
SPATIAL MODELLING OF AIR POLLUTION FOR
OPEN SMART CITIES

Shivam Gupta

2018

Geoinformatik

UNIVERSITY OF MÜNSTER

DOCTORAL DISSERTATION

SPATIAL MODELLING OF AIR POLLUTION FOR
OPEN SMART CITIES

Inaugural dissertation in fulfillment of the academic degree of

Doctor of Natural Sciences (Dr. rer. nat.)

at the Department of Geosciences
in the Faculty of Mathematics and Natural Sciences
of the Westfälische Wilhelms-Universität Münster

Submitted by
Shivam Gupta
from Jhansi, India

August, 2018

Dean: Prof. Dr. Harald Strauß

First supervisor: Prof. Dr. Edzer Pebesma (First Reviewer)

Second supervisor: Prof. Dr. Jorge Mateu (Second Reviewer)

Third supervisor: Prof. Dr. Ana Cristina Costa

Day of defence: 15.11.2018

Day of promotion: 15.11.2018

GEO-C

European Joint Doctorate in Geoinformatics: Enabling Open Cities
Marie Skłodowska-Curie Action (ITN-EJD)



Abstract

Half of the world's population already lives in cities, and by 2050 two-thirds of the world's population are expected to further move into urban areas. This urban growth leads to various environmental, social and economic challenges in cities, hampering the Quality of Life (QoL). Although recent trends in technologies equip us with various tools and techniques that can help in improving quality of life, air pollution remains the 'biggest environmental health risk' for decades, impacting individuals' quality of life and well-being according to World Health Organisation (WHO). Many efforts have been made to measure air quality, but the sparse arrangement of monitoring stations and the lack of data currently make it challenging to develop systems that can capture within-city air pollution variations. To solve this, flexible methods that allow air quality monitoring using easily accessible data sources at the city level are desirable. The present thesis seeks to widen the current knowledge concerning detailed air quality monitoring by developing approaches that can help in tackling existing gaps in the literature. The thesis presents five contributions which address the issues mentioned above. The first contribution is the choice of a statistical method which can help in utilising existing open data and overcoming challenges imposed by the bigness of data for detailed air pollution monitoring. The second contribution concerns the development of optimisation method which helps in identifying optimal locations for robust air pollution modelling in cities. The third contribution of the thesis is also an optimisation method which helps in initiating systematic volunteered geographic information (VGI) campaigns for detailed air pollution monitoring by addressing sparsity and scarcity challenges of air pollution data in cities. The fourth contribution is a study proposing the involvement of housing companies as a stakeholder in the participatory framework for air pollution data collection, which helps in overcoming certain gaps existing in VGI-based approaches. Finally, the fifth contribution is an open-hardware system that aids in collecting vehicular traffic data using WiFi signal strength. The developed hardware can help in overcoming traffic data scarcity in cities, which limits detailed air pollution monitoring. All the contributions are illustrated through case studies in Muenster and Stuttgart. Overall, the thesis demonstrates the applicability of the

developed approaches for enabling air pollution monitoring at the city-scale under the broader framework of the open smart city and for urban health research.

Acknowledgement

During the last three years I worked on this thesis, I received a lot of care and friendship of many people. I would like to use this opportunity to thank all of them.

First, I would like to thank my supervisor Prof. Dr. Edzer Pebesma, you have been a tremendous mentor to me. I want to thank you for encouraging my research and for allowing me to grow as a researcher. Your advice on both research as well as on my career have been invaluable. Furthermore, I would like to thank my committee members, Prof. Dr. Christian Kray, Prof. Dr. Jorge Mateu, Prof. Dr. Ana Christina Costa for serving as my second and third supervisors along with committee members. I also want to thank you for your brilliant support and suggestions during the PhD.

I am very grateful to Dr. Auriol Degbelo for his time, discussions and work he invested in guidance and support. I have always enjoyed being part of ifgi, especially at the Spatio Temporal Modelling and occasionally sitcom lab (thanks Chris). Thanks to all my colleagues and friends, who had been there to support me in finishing my PhD thesis. Sincere thanks also go to the open source communities for developing various tools that helped me during PhD. I would also like to express my gratitude to the members of the espaitec lab at Universitat Jaume I and Hansa luftbild for their support during my research visits. I would also like to thank numerous anonymous reviewers for their often helpful suggestion and remarks on my papers.

A special thanks to my family. Words can not express how grateful I am to my father and late mother for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. Thank you for supporting me for everything, and especially I can't thank you enough for encouraging me to follow my interest in geohealth studies and always believing in me.



This dissertation is funded by the European Commission within the Marie Skłodowska-Curie Actions (ITN-EJD). Grant Agreement num. 642332 - GEO-C - H2020-MSCA-ITN-2014.

List of Publication

1. Gupta, Shivam, Jorge Mateu, Auriol Degbelo, and Edzer Pebesma. "Quality of life, big data and the power of statistics." *Statistics & Probability Letters* 136 (2018): 101-104.
2. Gupta, Shivam, Edzer Pebesma, Jorge Mateu, and Auriol Degbelo. "Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models." *Sustainability* 10 (2018): 2071-1050.
3. Gupta, Shivam, Edzer Pebesma, Auriol Degbelo, and Ana Cristina Costa. "Optimisation of VGI based Air Quality Monitoring Networks for Cities." *ISPRS International Journal of Geo-Information*, (2018): (Under Review).
4. Gupta Shivam, Auriol Degbelo, and Edzer Pebesma. "Connecting Citizens and Housing Companies for Fine-grained Air Quality Sensing." *GI_Forum 2018 Journal*: (Under review).
5. Gupta, Shivam, Albert Hamzin, and Auriol Degbelo. "A low-cost open hardware system for collecting traffic data using WiFi signal strength." *Sensors* (2018): (Under Review)

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Scope	3
1.3	Problem Statement	3
1.4	Research questions	6
1.5	Approach	7
1.5.1	A statistical method to address data challenges for detailed air pollution monitoring	7
1.5.2	An optimisation method for the systematic placement of monitoring stations for air quality monitoring	7
1.5.3	An optimisation technique to utilise the public participation opportunities for detailed air pollution monitoring	8
1.5.4	Involving housing companies stakeholders for detailed air quality data collection and to overcome challenges in participatory sensing approaches	9
1.5.5	WI-FI based road traffic data collection in cities	9
1.6	Thesis Structure	10
2	Background	13
2.1	Open Smart City	13
2.1.1	Open City Toolkit (OCT)	14
2.2	Sensing Quality of Life (QoL)	15
2.3	Air pollution	16
2.4	Air pollution monitoring techniques	19
2.4.1	Spatial Proximity Models	20
2.4.2	Interpolation Models	20
2.4.3	Dispersion Models	21
2.4.4	Land Use Regression Models	22
2.4.5	Hybrid Models	23
2.5	Location optimisation techniques	24
2.5.1	Deterministic Approaches	25
2.5.2	Stochastic Approaches	25
3	Quality of life, big data and the power of statistics	27

3.1	Introduction	28
3.2	Environmental monitoring and big data challenges	28
3.3	Statistics and environmental monitoring	30
3.3.1	Land use regression (LUR)	30
3.3.2	Spatial simulated annealing for optimizing monitoring network	32
3.4	Conclusions	33
3.5	Acknowledgement	34
4	Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models	35
4.1	Introduction	36
4.2	Related Work	37
4.3	Material	39
4.4	Method	43
4.4.1	Spatial Simulated Annealing (SSA)	43
4.4.2	Optimisation Criterion Estimation	44
4.4.3	The Optimisation Procedure	47
4.5	Results	49
4.5.1	Optimisation without a Weighted Function for the Study Area	49
4.5.2	Optimisation with a Population Weighted Function for the Study Area	52
4.5.3	Sensitivity of the Optimisation Methods	54
4.5.4	Comparative Analysis	57
4.6	Discussion	57
4.6.1	Limitations and Future Work	60
4.7	Conclusions	61
5	Optimisation of VGI based Air Quality Monitoring Networks for Cities	63
5.1	Introduction	64
5.2	Related work	67
5.2.1	Citizen participation/ VGI	67
5.2.2	Air quality monitoring methods	69
5.3	Material	70
5.3.1	Study Area	70
5.3.2	Data	72
5.4	Method	74
5.4.1	Optimisation objective function	74
5.5	Results	78
5.5.1	Starting a new VGI campaign	78
5.5.2	How many sensors should we deploy?	81
5.5.3	Location significance	84
5.5.4	Where to place new VGI sensors?	85

5.6	Discussion	89
5.6.1	Significance	89
5.6.2	Limitation and Outlook	91
5.7	Conclusion	92
6	Connecting Citizens and Housing Companies for Fine-grained Air Quality Sensing	95
6.1	Introduction	96
6.2	Background	98
6.2.1	Data Completeness Challenges	98
6.2.2	Privacy Concerns	100
6.3	Method	100
6.3.1	Involving housing companies	101
6.3.2	Would housing companies want to join participatory sensing?	102
6.4	Results	103
6.5	Discussion	108
6.5.1	Addressing data completeness challenges	108
6.5.2	Addressing privacy concerns	109
6.5.3	Further opportunities for GIScience	110
6.6	Conclusion	110
7	A low-cost open hardware system for collecting traffic data using WiFi signal strength	113
7.1	Introduction	114
7.2	Related work	115
7.2.1	Traffic monitoring techniques	116
7.2.2	Privacy and Traffic Monitoring	119
7.3	Materials and Methods	119
7.3.1	System Design	120
7.3.2	System Implementation	121
7.4	Results	123
7.4.1	Vehicle Detection Algorithm	124
7.4.2	Vehicle detection	125
7.4.3	Vehicle count	128
7.4.4	Precision, Recall and F Measure	129
7.5	Discussion	130
7.6	Outlook	133
7.7	Conclusions	133
8	Synthesis	135
8.1	Summary	135
8.2	Summarised Results	136

