equation 3.4 – are relevant related environmental phenomena

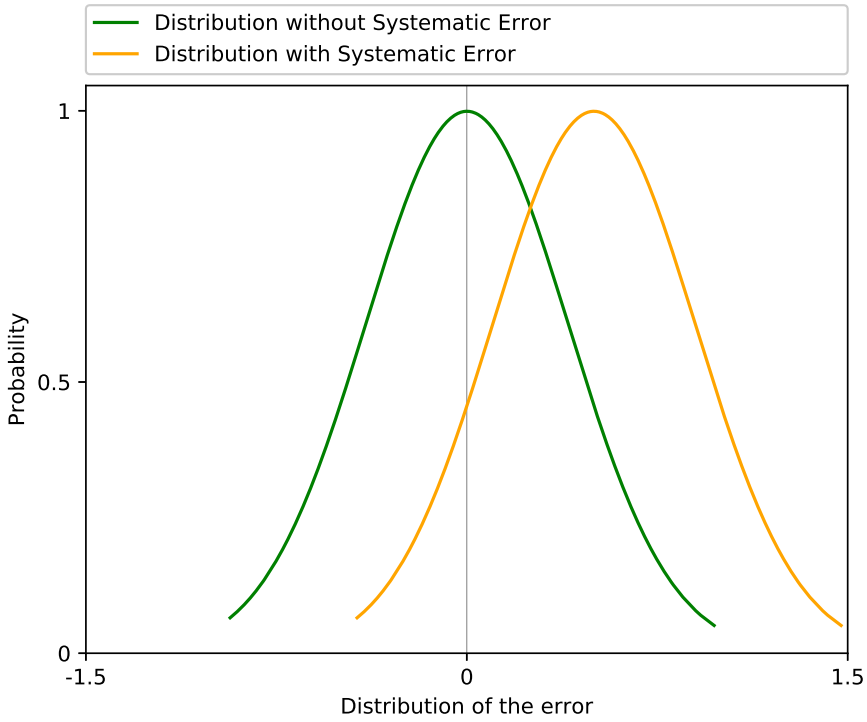


Figure 3.2: Noise models with and without Systematic Error

selected. The candidate phenomena which are suggested in the related work and for which data is collected are relative humidity, temperature, air pressure, and wind speed.

3.7 CALCULATE CORRECTION MODEL PARAMETERS FOR VARIOUS SETTINGS

Before creating a correction model it is useful to have a specification of the type of correction model beforehand. In this paragraph four typologies for the correction models are proposed – A, B, C and D – which vary in two dimensions: the amount of parameters and the domains (see figure 3.3).

Proposed typologies for correction models

First, the variables that can affect the PM concentrations are a dimension in which correction models vary in this research. This can be expressed in the amount of parameters that are included.

For the most basic correction model in this study are influences of other environmental phenomena neglected. Besides a constant value, there is then only one parameter introduced that corrects the given PM concentration. This is a correction model of type A or C as shown in figure 3.3. Introducing one or more environmental phenomena to the correction model results in more parameters that are included in the mathematical correction model, resulting in a model of type B or D. Those parameters correct for the collected quantitative environmental phenomena humidity, temperature, air pressure or wind speed. A particular implementation of a correction model can consist of various combinations, e.g. a parameter for PM and humidity only but not for the other variables, or only for PM and air pressure, or parameters for both of the PM sensors on a sensor node and temperature, etcetera. As long as there is at least one PM time series included in the training dataset.

Another dimension in which a correction model varies is the amount of domains for which the correction model is specified. The domains are for example a time domain (peak hours and off-peak hours, seasons), the wind direction, or a specific temperature domain. The basic correction model is a generic model that is used on the whole dataset: there are no

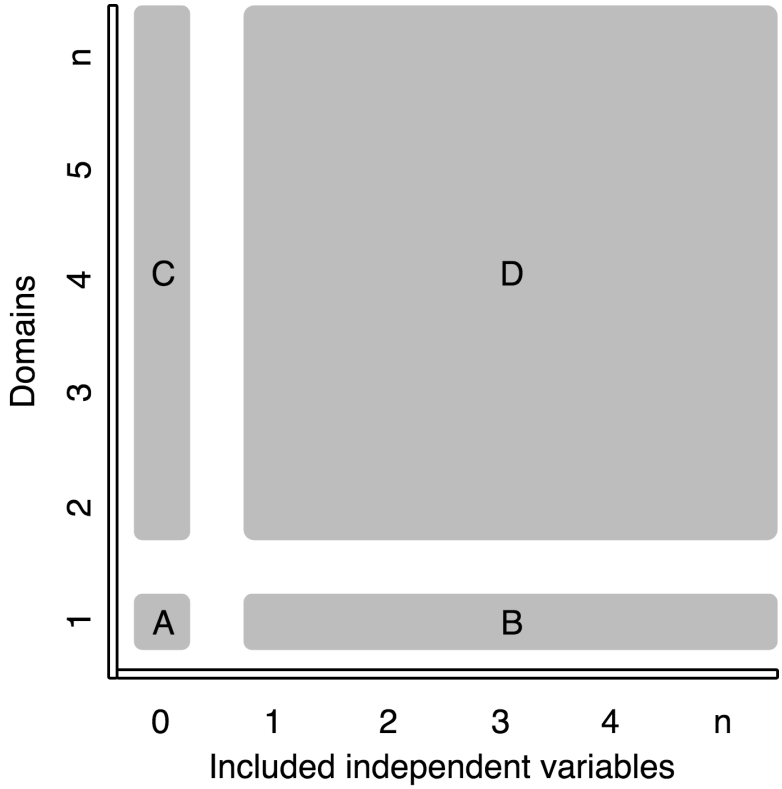


Figure 3.3: Typologies for the correction models

domains of the other variables where the correction model is not used. This would be a *generic* correction model of the proposed type A or B (figure 3.3). On the other hand, a *specific* correction model takes account of those domains (type C or D). It is then possible that there will be different parameters used in the correction model: there are basically more correction models and which one to use depends on the value of an external variable. Like in the work of Pianosi et al. [2014], the current mode of the system can then be classified, in this case the current mode for air quality in the built environment. For example, when it is peak hour the correction model could use different parameters than during off-peak hour, or depending on the wind direction another set of parameters could be used. The domains are empirically determined. Using more of those domains introduces more complexity to the model.

Finally, it is not necessary that the domain to select is classified based on quantitative data from variables in the correction model. The classes can also be *categorical*. For example, the moment on the day or the wind direction can be a classifier, although it is not included as variable in the correction model. Hence a "Type C" model can exist.

Calculating parameters

Parameters for all four model types are calculated using the *polynomial regression method*. These polynomial regression models vary based on (I) the amount of environmental phenomena that are taken into account, and (II) the degree of the polynomial regression.

First, for a correction model there could for example only one parameter be calculated, which takes only the PM2.5 value of the low-cost sensor into account: type A or type C. This is polynomial regression with degree 1 and thus simple linear regression. On the other hand, also a parameter for other candidate environmental variables can be introduced in the correction model, for example parameters for humidity and air pressure (type B or D). In that case, the method used is *Multiple Linear Regression* MLR. The MLR method is chosen since it is the most used linear model in the field of air quality forecasting, and MLR models have high interpretation of accuracy [Pires and Martins, 2011; Ausati and Amanollahi, 2016]. "Multiple" in multiple regression relates to the the amount of independent variables. "Multiple" thus not refers to the amount of dependent variables. The method which takes more than

one independent as well as more than one dependent variables into account is Multivariate Regression.

Moreover, there are various versions of correction models created, varying in amount of independent variables included and the polynomial degree. These models are created according to the *stepwise* method: each time when a new variable is included or another polynomial degree is used, a new correction model is created which could result in the "best" performing correction model. With this method, the various correction models and their results are stored, whereafter the performance of the correction models is evaluated.

In this research, two possible correction models for improving the accuracy of low-cost PM sensor data are distinguished. Those are MLR models with a polynomial degree of 1 and MLR models with higher polynomial degrees.

The first has the form of equation 3.6:

$$Y = a + p_1X_1 + p_2X_2 + ... + p_kX_k \tag{3.6}$$

With the dependent variable Y as the predicted value for PM2.5, the intercept a, the parameters p_1, p_2 ... p_k from the linear fit, and the predictor variables X_1, X_2 ... X_k, i.e. values from the dataset for the phenomena taken into account (PM, humidity, air pressure, etcetera).

A MLR correction model take would take higher polynomial degrees into account has the form of equation 3.7:

$$Y = a + p_1X_1 + p_2X_1^2 + ... + p_kX_1^k \tag{3.7}$$

Where X_1 is PM or a related environmental variable – e.g. humidity, wind speed, etcetera. The equation 3.7 can then be extended with X_2, X_3 ... X_k, representing data from from related environmental variables. Thus in those MLR models are the terms – values from the dataset – in some cases squared.

How to achieve the values for the parameters? There is no exact solution to calculate the unknown parameters. Namely, for there is no line of the form $Y = a + bX$ that goes trough all data points in the time serie datasets. Since there are many observations available compared to the amount of unknown parameters it is an over-determined system. Then a least squares fit can be used to calculate the unknown parameters. Besides, the samples in the dataset are collected with instruments that are sensing in the outdoor environment and can thus contain some error. The least squares method allows the data – observations – to change a small amount. The least squares approximation is the best estimate for a line which goes through nearly all data points [Lemmens, 2016c].

As described in Lemmens [2016b], Lemmens [2016c] and Jazaeri and Amiri-Simkooei [2013], the general requirement of the leas squares approximation is to minimize the sum of squares of the residuals. To find the unknown parameters of the least squares fit, find the values for x in equation 3.8.

$$x = (A^TA)^{-1}A^Ty \tag{3.8}$$

Where the result x is a vector with the intercept and coefficients from the linear fit, A is a matrix that represents the observations from the included time series and y a vector representing the ground truth data. The intercept and coefficients will be used to calculate new values – "corrected values" – for PM2.5.

Algorithm

Various correction models will be investigated. To evaluate the performance of each correction model and to validate the performance on data from a sensor node on another location is an algorithm designed and implemented (see figure 3.4 and algorithm 3.1). This algorithm takes various datasets and settings as input, such the normalized and cleaned sensor data, reference data, the amount of polynomial degrees, how many parameters to use and for which variables. First, the baseline measurement is conducted: the RMSE of the PM data is calculated. Then, using the various inputs, the parameters are calculated according to the MLR method. Consequently, using those parameters, new values for PM are calculated. Finally, the RMSE of the new dataset is computed. Algorithm 3.1 is the proposed algorithm in pseudocode.

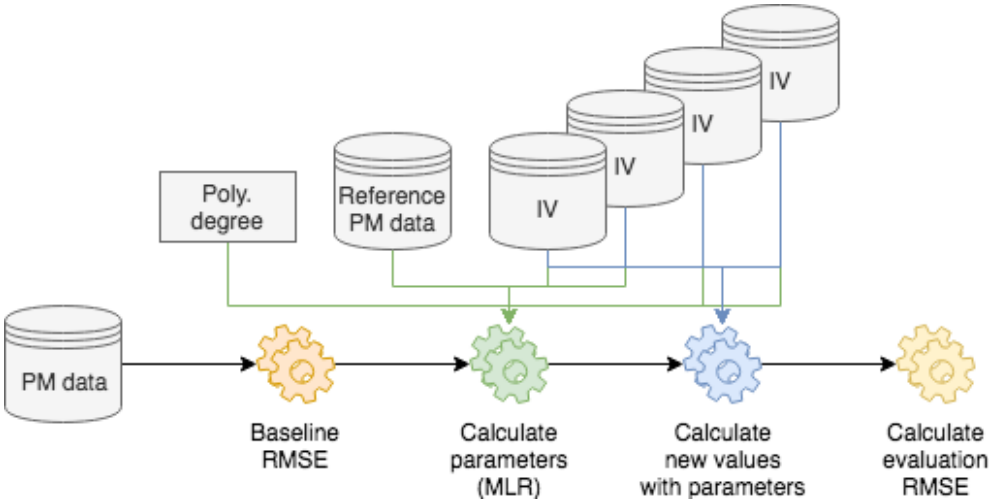


Figure 3.4: Schematic overview of the proposed algorithm

Algorithm 3.1: Algorithm for creating and evaluating correction models

Data: List with normalized sensor datasets
Data: List with normalized reference monitor datasets
Data: List with normalized environmental sensor datasets
Data: List with polynomial degrees to use for the iteration
Result: Evaluation statistics showing performance of correction model

```
initialization;
for dataset in list with normalized sensor datasets do
    Set currentSensorDataset;
    Set currentReferenceMonitorDataset;
    Set currentEnvironmentalSensorDatasets;
    Set degrees;
    Calculate evaluation metric RMSE;
    for c do
        item in degrees
    A = matrix with sensor observations;
    B = vector with ground truth;
    Param = solve linear system with LSA;
    for c do
        i in len(Parameters)
    newSensorDataset = Param[i]*currentSensorDataset[i]**degrees[i]
    residualsCurrentSensor = currentSensorDataset - newSensorDataset;
    Calculate evaluation metric RMSE again;
```

Example of the implementation

Finally, the last paragraph of the experimentation chapter describes an example given where all various correction model parameters are logged in a table. The example elaborates on the situation of the most basic correction model: only one parameter for the sensed PM2.5 value and no parameters for other eventually related environmental phenomena. The table includes the parameters, the sum of the squared residuals, $RMSE$ and the correlation coefficient R. Further, the parameters that are achieved with the polynomial fit. Besides, the noise model that represents the random instrument error is included in the example implementation.

3.8 PERFORMANCE OF THE CORRECTION MODEL

Using the best behaving correction models from the previous step, the statistics for the corrected datasets after applying those correction models are calculated. Those statistics are calculated using MLR correction models from the same sensor location as for which those correction models are designed. The evaluation metric for the performance of the model is RMSE, as discussed in paragraph 3.5.

3.9 VALIDATION OF CORRECTION MODELS

In this final step are the parameters in the MLR correction model from the dataset of sensor node location 1 used for the dataset of sensor node location 2, and vice versa. The correction model that produces at the other locations datasets with the lowest RMSE is most reliable.

3.10 CONCLUSION METHODOLOGY

Figure 3.1 shows a systematic overview of this methodology. When all those steps are executed successfully the main research question and its subquestions can be answered.

4 IMPLEMENTATION OF THE METHODOLOGY

Details of the implementation of the presented methodology and the various experiments that have been conducted are discussed in this chapter. It follows the same structure as the previous chapter.

4.1 CREATE SENSOR NODES

Since this research involves accuracy analysis of the low-costs PM sensor that is the most important hardware component. Further, the microcontroller which reads and stores the data is also of high importance. Other hardware components, i.e. the temperature and humidity sensors, are of less importance since this data can eventually be acquired via external sources – like with air pressure, wind speed and wind direction.

Particulate Matter sensor

The chosen low-cost air quality sensor is the "Plantower" PMS5003[1] optical particle counter. This PM sensor measures scattered laser light from a stream of aerosol particles from which the particulate mass concentration is reconstructed, as discussed in section 2.2.2.

This sensor is chosen because it uses the laser scattering method, there is an exhaustive manual available, the sensor itself is already in stock at the University, and the price of €20 makes it a low-cost sensor. The PMS5003 works on an electrical voltage of 5V and contains a fan which creates the air flow through the sensor. One disadvantage is that at the beginning of this research there was no reliable (open source) software library available in MicroPython. Namely, the used microcontroller needs to be configured with that programming language.

For this project, one sensor node will contain two PMS5003 sensors. Namely, two sensors are the minimum amount to provide data for the precision measurement of this instrument. Given that the PMS5003 uses the UART interface for the data output, two UART interfaces on the microcontroller is a must.

Microcontroller

For reading, storing, and transmitting the sensor data is the Pycom LoPy[2] used. This microcontroller is chosen because it contains an Espressif ESP32 chipset which features Wi-Fi and Bluetooth communication. The controller can be scripted with the MicroPython programming which is a lightweight version of the Python language. This controller has 520KB RAM processing memory and 8MB flash memory for storage and works on input voltages ranging from 3.5V to 5.5V. Moreover, the LoPy features LoRaWAN communication, the availability of the two required UART interfaces, two I2C interfaces, and ample analogue and digital input and output pins for hardware. The LoPy is available for a price of around €35 per unit.

Temperature and humidity

The chosen temperature and humidity sensor is the digital AM2302[3] sensor, available for around €8 but already in stock at the University. The AM2302 works on input voltages ranging from 3.3V to 5.5V. This sensor is chosen because it is pre-calibrated, has low power consumption, and there is a MicroPython library available for reading the AM2303 sensor data.

1 http://www.aqmd.gov/docs/default-source/aq-spec/resources-page/plantower-pms5003-manual_v2-3.pdf
2 https://pycom.io/wp-content/uploads/2018/08/lopy-specsheet.pdf
3 https://www.adafruit.com/product/393

Assembly

Each of the two sensor nodes contain additional components next to the individual sensors and microcontrollers. These are: a breadboard and wires for physically connecting the sensors and microcontroller; a LM2596[4] step-down converter to decrease the voltage in cases of sudden peaks; a MT3608[5] step-up converter to increase the voltage to a stable 5V which is needed for the PMS5003; a plastic waterproof box for the microcontroller compartment (€6 at local hardware store); half-open cover material for the sensor compartment (sponsored by the regional environmental agency); a plastic waterproof box (€3) for the electricity cable (€8); and a mobile phone charging adapter (€12). The total costs for hardware of one sensor node is thus €123.50. Figure 4.1 is a schematic overview of the sensor and microcontroller on the sensor node.

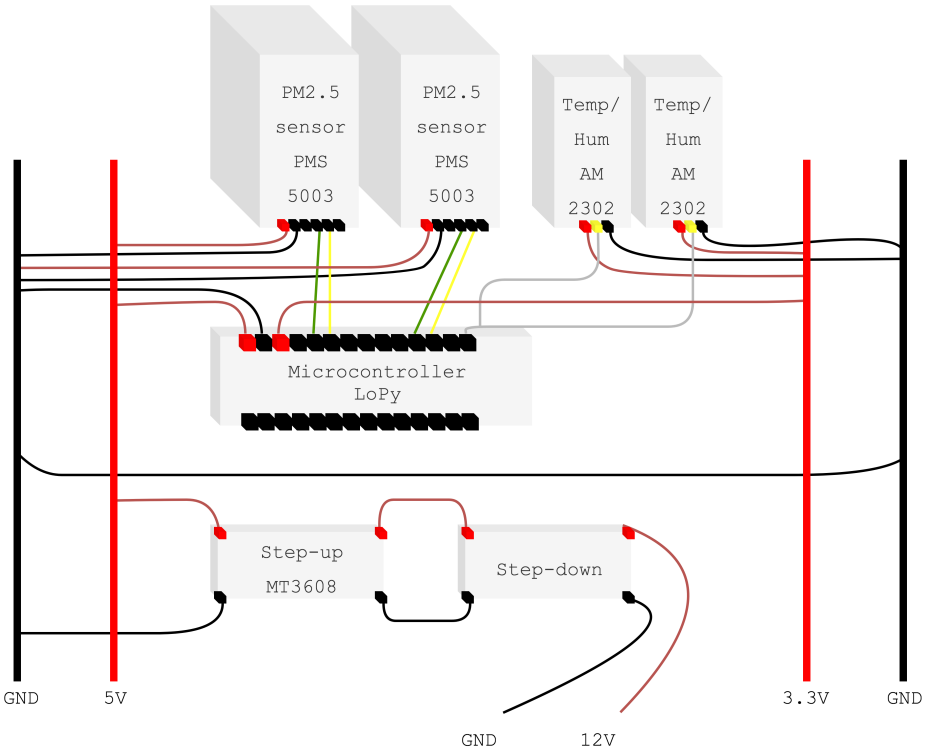


Figure 4.1: Schematic overview of the sensors and microcontroller on a sensor node.

Configuring software for sensor nodes

In parallel with the process of making two sensor nodes, each individual sensor was also configured with scripting software in the MicroPython language. The scripts can be found on the accompanying GitHub page[6]. The flowchart for the software on each sensor node is depicted in figure 4.2.

There are three main functions for the software on the LoPy microcontroller. The first main function is reading raw data from sensors. Each sensor has an own MicroPython library that fulfills this task. Second, in the *main* function are the sensors' sample intervals configured. This *main* is an ongoing loop – which would restore if there is an error – and creates one long string with all the data for a measuring round. This string of data is stored in a .csv file on the internal storage of the LoPy. Third, software on the LoPy connects via the Telnet protocol over Wi-Fi with the personal computer to upload the data from the internal storage of the LoPy to the harddisk storage of the personal computer. This uploading is done once a week, so the LoPy will not be overloaded with data. Next to that, by visiting the physical sensor

4 http://www.ti.com/lit/ds/symlink/lm2596.pdf
5 https://www.olimex.com/Products/Breadboarding/BB-PWR-3608/resources/MT3608.pdf
6 https://github.com/NiekB4/aqs

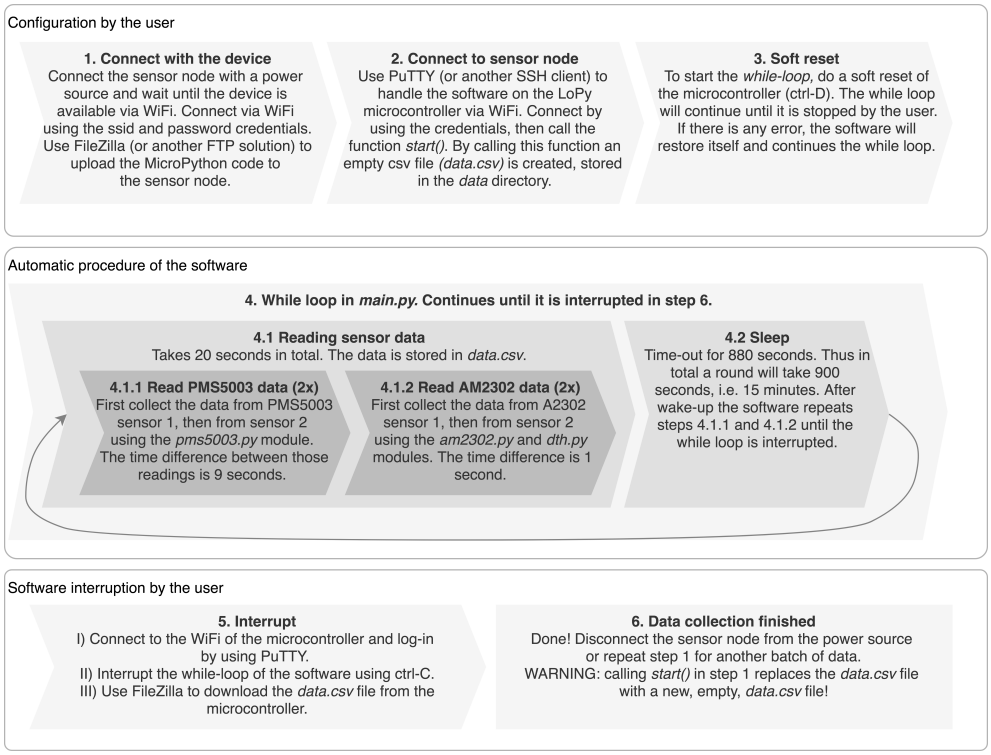


Figure 4.2: Flowchart of the software on the sensor node. The *main.py* is one continuous loop that would restore itself in case of an error.

node location the state of the sensor nodes can be checked visually and it can be validated that the data is collected correctly.