8.2.1	<i>How can we use statistical methods like LUR and SSA for detailed air quality monitoring in an open smart city?</i>	136
8.2.2	<i>How to systematically place stations for an air quality monitoring network to maximally reduce land use regression prediction errors?</i>	137
8.2.3	<i>How can citizen participation curb the air pollution data sparsity constraint for air quality monitoring?</i>	140
8.2.4	<i>How can housing companies act as stakeholders in participatory processes for air pollution monitoring to address data gaps?</i>	142
8.2.5	<i>How can we take advantage of WiFi networks to collect detailed traffic data in the city?</i>	144
8.3	Contribution to Open City Toolkit and its Significance	146
8.4	General Discussion	148
8.5	Outlook	151
9	Conclusions	157
	Bibliography	159
A	Supplementary figures from Chapter 4	187
A.1		188
A.2		189
B	Supplementary figures from Chapter 5	193
B.1		196
C	List of Abbreviations	197

List of Figures

2.1	Mind map reflecting the scope of this thesis, and its relationship to Quality of Life (QoL)	17
4.1	Study area: City of Münster.	42
4.2	NO_2 concentration ($\mu g/m^3$) map predicted by CHIMERE model as of 20 October 2017 for Münster.	43
4.3	Schematic overview of the proposed optimisation method. Since no LUR regression model was available for the study area at the moment of the analysis, the LUR model from the ESCAPE study was used in this paper.	48
4.4	Spatial mean prediction error achieved by SSA at different probabilities of acceptance using the optimisation method without weights.	50
4.5	Energy transition while running optimisation in SSA using parameters of 0.3 probability of acceptance after removing five higher values.	50
4.6	Monitoring network designs realised after using the first optimisation criterion.	51
4.7	Spatial mean prediction error achieved by SSA at different probability of acceptance using optimisation method with population weighted criterion.	52
4.8	Monitoring network designs obtained using a population weighted optimisation criterion.	53
4.9	Deviation of the spatial mean prediction error values from mean value obtained after 15 repetitions with same parameters.	55
4.10	Summary of least spatial mean prediction error values obtained for different numbers of monitoring stations.	56
5.1	Stations in Stuttgart (Source: Umwelt Bundesamt)	71
5.2	Study area: City of Stuttgart and the existing citizen sense air quality network	72
5.3	Optimisation outcome without using the spread aspect of the objective function (N=116)	80
5.4	Optimisation outcome considering the equal weight on both wide-spread as well as prediction error aspect of the objective function (N=116)	80

5.5	Annealing energy transition during the optimisation with objective function laying equal weight to prediction error and wide-spread aspect (N=116)	81
5.6	Influence of number of monitoring sensors on the decreased prediction error aspect of objective function with equal weights on both the aspects.	82
5.7	Optimal location identified for specific number of sensors to initiate VGI campaign.	83
5.7	Optimal location identified for specific number of sensors to initiate VGI campaign (Cont.).	84
5.8	Plot collectively representing all the configurations obtained by running objective function using different numbers of monitoring sensors which can be deployed for initiating VGI campaign to identify the location of significance.	85
5.9	Diagnostic study to capture the impact of extending the number of monitoring sensors into the existing VGI based monitoring network . .	87
5.10	Optimal location identified for extending the existing crowd sourcing monitoring network using proposed objective function.	87
5.10	Optimal location identified for extending the existing crowd sourcing monitoring network using proposed objective function (Cont.).	88
6.1	Have you ever considered quality of life indicators for the development and planning of housing?	103
6.2	How informed do you feel about quality of life of residents in your housing space?	104
6.3	How important is "Health" as QoL indicator for your company planning and development?	105
6.4	How important is "natural & living environment" as QoL indicator for your company planning and development?	105
6.5	How crucial is "health" and "natural & living environment" for housing space development plans for residents?	106
6.6	Would you like to use low-cost sensors to measure air quality around housing space, so that you can control it and residents can take measures to breath safe?	106
6.7	What would be your take on sharing the air quality monitoring data with the institutions which can help in data analysis and air quality monitoring and prediction, so that residents can also get forecast of bad air quality with monitoring information?	107
6.8	Would you like to provide air quality information to the residents like normal other services so that they can save them self from harmful pollutant impact?	107
7.1	Illustration of deployment plan for the proposed hardware system . . .	120

7.2	Experimental setup: Scenario 1 (Low traffic road)	122
7.3	Experimental setup: Scenario 2 (Heavy traffic road)	122
7.4	Illustration of the web-application developed for video stream analysis	123
7.5	Illustration of time window and associated signal fluctuation pattern identification.	125
7.6	Parameters summary statistics for Heisenbergstrasse.	126
7.7	Parameters summary statistics for Steinfurterstrasse.	127
8.1	Mind map shredding light on the possible extensions of the outcomes of this thesis for developing sustainable cities	155
A.1	Populated housing area map with initial monitoring station locations (red plus signs) for study area	187
A.2	Histogram of predictor variables for LUR used in the study	188
B.1	Configuration with various weights on prediction error (W1) and wide-spread(W2) aspect of developed objective function.	194
B.1	Configuration with various weights on prediction error (W1) and wide-spread(W2) aspect of developed objective function.	195
B.2	Optimal locations considering prediction error constrain for Stuttgart LUR model developed using low-cost sensor network data.	196
B.3	Optimal locations considering prediction error and wide-spread aspect in the objective function for Stuttgart LUR model developed using low-cost sensor network data.	196

List of Tables

1.1	Outline of the Thesis	11
3.1	Challenges of big data, and potential of the combined use of the proposed methods.	33
4.1	LUR variables selected	41
7.1	Threshold rules for vehicle identification using Heisenbergstrasse data	127
7.2	Threshold rules for vehicle identification using Steinfurterstrasse data .	128
7.3	Vehicle classification using algorithm and video	128
7.4	Vehicle classification using algorithm and video	129
7.5	Precision, recall and F Measure using Heisenbergstrasse data	130
7.6	Precision, recall and F Measure using Steinfurterstrasse data	130
A.1	All the configuration realised for optimisation at different probability of acceptance	192

Introduction

” *Sharing is good, and with digital technology,
sharing is easy.*

— **Richard Stallman**

1.1 Motivation

Over the last few years, there has been an explosion in research on smart cities (Ojo et al., 2016), sustainability and policies related programs for cities (Foley et al., 2017). Approaches like these capture the collective imagination with a promise to help in addressing various problems of ongoing economic crisis, of choices from the rigidity of urban systems, of opening up public decision making, and of greater respect for the environment, all through the use of new technologies. It is not difficult to appreciate the advantages of quickly advancing technologies on our day to day life. These new technologies propose solutions that can contribute to improving our Quality of Life (QoL) in many ways. QoL is tied to the perception of ‘meaning’. The quest for meaning is central to the human condition, and we are brought in touch with a sense of meaning when we reflect on that which we have created, loved, believed in or left as a legacy (Barcaccia, 2013). For some, it may mean safety and security, employment opportunity, a clean environment, ease of travel, adequate health care, good school, etc. Every individual in the society holds their own perception of quality of life (Ariely et al., 2008). QoL has been influenced by the multifaceted and complicated characteristics of multi-dimensional issues and features such as environmental pressure, total water management, total waste management, noise pollution and the level of air pollution (Feneri et al., 2013; Eusuf et al., 2014). QoL studies usually pertain to the analysis of more subjective factors, such as the quantity and quality of natural amenities (e.g. climate and physical beauty) as well human-created amenities (e.g. recreation/entertainment opportunities, education and health services) and other ‘objective’ factors (e.g. unemployment rate and human capital). In the past decade, there has been an increased interest in studying both objective and subjective measures of QoL such as happiness, socioeconomic, demographic as well as possible geographical determinants.

More than half of the world's population now lives in urban areas (Heilig, 2012). Several environmental constraints affect the life in urban cities, which are related to traffic congestion, overcrowding, environmental quality, waste management, health facilities, criminality and other factors such as the well-being of individuals, and spatial dimension (Royuela et al., 2007). The healthy environment for a living is an essential need for each human being living on the Earth. The outdoor environment is one important aspect, which directly affects the quality of people's lives (WHO, 2015). Pollution of any kind, be it noise, aerosol, smoke, smog, haze, oil spills, and unclean water reduces the quality of life in the city, its livability, its attractiveness, and most importantly the health of its inhabitants. The air we breath is the basis for our existence. On average, an adult breathes over 11,000 litres of air per day; children breathe even more air relative to the body surface area, breathing frequency, and heart rate. Any contaminant in the air will, therefore, take in and will be absorbed in the body, more in the case of children. Air quality—or its converse, air pollution—is a significant risk factor for human health. Numerous diseases may be caused by air pollution such as respiratory infection, lung cancer, cardiovascular disease, chronic obstructive pulmonary disease (COPD), and asthma (Prüss-Üstün and Corvalán, 2006; Sadalla et al., 2005). With an increasing number of humans now living in urban space, there are urgent needs of examining what the rising number of people in the cities means for air pollution, local climate and the effects these changes have on QoL.

Currently, air quality modelling approaches based on proximity, interpolation, dispersion, land use regression (LUR) and other sophisticated methods, integrated with modern statistical modelling techniques (e.g. neural networks, independent component analysis, boosting, and random forests) are utilised both as diagnostic models to explore the relationship between responses and influential factors and/or as predictive models (Sayegh et al., 2016; De Nazelle et al., 2013). Yet, because of the limited availability of air pollution monitoring data, proper results cannot be inferred (Shaddick et al., 2018). Most of the cities' or nations' air pollution measurements are usually collected using fixed monitoring sites which are limited in number, leading to a lack of spatial representativeness and temporal coverage. This poses demands on the development and deployment of tools and techniques that can help in collecting data from various sources for multi-scale integrated models and integrated urban services. Therefore, helping in adapting to the responsibilities associated with fast-growing cities along with changing climate and global challenges associated with our environment (Baklanov, 2012). During the writing of this thesis, in 2015-2017 Beijing and Delhi suffered from some of the worst air pollution issues worldwide, leading to the pollution and health emergency (Independent, 2017; BBC, 2015). The European Environment Agency (EEA) in 2017 has also warned that people living in European cities are being exposed to high concentrations of air pollution, with particulate matter (PM), nitrogen dioxide

(NO₂) and ground-level ozone (O₃) causing the most damage to human health (European Respiratory Society, 2017). Consequently, people suffer from bad air quality but continue their day to day activity. These episodes showed the necessity of having systems in place that can help in controlling air pollution impact, especially in growing cities. Extensive efforts are required for monitoring and controlling the harmful impact of bad air quality on human health and well-being. The primary goal of this thesis is to contribute to air pollution monitoring research by developing and investigating various approaches that can help address the issues caused by limited data available to enable air pollution monitoring at intracity level.