Configuring software for reference datasets

The "ground truth" reference data for PM must be of high quality to create a correction model for the low-cost sensors. DCMR is the regional institution that installs, operates and manages those kind of high quality air quality measurement stations in the region of Rotterdam. One manager of the Air Quality Department of DCMR was the main contact person during the design and data collection phases of this research. DCMR also provided transportation of the sensor nodes to and from both locations in Rotterdam.

The measuring network of DCMR for monitoring PM consists of BAM-1020 monitors – discussed in section 2.2.1 – on ten different locations in the Rijnmond region. Those BAM-1020 monitors are configured for an hourly cycle. The microcontrollers are configured for a 15 minute cycle, thus for every data record from the BAM-1020 there are four data records from the sensor nodes. This redundancy in the dataset allows to smooth the data after afterwards: it allows to execute a resampling operation whereby outliers could be removed and random errors can be smoothened using the median or average value for the hour.

DCMR publishers their data from the BAM-1020 monitors in three ways as open data. One possibility is to use the online viewer[7] on a website of RIVM, which is live and unvalidated data. Another possibility is to gather the data in .csv format from the website. That is validated by extra reference stations using the gravimetric weight method and if necessary extra corrected for weather influences.

In this research is chosen for the third way of retrieving the data, which is via an Application Program Interface (API), owned by the initiators of www.luchtmeetnet.nl and free to use under the CC BY-ND 4.0 license. With HTTP GET requests one can retrieve live data from all monitoring stations that are connected. The data is updated every hour with the newest data – thus from the previous hour. The HTTP response is in the JavaScript Object Notation (JSON) format. Because the API uploads during the whole day and night the PostgreSQL database must also be 24/7 online and is therefore hosted on a virtual machine of the University. With

7 www.luchtmeetnet.nl/

the programming tool Node-RED is this JSON response stored in a PostgreSQL database. The settings for Node-RED gathering and storing Luchtmeetnet data are included in Appendix B of this report.

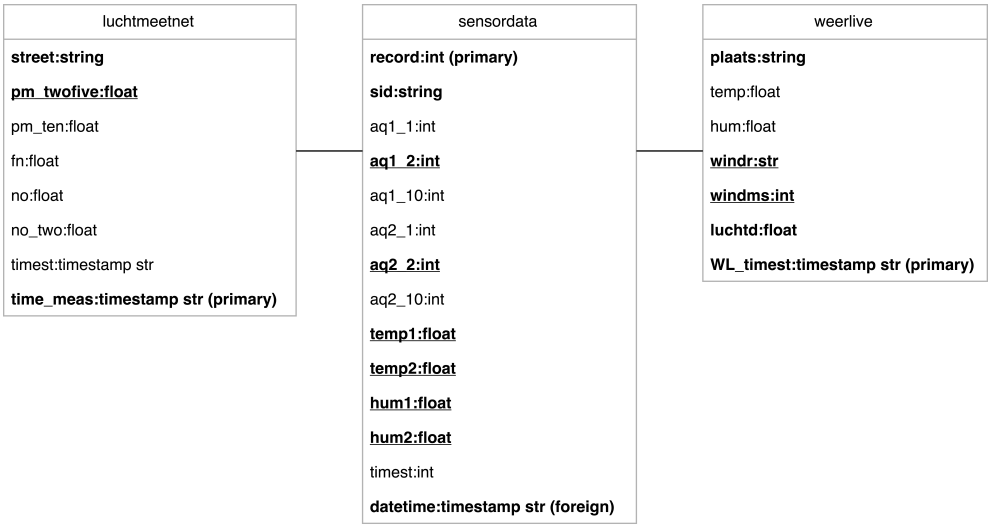


Figure 4.3: The variables, their data types, and relations between the datasets from different sources. Only the **bold** attributes are used in this research. The **underlined** attributes are the dependent and independent variables.

While the Luchtmeetnet API provides near-live data on PM2.5 and PM10 for the monitoring stations, another API provides data on the meteorological conditions air pressure, wind speed and wind direction. That is the Weerlive API which forwards data from the Dutch meteorological institute KNMI. As with the Luchtmeetnet API, HTTP GET requests are used and the response is in JSON format, and the data is gathered using Node-RED, see Appendix C. Figure 4.3 shows the relationships between the different dataset in a UML diagram. Data for the *luchtmeetnet* and *weerlive* tables are both stored in the PostgreSQL database. The *sensordata* is stored in .csv files.

4.2 LOCATION OF THE SENSOR NODES AND DATA COLLECTION

Since the goal of this research is to create a correction model for PM measurements, preferably PM2.5, the created sensor nodes should be located close to the monitors that supply the *ground truth*. At 7 of ten locations where DCMR monitors particulate matter where also PM2.5 concentrations monitored. Of those locations are two sufficient locations chosen, based on two requirements. The locations must be relatively close to each other but must have two distinct profiles.

The chosen monitoring stations are those at Pleinweg and Zwartewaalstraat, both in Rotterdam, the Netherlands, see figure 4.4. The profiles of those locations are distinctive since the *Pleinweg* station located between a busy inner-city road and a service road. The *Zwartewaalstraat* monitor, on the other hand, is a "background" monitor. It is located next to a small park and a – by then – vacant building in a residential district (figure 4.5). The historic key figures from the DCMR data collection at both locations for the year 2017 are shown with a yellow background in figure 2.4.

Figure 4.6 shows a schematic overview of the placement of the sensor node and reference monitor. The air inlets are placed at a height of 3 and 2.2 meters, for the reference monitor and sensor node respectively. These heights correspond to the guideline from the European Union, which states that those heights should be between 1.5 meters and 4 meters [EU, 2008]. The maximum distance between the air inlets of sensor node and reference monitor is 0.8 meters. Since the inlets are not exactly on the same place it cannot be guaranteed that the instruments will sense the same concentrations at the same moments. However,

Figure 4.4: Locations of the sensor nodes in the Rotterdam study area. Dienst Centraal Milieubeheer Rijnmond (DCMR) has reference monitors on those two locations. These locations are chosen because they are close to each other and have a distinctive profile.

Figure 4.5: Sensor node location "Pleinweg" (left) and "Zwartewaalstraat" (right). One is located near a busy inner-city road while the other is located in a calm area of a residential district. Both are located close to each other.

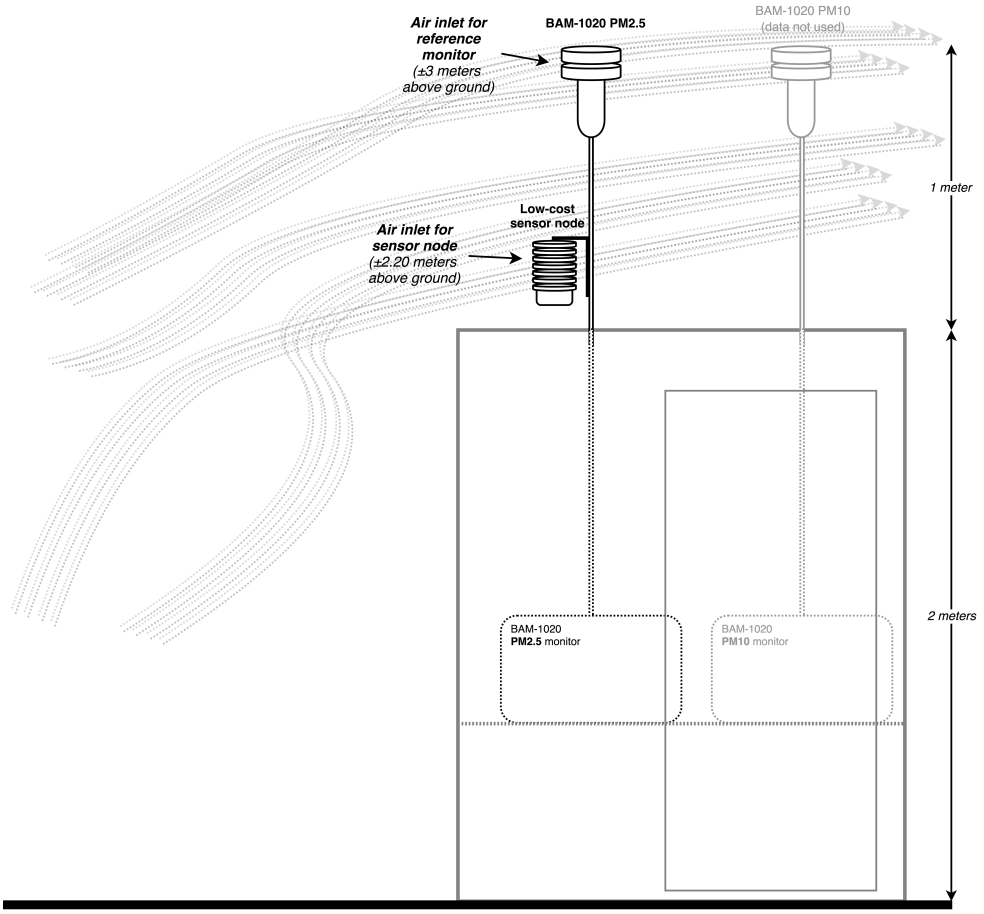


Figure 4.6: Schematic overview of the placement of the sensor node and the reference monitor. The reference monitor from DCMR is placed in accordance with the 2008/50/EG guideline [EU, 2008]. The low-cost sensor node is also placed in accordance with that guideline, but placed on .

Sensor node location	Data collection period	Observations in dataset	Missing observations	Uptime
Pleinweg	16-05-2018 09:30 until 10-06-2018 23:30	635	4	99.3%
Zwartewaalstraat	16-05-2018 10:30 until 10-06-2018 23:30	657	0	100%

Table 4.1: Information regarding the data collection of the sensor nodes at Pleinweg and Zwartewaalstraat.

the concentration from the BAM2-1020 reference monitor is valid for at least the immediate surrounding, thus also for the location where the low-cost sensor node is located.

The casing of the low-cost sensor is open, allowing the air samples to flow through the sensor node. Possible ways in which the casing can affect the air stream through the sensor node and the performance of the sensors is not the focus of this research. The goal of this research is to assess and improve the data quality from the low-cost sensors. Both sensor nodes are identical thus it is assumed that at both sensor nodes the casing will affect sensor performance in the same way. So it is assumed that the experiment setup will not affect the data quality for this research.

During one month – from 16^{th} of May 2018 until 19^{th} of June 2018 – data was collected. The two sensor locations at Pleinweg and Zwartewaalstraat were visited weekly in order to download the data from the microcontroller. When the data was downloaded new .csv files are stored on the harddrive of the personal computer. Next to that, the data from this new file is copied to a .csv file that contains all data from the corresponding street which has been collected until that particular moment.

For the sensor location at the Pleinweg are in total 3258 data records collected. The accompanying .csv file is 171kB. At the Zwartewaalstraat the sensor node yielded 3273 data records in total, stored in a .csv file of 176kB.

4.3 SENSOR NODE RELIABILITY

Concentrations of PM, humidity, and temperature are monitored with the engineered sensor nodes. Next to that, the reference monitors also measure PM. And the Weerlive reference monitor measures humidity, temperature, wind speed, wind direction and air pressure. The raw data originating from the sensor nodes is stored in separate .csv files, while the raw reference data is stored in a PostgreSQL database. Due to an error with the database storage the reference data was missing from the 11th of June 2018 onwards. Therefore, only the data from 16th of May 2018 11:00 until 10th of June 2018 24:00 is selected for further analysis. During preprocessing are the sensor node and reference datasets combined based on time and location and stored in two datasets: one for the location Pleinweg and one for the location Zwartewaalstraat. Each of this combined and resampled dataset contains one sample per hour, during 26 days. During the data collection period, the engineered sensor node at the Zwartewaalstraat worked without errors: there are no time gaps in the time series plots. On the other hand, the engineered sensor node at Pleinweg malfunctioned. At the morning of 25th of May the data samples of 7:30 and 8:30 are missing, and the first two samples of 28th of May (0:30 and 1:30) are missing, resulting in a sensor node reliability of 99.3%.

4.3.1 Data quality assessment of low-cost Particulate Matter sensors

Both PM sensors mounted on the nodes at two measurement locations had no remarkable shortcoming regarding reliability. These scores for the PM sensors correspond to the sensor nodes general reliability scores (table 4.1).

Figures 4.7 and 4.8 are time series plots of the original PM datasets. Red and orange lines represent the two PM sensors on the nodes, while the green line represents reference data. Overall, at both locations the two sensors on the sensor node follow more or less the same trend compared to the reference monitors. However, the PM data from the sensor nodes are at most times a factor higher than the reference PM data. It seems that when the PM data from the reference monitors increases the PM data from the low-cost sensors increases with a factor. Moreover, both the original and smoothened datasets show high peaks at some

		T=1SD		T=2SD		T=3SD		T=4SD		T=5SD	
		Deg=2	Deg=3	Deg=2	Deg=3	Deg=2	Deg=3	Deg=2	Deg=3	Deg=2	Deg=3
PW sensor 1	Average	185	191	41	36	12	11	2	3	0	0
	Median	185	191	40	35	13	11	2	2	0	0
PW sensor 2	Average	155	155	26	23	6	6	0	0	0	0
	Median	154	154	27	23	4	4	0	0	0	0
ZW sensor 1	Average	294	298	57	50	16	14	3	5	0	0
	Median	291	296	52	48	15	15	2	5	0	0
ZW sensor 2	Average	239	266	41	34	11	9	2	2	1	1
	Median	234	263	41	35	10	8	2	2	1	1

Table 4.2: Amount of detected outliers per selected threshold and polynomial degree (N around 600)

moments. Those can be outliers, although in a large number of cases the peaks are observed for both sensors at the same time. This can imply that the sensors are sensitive to and affected by weather conditions or that the value is indeed high in those cases.

Data cleaning

Data cleaning starts with identifying the standard deviation, RMSE and systematic error of the original observations (table 4.4). Since the calculated systematic error is not zero for all four sensors, it can be concluded that there is a systematic error in the instruments.

Next to that, possible gross errors or outliers in the original dataset are identified. Following the data snooping method is first a threshold calculated. This threshold is based on the standard deviation – of the PM values in the sensor node dataset – multiplied by a heuristically determined factor. Consequently, the unknown parameters are calculated using least squares adjustment. The polynomial degree is also set based on heuristics. New values are calculated by using these parameters and the original observations. The residual is the difference between the original observation and this new value.

The results of table 4.2 indicate that the sensor at the Zwartewaalstraat node has most outliers. Various settings are used for the polynomial degree and threshold. Figures 4.9 and 4.10 show the outliers in the time series plot. These figures show that values above 200 μg/m3 are gross errors and should be removed. Other high values are not regarded as outliers and should therefore not be removed.

Instead of removing the whole record from the dataset, and therewith also the temperature, humidity etc. values, it is chosen to replace the outlier with the median outlier of its 2 neighbors on each side. The advantage of replacing instead of removing is that the arrays of both sensor 1 and sensor 2 keep the same length, which would not be the case if the outliers were removed.

The identified outliers, their value in the original dataset, and the replace values are depicted in table 4.3. Only the detected outliers with polynomial degree two and a threshold of 4 standard deviations are included in this table. Notice the differences between the original value and replace value: those are in most cases still relatively limited. The only exception is the outlier of Zwartewaalstraat sensor 2, where an outlier is detected of size 215.0 but replaced with 44.0, which is the median of the two neighbors at both sides. After the outlier removal the dataset is again saved as .csv file.

4.3.2 Assess quality of not-PM datasets

Temperature and humidity sensor

On both sensor nodes, one of the two temperature and humidity sensors yielded no data because of a hardware error. The other temperature and humidity sensor yielded data, though for Pleinweg only $\pm$42% and for Zwartewaalstraat $\pm$56% of the dat records are reliable. Namely, for around half of the data samples the values are "0". That means both temperature and humidity are really 0 degrees Celsius and 0% humidity at those moments, which is unlikely taking into account the season when the data is collected and the overall trend of the other data. This happened most of the times for only one observation: the next observation the sensor was performing well and gave results, so there is a gap of 15 minutes without data. However, in some cases the data ulaty of the temperature and humidity sensor was very poor, for example when it yields no data for six observations in a row, i.e. 1.5 hours no data.

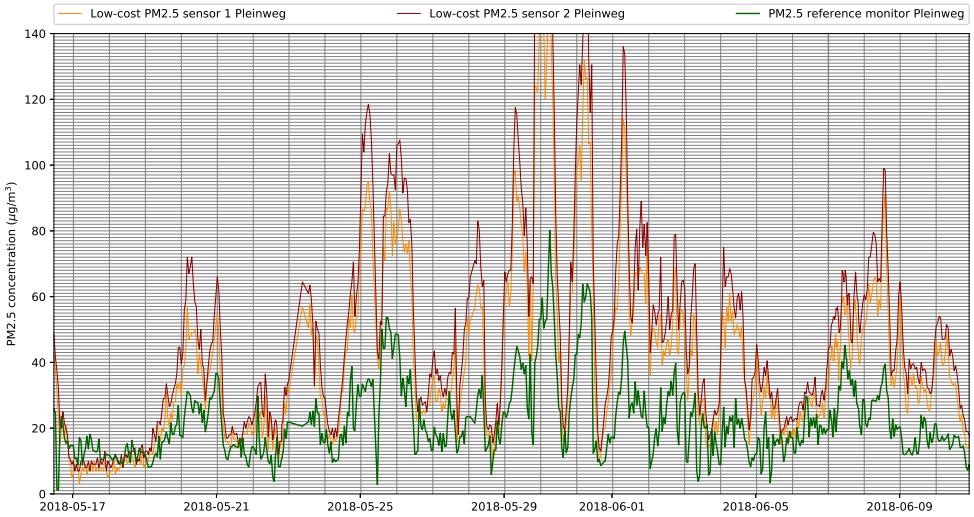


Figure 4.7: Time series plot for the PM data at Pleinweg, with air quality from PM sensor 1 (red), air quality from PM sensor 2 (orange), and the reference air quality (green). The time range is from the 16th of May 2018 10:30 until the 10th of June 2018 24:00.

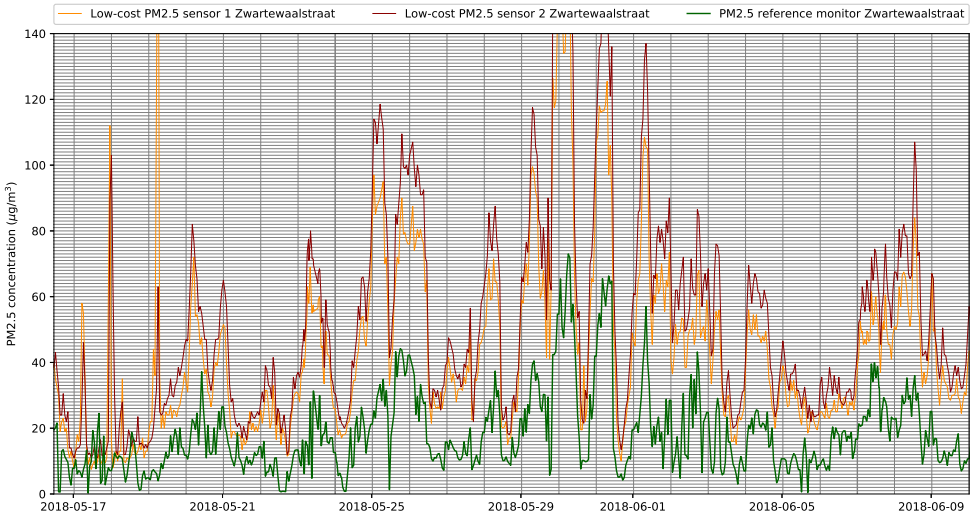


Figure 4.8: Time series plot for the PM data at Zwartewaalstraat, with air quality from PM sensor 1 (red), air quality from PM sensor 2 (orange), and the reference air quality (green). The time range is from the 16th of May 2018 11:00 until the 10th of June 2018 24:00.

		Index in dataset	Original value	Replace value
PW sensor 1	Average	321 322	205.5 198.0	166.5 167.63
	Median	321 322	205.0 199.0	167.75 168.63
PW sensor 2	Average	no outliers		
	Median	no outliers		
ZW sensor 1	Average	345 346	204.75 206.5	182.13 177.19
	Median	345 346 351	203.5 205.5 196.75	170.0 170.0 177.63
ZW sensor 2	Average	345 67	171.0 213.0	145.75 47.5
	Median	345 67	172.5 215.0	146.5 44.0

Table 4.3: Detected outliers, their original value and replace value (Threshold = 4 Standard deviations, Polynomial degree = 2).

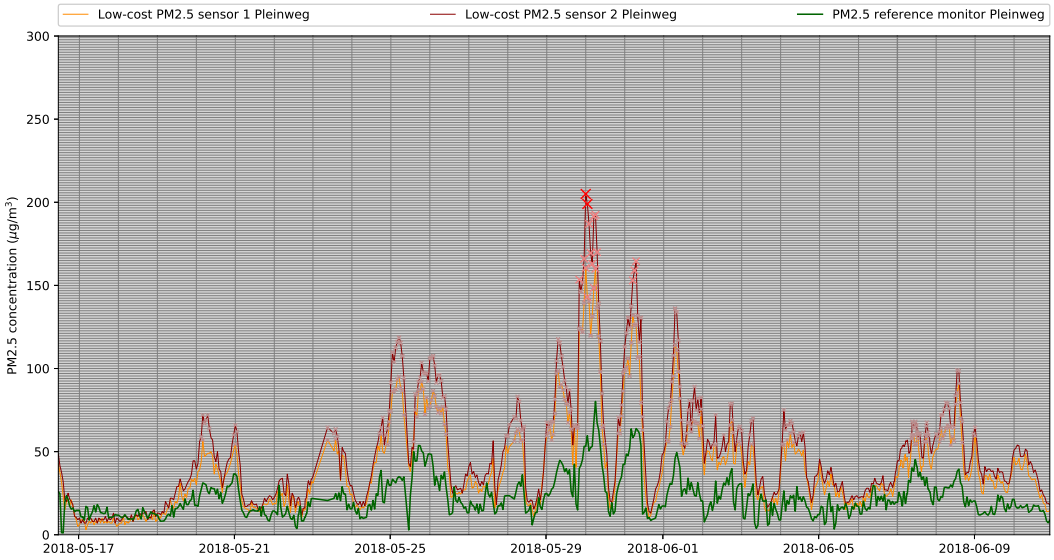


Figure 4.9: Timeseries of Pleinweg with outliers. The most extreme outliers are indicated with a red cross: those are removed from the dataset.