1.2 Scope

The spur for this thesis came from the topic of "Sensing Quality of life (QoL)" in the open smart city. As discussed in the previous section, quantifying QoL requires consideration of various subjective and objective indicators which make it a too broader as a topic. This study focuses on the monitoring of one hazardous environmental factor, which is affecting the QoL of the cities in recent years, namely air pollution. Hence, the title of this thesis is "Spatial modelling of air pollution for the open smart city". The scope of the thesis can be expressed as follows:

1. The thesis focuses on addressing concerns related to air pollution monitoring in open smart cities. The focus on air pollution monitoring can help to quantify the two quality of life indicators used by European commission (Eurostat, 2015): 'health'; and the 'natural and living condition'. Issues related to other objective or subjective indicators of QoL fall outside the scope of this thesis.
2. The studies carried out in this thesis are focused on detailed air pollution monitoring and overcoming constraints limiting it. The developed tool and techniques can be adapted for application in other scientific fields, such as sound pollution, urban heat islands estimation and other adverse environmental variable assessment which can affect the well-being of cities.

With the scope as mentioned earlier of the present work, various approaches developed in the study will be integrated as tools in the OCT.

1.3 Problem Statement

The growing population in cities is associated with a significant increase in road vehicles and air pollution (Kumar et al., 2015b; Gurjar et al., 2010). The impacts

of highly spatiotemporally restricted pollution and its exposure are still poorly understood (Kumar et al., 2015b). Air quality varies over a relatively small scale since the resulting pollutant concentration in a particular place depends predominantly on local emission sources and atmospheric flow conditions (Britter and Hanna, 2003). It is crucial for the local governing authorities, stakeholder, planners and the public to have precise data to take actions against unhealthy exposure to polluted air for improving quality of life in cities. Many countries have real-time air quality forecasting (RT-AQF) programs in place to forecast the concentrations of pollutants of particular health impact such as O₃, NO₂, PM_{2.5}, and PM₁₀ (Manins and Committee, 2001; Dye et al., 1999; Pudykiewicz and Koziol, 2001). However, RT-AQF pollutant concentration maps still lack spatial granularity for large urban areas at present because they require a significant amount of data, computing facilities and various other inputs that are not readily available for many cities. This complexity restricts the overall approaches for assessing human exposure to harmful pollutants. Getting hold of data and converting it into knowledge is one of the valuable steps for making decisions and forming policies for air pollution control. Despite local variations in air pollution concentrations (Nieuwenhuijsen et al., 2015), air pollution monitoring stations are few and sparsely installed in some cities. Monitoring station numbers are approximately 2-3 in a city with the population size of a million or more, as per the environmental directives from the EEA/EU directive of 2008 and USEPA (European Parliament Councils, 2008; Federal Register (OFR) and Government Publishing Office.OFR, 2018). Moreover, such monitoring systems are cumbersome in size to install, expensive and costly to maintain. Also, these existing sparsely organised expensive devices are also affected by highly localised factors, such as nearby trees, framing their measurements potentially not the representation of air quality even a few streets away.

For the measurement of such a complex behaviour of air quality distribution in cities, we need approaches which are well suited to collect spatiotemporal variability in air pollution in cities affected by local variables like traffic density, street topology, altitude, land use type and distance from sources of emission. We need to devise methods which are capable of adapting based on the available data sources for air quality monitoring considering the variability factors mentioned above. For the development of such methods, we need air pollution monitoring data sources. There is a need for low-cost, portable, and sensitive air pollution monitoring devices that are capable of long-term, continuous, routine, and real-world air pollution monitoring in cities (Su, 2018). Moreover, the rapid developments in information and communication technologies (ICT) can help us to collect real-time, localised data. These technologies, if used optimally can help spatial planners and decision-makers for gathering air pollution data and use it for modelling and exposure assessment. Increasing the number of monitoring stations could help to quantify and characterise air pollution gradients in the city.

Another complication of practical concern in air pollution monitoring is the selection of the locations of the air pollution measurement sensors. The application of advancing ICT low-cost air quality sensors can also help in enabling much denser air quality monitoring networks at a comparably lower cost than existing official stations. However, placing monitoring stations without considering their relevance may affect the outcome for air pollution monitoring, as one aspect of the significance of the monitoring data relies on “where” the data is collected (Kanaroglou et al., 2005a; Hao and Xie, 2018). Having a method which can systematically find the optimal locations for air quality monitoring network design, therefore, can promote the application of low-cost air quality sensors for robust air quality monitoring in cities.

Furthermore, the striking phenomena of urbanisation also create excessive demands for cities’ infrastructure. Increasing demands for transport combined with limited available urban road infrastructures can lead to traffic congestion. These traffic loads and congestion increase the vehicular emissions and degrade the ambient air quality in cities (Zhang and Batterman, 2013). The currently enforced 2008/50/EC EU Directive (European Union, 2008a) states that fixed urban traffic stations shall measure urban air pollution levels mostly influenced by road traffic emissions. Thus, to provide adequate information on air quality spatial distribution, traffic congestion data is required. Various air pollution monitoring methods rely on the traffic congestion data for emission inventory input (Thouren et al., 2018; Yu et al., 2018). However, traffic data is usually available for a very limited number of roads in a city (Ryan and LeMasters, 2007; Eeftens et al., 2012a). Currently, existing official traffic monitoring stations are large, expensive, power hungry and require the lane closure and traffic disruption for installation or regular maintenance (Balid et al., 2018). When the air pollution monitoring aims to develop a detailed map of air quality, traffic data in a detailed form is necessary. Having a low cost and easily installing devices for data sources can be a useful source to gather traffic data at higher resolution in cities. Besides, these low-cost sensor devices offer new opportunity to measure real-time data and provide immediate feedback, which serves as an opportunity to build the capacity of residents in the city to understand traffic impact, air pollution, spatial and temporal variability, and exposure patterns relevant to their locality (Clements et al., 2017a).

This research work aims to further widen the current knowledge about detailed air quality monitoring initiatives at the city level by devising methods for addressing the following problems:

- P1. Which approach could be well suited to collect air pollution data such that it is capable of representing the spatial variability caused by local geographical

variables like traffic density, street topology, altitude, land use type and distance from sources of emission?

P2. How to systematically find optimal locations for an air quality monitoring network design for robust air quality monitoring in cities?

P3. How can other data sources be well utilised for addressing scarcity and sparsity of air pollution measurement devices for air quality monitoring?

P4. How to gather traffic data at a higher resolution using low-cost sensors?

By addressing these four problems, we believe that we can address a few barriers to detailed air quality monitoring initiatives for the cities.

1.4 Research questions

Considering the problems stated above, the present thesis investigates the following research questions :

RQ 1. How can we use statistical methods like LUR and SSA for detailed air quality monitoring in an open smart city? (Addressing problem P1 and P2)

RQ 2. How to systematically place stations for an air quality monitoring network to maximally reduce land use regression prediction errors? (Addressing problem P2)

RQ 3. How can citizen participation curb the air pollution data sparsity constraint for air quality monitoring? (Addressing problem P3)

RQ 4. How can housing companies act as stakeholders in participatory processes for air pollution monitoring to address data gaps? (Addressing Problem P3)

RQ 5. How can we use the abundance of existing WiFi networks to collect detailed traffic data in the city? (Addressing problem P4)

The thesis aims to answer these five research questions and develop solutions for them. The approach pursued to answer these research questions are discussed in the following section.

1.5 Approach

1.5.1 A statistical method to address data challenges for detailed air pollution monitoring

The primary purpose of air quality monitoring is to identify areas where air quality is unhealthy for life in the city. To identify the impact of air pollution on quality of life of residents in the city, we require an air quality monitoring network that is capable of providing the continuous stream of data from certain locations. Generally, the map representing air pollution for cities are coarser in resolution, because of the sparse arrangement of monitoring sensors. Various alternate approaches ranging from downscaling to the utilisation of the latest sensor technologies were proposed in the literature to overcome the sparsity challenge. However, the deployment of new monitoring sensors for gathering spatiotemporal feed of air pollution data brings up new challenges concerning the handling of big data, represented as 5Vs: Volume, Velocity, Variety, Veracity and Value. Furthermore, sometimes the challenges are also caused due to the lack of representative data for air pollution from monitoring stations or not well-spread monitoring network to enable detailed air pollution monitoring at the desired scale in the cities. In order to address these challenges for open smart cities, an optimal setup of the data stream is desired. The combination of two well-established statistical methods: Land Use Regression (LUR) and Spatial Simulated Annealing (SSA) is proposed to optimise the selection of variables and locations for enabling timely spatial and temporal analysis of air pollution data sources, hence can help in addressing the data challenges. The approach presented in Chapter 3 as part of the thesis is helpful in addressing the challenges associated with big data.

1.5.2 An optimisation method for the systematic placement of monitoring stations for air quality monitoring

To capture the intraurban air pollution concentration variability, we need a network of monitoring stations that can represent the air pollution concentration across areas of interest. Barriers to collect the data within the city at a finer scale involve the sparse arrangement and limited amount of air quality monitoring devices. If we try addressing this limitation by adding new devices, the question of “where” to place the new data source is of great significance. Chapter 4 of the thesis presents the method that can help in identifying the “optimal” locations for placing new air pollution monitoring devices for robust air pollution estimation for LUR model. The optimisation considers the criterion implemented using SSA to find the specific

number of locations which together contribute to decreasing the spatial mean prediction error for the LUR estimation. Furthermore, the optimisation methods are extended to identify the optimal locations which together contribute to decreasing the spatial mean prediction error in the highly residential areas of the city by considering the population weight parameter in the optimisation process, as has been demonstrated in the case study for the city of Muenster. It, therefore, helps in addressing the data scarcity and limited data availability challenges by identifying the optimal locations which can improve the detailed air pollution monitoring efforts for the cities.