PM2.5 sensor	Resampling method	Standard deviation	RMSE	Systematic error	SE after outlier removal	RMSE after oulier and SE removal
PW sensor 1	Average	32.76	34.93	25.94	25.83	22.91
	Median	33.01	35.09	25.94	25.83	23.17
PW sensor 2	Average	26.81	25.42	18.31	18.31	17.63
	Median	26.86	25.40	18.26	18.26	17.65
PW reference	N.A.	11.76	N.A.	N.A.	N.A.	N.A.
ZW sensor 1	Average	32.20	40.63	33.83	34.03	22.16
	Median	32.45	40.65	33.70	33.56	22.18
ZW sensor 2	Average	26.87	31.39	25.23	25.13	17.08
	Median	26.92	31.23	25.01	24.71	17.07
ZW reference	N.A.	12.24	N.A.	N.A.	N.A.	N.A.

Table 4.4: Standard deviation, RMSE and systematic error of the separate datasets, before outlier removal and normalization.

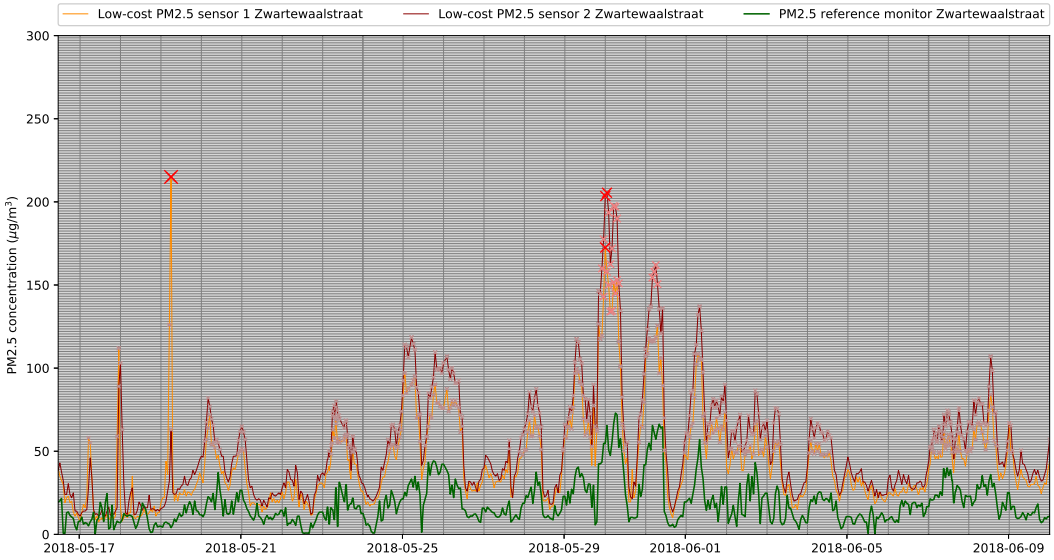


Figure 4.10: Timeseries of Zwartewaalstraat with outliers. The most extreme outliers are indicated with a red cross: those are removed from the dataset.

See the upper graph in figure 4.11, where all temperature and humidity data from the low-cost sensor at both locations is included. Since there is missing data those missing values for temperature and humidity must be interpolated. Formula 4.1 is used for interpolation.

$$Replacevalue_i = \frac{\sum[s_{i-w}; s_{i-1}] + \sum[e_{i+1}; e_{i+w}]}{2 + \frac{w}{2}} \tag{4.1}$$

Where s_{i-1} is the last known value before a missing value i; e_{i+1} is the first known value after that missing value; and w represents the user-defined minimal search window size, i.e. how many known values before and after the missing value are used for the averaging. The Python implementation is on the GitHub page[8]. The middle graph of figure 4.11 shows the effect of applying the interpolation method on the data. There are no zero-values any more and the temperature and humidity values follow a clear trend.

However, there are still some peaks in the humidity data. Those peaks are partly removed with the resampling method of paragraph 4.4, which acts as a median filter on the low-cost sensor data. The effects of this filter are depicted in the bottom graph of figure 4.11.

4.4 COMBINE VARIOUS DATASETS

Combination

Datasets from the sensor nodes are saved in Comma Separated Values (CSV) files, while reference data from Luchtmeetnet and Weerlive are saved in a PostgreSQL database. This data from different sources is combined in two separate CSV files: one for each sensor node location. The data is matched based on date and time. For the matching operation is the *combineDatasets.py*[9] module used.

Resampling

A resampling procedure is needed since the sensor node datasets contains records collected with an interval of 15 minutes; the DCMR reference dataset has an interval of 60 minutes;

8 https://github.com/NiekB4/aqs
9 https://github.com/NiekB4/aqs

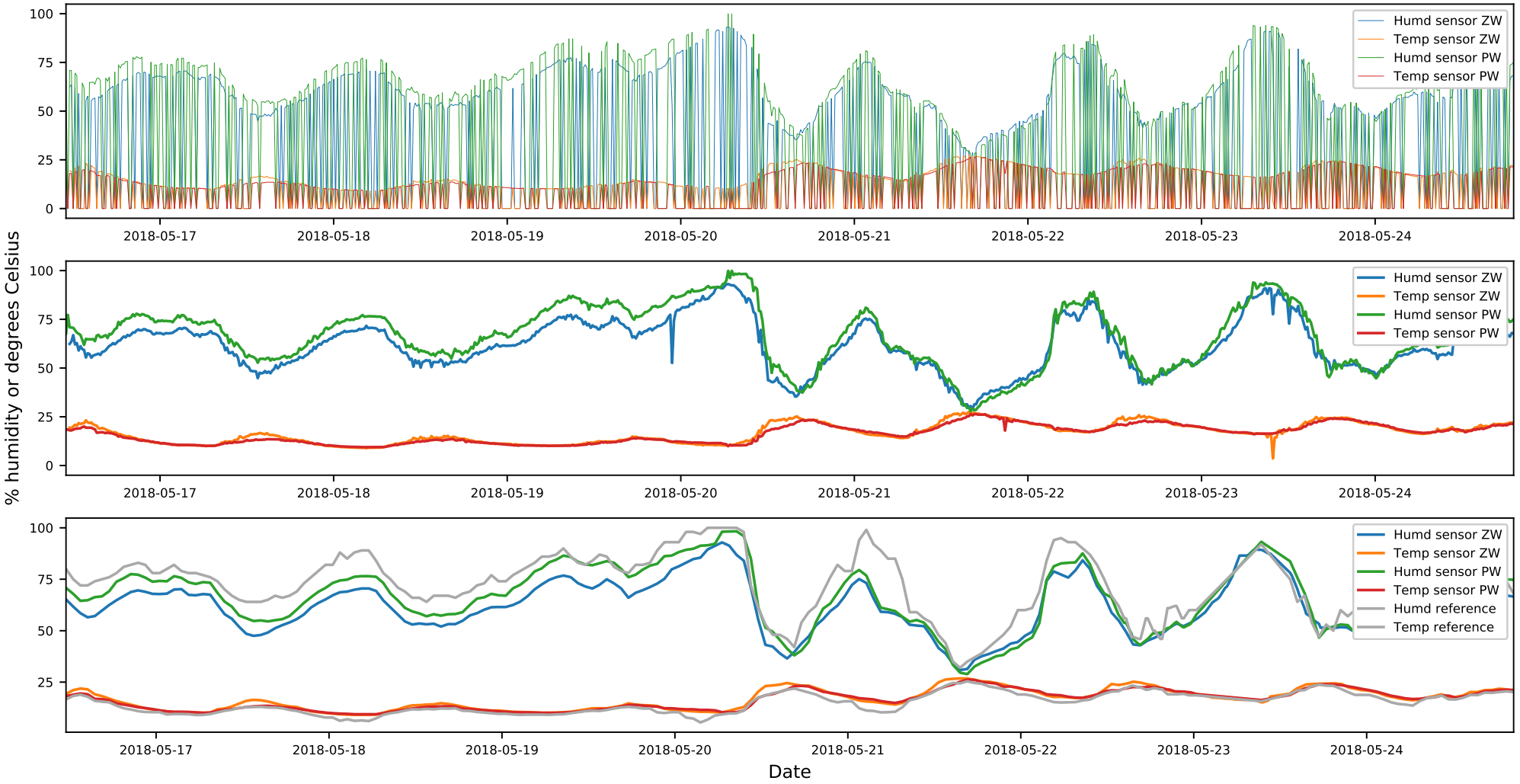


Figure 4.11: Top: missing values for temperature and humidity sensor at Pleinweg and Zwartewaalstraat. Middle: effect of the interpolation algorithm of 4.1 on all raw data (N=3272). Bottom: effect of the resampling algorithm which utilizes a median filter (N=635).

and the Weerlive reference dataset an interval of 20 minutes. The moments that the reference data is collected are also included in figure 4.12, depicted with red bars. Ideally, the sampling interval is brought back to 60 minutes, which is done with interpolation. The blue vertical bar in figure 4.12 indicates the place of the resampled values in the time domain. For data that originates from the sensor nodes is this resampled value an average of four measurements and for weather reference data an average of three measurements. Those averages are coupled with every second measurement of the hour from the sensor node, i.e. the measurement that occurs at the 30th minute.

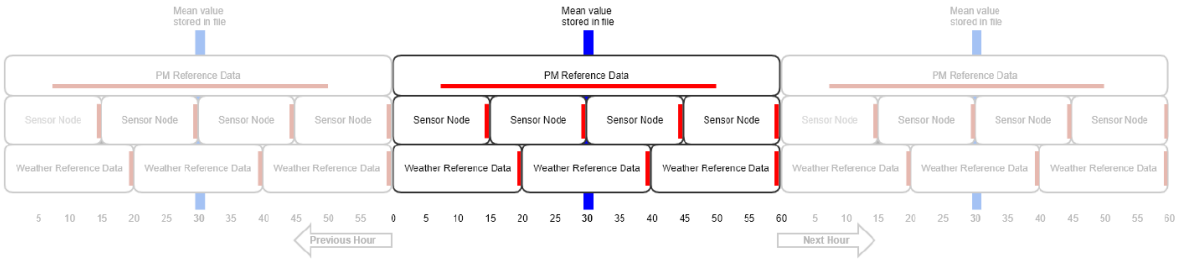


Figure 4.12: Time intervals and moments of data collection of the datasets (red bars); resampled values (blue bars)

Since the resampling procedure aggregates the data from the low-cost sensor nodes it behaves as a filter. Choosing the aggregation method therefore results in different outcomes. In this case there are two options considered: using the median as aggregate and using the average as aggregate.

Normalization

When different types of time series are compared with each other the normalized time series have to be used. Normalization equation 3.2[10] is then used. The normalization is performed after the data cleaning process, thus after removing the gross errors and systematic instrument error. Therefore, these errors does not affect the scale of the normalized PM2.5 time series.

4.5 CALCULATE STATISTICS FOR BASELINE MEASUREMENT

This section elaborates on the implementation of the baseline measurement statistics for the low-cost PM sensors. First, the scatterplots of the low-cost PM sensors plotted against the reference sensors are discussed.

Baseline measurement of RMSE

The Root Mean Square Error before outlier removal was calculated in paragraph 4.3.1. However, that was not-normalized data, while from now on normalized data is used. After outlier removal, the RMSE is calculated again and after removing systematic instrument error it will be calculated for a third time. Table 4.5 contains these calculated RMSE scores, together with the Standard Deviation and Systematic Error of the normalized sensor data. Together they represent the baseline measurement statistics to which the results from the correction model will be evaluated.

Noise models

Noise models represent the mean and deviation of the random measurement error. Random error of the sensors – represented by random variations around the true reference value [Fisher and Tate, 2006] – is shown in two various ways in figure 4.14: using histograms and using a Gaussian PDF. Besides, there are three representations for random error includes in those plots.

10 https://github.com/NiekB4/aqs

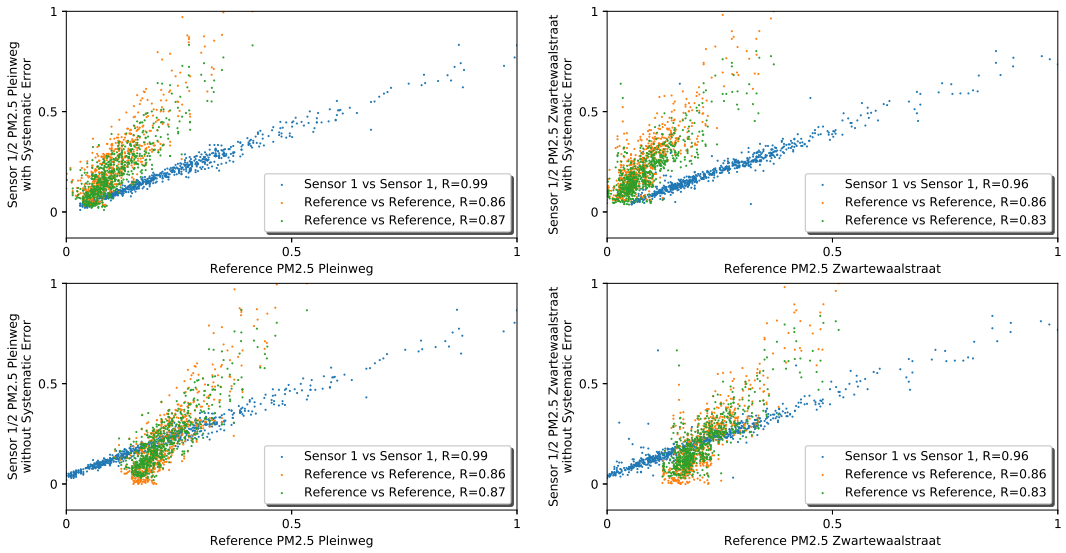


Figure 4.13: Top row: scatterplots of the normalized PM datasets before Systematic Error is removed. Bottom row: after removal of Systematic Error from the original observations.

	RMSE normalized data	Standard Deviation normalized data	R	RMSE after SE removal data	Standard Deviatio after SE removal
PW sensor 1	0.1813	0.1702	0.86	0.1249	0.1756
PW sensor 2	0.1327	0.1403	0.87	0.0918	0.1448
ZW sensor 1	0.2045	0.1620	0.86	0.1706	0.1714
ZW sensor 2	0.1527	0.1316	0.83	0.1256	0.1392

Table 4.5: Statistics for the normalized data after the removal of outliers from table 4.3. Consequently, after subtracting the Systematic Error RMSE was calculated again.

First, the noise model for both low-cost PM sensors on the sensor nodes, which show the deviation around the mean value of the random error and therefore indicating the precision of the low-cost instruments. Since the histogram and PDF plots are relatively tight – they fall within two standard deviations σ after error removal – the precision of the instruments seems to be relatively good.

Second and third, the noise models for PM sensor 1 respectively sensor 2 on the sensor node. In these cases are the values from the DCMR reference dataset used. These noise models show the deviation around the true reference values and therefore indicate the accuracy of the low-cost instruments. The histograms and PDF plots in the figure are relatively broad thus the accuracy of the instruments is relatively poor.

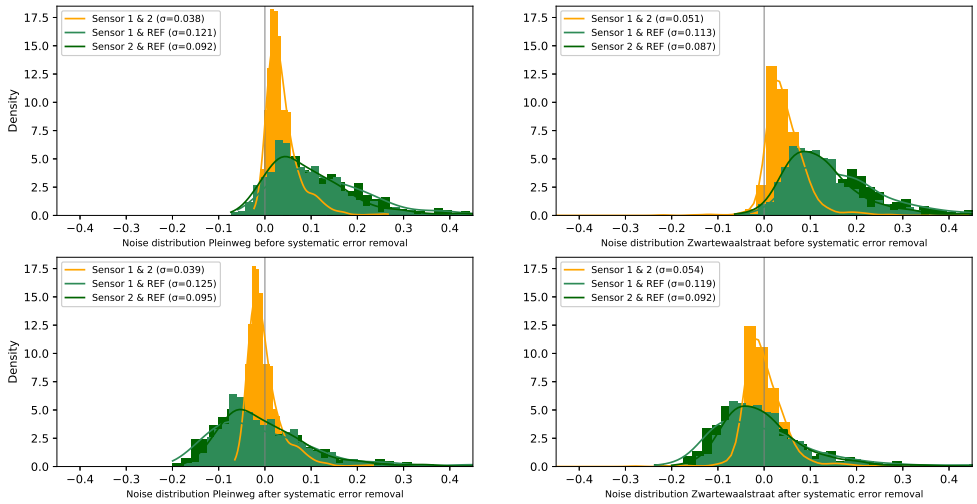


Figure 4.14: Noise models for the low-cost PM2.5 sensors. The orange histograms and Gaussian PDF plots represent the noise of the low-cost sensors compared with each other. The green and dark green ones show the noise models of the low-cost sensors against the reference monitors.

4.6 RELATIONSHIPS BETWEEN THE VARIABLES

This section is a discussion on the relations between the PM datasets and the other environmental phenomena: Which of those environmental phenomena should be included in the correction model? First, each of the candidate environmental variables are compared with each other to check for multicollinearity. Thereafter, the correlations with PM2.5 are investigated.

4.6.1 Relationships between the independent variables

The scatterplots in figure 4.15 show the relations between each of the candidate environmental variables, for Pleinweg (left) and Zwartewaalstraat (right). They show no strong relationships between the independent variables. Using equation 3.5 the VIF between each of the candidate independent variables are calculated, in order to check for the multicollinearity between the independent datasets. The VIF and correlation coefficient between each independent dataset is shown in table 4.6. The table shows that none of the VIF values are above 5. Moreover, the p-values are all under 0.05 (5%). Thus the null hypothesis that there is no difference between the independent datasets can be rejected: a significant difference exists.

So none of the independent variables are dependent on each other, therefore for now all independent variables could be used for the MLR correction models.

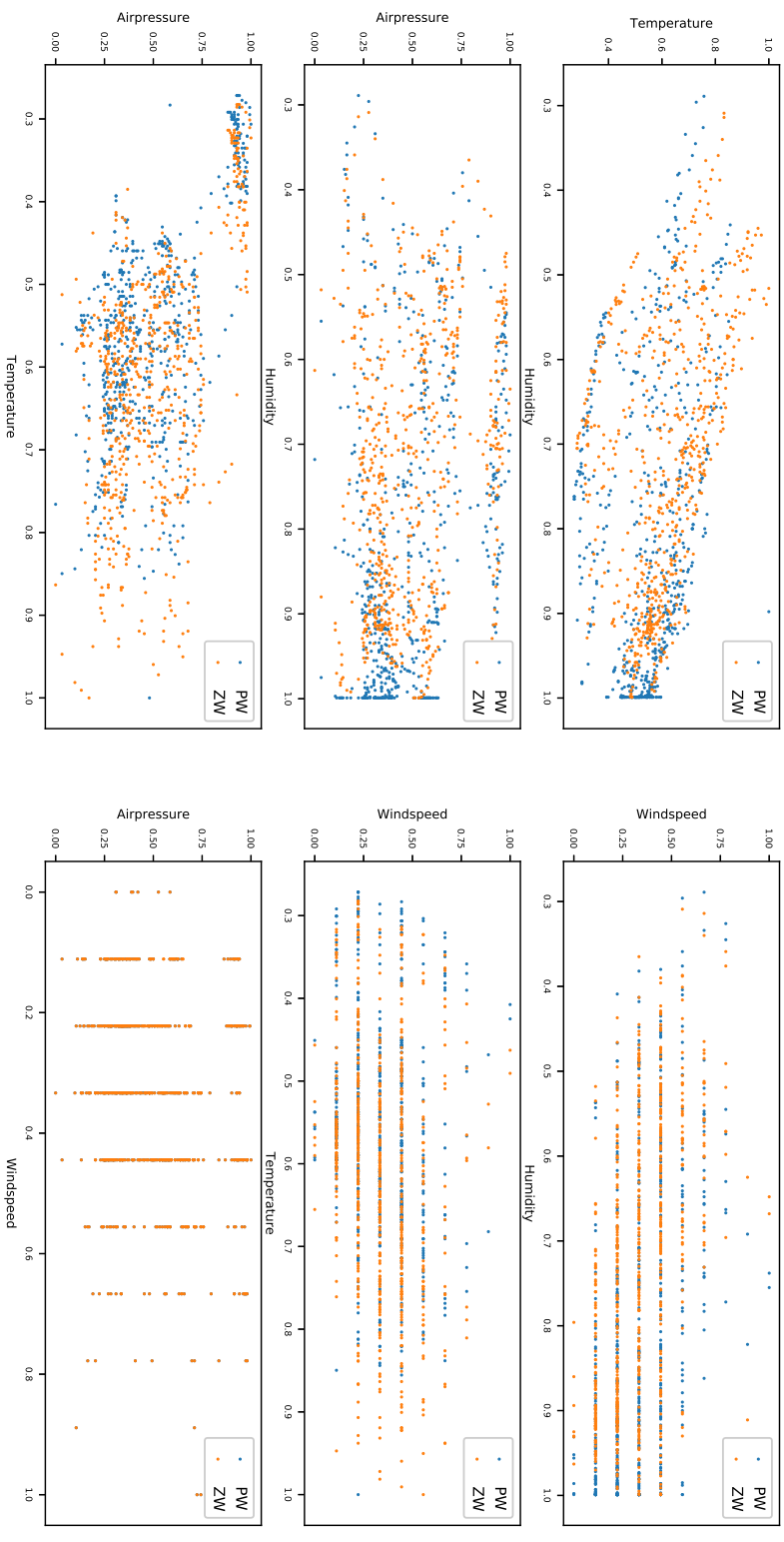


Figure 4.15: Scatterplots for data of each of the candidate variables – humidity, temperature, air pressure and wind speed (data for Pleinweg is indicated with blue dots and for Zwartewaalstraat with orange dots).

Set	VIF	R (Pearson)	p value
Humidity - Temperature	1.0848	-0.28	< 0.05
Humidity - Windspeed	1.2312	-0.43	< 0.05
Humidity - Airpressure	1.0767	-0.27	< 0.05
Temperature - Windspeed	1.0114	0.11	< 0.05
Temperature - Airpressure	1.5648	-0.60	< 0.05
Airpressure - Windspeed	1.0309	0.17	< 0.05
Set	**VIF**	**R (Pearson)**	**p value**
Humidity - Temperature	1.2296	-0.43	< 0.05
Humidity - Windspeed	1.3206	-0.49	< 0.05
Humidity - Airpressure	1.0695	-0.25	< 0.05
Temperature - Windspeed	1.0503	0.22	< 0.05
Temperature - Airpressure	1.3781	-0.52	< 0.05
Airpressure - Windspeed	1.0316	0.17	< 0.05

Table 4.6: The Variance Inflation Factor (VIF) multicollinearity metric per set of independent variables. Above: Pleinweg, below: Zwartewaalstraat.

4.6.2 Relationships between the independent and dependent variables

Humidity

The bottom plot in figure 4.11 is a time series plot of the interpolated humidity data from the sensor nodes and from the reference dataset – depicted with grey lines. Clearly, the data from both locations as well as from the reference location follow the same trend. However, the extreme values in some cases differ up to 15%. Next to that, the humidity data from the sensor node at Pleinweg has a high amount of maximum (100%) values which is caused by the malfunctioning hardware and the interpolation. It is clear that this type of low-cost sensor is more likely to yield errors when humidity is high ($\pm$90%) over a longer period. The sensor at Zwartewaalstraat, on the other hand, shows less extreme high humidity values.

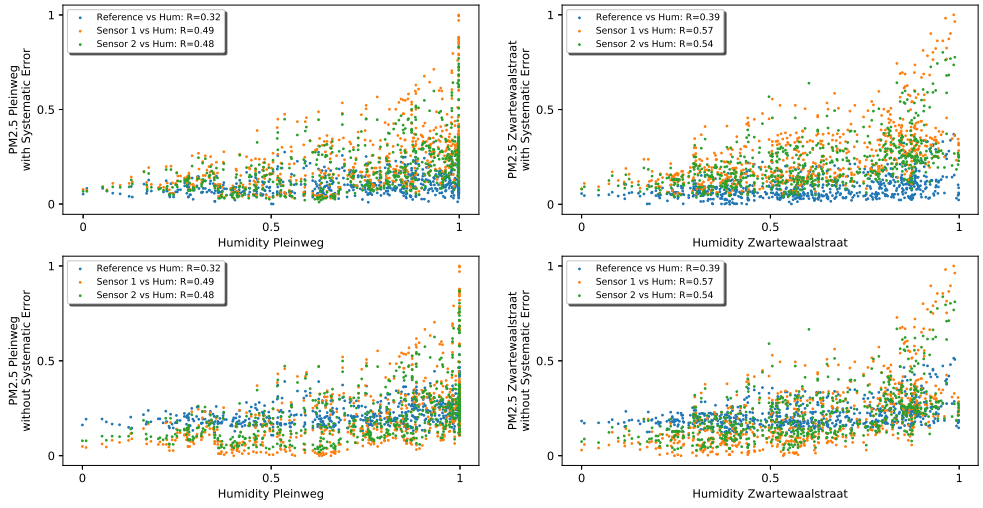


Figure 4.16: Top row: Scatterplots for humidity versus PM2.5 at Pleinweg and Zwartewaalstraat before systematic error is removed. Bottom row: Scatterplots for humidity versus PM2.5 at Pleinweg and Zwartewaalstraat after systematic error removal. The plot includes the deviation of the low-cost PM sensors – orange and green – as well as the high-quality BAM monitors – blue – against humidity from the sensor node.

Figure 4.16 shows two scatterplots of humidity versus PM concentrations in the Pleinweg and Zwartewaalstraat. Remarkable regarding those scatterplots is the moderate positive relationships between humidity and PM2.5 from the low-cost sensor nodes – ranging from R=0.48 to R=0.57. At the same time, the correlation coefficient between humidity and PM2.5 from the high-quality reference stations is weaker: R=0.39 and R=0.32 for Pleinweg and Zwartewaalstraat, respectively. This finding from the collected data is in line with the suggestions from Postolache et al. [2009] and Cross et al. [2017]: that low-cost PM sensors are affected by cross interference with humidity. This indicates that in this research the correction model should include a component that corrects PM when humidity is within a specific range.

Temperature

In the original dataset there were three extreme values in the temperature dataset. Those occured at Pleinweg on the 27th of May 22:30 and at Zwartewaalstraat at 29th of May 18:30 and 5th of June 6:30. Remarkably, the extreme maximum value at Pleinweg occured just before the sensor node would quit collecting data for around two hours, i.e. there was an error during those two hours.

Except for those three extreme values, the temperature data from the low-cost sensor nodes and the reference data follow the same trend, although the minimum values for the reference dat are most of the times around 3 °C lower than from the sensor nodes. Also, the temperature data from Zwartewaalstraat is often around 3 °C higher than at Pleinweg, which can be caused by the placement of the sensor nodes: the first is placed at a sunny location, the latter under the canopy of trees.

The scatterplots of figure 4.17 show the correlation between the collected temperature and PM data. Clearly, there is no correlation between temperature and PM since the coefficients of correlation are all below 0.20. There is however a high peak at the 18 °C to 20 °C range. Those peaks are visible for temperature plotted against data from the low-cost sensor nodes, as well as plotted against the reference data. The peak around those temperature values makes sense since during daytime the temperature is higher and also the amount of traffic and industry activities is higher, therefore a higher concentration of PM is a logical consequence. This indicates that the relationship between temperature and PM from the low-cost sensors could be non-linear.

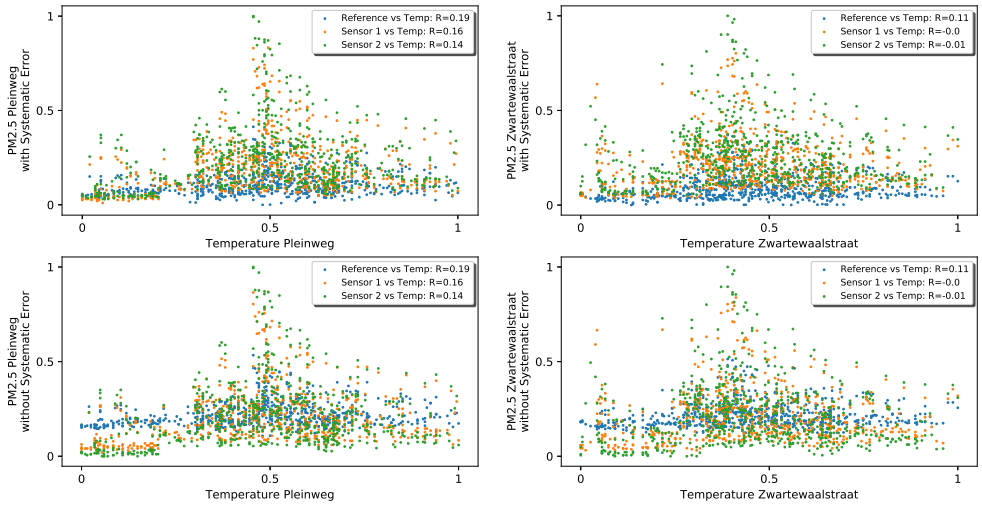


Figure 4.17: Top row: Scatterplots for temperature versus PM2.5 at Pleinweg and Zwartewaalstraat before systematic error is removed. Bottom row: Scatterplots for temperature versus PM2.5 at Pleinweg and Zwartewaalstraat after systematic error removal.

Air pressure

The air pressure data is from the Weerlive reference dataset, so for the whole city of Rotterdam. A scatterplot of this data versus the various PM datasets is shown in figure 4.18. That figure reveals that both the reference and sensor node PM data yield higher values when air pressure is in the range of 1012 to 1015 hPa – 0.1 to 0.3 on the normalized scale of the scatterplot. Moreover, at Zwartewaalstraat are relatively high values at the 2024-2025 hPa range – 0.9 to 1.0 on the normalized scale. These high values are not shown in the reference dataset, neither in the Pleinweg dataset.

Overall, there is only a weak negative relationship between air pressure and PM. However, at Pleinweg the relationships between this independent variable and PM from the low-cost sensors (R=-0.34 and R-0.32) are stronger than between the independent variable and PM from the reference monitor (R=-0.22). The R-value for Zwartewaalstraat does not show this difference, although the scatterplot indicates look similar to the scatterplot of Pleinweg.

Therefore, the independent variable air pressure will be included in the correction model, despite the weak relationship with the dependent variable.

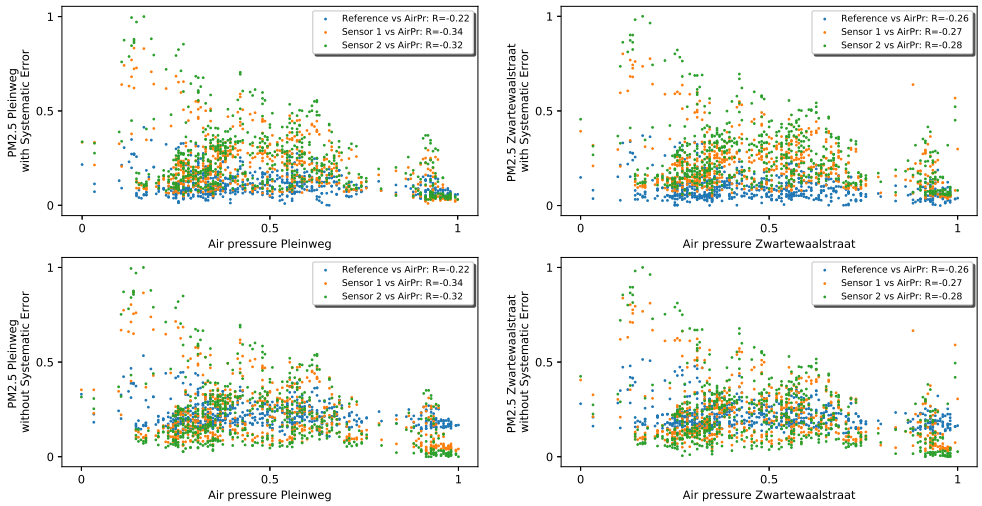


Figure 4.18: Scatterplots for air pressure versus PM2.5 at Pleinweg and Zwartewaalstraat. The plot includes the deviation of the low-cost PM sensors – orange and green – as well as the high-quality BAM monitors – blue – against air pressure from the reference dataset from KNMI.

Wind speed

Wind speed is included in the search for related environmental variables since the sensitive particles sensors could be affected by high wind speeds: the sensor would yield high PM values. On the other hand, high wind speeds could "clean" the air, i.e. remove the particles from the air, and therefore the PM sensor would sense lower concentrations.

The results from the scatterplot of the data collection at both study locations are in line with the second theory. It shows that the low-cost sensors on the node are not necessarily heavier affected by wind than the high-quality BAM monitors used for the reference data. When wind speed increases, it seems that with both type of PM monitoring techniques lower particle concentrations are found. This is also reflected with the – weak – negative correlation coefficient for both monitoring techniques.

So the relationship of wind speed with PM look similar to the relationship of air pressure with PM. Namely, despite the relationship is weak, the relationship between this independent variable and the low-cost PM sensor data is still stronger than the relationship between this independent variable and the data from the reference PM monitor. Therefore, for the same reason the independent variable of wind speed will not be excluded from the correction model.

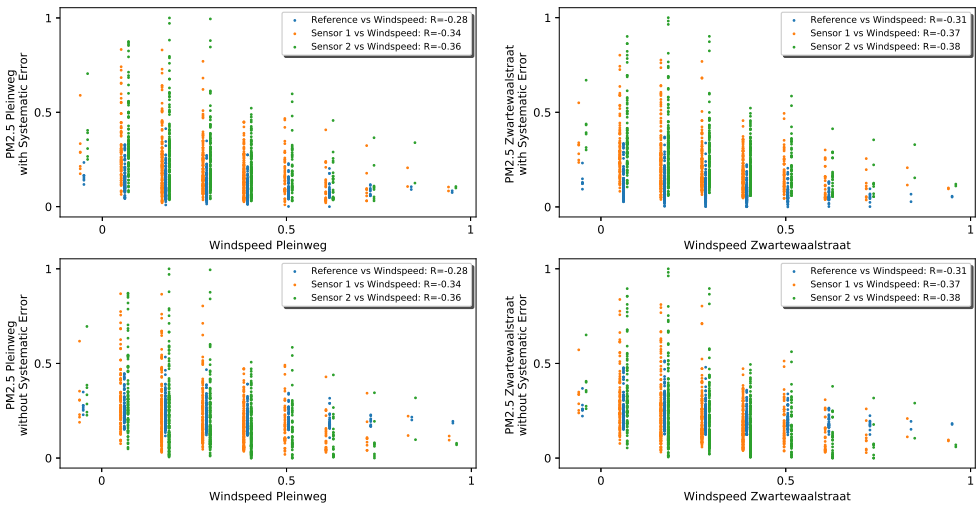


Figure 4.19: Scatterplots for wind speed versus PM2.5 at Pleinweg and Zwartewaalstraat. The plot includes the deviation of the low-cost PM sensors – orange and green – as well as the high-quality BAM monitors – blue – against wind speed from the reference dataset from KNMI.

4.6.3 Conclusion for relationships between the variables

The analysis of the relationships between the independent variables shows that there is no multicollinearity among the independent variables, therefore all variables may be included in the correction model. However, there is also no strong relationships between any of the independent and dependent variables. There exist only moderate relationships between humidity and PM, between windspeed and PM, and between air pressure and PM. However, despite those weak relationships, the relationships with PM from the low-cost sensors are still stronger than the relationships with PM from the reference monitors. Therefore, they are included in the next section, where the parameters for the correction models are calculated. Finally, Temperature shows no linear relationship with PM from the low-cost sensor, though only a weak non-linear relationship when investigating the scatterplots.

4.7 CALCULATE CORRECTION MODEL PARAMETERS FOR VARIOUS SETTINGS

This paragraph elaborates on one example implementation of a correction model for type A and B models. Then the proposed domains for Type C and D models are discussed.

Example implementation of type A and B models

The example for type A and B models is the most basic correction model, containing one parameter for the sensed PM2.5 value and no parameters for other eventually related environmental phenomena. The results are shown in table 4.7.

In this example are the parameters for the dataset of PM2.5 sensor 1 at Pleinweg calculated. The least squares fit is implemented with ready-to-use Python software[11]. For this example, the constant value is 0.075657 and the parameter 0.312259. Consequently, using these parameters are new values for PM2.5 calculated for the same time series data. Obviously, this would benefit the data quality of this particular time series significantly. The evaluation statistics column in the table indeed shows an increase of the data quality. Then, the same paramaters

11 The *numpy linalg.lstsq* package, see https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.linalg.lstsq.html

Correction Model Implementation and Evaluation				
Sensor #1 Pleinweg	**Constant: 0.151969... Parameter: 0.312259...**	**RMSE**	**Standard Deviation**	**R**
Baseline Measurement sensor#1 Pleinweg		0.1249*	0.1756	0.8643
After parameters applied on sensor data		0.0319	0.0548	0.8643
Baseline Measurement sensor#2 Pleinweg ("other")		0.0952*	0.1448	0.8674
After parameters applied on sensor data		0.0331	0.0452	0.8674
Baseline Measurement sensor#1 Zwartweaalstraat		0.1192*	0.1714	0.8639
After parameters applied on sensor data		0.0334	0.0535	0.8639
Baseline Measurement sensor#2 Zwartewaalstraat		0.0918*	0.1392	0.8344
After parameters applied on sensor data		0.0381	0.0435	0.8344

Table 4.7: An example of the implementation of a correction model. This is the basic correction model which includes only parameters for PM (* values from table 4.5).

are applied to the second sensor, but on the same sensor node at Pleinweg. The evaluation statistics again show an increase for the data quality.

Finally, for the validation part of this implementation of the correction model are the calculated parameters used on the time series for the low-cost PM sensors on the other sensor node: the one at Zwartewaalstraat. The evaluation statistics indicate that the data quality increased, therefore this correction model is valid. So the basic correction model which uses only correction parameters for PM already performs well. The resulting noise models are shown in figure 4.20.

Proposed domains for Type C and D models

In this research are two domains proposed for type C and D models: wind direction and the time interval. Wind direction is chosen because as categorical data type it can be distinguished in groups (North, East, South, West) and data on the wind direction for Rotterdam is available in the Weerlive API. Next to that, different wind directions could result in different concentrations of PM2.5, since high-pollutant industrial activities in Rotterdam are placed in various parts of the city. For example, in the harbor area are most industrial activities and that area is located in the west part of the city. Therefore, it can be expected that if the wind direction is West, the PM2.5 concentrations are higher.

Time interval is chosen since environmental conditions change during the 24 hours of the day: splitting the day in subsets of equal size having from the same hours could yield better parameters for those subsets. For this research the data will be collected during one month and with an sample interval of one observation per hour – the sample of the DCMR reference data – 744 observations are expected. Therefore, it is chosen to split to use time periods of four hours, thus six subsets, where each "time period subset" will have 124 observations. That is enough to use RMSE as evaluation metric, as argued by Chai et al. [2014].

- 00:00 - 04:00
- 04:00 - 08:00
- 08:00 - 12:00
- 12:00 - 16:00
- 16:00 - 20:00
- 20:00 - 00:00

4.8 CONCLUSION FOR IMPLEMENTATION CHAPTER

The performance and validation of datasets that are modified using the correction models part of the results of this research. Therefore, they are included in the next section.

This chapter described the implementation of the methodology that was proposed in chapter 3. The PM data is normalized, outlier are removed and systematic errors are removed. All relationships between the independent variables and between the independent variables and dependent variable are investigated. All quantitative candidate independent variables – humidity, temperature, wind speed, air pressure – are sufficient to be included variants of the

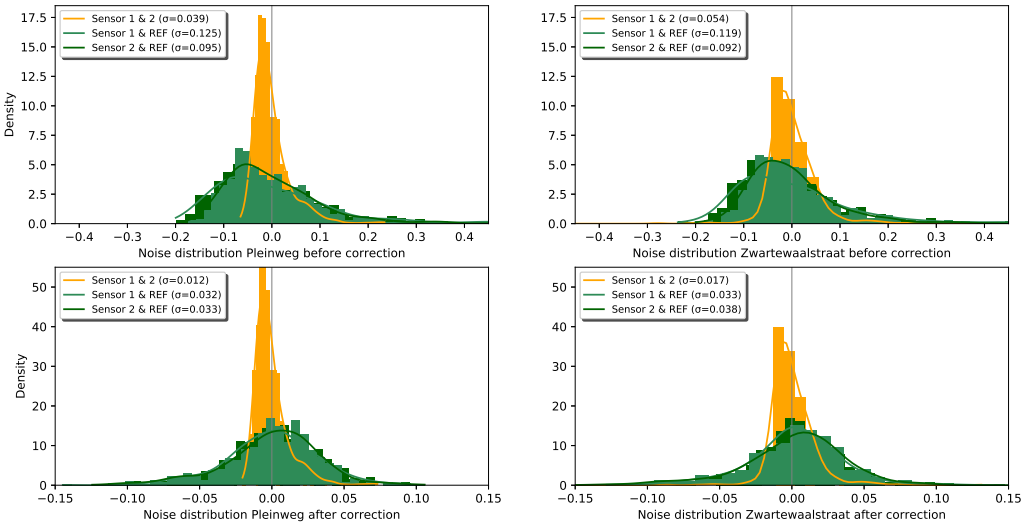


Figure 4.20: Noise models before and after applying the correction model with parameters for PM only. Notice the different ranges on the axes.

correction model. So the data is now prepared for creating correction models with the – stepwise – MLR method, applying those correction models and assessing the performance.

5 RESULTS AND DISCUSSION

5.1 INTRODUCTION

The previous chapter ended with the table containing evaluation statistics and the noise model of only one variant of a correction model. That correction model included only parameters for PM, and was calculated only for low-cost sensor #1 at Pleinweg. This chapter shows the results of the implemented stepwise MLR method. The proposed algorithm is therefore applied. All the correction models and their RMSE evaluation metrics are shown and discussed. First, the correction models of Type A and B are discussed. Thereafter, Type C and D models are discussed.

5.2 CORRECTION MODELS OF TYPE A AND B

Figures 5.1 and 5.2 show the results of applying various correction models on the PM datasets. These are correction models for only one domain, but with varying amounts of parameters. Therefore, it are "type A" or "type B" correction models as depicted in figure 3.3.

The figures show the evaluation criterium – RMSE – per implemented correction model. Besides, the baseline measurement of RMSE is included in the bar chart (grey bar). In each bar chart are the correction parameters calculated for the first location, applied on the data for that location, applied on the data for the other sensor on the same location, and applied on data originating from the two sensors at the other location. The "other" sensor location is the location where the performance of the correction model is validated.

For example, for the top bar chart of figure 5.1 are the correction parameters calculated with the data from Pleinweg sensor 1; the parameters are applied on the data from this sensor ("PW#1", the red bars); the parameters are applied on the data from the other sensor at the same location ("PW#2", orange bars); and for validation are the parameters applied on the data from sensors at the other location ("ZW#1", blue bars and "ZW#2", purple bars).

Both figures 5.1 and 5.2 include various correction models. The first correction model that is included contains parameters for only PM ("PM"), the second correction model parameters for PM and humidity ("PM+H"), then PM and temperature ("PM+T"), until all external datasets are included in the model ("PM+H+T+WS+AP").

Furthermore, per type of correction model there are three versions calculated, depending on the chosen polynomial degree. The polynomial degrees are one, two or three. Therefore there are always three bars included. In order to improve the visualization, the transparency of the bars in the bar chart is alternated per set of three. Besides, the results of applied with correction models using polynomial degrees > 3 are excluded from figures 5.1 and 5.2. Namely, while they improve the data quality at the sensor location for which such a correction model is created, it does not further improve the data quality on the other sensor location.

Overall, both figures indicate that the correction model decreases the RMSE error for the observations. Therewith it seems to improve the accuracy of the data. Also when the correction parameters calculated for one street are transferred to the other street, the RMSE decreases. So the parameters can be transferred to other locations and improve the data quality at that other location.

However, the figures already indicate that including parameters for more variables than only PM does not really improve the data quality. That is a logical results since in the previous chapter 4 was already found that the independent variables have weak relationships with the dependent variable PM. On the other hand, including more independent variables neither decreases the data quality on the validation location. Which are good performing correction models for the two PM sensors on each of the two locations and which independent variables are included?