1.5.3 An optimisation technique to utilise the public participation opportunities for detailed air pollution monitoring

As discussed above, we developed an optimisation method to identify the set of locations where the air pollution monitoring sensors can be placed for applications in detailed air pollution monitoring. Although governmental organisations are collecting data, the limitations imposed by cost and hardship to maintain bulky sensor generally lead to a sparsely arranged limited number of monitoring stations. With the advancement in ICT, the application of low-cost and small devices for measuring air quality is under discussion recently for addressing the sparsity, and limited data availability challenges (Borrego et al., 2016; Spinelle et al., 2017). For the fruitful application of these new technologies for air pollution monitoring, some crucial aspects need to be considered. In Chapter 5, we discuss these aspects which are essential for planning the participatory initiatives for air pollution monitoring by systematically identifying the locations for the participatory nodes. The systematic placement helps in reducing the flow of redundant and less useful data into the air pollution monitoring process and increase the resolution of air quality monitoring initiatives. It used the optimisation method from the previous step with certain modifications based on participatory data constraints and demonstrated in the case study of the city of Stuttgart. In all, the systematic placement of participatory air pollution monitoring nodes helps in collecting more data and reduces the barrier of sparsity and limited data source availability. Hence, improving the data collection procedures to develop detailed air pollution monitoring with equal public participation.

1.5.4 Involving housing companies stakeholders for detailed air quality data collection and to overcome challenges in participatory sensing approaches

As discussed in the above section, the scientific community can benefit by democratising the collection of air pollution data. However, there are a few instances where these approaches fail to achieve as desired. Since participatory data collection is a voluntary and open contribution, there is room for inclusion of corrupt data in the data collection process. The literature suggests participatory sensing is prone to various inimical behaviours, such as false values, misuse of the sensor, lurker phenomena, calibration, maintenance and ethical concerns. We briefly elaborate on these aspects in Chapter 6 of the thesis, where we also highlighted the two significant challenges of public participation approach for air pollution monitoring: Data challenge and privacy challenge. The chapter discusses these two challenges and how it can influence the data collection process for detailed air pollution monitoring.

Furthermore, to solve the discussed limitations of the participatory approach, we proposed the involvement of housing companies as a stakeholder in the participatory sensing approach for air pollution data collection. In order to recognise the practicality of the proposed approach, a survey was conducted with 71 housing companies in Germany for understanding their prospective to be part of air pollution monitoring using low-cost sensors. Broadly, 78% housing companies shared their interest to use low-cost sensors for air pollution monitoring in their property. We believe that the proposed approach will aid in reducing the corrupt data, increasing the spatial spread of the large number of low-cost well-maintained sensors in the city, complemented with its advantage to confront challenges of participatory sensing.

1.5.5 WI-FI based road traffic data collection in cities

For monitoring the air pollution in cities, detailed information about the input variables required to produce a detailed output for air pollution monitoring. Considering the importance of road traffic data for air pollution estimation, we developed a hardware device which is capable of reducing barriers in comprehensive traffic flow data collection in cities. The hardware is low-cost, small in size, consumes less power than traditional devices and easy to install. The device utilises the power of WiFi signals existing around us to identify the type and count vehicles on the road. The primary advantage of this hardware is its ability to work efficiently in situations like rain, storms or in dark backgrounds where some time traditional systems like video surveillance system fails. We evaluated the effectiveness of the developed hardware by deploying it in two different scenarios. Chapter 7 provides the detailed results of

the developed hardware and its application. By developing these open hardware solutions, we also attempt to support detailed open data collection for improving the air pollution monitoring in cities.

1.6 Thesis Structure

The following Chapter 2 provides background information about the developed methods, hardware, and on the application of the developed outcomes. It also briefly describes various existing tools and techniques for monitoring the air pollution in the smart city. The structure of the six subsequent chapters follows the five questions, and the applied approaches as described in the previous sub-section. These chapters consist of manuscripts which were published or submitted for publication, listed in Table 1.1 below.

Chapter 3 describes briefly existing methods of air quality monitoring and application of two well known statistical methods for overcoming challenges in environmental monitoring with the particular focus on air quality monitoring for improving quality of life in the city. The method is also helpful in overcoming the challenges of big data for air quality monitoring. The advantages and limitations of the proposed method are also discussed.

Chapter 4 presents the optimisation method developed to systematically place air pollution monitoring devices for addressing the data sparsity and limited data availability challenges of air quality monitoring. The optimisation method developed in this chapter demonstrates the ecological validity of the method selected in Chapter 3.

Chapter 5 demonstrates the application of the developed optimisation method in Chapter 4 for crowdsourcing air pollution by public participation for further addressing sparsity and limited air pollution monitoring station data availability for detailed air pollution monitoring in the city of Stuttgart.

Chapter 6 proposes the involvement of housing company as another stakeholder participatory sensing framework for air pollution data collection. The proposed approach is supported by the survey results from housing companies concerning their interest to be part of the proposed approach.

Chapter 7 discusses the open hardware device we developed for collecting the detailed road traffic data in cities. Finally, the summarised results, their implications

and outlook for the complete research will be presented in Chapter 8. Chapter 9 presents the conclusion of the thesis.

Table. 1.1. Outline of the Thesis

Outline			
Chapters	Manuscripts	Research Question	Approach
Chapter 3	Gupta, Shivam, Jorge Mateu, Auriol Degbelo, and Edzer Pebesma. "Quality of life, big data and the power of statistics." <i>Statistics & Probability Letters</i> 136 (2018): 101-104.	1	1.4.1
Chapter 4	Gupta, Shivam, Edzer Pebesma, Jorge Mateu, and Auriol Degbelo. "Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models." <i>Sustainability</i> 10 (2018): 2071-1050.	2	1.4.2
Chapter 5	Gupta, Shivam, Edzer Pebesma, Auriol Degbelo, and Ana Cristina Costa. "Optimisation of VGI based Air Quality Monitoring Networks for Cities." <i>ISPRS International Journal of Geo-Information</i> , (2018): (Under Review).	3	1.4.3
Chapter 6	Gupta Shivam, Auriol Degbelo, and Edzer Pebesma. "Connecting Citizens and Housing Companies for Fine-grained Air Quality Sensing." <i>GI_Forum 2018 Journal</i> : (Under review).	4	1.4.4
Chapter 7	Gupta, Shivam, Albert Hamzin, and Auriol Degbelo. "A low-cost open hardware system for collecting traffic data using WiFi signal strength." <i>Sensors</i> (2018): (Under Review)	5	1.4.5

Background

This chapter presents the related work and provides more background for the core chapters (Chapter 3-7) of the thesis. As stated in the motivation, the underlying theme for this thesis is from the main topic of sensing Quality of Life (QoL) in open smart cities. The mind map in Figure 2.1 visually organises the path we chose while researching for the thesis. The lines in red colour represent the path we took from the inception to the final stage of the research work for the thesis. This chapter will briefly discuss in Section 2.2 the general overview on the topic: *Sensing Quality of Life* and its relation to air pollution. The next Section 2.3 phrase the current concerns in cities related to air pollution and why it is important to focus on solving this problem. The Section 2.4 catalogues the background of various existing approaches used for modelling air pollution in cities and their significance to the research we conducted in this thesis. Furthermore, Section 2.5 reviews the application of Spatial Simulated Annealing (SSA) for optimisation of air pollution monitoring network studies, which will be useful for Chapters 3, 4 and 5.

2.1 Open Smart City

The increasing demographic pressure caused by urbanisation, coupled with the crises of climate change and economic instabilities lead to a range of new concepts that centre cities to solve these problems. Over the last 20 years, the concept of smart cities has gained considerable attention among businesses, governments and academia (Ojo et al., 2016). With the advancements in information and communication technologies (ICT), the smart cities vision represents a concept of urban development by the utilisation of human, collective and technological capital (Angelidou, 2014). The concept is still evolving, and there are differences in the conception of a smart city (the academic, business and government envision differently for the economic, ideological and theoretical orientations). Academic institutions working on smart cities, especially in the computational sciences position their work as pragmatic and not ideological. They endure the idea to produce technologies and solutions that can improve governance and economic advancements. At the same time, businesses aim to present their ideas as being city and citizen-oriented. They are vested in interest to push market lead technological solutions to city administration. Similarly, governments along with supra-national states, such as European Union (EU), endorse

the smart city vision positively for socio-economic progress, with the goal of making cities more livable, safe, and sustainable for the better quality of life (Ahvenniemi et al., 2017; Viitanen and Kingston, 2014).

For this thesis, we define a smart city as suggested by Yin et al., 2015: “*As a system integration of technological infrastructure that relies on advanced data processing with the goals of making city governance more efficient, citizens happier, businesses more prosperous and the environment more sustainable*”. This definition considers the citizens as one main beneficiary along with other stakeholders, such as businesses and government. A smart city is a place where the exploitation of digital technology can solve previously untraceable social and environmental problems that are impacting the city. More interesting is the new trend of envisioning smart cities with the application of open data. Since massive data collection with the help of ICT had been a characteristic feature of smart cities, publishing such data as open data for city management and life is a relatively recent phenomenon (Ojo et al., 2015). Various studies of smart cities highlight the importance of enabling citizen participation by various open data initiatives (Graaf and Veeckman, 2014; Zubizarreta et al., 2015; Wijs et al., 2016). These open framework based smart cities can be referred to as “Open Smart Cities”. Opening up cities for citizen participation and publishing open data can foster innovation, creativity, and citizen-centric solution development which can help in improving QoL in smart cities (Oser, 2017; Degbelo et al., 2016).