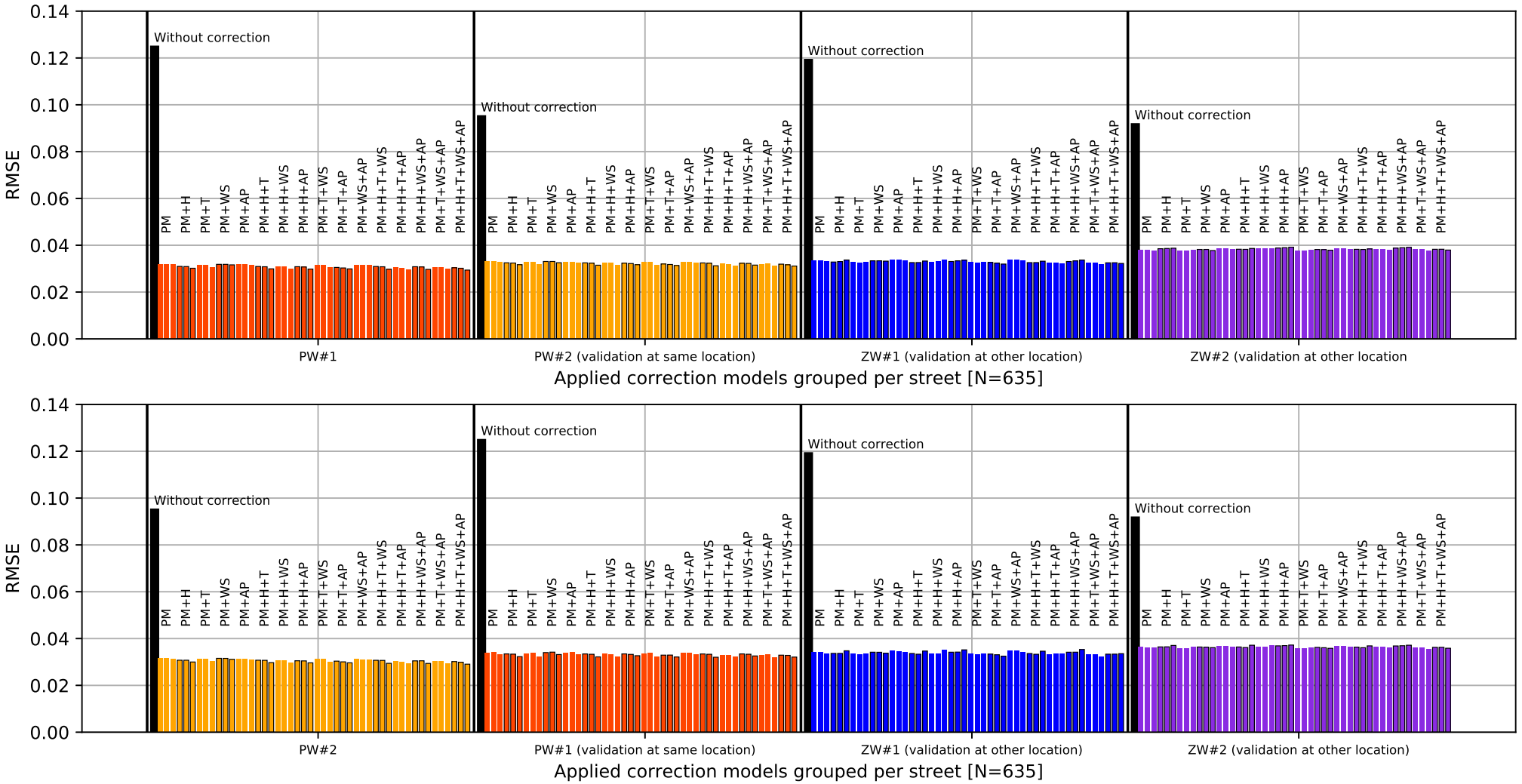


Figure 5.1: Resulting RMSE for various correction models. These models are created with parameters for Pleinweg sensor 1 (top) and Pleinweg sensor 2 (bottom).

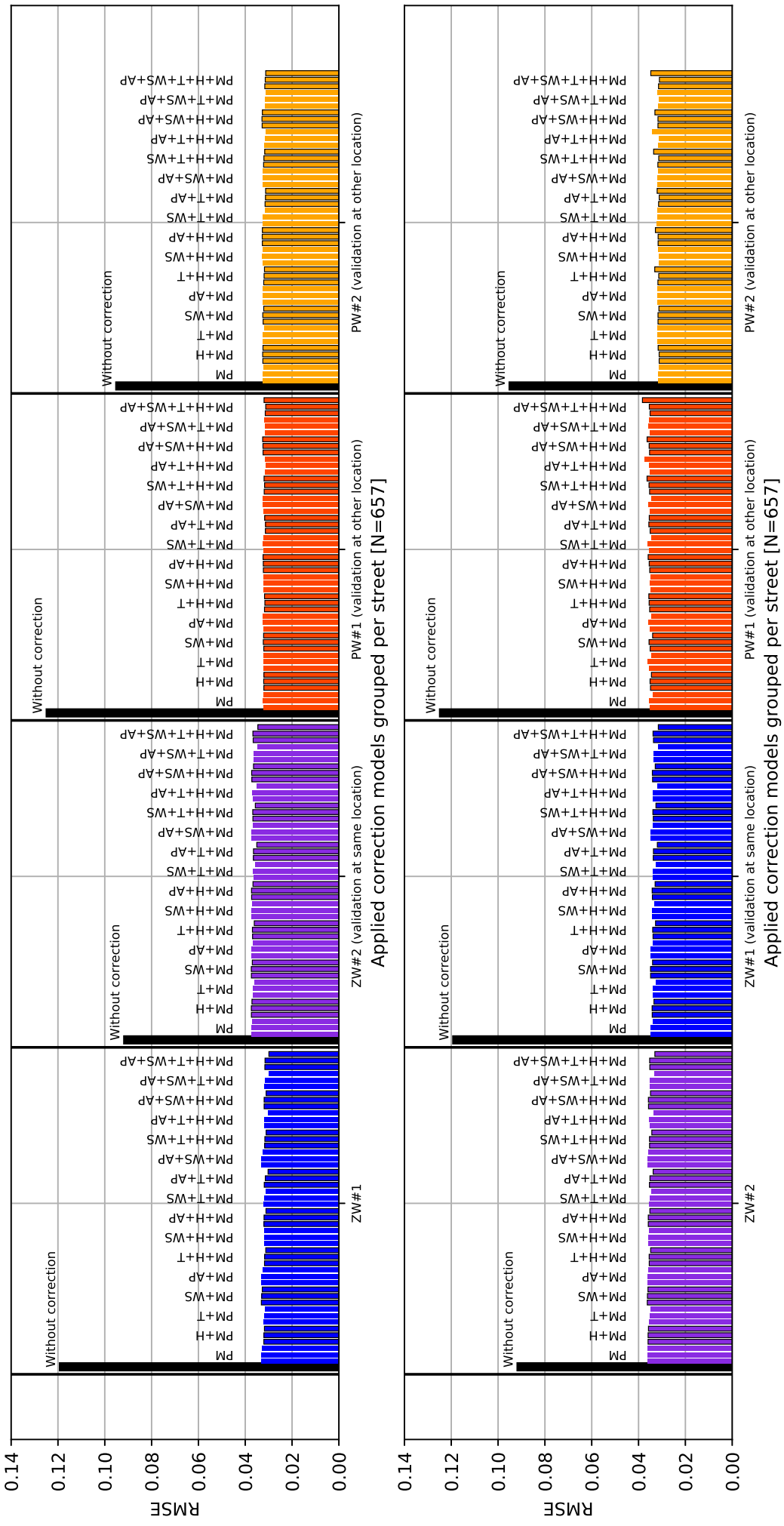


Figure 5.2: Resulting RMSE for various correction models. These models are created with parameters for Zwartewaalstraat sensor 1 (top) and Zwartewaalstraat sensor 2 (bottom).

Best correction models when the parameters are calculated for PW#1				
	Pleinweg sensor node		Validation: Zwartewaalstraat sensor node	
	PW#1	PW#2	ZW#1	ZW#2
1	PM+H+T+WS+AP, deg=3 RMSE: 0.02940	PM+H+T+WS+AP, deg=3 RMSE: 0.03112	PM+T+WS+AP, deg=3 RMSE: 0.03192	PM+T, deg=2 RMSE: 0.03759
2	PM+H+T+AP, deg=3 RMSE: 0.02955	PM+T+WS+AP, deg=3 RMSE: 0.03117	PM+T+AP, deg=3 RMSE: 0.03197	PM+T, deg=1 RMSE: 0.03759
3	PM+H+WS+AP, deg=3 RMSE: 0.02970	PM+H+T+WS, deg=3 RMSE: 0.03124	PM+H+T+WS+AP, deg=3 RMSE: 0.03223	PM, deg=3 RMSE: 0.03766
4	PM+H+T+WS, deg=3 RMSE: 0.02976	PM+H+T+AP, deg=3 RMSE: 0.03133	PM+H+T+AP, deg=3 RMSE: 0.03225	PM+T+WS, deg=1 RMSE: 0.03770
5	PM+T+WS+AP, deg=3 RMSE: 0.02977	PM+T+AP, deg=3 RMSE: 0.03135	PM+T+AP, deg=2 RMSE: 0.03239	PM+T+WS, deg=2 RMSE: 0.03773
Best correction models when the parameters are calculated for PW#2				
	Pleinweg sensor node		Validation: Zwartewaalstraat sensor node	
	PW#2	PW#1	ZW#1	ZW#2
1	PM+H+T+WS+AP, deg=3 RMSE: 0.02907	PM+H+T+WS, deg=3 RMSE: 0.03203	PM+T+WS+AP, deg=3 RMSE: 0.03237	PM+T+WS+AP, deg=3 RMSE: 0.03552
2	PM+H+T+AP, deg=3 RMSE: 0.02935	PM+T+WS+AP, deg=3 RMSE: 0.03204	PM+T+AP, deg=3 RMSE: 0.03250	PM+T, deg=2 RMSE: 0.03573
3	PM+H+WS+AP, deg=3 RMSE: 0.02938	PM+H+T+WS+AP, deg=3 RMSE: 0.03210	PM+T+AP, deg=2 RMSE: 0.3310	PM+T+WS, deg=2 RMSE: 0.03581
4	PM+T+WS+AP, deg=3 RMSE: 0.02938	PM+T+AP, deg=3 RMSE: 0.03215	PM+T+WS+AP, deg=2 RMSE: 0.03319	PM+T+AP, deg=3 RMSE: 0.03583
5	PM+H+T+WS, deg=3 RMSE: 0.02942	PM+H+T, deg=3 RMSE: 0.03218	PM+H+T+WS, deg=2 RMSE: 0.03330	PM+H+T+WS+AP, deg=3 RMSE: 0.03587
Best correction models when the parameters are calculated for ZW#1				
	Zwartewaalstraat sensor node		Validation: Pleinweg sensor node	
	ZW#1	ZW#2	PW#1	PW#2
1	PM+H+T+WS+AP, deg=3 RMSE: 0.02990	PM+H+T+WS+AP, deg=3 RMSE: 0.03460	PM+H+T+WS+AP, deg=2 RMSE: 0.03125	PM+T+WS+AP, deg=3 RMSE: 0.3116
2	PM+T+WS+AP, deg=3 RMSE: 0.03002	PM+T+WS+AP, deg=3 RMSE: 0.03464	PM+H+T+AP, deg=2 RMSE: 0.03128	PM+H+T+WS+AP, deg=3 RMSE: 0.03126
3	PM+H+T+AP, deg=3 RMSE: 0.03016	PM+T+AP, deg=3 RMSE: 0.03507	PM+T+AP, deg=2 RMSE: 0.03134	PM+T+AP, deg=3 RMSE: 0.03130
4	PM+T+AP, deg=3 RMSE: 0.03026	PM+H+T+AP, deg=3 RMSE: 0.03511	PM+T+AP, deg=1 RMSE: 0.03140	PM+H+T+AP, deg=3 RMSE: 0.03132
5	PM+H+T+WS, deg=3 RMSE: 0.03103	PM+T+WS, deg=3 RMSE: 0.03563	PM+H+T+AP, deg=1 RMSE: 0.03140	PM+T+AP, deg=2 RMSE: 0.03135
Best correction models when the parameters are calculated for ZW#2				
	Zwartewaalstraat sensor node		Validation: Pleinweg sensor node	
	ZW#2	ZW#1	PW#1	PW#2
1	PM+H+T+WS+AP, deg=3 RMSE: 0.03304	PM+H+T+WS+AP, deg=3 RMSE: 0.03157	PM+WS, deg=3 RMSE: 0.03400	PM+H+T+WS+AP, deg=2 RMSE: 0.03110
2	PM+T+WS+AP, deg=3 RMSE: 0.03330	PM+T+WS+AP, deg=3 RMSE: 0.03175	PM, deg=3 RMSE: 0.03400	PM+T+AP, deg=2 RMSE: 0.03119
3	PM+H+T+AP, deg=3 RMSE: 0.03358	PM+H+T+AP, deg=3 RMSE: 0.03185	PM+WS+AP, deg=3 RMSE: 0.03445	PM+H, deg=1 RMSE: 0.03120
4	PM+T+AP, deg=3 RMSE: 0.03378	PM+T+AP, deg=3 RMSE: 0.03199	PM+AP, deg=3 RMSE: 0.03447	PM+H+T+AP, deg=2 RMSE: 0.03121
5	PM+H+T+WS, deg=3 RMSE: 0.03435	PM+H+T+WS, deg=3 RMSE: 0.03258	PM+T, deg=3 RMSE: 0.03448	PM+H, deg=2 RMSE: 0.03121

Table 5.1: Top 5 for each variant of the correction models, type A and B

Top-5 best correction models for type A and B

Table 5.1 shows each top-five correction model per variant. This table reveals how good each correction model variant worked. Choosing a higher polynomial degree improves the data at the location for which the parameters are calculated: at all four locations the correction model with the highest polynomial degree (i.e. 3) results in the lowest RMSE and therefore scores best, when applied on the dataset for which the parameters are calculated. This was expected, since the least squares method tries to fit data from one dataset to another dataset and calculates the parameters that effectuate that "best" situation: in this case fitting data from the low-cost PM sensors to the high-quality PM monitors, with or without parameters for external environmental phenomena. And, the higher the polynomial degree, the more parameters, and thus the better the fit.

Also when applied to the datasets from the other sensor – but still on the same sensor node – the correction models that has included most parameters often yields the lowest RMSE values. This result was also expected since the low-cost PM sensors at the two sensor nodes have a similar noise model, see figure 4.14.

When transferred to the "other" street – for validation– the resulting RMSE scores do not decrease when more parameters are included. Also, for each top-5 the combination of independent variables is different, making it difficult to create one correction model for one domain. Thus only a factor for PM would be enough to improve the RMSE, and therewith the accuracy, already significantly. So when a correction model is created for one domain, it is not necessary to include parameters for external environmental phenomena in the correction model.

Influence of the parameters

What are the calculated parameters of each best performing correction model per sensor? Table 5.2 shows the parameters of those best performing correction models and the resulting RMSE evaluation metric for the validation locations. The parameters of the model reveal why including more independent variables does not increase the data quality significantly: the parameters for those independent variables are I) relatively small or II) they cancel each other out.

Figures 5.3 and 5.4 show the influence of the separate parameters for the best four correction models. Instead of using an observation from the empirical dataset, "dummy" data is now used, with a value of 0.5 for each "dummy" observation. The value 0.5 is chosen because it is the mean of a normalized dataset with observations ranging from 0 to 1 when it has a normal distribution. The parameters for the four best performing correction models are included in the figures, as well as the intercept. Consequently, the new (theoretical) values of the dummy observations are calculated using those parameters. Then, the figures show the absolute and relative influence of each of the included variables in the model.

The last column in the figures 5.3 and 5.4 shows the relative influence of the included variables. First, it seems that wind speed has a negative effect on the corrected value. Second, when air pressure is included as independent variable, its influence seems to be relatively high. Also temperature seems to have a high influence on the "dummy" observations.

However, the influence of those independent variables should not be overestimated, since the relationships between the independent variables and PM are relatively weak, as indicated in Chapter 4 (figures 4.16 to 4.19). Moreover, of those independent variables, humidity has the highest correlation with PM, while at the same time it is not included in the best correction models or only for a small degree. Finally, figures 5.1 and 5.2 show clearly that including more variables does not decrease RMSE – i.e. improve the data quality – when the parameters are applied on the empirical data.

Therefore, based on the findings showed in figures 5.1 and 5.2 is concluded that for Type A and Type B models it is not necessary to include parameters for other independent variables in order to improve the data quality. Namely, if those factors are included, it only has a negligible effect on the improvement of the data quality.

5.2.1 Applying the best performing correction models on the datasets

Finally, the figures 5.5 and 5.6 show the time series plots for each of the four datasets, with parameters from the best assessed correction models that are depicted in table 5.2. The blue and purple lines in the figures show each time the validation of the best correction model on

Parameters calculated with data from Pleinweg sensor 1					
Best correction model for data from:	Included variables	Poly. degree	Formula	RMSE baseline	RMSE corrected
Pleinweg sensor 1	PM + H + T + WS + AP	3	0.1052412 + 0.39386163*PM - 0.10467195*PM^2 + 0.05906324*PM^3 + 0.02345433*H + 0.02007415*H^2 - 0.05694579*H^3 - 0.0269377*T + 0.10067003*T^2 - 0.0463558*T^3 - 0.10344781*WS + 0.18950852*WS^2 - 0.08786302*WS^3 + 0.26917627*AP - 0.55468182*AP^2 + 0.35617762*AP^3		
Pleinweg sensor 2	PM + H + T + WS + AP	3	0.1052412 + 0.39386163*PM - 0.10467195*PM^2 + 0.05906324*PM^3 + 0.02345433*H + 0.02007415*H^2 - 0.05694579*H^3 - 0.0269377*T + 0.10067003*T^2 - 0.0463558*T^3 - 0.10344781*WS + 0.18950852*WS^2 - 0.08786302*WS^3 + 0.26917627*AP - 0.55468182*AP^2 + 0.35617762*AP^3		
Zwartewaalstraat sensor 1	PM+T+WS+AP	3	**0.10521865 + 0.32571039*PM + 0.02614938*PM^2 - 0.02604104*PM^3 - 0.05075344*T + 0.23718998*T^2 - 0.14359434*T^3 - 0.07727786*WS + 0.12505315*WS^2 - 0.03975862*WS^3 + 0.19756856*AP - 0.3963041*AP^2 + 0.27752979*AP^3**	0.11920	0.03192
Zwartewaalstraat sensor 2	PM+T	2	**0.14395276 + 0.29300555*PM + 0.02165059*PM^2 + 0.02158459*T**	0.09180	0.03759
Parameters calculated with data from Pleinweg sensor 2					
Best correction model for data from:	Included variables	Poly. degree	Formula	RMSE baseline	RMSE corrected
Pleinweg sensor 2	PM + H + T + WS + AP	3	0.10096747 + 0.40961805*PM + 0.04190396*PM^2 - 0.04531164*PM^3 + 0.00554755*H + 0.05702176*H^2 - 0.07533136*H^3 - 0.00899824*T + 0.06051606*T^2 - 0.02020797*T^3 - 0.12458092*WS + 0.20477308*WS^2 - 0.08450959*WS^3 + 0.23936033*AP - 0.49638581*AP^2 + 0.32622517*AP^3		
Pleinweg sensor 1	PM + H + T + WS	3	0.16476838 + 0.37458374*PM + 0.16242578*PM^2 - 0.1651006*PM^3 + 0.03161421*H + 0.01537049*H^2 - 0.0590057*H^3 - 0.11033826*T + 0.19645722*T^2 - 0.08609555*T^3 - 0.15359734*WS + 0.284778*WS^2 - 0.14488627*WS^3		
Zwartewaalstraat sensor 1	PM+T+WS+AP	3	**0.10308161 + 0.33042186*PM + 0.21356825*PM^2 - 0.17282954*PM^3 - 0.04008598*T + 0.20258153*T^2 - 0.11972426*T^3 - 0.09919916*WS + 0.14195414*WS^2 - 0.0372197*WS^3 + 0.1804492*AP - 0.36541455*AP^2 + 0.2614511*AP^3**	0.11920	0.03237
Zwartewaalstraat sensor 2	PM+T+WS+AP	3	**0.10308161 + 0.33042186*PM + 0.21356825*PM^2 - 0.17282954*PM^3 - 0.04008598*T + 0.20258153*T^2 - 0.11972426*T^3 - 0.09919916*WS + 0.14195414*WS^2 - 0.0372197*WS^3 + 0.1804492*AP - 0.36541455*AP^2 + 0.2614511*AP^3**	0.09180	0.03552
Parameters calculated with data from Zwartewaalstraat sensor 1					
Best correction model for data from:	Included variables	Poly. degree	Formula	RMSE baseline	RMSE corrected
Zwartewaalstraat sensor 1	PM + H + T + WS + AP	3	0.03507838 + 0.20432782*PM + 0.37726713*PM^2 - 0.24901711*PM^3 + 0.15644497*H - 0.29806265*H^2 + 0.16205383*H^3 + 0.24621526*T - 0.2637472*T^2 + 0.11191071*T^3 - 0.04089487*WS - 0.02085173*WS^2 + 0.06011192*WS^3 + 0.32861135*AP - 0.75922945*AP^2 + 0.53588285*AP^3		
Zwartewaalstraat sensor 2	PM + H + T + WS + AP	3	0.03507838 + 0.20432782*PM + 0.37726713*PM^2 - 0.24901711*PM^3 + 0.15644497*H - 0.29806265*H^2 + 0.16205383*H^3 + 0.24621526*T - 0.2637472*T^2 + 0.11191071*T^3 - 0.04089487*WS - 0.02085173*WS^2 + 0.06011192*WS^3 + 0.32861135*AP - 0.75922945*AP^2 + 0.53588285*AP^3		
Pleinweg sensor 1	PM + H + T+ WS + AP	2	**0.15058079 + 0.30209301*PM + 0.06611508*PM^2 - 0.01837796*H + 0.04144995*T -0.01775492*WS + 0.01195488*AP**	0.12490	0.03125
Pleinweg sensor 2	PM + T + WS + AP	3	**0.05299849 + 0.19726337*PM + 0.3741821*PM^2 - 0.23821822*PM^3 + 0.24289341*T - 0.24101784*T^2 + 0.09724556*T^3 - 0.02775145*WS - 0.0451983*WS^2 + 0.07310478*WS^3 + 0.31729333*AP - 0.72695767*AP^2 + 0.51855344*AP^3**	0.09520	0.03116
Parameters calculated with data from Zwartewaalstraat sensor 2					
Best correction model for data from:	Included variables	Poly. degree	Formula	RMSE baseline	RMSE corrected
Zwartewaalstraat sensor 2	PM + H + T + WS + AP	3	0.01199191 + 0.14270135*PM + 0.70301122*PM^2 - 0.54949727*PM^3 + 0.2275337*H - 0.45204605*H^2 + 0.28251691*H^3 + 0.33109401*T - 0.34690918*T^2 + 0.13506901*T^3 - 0.08418381*WS + 0.02324964*WS^2 + 0.05705981*WS^3 + 0.28603361*AP - 0.69737815*AP^2 + 0.52268564*AP^3		
Zwartewaalstraat sensor 1	PM + H + T + WS + AP	3	0.01199191 + 0.14270135*PM + 0.70301122*PM^2 - 0.54949727*PM^3 + 0.2275337*H - 0.45204605*H^2 + 0.28251691*H^3 + 0.33109401*T - 0.34690918*T^2 + 0.13506901*T^3 - 0.08418381*WS + 0.02324964*WS^2 + 0.05705981*WS^3 + 0.28603361*AP - 0.69737815*AP^2 + 0.52268564*AP^3		
Pleinweg sensor 1	PM + WS	3	**0.16373851 + 0.18891569*PM + 0.60009925*PM^2 - 0.47379366*PM^3 -0.10048174*WS + 0.19601084*WS^2 - 0.11051505*WS^3**	0.12490	0.03400
Pleinweg sensor 2	PM + H + T+ WS + AP	2	**0.1295599 + 0.32312096*PM + 0.10592578*PM^2 + 0.0041597*H + 0.05082098*T - 0.02241293*WS + 0.01180015*AP**	0.09520	0.03110

Table 5.2: Best correction models per location. The **bold** formulas yield lowest RMSE values at the other sensor location, i.e. at the validation location.