2.1.1 Open City Toolkit (OCT)

The thesis is the part of the GEO-C project that aims to enable open smart cities by increasing the transparency, facilitating collaboration among citizens and other stakeholders of cities, and enabling the participation of citizens in the improvement of their cities’ services and quality of life. The open city toolkit (OCT) is the resulting platform of the project bringing together the research contributions from various research studies conducted under the project. In essence, the OCT is a collection of tools, applications, services, datasets, specifications and guidelines to empower citizens and various stakeholders to participate in shaping the future of their cities by delivering services based on open data that can be useful for citizens, businesses and governing bodies. The complete toolkit was developed using open source tools, and all the new developments are regularly published as open source software components on GitHub. The present thesis outcome will also be contributed as part of OCT.

2.2 Sensing Quality of Life (QoL)

55% of the world population is already living in the urban areas, and urbanisation is expected to increase this proportions to 68% by 2050 (UNDESA, 2018). There are urgent needs of examining what this rising number of people means for cities' environment, resources and Quality of Life (QoL). A large and rapidly growing number of studies these days focuses on urban areas, building on the long tradition of analysing 'objective' QoL measures and combining them with subjective approaches to measuring well-being (Ballas, 2013; Marans and Stimson, 2011). The QoL indicators are the ways to measure the vital signs of a community. The conventional approach used for policy has been to use measures of gross domestic product (GDP) or regional gross valued added. Various EU policies are increasingly laying stress on the importance of equality, citizenship and public participation in decision-making, to the extent that economic measures are related to resource use and consumption. There are also various pressing issues about public goods and the sustainability of economic growth. Other facets of QoL besides income includes environment, freedom, health, working conditions, leisure, social and family relationships, and levels of satisfaction, amongst others, (Diener and Suh, 1997). Economists do not deny that these facets play a role in the quality of life. Their arguments have instead been that there is a trade-off between income and other facets of QoL.

The multifaceted and complicated characteristics of QoL is a multi-dimensional issue impacted by features such as environmental pressure, total water management, total waste management, noise and the level of air pollution (Feneri et al., 2013; Eusuf et al., 2014). The QoL index, which is shaped by a series of factors including safety, healthcare, consumer prices and purchasing power, traffic commute time, pollution and property price to income ratio is used to represent the QoL in a city (Eusuf et al., 2014). However, not much consideration has been given about the impact of environmental parameters and their impact on QoL. Several environmental constraints affect the QoL in urban spaces. Factors related to traffic congestion, overcrowded, environmental quality, waste management, health facilities, and criminality is still impacting the well being of individuals, with the spatial dimension.

Pollution of any kind, be it noise, aerosol, smoke, smog, haze, oil spills, and unclean water reduces the QoL in cities, its livability, its attractiveness, and most importantly the health of its inhabitants. One of the essential longing to survive on earth is "air". The air around us is one of the indicators which defines the existence. On average, an adult breathes over 3,000 gallons of air per day; children breathe even more air relative to body surface area, breathing frequency, and heart rate. Our bodies carefully regulate the oxygen-carbon dioxide exchange that occurs when we inhale and exhale. When the air is polluted, we breathe in pollutants along with the air.

At present, bad air quality is accounting for more pressing environmental concerns. As per the records of the World Health Organization (WHO), more than 80% of people living in urban areas that monitor air pollution are exposed to harmful air quality limits (WHO, 2018c). The pollutants in the air are responsible for chronic or non-communicable diseases (NCDs), causing over causing an estimated one-quarter (24%) of all adult deaths from heart disease, 25% from stroke, 43% from chronic obstructive pulmonary disease and 29% from lung cancer (WHO, 2018a). This issue is included by the European Union (EU) as one of the challenges for smart cities in its H2020 programme, recently debated in the European Forum on Eco-Innovation (European forum on eco-innovation, 2018). As an ever-increasing number of cities loom large, a new generation of multi-scale integrated detailed air pollution models and services are needed for helping to ensure that we can adapt to the coming responsibilities associated with faster growing cities. Hence, we decided to work on methods to address air pollution problems (see Figure 2.1).

2.3 Air pollution

Although clean air is considered as a vital requirement for living and maintenance of human health and well being, air pollution continues to pose a significant threat. Degradation of the environment through air pollution combined with lifestyle changes can contribute to increasing rates of obesity, diabetes, cardiovascular disease, nervous system problems and cancer-like disease. Air quality—or its converse, air pollution—is a significant risk factor for human health. Numerous diseases may be caused by air pollutants such as respiratory infection (Grigg, 2018), lung cancer (Gharibvand et al., 2017), cardiovascular disease (Hadley et al., 2018), chronic obstructive pulmonary disease (COPD) (Cohen et al., 2017), and asthma (Bowatte et al., 2017). In the year 2016, ambient air pollution was responsible for 4.2 million deaths worldwide. Ambient air pollution is estimated to cause about 16% of the lung cancer deaths, 25% of chronic obstructive pulmonary disease (COPD) deaths, about 17% of ischaemic heart disease and stroke, and about 26% of respiratory infection deaths (WHO, 2016). The top five causes of deaths globally are directly related to air pollution (WHO, 2018b). Furthermore, reproductive and mental health problems are also on the rise because of air pollution, which concerns the future generation. Children represent the largest subgroup of the population susceptible to the effects of air pollution (WHO, 2014).

Among children, air pollutants are associated with increased acute respiratory illness, increased incidence of respiratory symptoms and infections, episodes of longer duration, and lowered lung function. Asthma, the most common chronic disorder of childhood, is on the rise in the United States and other industrialised nations.

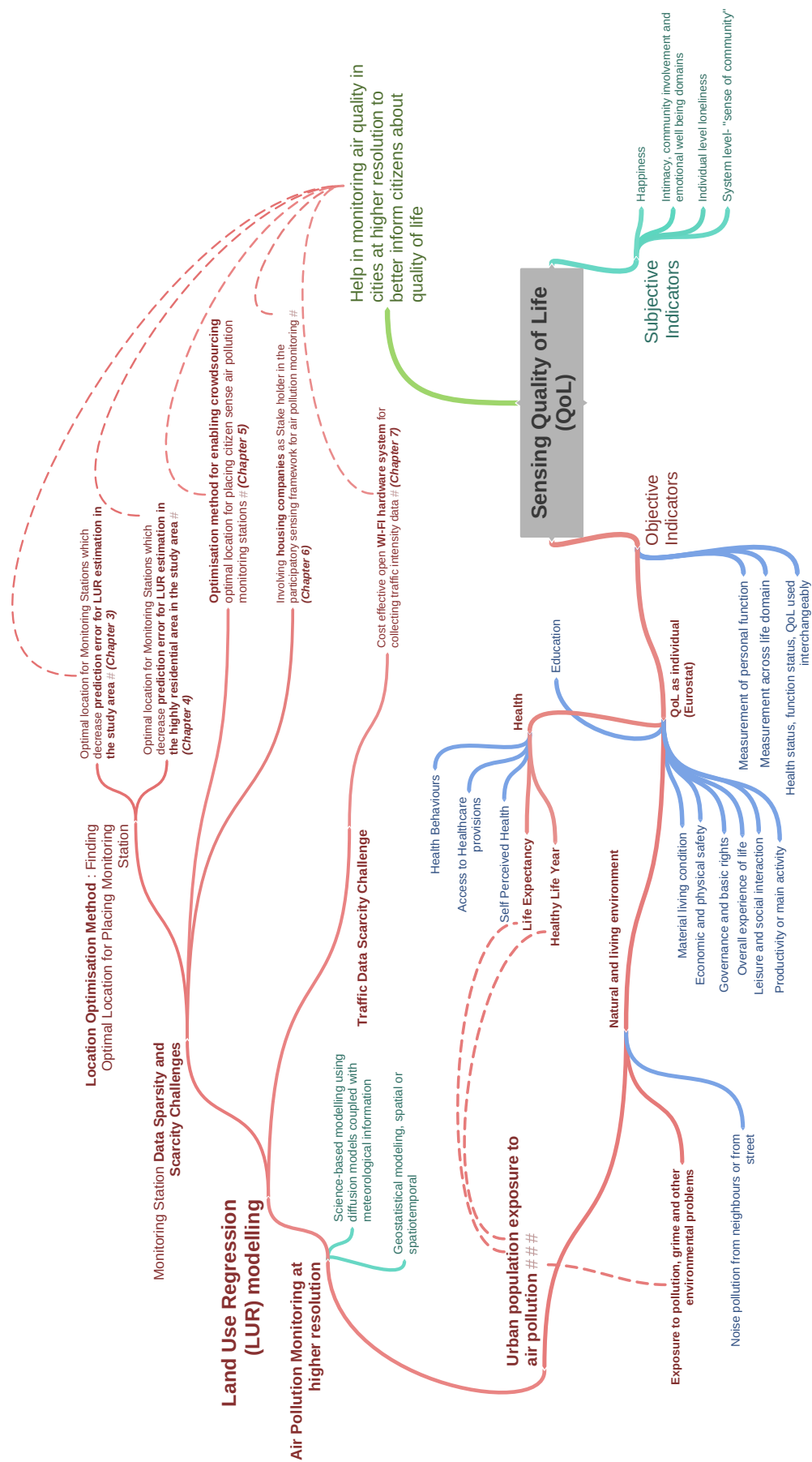


Figure. 2.1. Mind map reflecting the scope of this thesis, and its relationship to Quality of Life (QoL)

Several studies have linked ozone and particulate air pollution with exacerbation of asthma in children afflicted with the disease (Gauderman et al., 2015). Due to their greater respiratory rates, children breathe a proportionately greater volume of air than adults. As a result, children inhale more pollutants per pound of body weight. They also spend more time engaged in vigorous activity than adults. Besides, because of young children's height and play habits (crawling, rolling), they are more likely to be exposed to pollutants or aerosols that are heavier than air and tend to concentrate in their breathing zone near ground level (Goldizen et al., 2016). Children's physiological vulnerability to air pollution arises from their narrower airways and the fact that their lungs are still developing. Irritation caused by air pollutants that would produce only a slight response in an adult can result in a potentially significant obstruction in the airways of a young child. Various studies suggest that air pollutants are associated with a wide variety of adverse health effects in children (Gauderman et al., 2015; Gent et al., 2003), such as; increased death rates in very severe pollution episodes and increased mortality risks for those living in highly polluted areas, increased risk of acute respiratory illness, aggravation of asthma, increased respiratory symptoms, and increased sickness rates (as indicated by kindergarten and school absences), and decreases in lung function. Previous research also indicates that carcinogens in ambient air can be transferred transplacentally from the mother to the fetus (Fucic et al., 2017). Such impacts can lead to genetic damage to the fetus, higher than damage to mothers, indicating the increased sensitivity of the developing fetus to the effects of carcinogenic air pollutant exposures. In all, it is evident that air pollution has a direct impact on individuals' health, and an economic impact induced by health costs and missed days at work and school (Sadalla et al., 2005).