	Parameters from Pleinweg sensor 1						
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Zwartewaalstraat sensor 1	*intercept*	0.10521865	N.A.	N.A.	N.A.	0.10521865	**34%**
	PM	0.32571039	0.5	0.5	0.162855195	0.16613741	**54%**
	PM**2	0.02614938	0.5	0.25	0.006537345		
	PM**3	-0.02604104	0.5	0.125	-0.00325513		
	T	-0.05075344	0.5	0.5	-0.02537672	0.015971483	**5%**
	T**2	0.23718998	0.5	0.25	0.059297495		
	T**3	-0.14359434	0.5	0.125	-0.01794929		
	WS	-0.07727786	0.5	0.5	-0.03863893	-0.01234547	**-4%**
	WS**2	0.12505315	0.5	0.25	0.031263288		
	WS**3	-0.03975862	0.5	0.125	-0.00496983		
	AP	0.19756856	0.5	0.5	0.09878428	0.034399479	**11%**
	AP**2	-0.3963041	0.5	0.25	-0.09907603		
	AP**3	0.27752979	0.5	0.125	0.034691224		
	TOTAL					**0.309381551**	
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Zwartewaalstraat sensor 2	*intercept*	0.14395276	N.A.	N.A.	N.A.	0.14395276	**48%**
	PM	0.29300555	0.5	0.5	0.146502775	0.151915423	**50%**
	PM**2	0.02165059	0.5	0.25	0.005412648		
	T	0.02158459	0.5	0.5	0.010792295	0.005396148	**2%**
	TOTAL					**0.30126433**	
	Parameters from Pleinweg sensor 2						
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Zwartewaalstraat sensor 1 and sensor 2	*intercept*	0.10308161	N.A.	N.A.	N.A.	0.10308161	**31%**
	PM	0.33042186	0.5	0.5	0.16521093	0.1969993	**60%**
	PM**2	0.21356825	0.5	0.25	0.053392063		
	PM**3	-0.17282954	0.5	0.125	-0.02160369		
	T	-0.04008598	0.5	0.5	-0.02004299	0.01563686	**5%**
	T**2	0.20258153	0.5	0.25	0.050645383		
	T**3	-0.11972426	0.5	0.125	-0.01496553		
	WS	-0.09919916	0.5	0.5	-0.04959958	-0.018763508	**-6%**
	WS**2	0.14195414	0.5	0.25	0.035488535		
	WS**3	-0.0372197	0.5	0.125	-0.00465246		
	AP	0.1804492	0.5	0.5	0.0902246	0.03155235	**10%**
	AP**2	-0.36541455	0.5	0.25	-0.09135364		
	AP**3	0.2614511	0.5	0.125	0.032681388		
	TOTAL					**0.328506613**	

Figure 5.3: Influence of the parameters on "dummy" observations, for best correction models from Pleinweg

Parameters from Zwartewaalstraat sensor 1							
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Pleinweg sensor 1	*intercept*	0.15058079	N.A.	N.A.	N.A.	0.15058079	**45%**
	PM	0.30209301	0.5	0.5	0.151046505	0.167575275	**50%**
	PM**2	0.06611508	0.5	0.25	0.01652877		
	H	-0.01837796	0.5	0.5	-0.00918898	-0.00918898	**-3%**
	T	0.04144995	0.5	0.5	0.020724975	0.020724975	**6%**
	WS	-0.01775492	0.5	0.5	-0.00887746	-0.00887746	**-3%**
	AP	0.01195488	0.5	0.5	0.00597744	0.00597744	**2%**
	TOTAL					**0.33598102**	
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Pleinweg sensor 2	*intercept*	0.05299849	N.A.	N.A.	N.A.	0.05299849	**18%**
	PM	0.19726337	0.5	0.5	0.098631685	0.162399933	**56%**
	PM**2	0.3741821	0.5	0.25	0.093545525		
	PM**3	-0.23821822	0.5	0.125	-0.02977728		
	T	0.24289341	0.5	0.5	0.121446705	0.07334794	**25%**
	T**2	-0.24101784	0.5	0.25	-0.06025446		
	T**3	0.09724556	0.5	0.125	0.012155695		
	WS	-0.02775145	0.5	0.5	-0.01387573	-0.016037203	**-6%**
	WS**2	-0.0451983	0.5	0.25	-0.01129958		
	WS**3	0.07310478	0.5	0.125	0.009138098		
	AP	0.31729333	0.5	0.5	0.158646665	0.041726428	**14%**
	AP**2	-0.72695767	0.5	0.25	-0.18173942		
	AP**3	0.51855344	0.5	0.125	0.06481918		
	TOTAL					**0.288746363**	
Parameters from Zwartewaalstraat sensor 2							
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Pleinweg sensor 1	*intercept*	0.16373851	N.A.	N.A.	N.A.	0.16373851	**49%**
	PM	0.18891569	0.5	0.5	0.094457845	0.18525845	**55%**
	PM**2	0.60009925	0.5	0.25	0.150024813		
	PM**3	-0.47379366	0.5	0.125	-0.05922421		
	WS	-0.10048174	0.5	0.5	-0.05024087	-0.015052541	**-5%**
	WS**2	0.19601084	0.5	0.25	0.04900271		
	WS**3	-0.11051505	0.5	0.125	-0.01381438		
	TOTAL					**0.333944419**	
Best correction model when applied on data from:	Variable and poly degree	Calculated parameter	Normalized dummy value	Dummy value times degree	New dummy value	Absolute influence of this variable	Relative influence of this variable
Pleinweg sensor 2	*intercept*	0.1295599	N.A.	N.A.	N.A.	0.1295599	**41%**
	PM	0.32312096	0.5	0.5	0.16156048	0.188041925	**59%**
	PM**2	0.10592578	0.5	0.25	0.026481445		
	H	0.0041597	0.5	0.5	0.00207985	0.00207985	**1%**
	T	0.05082098	0.5	0.5	0.02541049	0.02541049	**8%**
	WS	-0.02241293	0.5	0.5	-0.01120647	-0.011206465	**-4%**
	AP	0.01180015	0.5	0.5	0.005900075	0.005900075	**2%**
	TOTAL					**0.319681675**	

Figure 5.4: Influence of the parameters on ”dummy” observations, for best correction models from Zwartewaalstraat

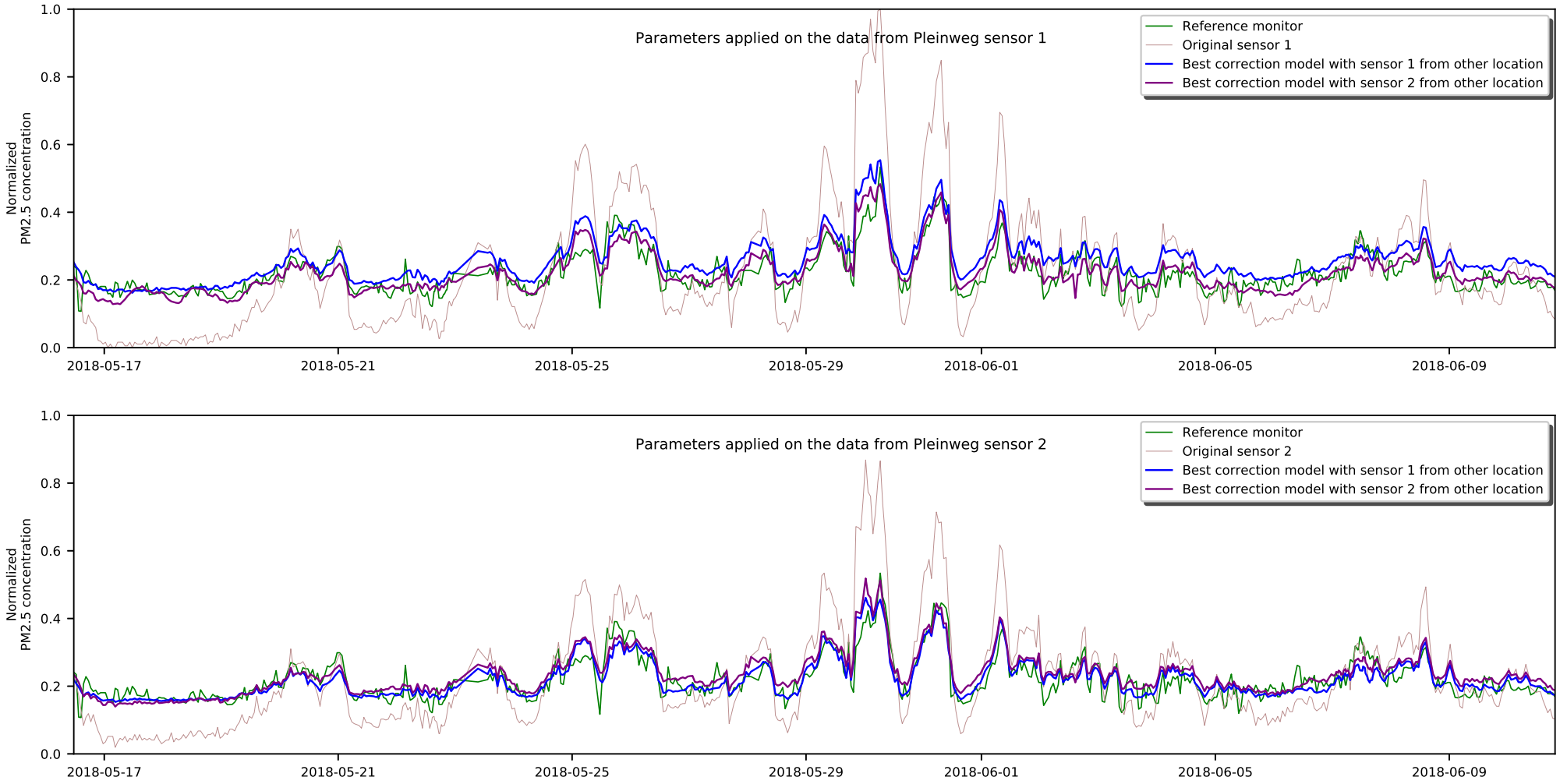


Figure 5.5: The best parameters from the Zwartewaalstraat dataset applied on the normalized data from the Pleinweg dataset.

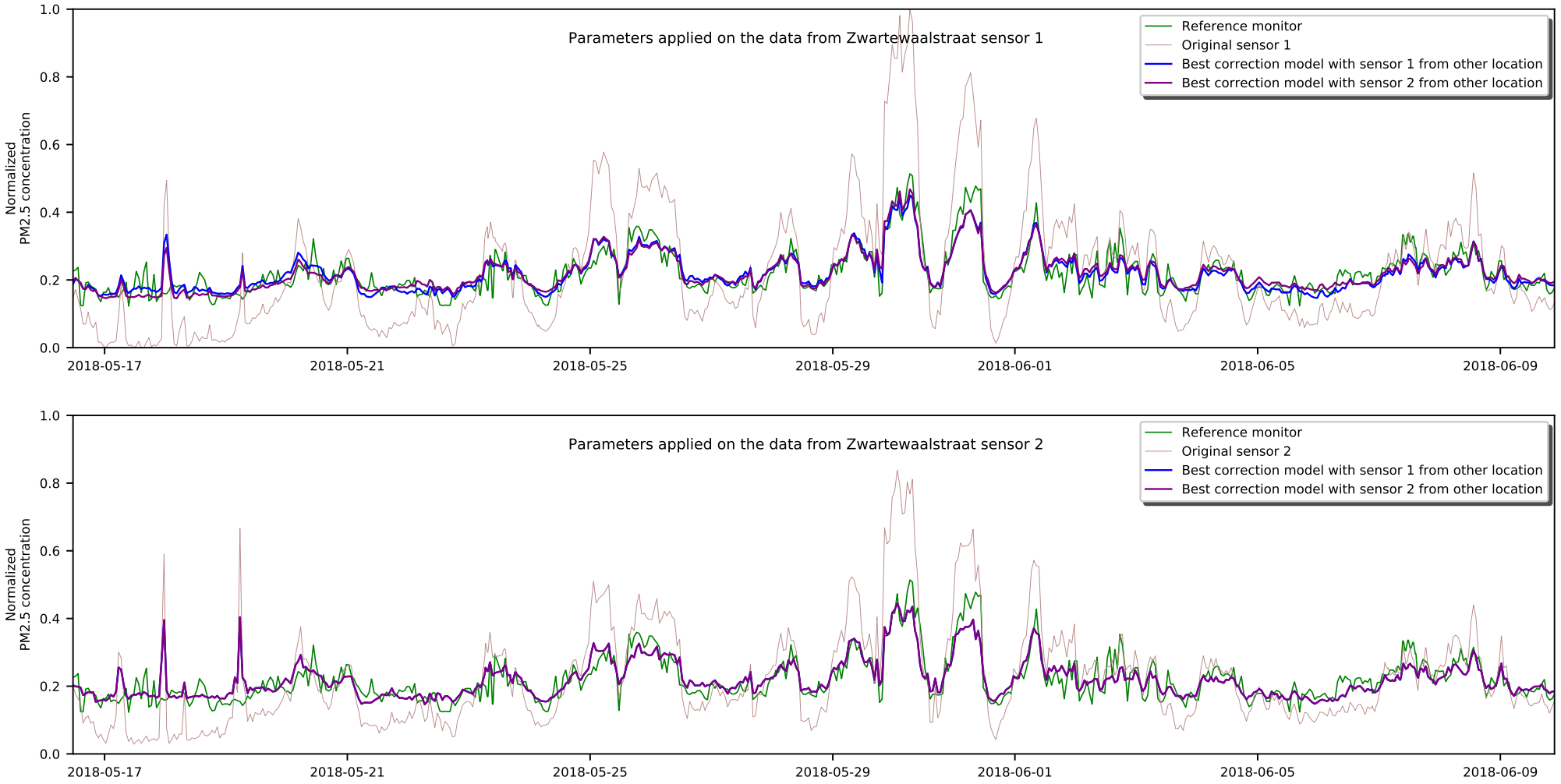


Figure 5.6: The best parameters from the Pleinweg dataset applied on the normalized data from the Zwartewaalstraat dataset.

that specific location. These time series plots reveal that the "best" correction models indeed perform relatively well: the data in the time series is corrected and it follows the same trend as the high quality reference data (green line).

5.3 CORRECTION MODELS OF TYPE C AND D

The suggested "Type C" correction models are correction models where only parameters for PM are used – no other environmental variables – but for various domains. "Type D" correction models on the other hand use parameters for a varying amount of environmental variables and take domains into account. In this research are the following domains analyzed:

- Wind direction
- Peak or off-peak

Wind direction

For the wind direction domain is the dataset filtered based on the wind direction group. Four groups are defined: North (n=389) East (n=74), South (n=44) and West (n=150). Thus, the amount of records per group is not equal. However, if N is large enough the parameters can be calculated.

For the North and East groups of Pleinweg the resulting RMSE scores are lower compared to the the scores when no wind direction groups are taken into account. The same is true for the North and East groups of Zwartewaalstraat. On the other hand, the parameters for the South and West groups applied to the sub-datasets yields worse RMSE scores. One reason could be the low n-value: using a larger dataset could improve correction capability of the model. Therefore, for the "wind direction" domain no conclusions can be drawn in this research.

Moment on the day

The "Moment on the day" domain consists of six groups. This amount of groups was chosen because they will have an equal size of around N=100, referring to the amount of observations for which RMSE can be used successfully as evaluation metric [Chai et al., 2014], as discussed in Chapter 4.

- 00:00 - 04:00
- 04:00 - 08:00
- 08:00 - 12:00
- 12:00 - 16:00
- 16:00 - 20:00
- 20:00 - 00:00

When for each of these groups and for each combination of variables the parameters are calculated and consequently applied on the sub-dataset, there are various results in RMSE values. For some time intervals the evaluation criterium value decreases, which would improve the dataset, while on the other hand for other time intervals the RMSE increase. Namely, for most of the time intervals on both locations the data quality for the dataset from the sensor wherefore the parameters are created improved. When transferred to the other location for validation, the RMSE results are sometimes higher and sometimes lower when compared to I) the baseline RMSE measurement and II) the RMSE of the type A and B correction models. Per time interval and per location are the five best performing correction models included in the following figure 5.7.

Subdividing the datasets in subgroups for the moment on the day – groups of 4 hours – yields interesting results. It is found that the data quality of four of these groups – 00:00 to 04:00, 12:00 to 16:00, 16:00 to 20:00 and 20:00 to 00:00 – improves for all four sensors. That is, it improves when compared to the results of the top-5 of the best correction models for that location. On the other hand, the data quality for the 08:00 to 12:00 group degenerated. The data quality of the 04:00 to 08:00 group alternates between improving and degenerating.

A reason for the better performance of the mentioned groups is that they are during off peak hours, while the 04:00 to 08:00 is during peak hour. The PM concentrations shows higher peaks during those moments, therefore the data is less suitable to fit a model to. Moreover,

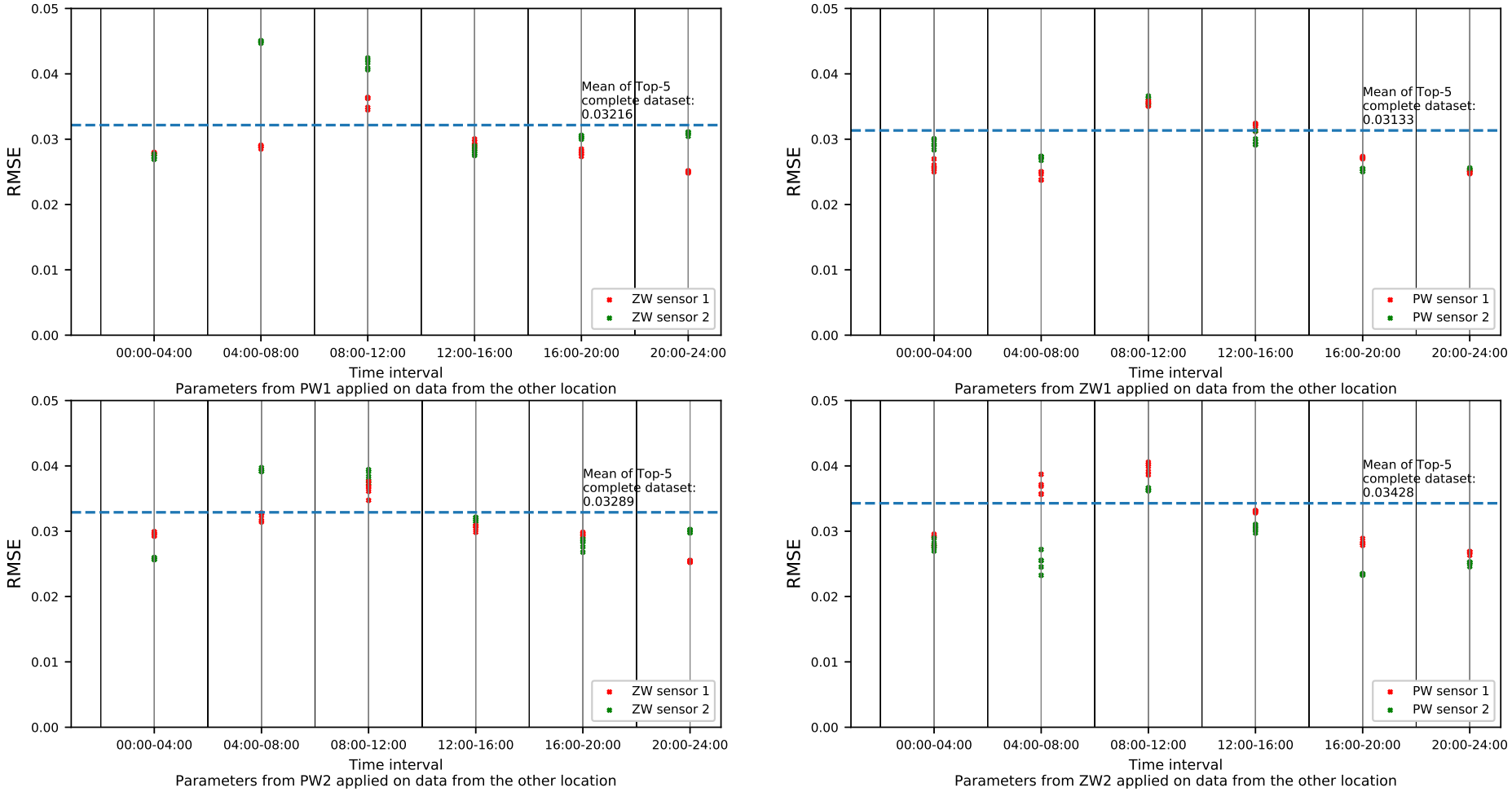


Figure 5.7: In these figures is the complete dataset subdivided in sets based on intervals of four hours. For each subset are the parameters calculated on one location and then applied on the data from the same interval though from the other location. Shown are the RMSE values of the five most improved PM sub-datasets per sensor location. This figure shows that a time-dependent subdivision improves the data quality (lower RMSE) except for some of the "morning" intervals (04:00-08:00 and 08:00-12:00).