These impacts of spatially restricted air pollutants and their exposure are still poorly understood (Carvalho, 2016). Considering the influence of air pollution on urban sustainability and health, therefore, requires measuring air pollution and utilisation of the gathered information to predict and discover relationships between different urban health problems. Hence, to better plan and understand the city needs for better QoL, it is crucial to comprehend the air quality perspective. City morphology and road traffic play a significant role in realising and dynamics of air quality in the streets of the city (Di Sabatino et al., 2018). Air quality varies over a relatively small scale since the resulting pollutant concentration in a particular place depends predominantly on local emission sources and atmospheric flow conditions (Britter and Hanna, 2003). It is difficult to predict and identify individual air pollutant as a sole cause for adverse health effect without sophisticated modelling approaches. The following part of this chapter briefly outlines the various air pollution monitoring approaches used for air quality monitoring and indicates the existing gaps in knowledge about the topic.

2.4 Air pollution monitoring techniques

With consideration to our discussions in the previous section, accurate air quality information can offer tremendous societal and economic benefits by enabling advanced planning for individuals, organisations, and communities in order to reduce adverse health impacts in the urban space and foster urban sustainability. Air pollution monitoring can provide crucial quantitative insight into the pollutants concentrations and their spatiotemporal disposition. However, air pollution monitoring can only describe air quality without imparting clear identification of the source of pollution. Interest in assessing the ambient air quality at the interurban scale has increased in the recent years. Several public health and epidemiological studies have reflected on the importance to account for spatial and temporal variation in air pollution within cities. However, it is not an easy task to estimate varying air quality at a city level as cities are very complex organisations. No two cities are identical, and hence need different approaches in order to assess the air pollution concentration.

Traditional approaches for monitoring air pollution involves setting up networks of fixed stations for precise measurements of air pollution, requiring significant investment that is mainly led by environmental or governmental authorities. The official fixed site suffer from the low spatial resolution constraint of the data, as the monitoring networks of fixed-site monitors have low spatial densities (i.e. the distance between monitors is generally 1–10 km). The concentrations of pollutants in the air can vary significantly within 10–100 m from local variables around like roadways or buildings (Snyder et al., 2013), which may lead to inaccurate assessment of air pollution over the study area. These fixed-site monitoring stations are often located away from roadsides and major traffic congestion areas, which can also cause a localised impact on emissions and pollutant concentration measurements. This sparse arrangement of monitoring stations usually provide a large quantity of data for a wide range of pollutants at detailed temporal resolution, but with limited spatial context. The monitoring stations cannot be placed at all the locations in cities and hence limiting it to few monitoring stations places at critical sites and thus making them the representation of specific microclimates rather than of the city. Hence, makes it difficult to compile thorough representative and reliable information for high-resolution monitoring for a city or area as a whole, and thereby, form a more macroscopic view of pollution field trends for the city. Detailed city level characterisation of air quality cannot be sufficiently attained using the sparse network of air pollution monitors (Mead et al., 2013).

Air pollution modelling approaches are essential tools that allow the determination of relationships between emissions and concentrations of pollutants. These can provide insights into the consequences of past and upcoming space and time scenarios for

air pollution management strategies. Air pollution modelling involves the use of mathematical and numerical approaches to simulate the physical and chemical processes which are endured by pollutants in air and to simulate their impacts due to geographical variables around the measurement site, dispersion and reaction in the atmosphere (Jerrett et al., 2005a; Amoako et al., 2005). Several modelling approaches exist in the literature for air quality modelling, but they are applicable in certain specific conditions. Hence, to understand these specific conditions, it is crucial to know their possible applications, impacts and advantages of each group of models. A better understanding of these modelling approaches and their locus will help in better management of the air quality and its health effects in cities. A variety of modelling approaches has been proposed in the literature to estimate the air pollutants concentration and personal exposure to them. The approaches include proximity measures, linear regression, geostatistical techniques, Gaussian models, artificial intelligence and compressed sensing. Taking into account the discussions in the air pollution monitoring community (Hystad et al., 2011), commonly mentioned techniques involve spatial proximity, geostatistical interpolation, land use regression (LUR) and dispersion models. These all techniques are discussed below, to briefly highlight their characteristics, advantages and limitations.

2.4.1 Spatial Proximity Models

Previous research suggests that the most basic approach to model spatiotemporal variability in intraurban air pollution are spatial proximity models (Jerrett et al., 2005a). Spatial proximity based models are being used in identifying relationships between air pollution and health outcome based on the assumption that spatial proximity of the emission source proxies for exposure in the human living environment (Hoffmann et al., 2006; Dadvand et al., 2014). The two advantageous aspects of proximity metrics are 'their clear relevance to policy' and not so strict requirement for air pollutant measurements (Allen et al., 2011). However, the major drawback of these models is the simple underlying assumption used. The method discards much of the exposure information by proxying exposure with the distance to the source. Also, the vehicle mix may have an influence on emission and this method completely ignore those parameters of air quality measurement. Proximity models take into account the same dispersion assumption in all directions (Ryan and LeMasters, 2007), which somehow also impact the final outcomes.

2.4.2 Interpolation Models

Another method to model air pollution is the spatial interpolation. The method, in general, relies on deterministic and stochastic geostatistical techniques. The models

estimate the value at unmeasured locations as a weighted average of the measurements at the surrounding air quality monitoring station. By applying interpolation methods, spatial maps of air pollutant concentrations are derived. The interpolation methods differ in their choice of sample weights and available monitoring stations measurements. Commonly used methods for air pollution estimation include: spatial averaging, nearest neighbour, inverse distance weights and kriging. These methods are simple to apply, and in this sense, estimates obtained may be more appropriate in an instance where the sampling network is sparse, and errors are assumed to be substantial. With the help of the derived estimation surface, pollutant concentrations can be spatially related to a population or a specific subpopulation. Interpolation methods are suitable to model regional pollution patterns but fail to capture the air pollution variation at a finer scale (Brauer et al., 2003). Problems in this approach can also come up while integrating factors like terrain or localised pattern in other possible predictors. Another problematic characteristic of the method is that a smooth varying concentration surface is created, which may lead to bad coverage of hot-spots in the city like roadways (Hoek et al., 2008). Lack of pollution measurement station data limits the application of interpolation methods (Adams and Kanaroglou, 2016; Singh and Gokhale, 2015).

2.4.3 Dispersion Models

Dispersion models take into account the insights related to the chemical and physical processes and assumptions of the dispersion for explaining the transformation of pollutant considering the emission sources to predict the concentrations, as well as pollutants spatiotemporal variability (Vardoulakis et al., 2003). The dispersion modelling approach requires a different kind of data to estimate pollutants concentration, to be specific topography, emission inventory, meteorological and environmental data. These models have been widely used in road traffic-related pollution prediction by using other road-related characteristics, such as traffic intensity, vehicle speed, terrain, obstruction height, meteorological conditions, etc. The dispersion models usually vary depending on the mathematical procedures involved in developing the model. Most commonly used models for pollutant dispersion modelling are Gaussian-based dispersion models (Lagzi et al., 2014), non-Gaussian dispersion models have also been developed to estimate the pollutant concentrations within street canyons but are less widely used (Mensink et al., 2003; Oftedal et al., 2008).

Dispersion models can help in representing the spatial and temporal differences in pollutant concentrations, without relying on the data from dense monitoring network for modelling. The models are also capable of modelling at different geographical scales, and they can easily be tailored for use in different study areas. However, the downfall of the method is the cost of the data needed for the study

with various assumptions about the dispersion patterns which may not be applicable in real-world scenarios and the need for monitoring station data for extensive cross-validation. Even though a dense monitoring network is not necessary, dispersion models are still very data intensive, and utilisation of these methods requires months of training (Ross et al., 2006; Hao and Xie, 2018). Other problems associated with the dispersion modelling approach include the data requirement of different time periods which can cause estimate errors and the need for extensive cross-validation. In all, these constraints make this modelling approach less suitable for research purposes, especially in the case of growing cities with limited resources.

2.4.4 Land Use Regression Models

The models are based on the principle that the pollutants concentration at any particular location depends on the geographical characteristics of the surrounding area. Land Use Regression (LUR) models are the one that uses least square regression modelling approach to predict the pollution surface based on pollution monitoring data and existing exogenous independent variables to predict the pollution concentration at a given site using information based on surrounding land use, traffic characteristic or topography (Beelen et al., 2013a). These models can then help in predicting the air pollution at the locations of importance like homes (Ryan and LeMasters, 2007), which can further be used for epidemiological and public health studies in the city. Often the predictor variables like traffic intensity, road length, distance to the major road, road type, population density, land cover, wind speed, altitude etc. help reasonably well in explaining the spatial variation in the intracity pollution concentration (Beckerman et al., 2013a; Korek et al., 2017).

One of the major advantages of the LUR models is its applicability to estimating the within-city variability of air pollutants (Poplawski et al., 2009; Kerckhoffs et al., 2017). Much focus is also given on this method in last few years to understand the air pollutant dynamics in the city by integrating it with a different kind of algorithms to make the prediction more efficient (Rahman et al., 2017; Dirgawati et al., 2015; Dons et al., 2013). The competence of a LUR model to estimate pollutant concentration can be further improved by adding more monitoring stations (Wang et al., 2014b). However, few studies reject this argument. They believe that the variability in monitoring station location is a more important factor than the number of monitoring stations. LUR models have often been used for epidemiological studies (Beckerman et al., 2013b; Wang et al., 2013a; Gilbert et al., 2005; Beckerman et al., 2012; Gulliver et al., 2011; Chen et al., 2010; Adam-Poupart et al., 2014). Another major advantage in comparison to dispersion models is that the LUR models are more favourable because of less data intensiveness. Besides, LUR models integrate more factors than spatial proximity models, typically performs better than, or equivalent

to, the geostatistical interpolation models and dispersion models (Ross et al., 2006; Hoek et al., 2008). It is also found that LUR is a powerful air pollution modelling approach and is relatively simple concerning data requirements and analysis (Ross et al., 2006). Regarding cost-effectiveness considering the transferability, LUR models perform well, especially when seen in the context of the budget for large epidemiological studies (Hoek et al., 2008). Hence, LUR models are considered as an important tool for incorporating traffic and geographical information in the air pollution impact assessment within a city (Oiamo et al., 2015).

The main limitation of these models is that it is hard to distinguish between the influence of the different air pollutants. The underlying reason for such limitations is that some of the harmful pollutants like NO_2 and PM_x are usually correlated. This limitation makes LUR models limited to address only one pollutant at a time. Another limitation is the inability of LUR models to characterise large variations in pollutant's concentration over a short distances (Hoek et al., 2008). Although transferability was discussed in the advantages, it can also be seen as a limitation to some extent as the transfer of the model is only possible up to a certain extent. Transferability of the model depends on the similarity between two areas concerning land use and the morphological structure of the city where the model is going to be used (Hoek et al., 2008; Johnson et al., 2010). It is also important to note that the quality of input data profoundly influences the LUR models outcomes.

2.4.5 Hybrid Models

Various studies have also started exploring the applicability of the combinations of two or more air pollution models considering the limitations of individual methods discussed above. Jerrett et al., 2005a gave the example of personal monitoring combined with regional monitoring. Beckerman et al., 2013a developed a hybrid model which combines LUR and Bayesian Maximum Entropy (BME) interpolation. Beelen et al., 2009 developed the methods combining kriging interpolation and regression to extend fine spatial scale air pollution monitoring. Similarly, Akita et al., 2014 developed a modelling framework based on the BME method that integrates monitoring data, the output of LUR, and chemical transport models. Recently, Wu et al., 2018 proposed a hybrid kriging/LUR model to improve the accuracy of air pollution estimation.

One of the objectives of this thesis is to support detailed air pollution modelling preferably by utilising datasets which are easily accessible, openly available. The precedence of LUR models over other air pollution modelling methods makes it an advantageous method for our objective as well as for air pollution impact assessments in general (Michanowicz et al., 2016). LUR models have been used to estimate the

intra-urban small-scale spatial variations in exposure to air pollution in one of the most comprehensive European Study of Cohorts for Air Pollution Effects (ESCAPE) study to understand the impact of ambient air pollution on health (Habermann et al., 2015; Lipfert, 2017). Taking into account the competence of ESCAPE to help in air pollution monitoring, we decided to build upon this method (see Figure 2.1).

2.5 Location optimisation techniques

The spatial representation of air pollution monitoring stations plays a decisive role in describing the impact of air pollution at a particular location. It is important to provide useful measurements from monitors which are placed to observe pollutants under the variety of pollutant level for modelling air pollution. Because monitoring stations cannot be placed everywhere, the monitoring stations network must be designed in such a way that it can be used to predict and forecast pollutant levels for the locations with no monitors. The air pollution concentrations from various monitoring stations are used in the modelling methods discussed before. The models then help in estimating the pollutant concentration at unsampled locations. However, the outcomes obtained by applications of various modelling approaches may inherit aleatoric and epistemic uncertainties due to rough approximations in input data sources (Kumar et al., 2015b). Lack of appropriate air pollution monitoring station measurements as input can impact the application and quality of outcome possible from specific models for detailed air pollution monitoring. The number and spatial representation of monitoring sites are of great importance in identifying the relevance for collected air pollution measurements, and the approaches it can be used in, as one critical aspect of the air pollution data is "*where*" the data is collected (Wang et al., 2014b; Ryan and LeMasters, 2007).

Focusing on a few highly populated locations in the city is an example of conditional sampling. For statistical inferences, sometimes the data collected from such sampling approach are regarded as biased data. Seldom, the air pollution model estimations for the areas with no monitoring stations are used to identify new monitoring station locations. Moreover, if the specific model is calibrated to a particular sample of monitors, then the outcomes that are utilised to inform changes to the future monitoring network design are likely to be biased. In practice, the search for optimal design of monitoring networks involves two possible situations: the design of a new monitoring network and the redesign of an existing monitoring network. Various regulation and actors govern the design of air pollution monitoring stations which also need to be considered with the objective of the study to find the optimal locations (Muller and Ruud, 2018). The optimal designing of a monitoring network can be performed by using one or more objective functions and assign a score

to each of the possible configurations of the designed monitoring network. The monitoring network configuration that minimises the objective function's assign score is considered the optimal network for air pollution monitoring. Several methods can be used to run through the set of locations for monitoring stations to optimise them concerning the objective function. Usually, optimisation algorithms contain random steps, which can lead to varying outcomes using the same data. Sometimes the optimisation algorithms are heuristics that do not provide certainty that the best optimal configuration can be achieved (Helle and Pebesma, 2015). Different approaches were explored to assess the spatial representativeness for monitoring network design and optimisation, for maximising the spatial resolution, coverage and avoiding redundant stations (Righini et al., 2014; Janssen et al., 2012; Santiago et al., 2013; Shi et al., 2018). The collection of various other optimisation approaches can be referred from Mateu and Müller, 2012. Here we are discussing algorithms distinguished as deterministic and stochastic, which were proposed in the literature for optimisation.

2.5.1 Deterministic Approaches

Deterministic optimisation approaches take advantage of the analytical properties of the criterion at hand to generate a sequence of configurations that converge to a global optimal solution. Approaches use the cost function to assess the difference between different configurations to identify the optimal once Kumar et al., 2015a. These approaches can provide general tools for solving the optimisation problems to obtain an optimum spatial configuration. Deterministic search approaches like greedy search have been used commonly for spatial optimisation (Holan and Wikle, 2012; Fussl et al., 2012; Pilz et al., 2012; Helle and Pebesma, 2015). Other approaches like neural network algorithm (Wang et al., 2015a) and several regularisation methods have also been exploited to realise the optimisation goals (Zhang et al., 2017; Ma et al., 2017).

2.5.2 Stochastic Approaches

Concerning stochastic optimisation approaches Behzadian et al., 2009; Kumral and Ozer, 2013 used genetic algorithms, which start with a set of possible sampling configurations, ranking them according to the objective and then later combine the single locations of the good configurations. Hu and Wang, 2011 used Particle Swarm optimisation which was further integrated with Monte Carlo simulations to identify optimal configuration. Spatial Simulated Annealing (SSA) is among the stochastic optimisation algorithm which was commonly used in recent years for optimisation. SSA approach was developed by Van Groenigen and Stein, 1998, which is based

on Simulated Annealing. A standard application of this approach is to minimise the objective function value, Heuvelink et al., 2012 used it for minimising kriging variance, Marchant et al., 2013 used it add samples to delineate areas which need soil remediation, Szatmári et al., 2018 used it for multivariate soil mapping and Barca et al., 2015 used it for redesigning the environmental monitoring networks. Most of the SSA optimisation was implemented for the soil science studies. In Chapters 3, 4 and 5 of this thesis, the SSA optimisation method is used to identify optimal locations for air pollution monitoring in cities.

Quality of life, big data and the power of statistics

Published as: Gupta, Shivam, Jorge Mateu, Auriol Degbelo, and Edzer Pebesma. "Quality of life, big data and the power of statistics." *Statistics Probability Letters* 136 (2018): 101-104.

Abstract

The digital era has opened up new possibilities for data-driven research. This paper discusses big data challenges in environmental monitoring and reflects on the use of statistical methods in tackling these challenges for improving the quality of life in cities.

Keywords: Air quality, Big data, Land use regression , Optimal location, Smart city

3.1 Introduction

Quality of life (QoL) is tied to the perception of ‘meaning’. The quest for meaning is central to the human condition, and we are brought in touch with a sense of meaning when we reflect on what we have created, loved, believed in or left as a legacy (Barcaccia, 2013). QoL is associated with multi-dimensional issues and features such as environmental pressure, total water management, total waste management, noise and level of air pollution (Eusuf et al., 2014). A significant amount of data is needed to understand all these dimensions. Such knowledge is necessary to realize the vision of a smart city, which involves the use of data-driven approaches to improve the quality of life of the inhabitants and city infrastructures (Degbelo et al., 2016).

Technologies such as Radio-Frequency Identification (RFID) or the Internet of Things (IoT) are producing a large volume of data. Koh et al., 2015 pointed out that approximately 2.5 quintillion bytes of data are generated every day, and 90 percent of the data in the world has been created in the past two years alone. Managing this large amount of data, and analyzing it efficiently can help making more informed decisions while solving many of the societal challenges (e.g., exposure analysis, disaster preparedness, climate change). As discussed in Goodchild, 2016, the attractiveness of big data can be summarized in one word, namely *spatial prediction* - the prediction of both the *where and when*.

This article focuses on the 5Vs of big data (volume, velocity, variety, value, veracity). The challenges associated with big data in the context of environmental monitoring at a city level are briefly presented in Section 3.2. Section 3.3 discusses the use of statistical methods like Land Use Regression (LUR) and Spatial Simulated Annealing (SSA) as two promising ways of addressing the challenges of big data.