04:00 to 08:00 and 08:00 to 12:00 groups are both during the morning, when the changes of environmental variables such as temperature, humidity and wind speed are less smooth, when compared to the changes during afternoon and evening. However, this theory should be validated with more experiments.

5.4 CONCLUSION RESULTS

In this chapter are the results of the correction models shown and discussed. The stepwise MLR method is successfully applied on the PM datasets. One finding is that for Type A and Type B models there is no need to include parameters for more independent variables in order to improve the data quality. Namely, when those factors are included there is a negligible effect on the improvement of the data quality.

Humidity, temperature or air pressure are no independent variables that improve the accuracy of the data from low-cost PM sensors, when the stepwise MLR method is applied as described in this research.

Finally, subdividing the datasets in subgroups for wind-direction yields no reliable correction models in this research, since the dataset contains not enough observations from the East and South directions.

6 CONCLUSION AND FUTURE WORK

This section is the conclusion of the research performed in this report. First, the research questions are answered, starting with the sub-questions and then the main research question. That is followed by a reflection on the research process and finally recommendations for future work.

6.1 RESEARCH QUESTIONS

How do temperature, humidity, air pressure, and wind speed affect Particulate Matter measurements?

The discussion in Chapter 4, paragraph 4.6, elaborated on this question. It is found that temperature does not affect PM from the low-cost sensors: there are no relationships found. Humidity, wind speed and air pressure have weak to moderate relationships with PM from the low-cost sensors. Postolache et al. [2009] and Cross et al. [2017] already indicated that low-cost PM sensors could be affected by cross interference with humidity. In this research the relationship between low-cost PM sensor data and humidity is investigated and range between R=0.48 and R=0.57. At the same time, the relationship between the data from the PM reference monitor and humidity ranges between R=0.32 and R=0.39. Thus, although the relationship is not strong, it is stronger than the relationship with the reference data and can therefore be included in the correction model.

The same is true for air pressure and wind speed. Air pressure has a weak negative correlation with PM from low-cost sensors, ranging from R=-0.32 to R=-0.34, which is higher than R=-0.22 to R=-0.26 for the reference monitors. For wind speed the correlations are also weak and negative: R=-0.34 to R=-0.38 versus R=-0.28 to R=-0.31.

Since for air pressure and wind speed the correlations with PM are also stronger for PM from the low-cost sensor nodes than for PM from the monitoring station, these independent variables are also included in the correction model as possibility.

What is a good experimental setup for calibrating air quality measurements and how to develop this sensor setup?

An air quality monitoring network, containing low-cost sensor nodes and a high-quality monitoring system on two different locations in Rotterdam, is used to acquire the data. Two locations – "Pleinweg" (PW) and "Zwartewaalstraat" (ZW) – for sensor nodes are chosen because then it is possible to create a correction model at one location and validate the correction model at the other location. So, the data from one location is used as training data to calculate the coefficients in the MLR correction model. Consequently, the correction model is applied to data from another location in order to assess the performance.

What can be acquired from existing correction models in the field of air quality monitoring and related fields of research?

The Multiple Linear Regression method is used because it is the most used linear model in the field of air quality monitoring [Pires and Martins, 2011; Ausati and Amanollahi, 2016]. Each correction model will contain a parameter for the original PM value. Parameters for more independent variables are also included. RMSE is a suitable metric to assess the performance of these correction models [Cross et al., 2017], as long as the error follows a Gaussian distribution, there is a sample size of n=100 or more, and the error is unbiased, i.e. there is no systematic error [Chai et al., 2014].

In the field of correction models for air quality datasets there are no papers found that utilize autoregressive models.

"How to create the new correction model?" and "How to validate the correction model?"

First, the relationships among the independent variables are investigated. The relations among the independent variables – multicollinearity – should be low, i.e. a VIF < 5 and low correlation coefficients [Ausati and Amanollahi, 2016]. Then, the experiments are conducted as follows:

- Preprocessing (outlier removal, systematic error removal, normalization).
- Baseline measurements of the RMSE metric.
- Create correction models with the (Stepwise) MLR method.
- Calculate RMSE for the corrected datasets at the location for which the correction model is created.
- Calculate RMSE for datasets from the other location, i.e. validation.
- Select the correction model with the best performance, i.e. the lowest RMSE at the other location.

An algorithm performs all these steps for various input datasets. A schematic overview of this algorithm is shown in figure 3.4 and written in pseudocode in algorithm 3.1. The GitHub repository contains the Python implementation of this algorithm.

Main research question

The main research question was defined as follows:

How can accuracy and precision of Particulate Matter measurement results from a low-cost outdoor sensor network be improved by using a correction model, using data from reference sensors and additional sensors measuring inferencing phenomena?

Without applying a correction model the normalized RMSE ranged from 0.0918 to 0.1249, after removing the systematic instrument error (see table 4.5). The correlations between various candidate independent variables and the dependent variable PM2.5 are investigated and it is found that humidity correlates strongest with PM from the low-cost sensors. The method used in this research to create the correction model is the stepwise MLR method, which uses LSA to find the values for the intercept and parameters in the model. Since the correlations between the independent variables and PM are all moderate to weak, it is unlikely that the best correction model would include those variables.

Indeed, for the proposed Type A and Type B models, that take only one domain into consideration, it is found that the "best" correction models are those that include only the original PM data and the effect of adding more independent variables is limited. Figures 5.1 and 5.2 show the results of the correction models when applied on the empirical data. All correction models are able to decrease the RMSE of the observations: the original normalized values ranged from 0.0918 to 0.1249, while the corrected normalized values range from 0.03110 to 0.03759, see table 5.2. So, it is possible to improve the data quality of low-cost PM sensors with the stepwise MLR method and setup as shown in this research. However, including parameters for independent variables humidity, temperature, air pressure or wind speed does not improve the data quality significantly.

For the proposed Type C and Type D correction models, the "moment on the day" subdivision – into groups of 4 hours – it is found that the data quality for four groups improved (00:00 to 04:00, 12:00 to 16:00, 16:00 to 20:00 and 20:00 to 00:00). For one group (08:00 to 12:00) the data quality degenerated, and for one group (04:00 to 08:00) the effect of the correction model for the sub-group alternates between improvement and degeneration. These findings are included in figure 5.7 and are valid for all four sensors. An explanation for those differences could be that the groups that increase in performance are during off peak hours, while the 04:00 to 08:00 is during peak hour. Next to that, the 04:00 to 08:00 and 08:00 to 12:00 groups are both during the morning, when temperature, humidity, air pressure and wind speed are less smooth when compared to the afternoon and evening time slots. That results in more varying data in these time series, with more peaks, thus fitting a linear model as with the used MLR method is less suitable. However, this theory should be validated with more experiments.

Finally, like Castell et al. [2017] and Mukherjee et al. [2017] also mentioned, it is necessary to calibrate each individual low-cost sensor before adding it to an quality measuring network

of the type as described in this research, and to quantify the performance of each sensor. Namely, each sensor will have different parameters in the correction model and in some cases the independent variable wind speed is included in the correction model.

6.2 REFLECTION

The objective of this research was to create a correction model for air quality measurements using a low-cost sensor network. This objective consisted of two parts: assessing the data quality and improving the data quality. An error correction model is used to improve the data quality. Although the proposed methodology yielded satisfying results – it is validated that the correction model improved the data quality – there are some remarks regarding the research process.

For a long time during the research process it was not clear that there were actually two research objectives: assessing the data quality and improving the data quality. If this division of research objective was clear from the beginning onwards, conducting the research would have been easier. For example, the evaluation metric (RMSE) was chosen relatively lately in the research process, while it should be clear that the aim of the methodology is to improve the data quality expressed with that metric.

On the other hand, the choice to use a correction model as the means to improve the data quality was made relatively early in the research process. Alternatives were not considered at that moment and in the final report only described in the introduction section – referring to the work of Batini et al. [2009]. At the beginning of the research process most time was spent on creating the sensor nodes: not on justifying the choices regarding the methodology and the describing the steps of the methodology. Especially the assembly of the sensor nodes took a significant amount of time in the early research process. The major reason why that took a relatively large amount of time was that wireless communication protocols – LoRa, WiFi, and MQTT – to transmit the data were considered but did not yield satisfying results. In the end, it was chosen to store the data locally – on microcontroller of the sensor node – which was satisfactory. However, using a wireless communication protocol to transmit the data is favorable when a project like this is scaled up, and therefore recommended for future work. Moreover, real-time sensor data can then be used, allowing to correct the sensor data real-time.

Finally, instead of first focusing on creating sensor nodes, spending a month on data collection, and then creating the methodology, this process should have been turned around. Thus, first creating the **complete** methodology as a proof of concept, possibly using random generated "dummy" data, then creating the sensor nodes and data collection. That way, the results of the raw data can not influence how the researcher creates the methodology. Namely, I can imagine – and experienced – that results of the raw data fascinates, overwhelms or harms the objectivity of the researcher in another way.

6.3 FUTURE WORK

Besides redoing the research but with a larger dataset, there are various suggestions for taking this research to the next step.

One suggestion is to use real-time sensor data, where the parameters can be applied to the new data immediately and whereby the parameters can be updated at regular intervals. Although the parameters from this research could be applied to new data, it does not have real-time calibration. How can the proposed methodology for correcting low-cost PM data be integrated in an infrastructure that supplies real-time air quality data?

Another recommendation for future work would be to include the data from the reference stations as extra input variables in the correction model. Then the reference data will have an extra "task" besides functioning as "ground truth" for creating the correction models. So while in this research – in some cases – data for the input variables humidity, temperature, wind speed and air pressure is included, this could be extended with also using the high-quality reference data as input variable. In that case, the low-cost sensor data would contribute to interpolation of the high-quality monitoring network. Besides, the distance of a low-cost sensor node to the set of high-quality monitoring stations in the vicinity can be expressed in a weight parameter.

BIBLIOGRAPHY

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., and Hamburg, S. P. (2017). High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environmental Science and Technology*, 51:6999–7008.

Ausati, S. and Amanollahi, J. (2016). Assessing the accuracy of ANFIS , EEMD-GRNN, PCR, and MLR models in predicting PM2.5. 142:465–474.

Barrosso, J. M. (2014). Commission Delegated Regulation (EU) No 522/2014.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41(3):52.

Bentayeb, M., Wagner, V., Stempfelet, M., Zins, M., Goldberg, M., Pascal, M., Larrieu, S., Beaudeau, P., Cassadou, S., Eilstein, D., Filleul, L., Le Tertre, A., Medina, S., Pascal, L., Prouvost, H., Quénel, P., Zeghnoun, A., and Lefranc, A. (2015). Association between long-term exposure to air pollution and mortality in France: A 25-year follow-up study. *Environment International*, 85:5–14.

Berninger, A., Lohberger, S., Stängel, M., and Siegert, F. (2018). SAR-Based Estimation of Above-Ground Biomass and Its Changes in Tropical Forests of Kalimantan. *Remote Sensing*, 10:1–22.

Borrego, C., Costa, A. M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., Sioumis, T., Katsifarakis, N., Konstantinidis, K., Vito, S. D., Esposito, E., Smith, P., G, P., Andr, N., Reimringer, W., Otjes, R. P., Sicard, O. V., Pohle, R., and Elen, B. (2016). Assessment of air quality microsensors versus reference methods : The EuNetAir joint exercise. *Atmospheric Environment*, 147(2):246–263.

Carminati, M., Sampietro, M., and Carminati, G. (2011). Analysis of instrumentation performance for distributed real-time air quality monitoring. *2011 IEEE Workshop on Environmental Energy and Structural Monitoring Systems*, pages 1–6.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., and Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99:293–302.

Chai, T., Draxler, R. R., and Prediction, C. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7:1247–1250.

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T. (2017). Use of electrochemical sensors for measurement of air pollution: Correcting interference response and validating measurements. *Atmospheric Measurement Techniques*, 10(9):3575–3588.

De Vries, J., Voors, R., Ording, B., Dingjan, J., Veefkind, P., Ludewig, A., Kleipool, Q., Hoogeveen, R., and Aben, I. (2016). TROPOMI on ESA's Sentinel 5p ready for launch and use. In *Fourth International Conference on Remote Sensing and Geoinformation of the Environment*, number 9688, page 13. Proceedings of SPIE.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 74(366):427.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction : representation , estimation , and testing. *Econometrica*, 55(2):251–276.

Environmental Protection Agency (2016). Particulate Matter (PM) Basics.

EU (2008). Guideline 2008/50/EG.

Fisher, P. F. and Tate, N. J. (2006). Causes and consequences of error in digital elevation models. *Progress in Physical Geography*, 30(4):467–489.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons, Inc., Hoboken, New Jersey, 4 edition.

Gardner, W. A., Napolitano, A., and Paura, L. (2006). Cyclostationarity: Half a century of research. *Signal Processing*, 86(4):639–697.

Hendriks, C., Kranenburg, R., Kuenen, J., Van Gijlswijk, R., Van Denier Der Gon, H., and Schaap, M. (2012). Establishing the Origin of Particulate Matter Concentrations in the Netherlands. Technical report, TNO, Utrecht.

Jazaeri, S. and Amiri-Simkooei, A. R. (2013). Data-snooping procedure applied to errors-in-variables models. *Studia Geophysica et Geodaetica*, 57(July 2013):426–441.

Juliana, W., Nederlanden, K. D., and Oranje-nassau, P. V. (2009). Wet milieubeheer.

Kennedy, C., Pincetl, S., and Bunje, P. (2011). The study of urban metabolism and its applications to urban planning and design. *Environmental Pollution*, 159(8-9):1965–1973.

Kilian, L. and Lütkepohl (2017). Chapter 3: Vector Error Correction Models. In *Themes in Modern Econometrics*, chapter 3, pages 75–108. Cambridge University Press, Cambridge.

Kumar, A., Kumar, A., and Singh, A. (2017). Energy Efficient and Low Cost Air Quality Sensor for Smart Buildings. *Computational Intelligence and Communication Technology*, pages 17–20.

Lemmens, M. (2016a). Electromagnetic Spectrum (lecture notes on Sensing Technologies, available upon request).

Lemmens, M. (2016b). Least Squares Adjustment Linear Relationship (lecture notes on Sensing Technologies, available upon request).

Lemmens, M. (2016c). Least Squares Adjustment Non-Linear Relationship (lecture notes on Sensing Technologies, available upon request).

Lemmens, M. (2017). Quality of Geo-information 1 (lecture notes on Geo Data Quality, available upon request).

Lewis, A. and Edwards, P. (2016). Validate Personal Air Pollution Sensors. *Nature*, 535:29–31.

Li, J. and Biswas, P. (2017). Optical characterization studies of a low-cost particle sensor. *Aerosol and Air Quality Research*, 17(7):1691–1704.

Liang, S., Huang, C., and Khalafbeigi, T. (2016). OGC SensorThings API Part 1: Sensing.

Liao, Z. and Phillips, P. C. (2014). Automated estimation of vector error correction models. *Econometric Theory*, 31(3):581–646.

Marshall, A. (2017). Alphabet Is Trying To Reinvent the City , Starting With Toronto. *Wired Magazine*.

Mead, M. I., Popoola, O. A., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J. J., McLeod, M. W., Hodgson, T. F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J. R., and Jones, R. L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186–203.

Met One Instruments (2010). Bam 1020 Particulate Monitor With BX-970 Touch Screen Display Operation Manual - BAM-1020-9803 REVISION K. Technical Report Revision K, Met One Instruments Inc., Grant Pass, OR, USA.

Met One Instruments (2016). BAM1020 Particulate Monitor Operation Manual BAM1020-9800 Rev W. Technical report, Met One Instruments Inc., Grant Pass, OR, USA.

Ministerie I&M (2015). Smart Cities; Naar een 'smart urban delta'.

Monks, P. S., Granier, C., Fuzzi, S., Stohl, A., Williams, M. L., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R. S., Carslaw, K., Cooper, O. R., Dentener, F., Fowler, D., Fragkou, E., Frost, G. J., Generoso, S., Ginoux, P., Grewe, V., Guenther, A., Hansson, H. C., Henne, S., Hjorth, J., Hofzumahaus, A., Huntrieser, H., Isaksen, I. S., Jenkin, M. E., Kaiser, J., Kanakidou, M., Klimont, Z., Kulmala, M., Laj, P., Lawrence, M. G., Lee, J. D., Liousse, C., Maione, M., McFiggans, G., Metzger, A., Mieville, A., Moussiopoulos, N., Orlando, J. J., O'Dowd, C. D., Palmer, P. I., Parrish, D. D., Petzold, A., Platt, U., Pöschl, U., Prévôt, A. S., Reeves, C. E., Reimann, S., Rudich, Y., Sellegri, K., Steinbrecher, R., Simpson, D., ten Brink, H., Theloke, J., van der Werf, G. R., Vautard, R., Vestreng, V., Vlachokostas, C., and von Glasow, R. (2009). Atmospheric composition change - global and regional air quality. *Atmospheric Environment*, 43(33):5268–5350.

Mukherjee, A., Stanton, L. G., Graham, A. R., and Roberts, P. T. (2017). Assessing the utility of low-cost particulate matter sensors over a 12-week period in the Cuyama valley of California. *Sensors (Switzerland)*, 17(8):1–16.

Naafs, S. (2017). De Groene Amsterdammer - De Muren Hebben Sensoren.pdf.

Nelson, C. R. and Plosser, C. I. (1982). Trends and random walks in macroeconmic time series: Some evidence and implications. *Journal of Monetary Economics*, 10(2):139–162.

NEN (2014). Nen EN 12341. Technical report, Nederlands Normalisatie-instituut (NEN), Delft.

Pianosi, F., Castelletti, A., Mancusi, L., and Garofalo, E. (2014). Improving fl ow forecasting by error correction modelling in altered catchment conditions. 2534(April 2013):2524–2534.

Pijpers-van Esch, M. (2015). *Designing the Urban Microclimate*. Architecture and the Built Environment, Delft.

Pires, J. C. M. and Martins, F. G. (2011). Correction methods for statistical models in tropospheric ozone forecasting. *Atmospheric Environment*, 45(14):2413–2417.

Pope, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air and Waste Management Association*, 56(6):709–742.

Postolache, O., Pereira, J., and Girao, P. (2009). Smart Sensors Network for Air Quality Monitoring Applications. *Ieee Transactions On Instrumentation And Measurement*, 58(9):3253–3262.

Salim, F. D. (2012). Probing Streets and the Built Environment with Ambient and Community Sensing. *Journal of Urban Technology*, 19(2):47–67.

Tseng, C. H., Lu, L. C., Lan, S. H., Hsieh, Y. P., and Lan, S. J. (2017). Relationship between emergency care utilization, ambient temperature, and the pollution standard index in Taiwan. *International Journal of Environmental Health Research*, 27(5):344–354.

US EPA (2016). *National Exposure Research Laboratory Exposure Methods and Measurement Division (MD-D205-03)*. United States Environmental Protection Agency.

Van Alphen, A. and Pot, J. (2014). Monitoringsysteem luchtkwaliteit in perspectief - achtergrondrapport. Technical report, Rijksinstituut voor Volksgezondheid en Milieu (RIVM).

Van Breugel, P. and Van den Elshout, S. (2018). Lucht in cijfers 2017. Technical report.

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J., and Biswas, P. (2015). Laboratory Evaluation and Calibration of Three Low-Cost Particle Sensors for Particulate Matter Measurement. *Aerosol Science and Technology*, 49(11):1063–1077.

WHO (2006). WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. Technical report, World Health Organization (WHO), Geneva, Switzerland.

Xiong, L. and Connor, K. M. O. (2002). Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal*, 47(4):621–639.

Yu, X., Shi, Y., Wang, T., and Sun, X. (2017). Dust-concentration measurement based on Mie scattering of a laser beam. *PLoS ONE*, 12(8):1–15.

Zivot, E. and Wang, J. (2006). Vector Autoregressive Models for Multivariate Time Series. In *Modeling Financial Time Series with S-PLUS®*, chapter 11, pages 383–427. Springer, New York, NY.

A RESULTS OF VARIOUS CORRECTION MODELS

COMPLETE DATASETS										
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**						
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 without other env. variables: 3 degrees	0.03407	0.07494487	0.28160906	0.19865123	-0.21872899			
	Zwartewaalstraat 2	PWSensor1 without other env. variables: 3 degrees	0.03793	0.07494487	0.28160906	0.19865123	-0.21872899			
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 without other env. variables: 3 degrees	0.03422	0.0686642	0.30546904	0.43030244	-0.5000271			
	Zwartewaalstraat 2	PWSensor2 with parameters for windspeed: 3 degrees	0.03628	0.08380234	0.31136819	0.39465268	-0.46133743	-0.13097859	0.28223335	-0.15711535
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for windspeed: 1 degrees	0.03363	0.05779878	0.3347218	0.00732436				
	Pleinweg 2	ZWSensor1 with parameters for windspeed: 1 degrees	0.03363	0.05779878	0.3347218	0.00732436				
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 without other env. variables: 3 degrees	0.03492	0.05480273	0.31993428	0.46399272	-0.55839156			
	Pleinweg 2	ZWSensor2 with parameters for windspeed: 3 degrees	0.03264	0.06839556	0.31482591	0.45017594	-0.53004038	-0.09500081	0.18531912	-0.10448684

Figure A.1: Results of various correction models.

Datasets subdivided in time intervals of 4 hours.													
SUBDATASET 00:00 - 04:00													
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**									
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*			
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 with parameters for temperature+windspeed: 3 degrees	0.02865	0.13323457	0.14016075	0.72964621	-0.75890357	-0.32313478	0.77366054	-0.46994346	-0.21988657	0.27387571	0.02334233
	Zwartewaalstraat 2	PWSensor1 with parameters for temperature+windspeed: 3 degrees	0.02708	0.13323457	0.14016075	0.72964621	-0.75890357	-0.32313478	0.77366054	-0.46994346	-0.21988657	0.27387571	0.02334233
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 with parameters for temperature+windspeed: 2 degrees	0.03034	0.06662995	0.29431684	0.08600259	0.01930581	-0.03052773					
	Zwartewaalstraat 2	PWSensor2 with parameters for windspeed: 3 degrees	0.02667	0.11125885	0.18522188	1.04121328	-1.28023495	-0.33292792	0.65347584	-0.3341339 8			
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for windspeed: 3 degrees	0.02843	0.06851163	0.2209984	0.49332285	-0.5195832	0.00925264	-0.24075177	0.32902276			
	Pleinweg 2	ZWSensor1 with parameters for windspeed: 3 degrees	0.03167	0.06851163	0.2209984	0.49332285	-0.5195832	0.00925264	-0.24075177	0.32902276			
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.0284	0.09782665	0.205247	0.83876022	-0.96802739	0.04010303	-0.44078534	0.54879438	-0.21444961	0.47308686	-0.30961098
	Pleinweg 2	ZWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.02998	0.09782665	0.205247	0.83876022	-0.96802739	0.04010303	-0.44078534	0.54879438	-0.21444961	0.47308686	-0.30961098

SUBDATASET 04:00 - 08:00																	
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**													
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*	*C11*	*C12*	
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 with parameters for humidity+temperature: 3 degrees	0.02973	0.55822835	0.266829	0.07558872	0.12777762	-2.13431766	2.34164221	-0.81492402	1.53544042	-4.74868984	4.40766989				
	Zwartewaalstraat 2	PWSensor1 with parameters for humidity+temperature+windspeed+airpressure: 2 degrees	0.04225	0.00221451	0.17330989	0.24387303	0.04911441	0.14940657	-0.03705963	0.04258258							
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 without other env. variables: 1 degree	0.03048	0.07456945	0.39211557												
	Zwartewaalstraat 2	PWSensor2 with parameters for humidity+temperature: 3 degrees	0.03778	1.24584978	0.25781478	0.0200435	0.42293638	-4.21442332	5.00851767	-1.92716269	0.3625388	-2.06600886	2.4058894				
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for humidity+temperature+windspeed: 3 degrees	0.02547	2.12633291	0.18182029	0.29854641	-0.0635411	-6.49534322	8.02849789	-3.20070512	-1.87810555	2.69335291	-0.82943396	-0.17633075	0.54074033	-0.54852914	
	Pleinweg 2	ZWSensor1 with parameters for windspeed: 3 degrees	0.02755	0.08008285	0.26706055	0.2868738	-0.12549994	-0.14446495	0.31038976	-0.1795161							
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for humidity+temperature: 3 degrees	0.03393	3.13623935	0.11685641	0.621756	-0.27715744	-9.26446494	11.38533745	-4.48384666	-4.0765635	7.44298089	-4.04257112				
	Pleinweg 2	ZWSensor2 with parameters for windspeed: 3 degrees	0.0258	0.09435256	0.26947464	0.32462035	0.08222305	-0.25140826	0.46255236	-0.25278313							

SUBDATASET 08:00 - 12:00																					
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**																	
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*								
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 with parameters for windspeed+airpressure: 3 degrees	0.03555	0.06511184	0.26398702	0.62406238	-0.774629	-0.3267886	0.74033674	-0.5412499	0.42328277	-0.9433119	0.59410448								
	Zwartewaalstraat 2	PWSensor1 with parameters for windspeed+airpressure: 3 degrees	0.03921	0.06511184	0.26398702	0.62406238	-0.774629	-0.3267886	0.74033674	-0.5412499	0.42328277	-0.9433119	0.59410448								
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.03612	0.06095038	0.30987315	0.83179379	-1.092426	-0.2691779	0.62423805	-0.4726165	0.34956541	-0.8132512	0.53024907								
	Zwartewaalstraat 2	PWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.03731	0.06095038	0.30987315	0.83179379	-1.092426	-0.2691779	0.62423805	-0.4726165	0.34956541	-0.8132512	0.53024907								
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for windspeed+airpressure: 3 degrees	0.03572	-0.0220553	0.28089388	0.57908389	-0.6862582	-0.1351598	0.15809261	-0.0932623	0.87549872	-1.8477604	1.11615538								
	Pleinweg 2	ZWSensor1 with parameters for airpressure: 3 degrees	0.03657	-0.0451037	0.3508789	0.29230092	-0.3373615	0.85021678	-1.8478328	1.14024286											
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for windspeed: 3 degrees	0.03764	0.07432978	0.22539116	1.68470191	-2.5096072	0.04756884	-0.4395996	0.41724904											
	Pleinweg 2	ZWSensor2 with parameters for humidity+temperature: 3 degrees	0.03671	-0.0795319	0.28021137	1.56232592	-2.486553	1.49142579	-2.1799378	1.02786345	-1.2507744	2.39032832	-1.3982863								

SUBDATASET 12:00 - 16:00																					
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**																	
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*											
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 with parameters for airpressure: 3 degrees	0.03129	-0.0026886	0.29661539	1.10741534	-2.4863301	0.4146369	-0.7768882	0.44437296											
	Zwartewaalstraat 2	PWSensor1 with parameters for airpressure: 3 degrees	0.03254	-0.0026886	0.29661539	1.10741534	-2.4863301	0.4146369	-0.7768882	0.44437296											
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 with parameters for airpressure: 3 degrees	0.03171	-0.0143023	0.28947454	1.92202159	-4.5517082	0.4688267	-0.9018833	0.52389049											
	Zwartewaalstraat 2	PWSensor2 with parameters for airpressure: 3 degrees	0.03228	-0.0143023	0.28947454	1.92202159	-4.5517082	0.4688267	-0.9018833	0.52389049											
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for temperature+windspeed: 1 degrees	0.03442	0.06669132	0.33736831	-0.0073104	-0.0091001														
	Pleinweg 2	ZWSensor1 with parameters for airpressure: 3 degrees	0.03357	0.0047601	0.23548844	1.79825248	-4.0988126	0.3043017	-0.6461427	0.42009788											
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for temperature+windspeed: 1 degrees	0.03399	0.06937572	0.4002339	-0.0143989	-0.0163458														
	Pleinweg 2	ZWSensor2 with parameters for temperature+windspeed: 1 degrees	0.03255	0.06937572	0.4002339	-0.0143989	-0.0163458														

SUBDATASET 16:00 - 20:00																					
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**																	
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*								
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 without other env. variables: 3 degrees	0.03157	0.06263398	0.47239855	-1.0729192	2.26095892														
	Zwartewaalstraat 2	PWSensor1 with parameters for airpressure: 2 degrees	0.03197	0.04901145	0.3775414	0.00702714	0.02776365														
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 with parameters for airpressure: 2 degrees	0.03277	0.04539178	0.38502658	0.20133454	0.02543471														
	Zwartewaalstraat 2	PWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.03034	0.04086005	0.69507348	-2.0761577	4.22107328	-0.1043304	0.24962155	-0.1499121	0.08981081	-0.1630869	0.11726308								
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for airpressure: 3 degrees	0.03169	0.01091465	0.38392554	-0.8692184	2.55044397	0.26565815	-0.54509	0.34680351											
	Pleinweg 2	ZWSensor1 with parameters for windspeed+airpressure: 3 degrees	0.03043	0.03511337	0.38445672	-0.8954071	2.65355319	-0.1878624	0.37342016	-0.2166623	0.28346754	-0.5710785	0.35639429								
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for airpressure: 1 degrees	0.03094	0.02908251	0.42278848	0.03060878															
	Pleinweg 2	ZWSensor2 with parameters for windspeed+airpressure: 3 degrees	0.02732	0.02307702	0.56392559	-1.569126	4.36058723	-0.2424587	0.45478404	-0.2464526	0.37184874	-0.7606853	0.48269878								

SUBDATASET 20:00 - 00:00																					
Parameters from:	**Correction model applied on data from:**	**Name of "best" correction model:**	**RMSE value**	**Parameters**																	
				Constant	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*											
Pleinweg 1	Zwartewaalstraat 1	PWSensor1 with parameters for windspeed: 2 degrees	0.02814	0.07558529	0.344702	-0.1350055	-0.0118469														
	Zwartewaalstraat 2	PWSensor1 with parameters for windspeed: 3 degrees	0.0326	0.09711325	0.23586652	0.60763429	-0.8820257	-0.1717165	0.2947247	-0.1352925											
Pleinweg 2	Zwartewaalstraat 1	PWSensor2 without other env. variables: 2 degrees	0.02774	0.06599411	0.39595465	-0.1420089															
	Zwartewaalstraat 2	PWSensor2 with parameters for windspeed: 2 degrees	0.03137	0.07414778	0.37922827	-0.123054	-0.0230782														
Zwartewaalstraat 1	Pleinweg 1	ZWSensor1 with parameters for windspeed: 3 degrees	0.02854	0.0726457	0.27336267	0.17336363	-0.2495035	-0.1169263	0.2598103	-0.2025586											
	Pleinweg 2	ZWSensor1 with parameters for windspeed: 3 degrees	0.02887	0.0726457	0.27336267	0.17336363	-0.2495035	-0.1169263	0.2598103	-0.2025586											
Zwartewaalstraat 2	Pleinweg 1	ZWSensor2 with parameters for windspeed: 3 degrees	0.02875	0.0800388	0.36518239	-0.1036497	-0.023042	-0.1553182	0.22466153	-0.0884956											
	Pleinweg 2	ZWSensor2 with parameters for windspeed: 3 degrees	0.02764	0.0800388	0.36518239	-0.1036497	-0.023042	-0.1553182	0.22466153	-0.0884956											

Figure A.2: Results of various correction models (continued).

B NODE-RED SETTINGS FOR LUCHTMEETNET DATA

B.1 OVERVIEW OF THE LUCHTMEETNET NODES

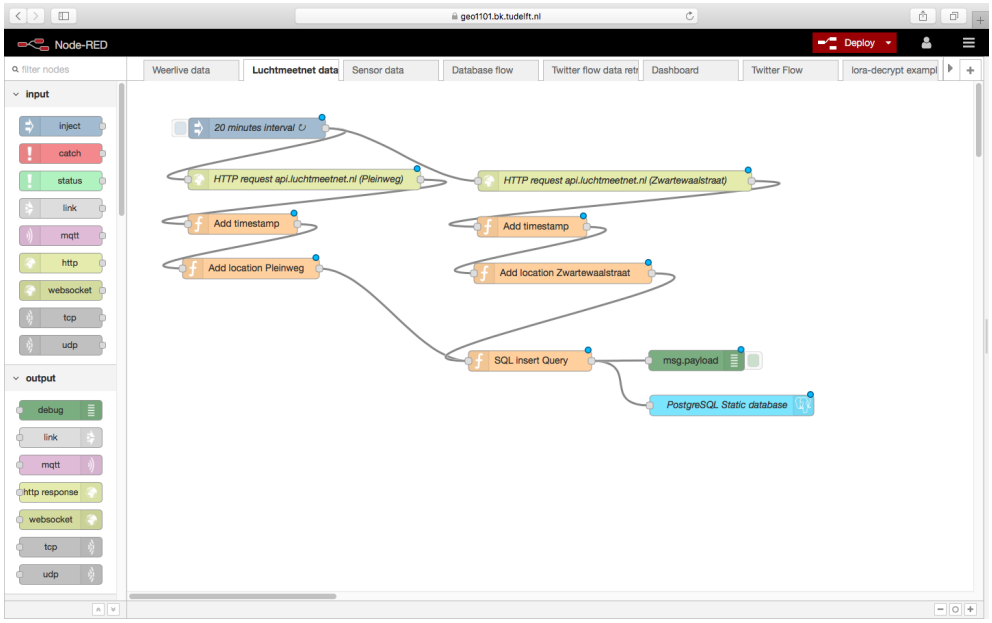


Figure B.1: Overview of the Node-RED nodes for the Luchtmeetnet API

B.2 SEPARATE LUCHTMEETNET NODES

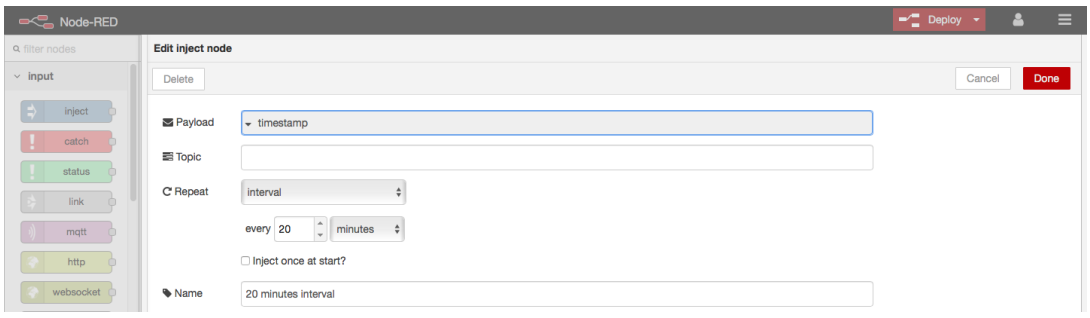


Figure B.2: Settings for the "inject" node **20 minutes interval**

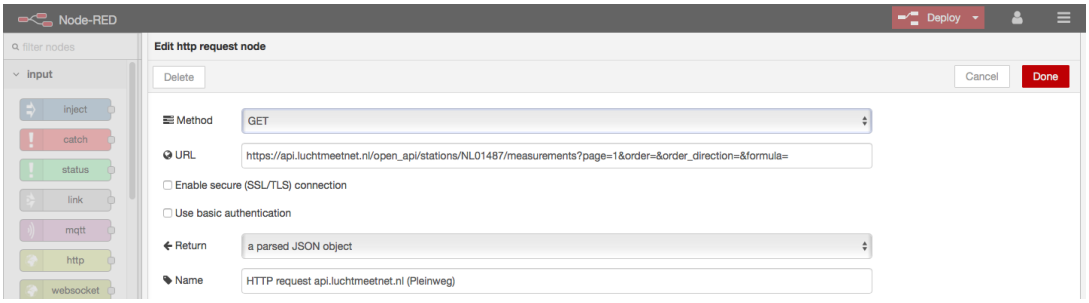


Figure B.3: Settings for the "http" node **HTTP request api.luchtmeetnet.nl (Pleinweg)**

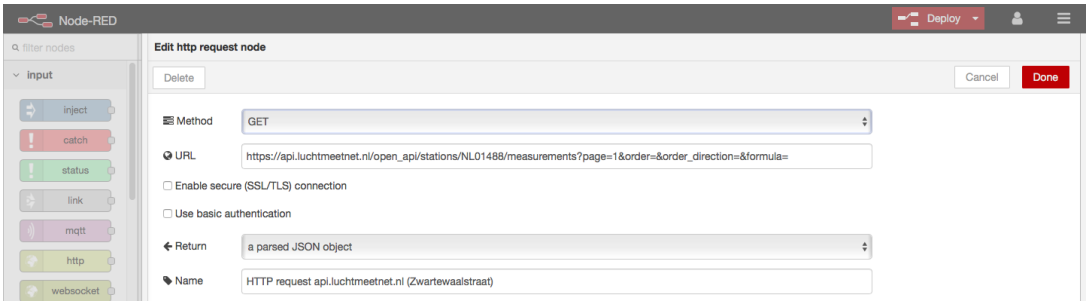


Figure B.4: Settings for the "http" node **HTTP request api.luchtmeetnet.nl (Zwartewaalstraat)**

Figure B.5: Settings for the "function" node **Add timestamp**

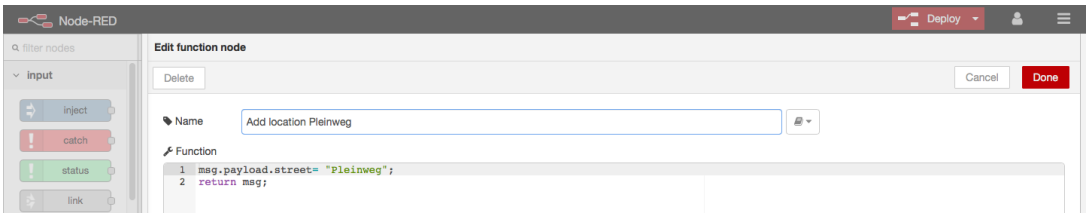


Figure B.6: Settings for the "function" node **Add location Pleinweg** (same for Zwartewaalstraat)

Figure B.7: Settings for the "function" node **SQL inset Query**

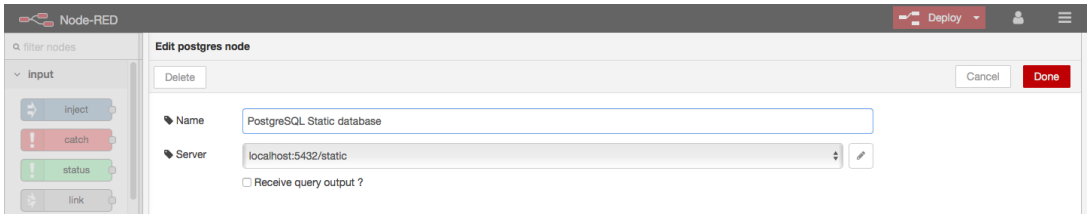


Figure B.8: Settings for the "Postgres storage" node **PostgreSQL Static database**

C NODE-RED SETTINGS FOR WEERLIVE DATA

C.1 OVERVIEW OF THE WEERLIVE NODES

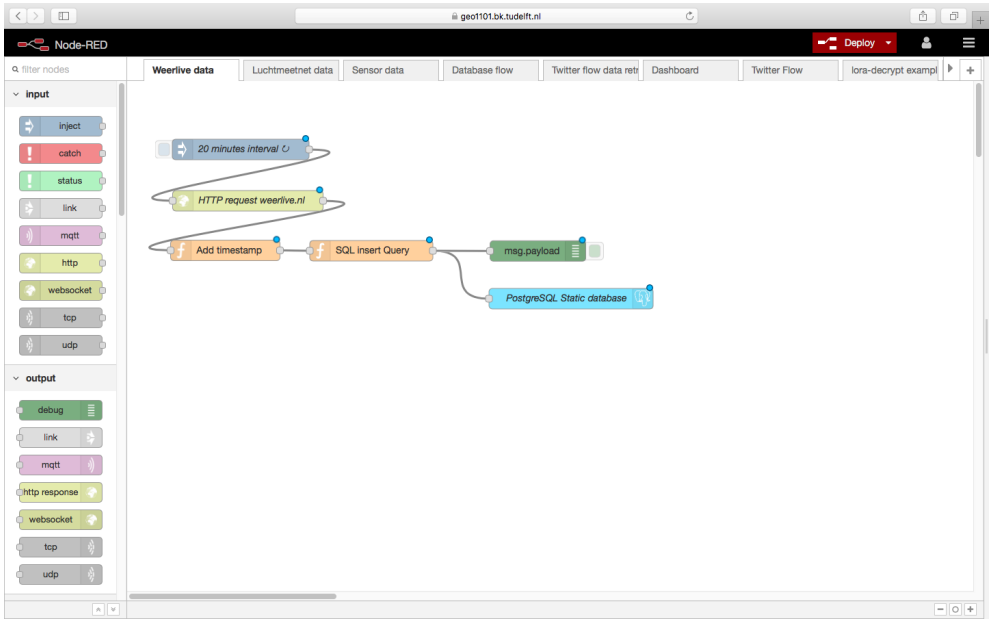


Figure C.1: Overview of the Node-RED nodes for the Weerlive API

C.2 SEPARATE WEERLIVE NODES

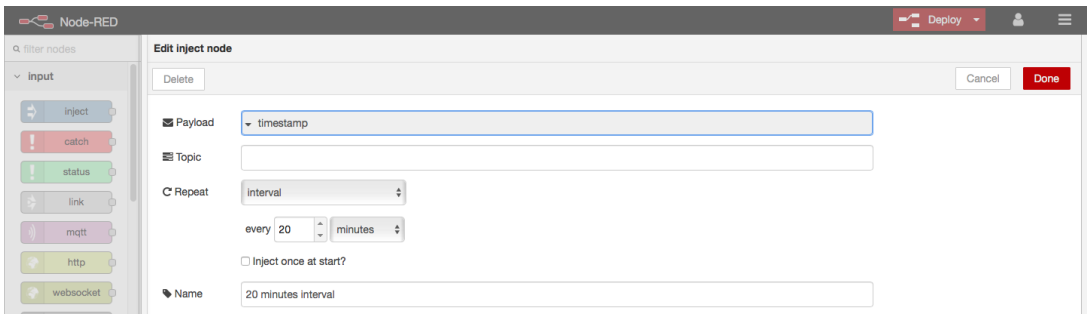


Figure C.2: Settings for the "inject" node **20 minutes interval**

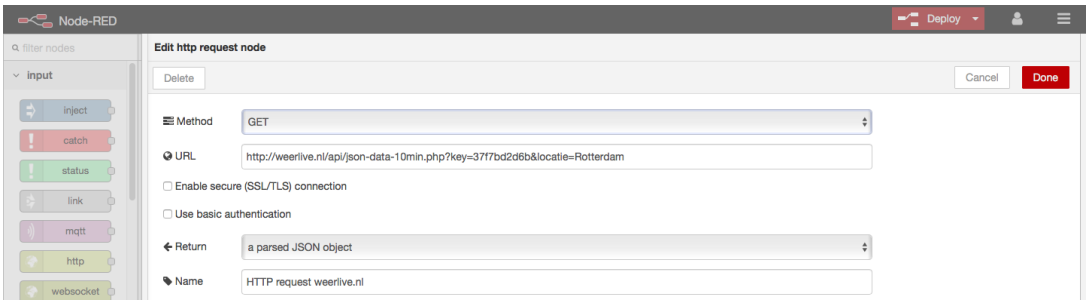


Figure C.3: Settings for the "http" node **HTTP request weerlive.nl**

Figure C.4: Settings for the "function" node **Add timestamp**

Figure C.5: Settings for the "function" node **SQL insert Query**

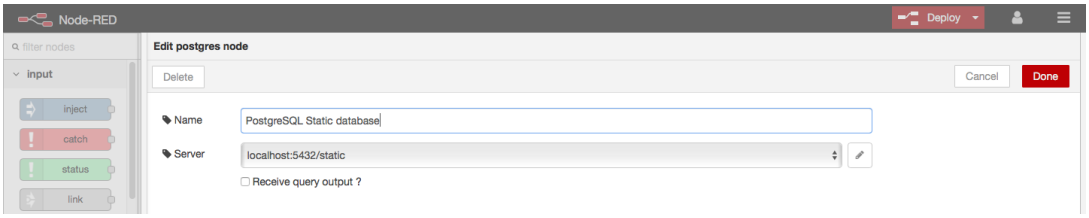


Figure C.6: Settings for the "Postgres storage" node **PostgreSQL Static database**

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the `arsclassica` package by Lorenzo Pantieri, which is an adaption of the original `classicthesis` package from André Miede.