3.2 Environmental monitoring and big data challenges

With an increasing number of people moving in (and to) urban areas, there is an urgent need of examining what this rising number means for the environment and QoL in cities. Air quality has an effect on the population’s QoL (Darçın, 2014), which is also the major environmental risk factor for health. In 2012, one in eight deaths could be attributed to exposure to air pollution according to the World Health Organization¹. Air quality has high fluctuation at a fine scale due to its very complex

¹See http://www.who.int/phe/health_topics/outdoorair/global_platform/en/ (last accessed: December 4, 2017).

distribution, the structure of the city, and dispersion processes. Institutions such as the European Environmental Agency have produced maps of air quality across Europe. Nonetheless, these maps have two drawbacks: first, their spatial resolution is coarse (i.e., they are usually available for the member state level), and second, they do not give a real-time account of the situation. Projects such as the World Air Quality Index provide real-time air quality maps (see <http://aqicn.org/>), but again they have a relatively coarse spatial resolution.

Data for environmental and meteorological analysis are not only of a significant volume but are also complex in space and time. Formats and types of data are also very diverse (e.g., netCDF, GDB, CSV, GeoTIFF, shapefile, JSON, etc.), and many interconnections prevail within data, which make it complicated for traditional data analysis procedures. Fusing official monitoring stations data with methods like IoT based crowd-sourced data sources can increase redundancy and make data management a serious challenge. Using this example, challenges associated with big data can be illustrated as:

Volume: The large data volume is induced by fusing data from monitoring stations, with crowd-sourcing sensors which can further be integrated with significant environmental data, city dynamics data and other parameters like city land use information. The data size for some variables varies from MBs to TBs (e.g., a single data file for atmospheric data is around 2GB's for a single point of interest). Handling this amount of data needs proper planning; otherwise, the analysis may take longer time because of the mixture of redundant or less relevant data.

Velocity: The speed at which the data from monitoring stations, added sensors and other data sources are created, captured, extracted, processed and stored also needs to be dealt with appropriately. Statistical issues arise from fusing together different data source streams at different spatiotemporal scales. Delay in data fetching from remote storage devices or geographical constraints may also impact the process. Velocity is one crucial characteristic that defines the kind of outcomes we can develop from the data sources.

Variety: Environmental data are in various formats (e.g., NetCDF files for environmental variables, GeoTIFF files for land use, shapefiles of the city for road networks and traffic congestion), which represents heterogeneity challenges, entity resolution issues arising by merging data from different data sources and interaction challenges between big data and data applications.

Veracity: With the variety of data pouring in the analysis, the level of uncertainty also increases. Outcomes expected from the analysis may be affected by some offsets and origin errors of data sources. To maintain data veracity, it is sometimes advised

to discard noisy sources and include only reliable sources. However, ignoring some data points may lead to missing some air quality pattern in the city.

Value: A large amount of data is of no use until it is converted into value. For air quality, the value can be considered as the extraction of intelligence to improve QoL in the city through the development of applications which help city dwellers become aware of their air quality exposure. However, issues such as inefficient handling of large amounts of data, inability to provide quality results on a timely basis, the bottleneck in sharing processed data, high computational cost of big data processing hinder the provision of efficient, easy outcomes for public use.

3.3 Statistics and environmental monitoring

As Scott (2017) said, statistics remains highly relevant irrespective of ‘bigness’ of data. It provides the basis to make data speak while taking into account the inherent uncertainties. Statistical analysis involves developing data collection procedures to further handle different data sources and to propose formal models for analysis and predictions. There are a number of statistical methods varying from sophisticated data requirement (e.g., dispersion models) to simple inference models (e.g., proximity-based models) for air quality prediction. Each of the methods has their specific data and computational requirements. Some methods cannot always be implemented due to the cost, time and resources involved. Notable air quality modeling methods, such as dispersion models, are very sophisticated and require deep insight into the chemical and physical assumptions of the pollutant along with pollutant monitoring sites in the city at a very fine spatiotemporal resolution. The downfall of these methods also includes the cost of the data needed for the study with disputable assumptions about the dispersion pattern (i.e., Gaussian dispersion) and extensive cross-validation with monitoring station data (Jerrett et al., 2005a). The next subsections highlight the potential of land use regression and spatial simulating annealing in addressing both big data challenges, and shortcomings of previous work.

3.3.1 Land use regression (LUR)

Land use regression requires simple geographical variables for predicting environmental factors such as air pollution or sound pollution in the city. It is one of the standard methods used by epidemiologists and health care researchers for exposure analysis. LUR helps in breaking the limitations in developing the models while offering the flexibility to use already available data sources. Regarding performance, LUR-models have been outperforming geostatistical methods and may perform

equally, or sometimes better, than dispersion models (Gulliver et al., 2011). With LUR, researchers can estimate individual exposures from statistical models that combine the predictive power of several surrogates based on their relationship with measured concentrations.

Advantages. The advantage of the LUR approach is the flexibility of incorporating more theoretical knowledge about the process governing the spatial and spatiotemporal variation. This way the challenges due to the addition of new data (e.g., IoT data) can be handled with the context-based variable selection. This restricts the amount of input in the analysis and hence can help in tackling the volume, variety, veracity and velocity data challenges. The perk of LUR also is its ability to run models within raster spatial environments, which allows for a rapid computation. Hence, it can help with challenges related to the value aspect of big data analysis. Another major advantage of the LUR-model over dispersion and interpolation models is to gain the spatial scale desired at the city level. LUR-models are better at describing hot-spots in cities, unlike aforementioned methods which provide smoother concentration maps (Marshall et al., 2008).

Drawbacks. Compared to dispersion models, the LUR method requires less detailed input data at the expense of the need to obtain monitoring data for a sufficiently large number of sites. Moreover, LUR-models have limited capacity to separate the impact of some pollutants because they are collinear to each other, which is the same case of other exposure study methods. LUR methods can benefit from a more systematic selection and description of space-time attributes of monitoring locations.

Building reliable models for big data requires strengthening the sampling process towards locations and time points which can improve predictability. The reliability of methods always depends on the quality of input data. The selection of monitoring sites to develop the air quality models has been identified as one of the factors affecting the quality of models' outcomes. We still lack rigorous methods to determine the number and distribution of monitoring sites (Hoek et al., 2008). Using a large number of monitoring sites to build a model improves its ability to estimate the pollutants. However, improvement in models' predictive power can be achieved by a certain number and specific distribution of monitoring stations. Selecting optimal locations may assist in minimizing data redundancy and can enhance the computational time. Various statistical methods exist for optimizing the sampling process. Here, we discuss the method called "Spatial Simulated Annealing (SSA)" for optimization of an air quality monitoring network.

3.3.2 Spatial simulated annealing for optimizing monitoring network

Placing sensors at certain locations often should fulfill several purposes, and these can be achieved by combining the respective cost functions. By defining the cost function, we try moving each sensor in their neighbourhood cells/locations and find the best places where a cost function can be achieved so that the purpose of placing the sensor is worth. SSA takes into account the spatial neighborhood to optimize spatial sampling schemes based on a defined cost function. During the process, both the size of the movement of sensors around the specified area of interest, and the probability to agree to the worst results decrease with a decreasing annealing temperature. By using this approach, we can decrease the amount of data needed to perform the analysis with optimal results.

Recent works regarding spatial sample configuration optimization using SSA, emphasized on the following aims: (a) conditioned latin hyper cube sampling (Roudier et al., 2012); (b) variogram identification and estimation using constraints like pairs contributing to each lag-distance class (Truong et al., 2013); (c) spatial interpolation using constraints like minimization of the kriging variance in a space-time setting (Delmelle, 2014).

Advantages. SSA enables the specification of various types of optimization goals during the spatial analysis. Once the goal is decided, we can limit the area of interest along with the geostatistical criterion, i.e., we magnify our research goals at certain spatial vicinity for a proposed outcome. This method can take into account the weight of the area we are more keen on collecting knowledge about. By limiting the area and incorporating context-aware goals in location selection processes, we also curb the creation of a large amount of nonessential data, thus helps in overcoming aforementioned big data challenges.

Drawbacks. Based on convergence analysis, different forms of temperature updating functions are followed concerning different kinds of probability density functions employed. The convergence of the objective using SSA depends on the input of appropriate conditions for both the probability density function and the temperature updating function. Calculating these inputs for SSA can be a time-consuming process and needs practical experience. Depending on the objective and size of an area, the processing of the algorithm is intermittently time-consuming too. However, time-consuming processes will pay off at a later phase as they help in selecting the possible best locations for data collection and hence improve the overall flow of the analysis. Table 3.1 summarizes the key points of the combined use of both methods discussed earlier to tackle challenges of big data for environmental monitoring.

Dimension	Challenge	Solution
Volume	Data reduction techniques (Koh et al., 2015)	LUR selects variables and SSA selects the optimal locations, hence reducing data for analysis
Velocity	Quick and constant access of data (Namiot and Sneps-Sneppe, 2012)	By finding optimal locations we can decrease the data to process and accelerate data access
Variety	Creating a knowledge base from different data formats (Koh et al., 2015)	LUR reduces the number of variables to process, possibly reducing the variety of data sets for analysis
Veracity	Avoid inessential data (Koh et al., 2015)	LUR variable selection and SSA optimal location can avoid the inessential data
Value	Complexity restricts timely processing (Villanueva et al., 2014)	LUR selected variables provide context and SSA puts cost function to achieve context-aware outcomes timely

Table. 3.1. Challenges of big data, and potential of the combined use of the proposed methods.

3.4 Conclusions

In this paper, we have focused on the role of statistics in handling the five Vs of big data, and the challenges posed. Big data analytics demands different methodologies from traditional statistical approaches which can enable efficient computer processing and timely outcomes for the efficient use of data. We propose to combine two well-established statistical methods to optimize the selection of variables and locations for spatial and temporal analysis of environmental data sources. The combined use of both methods will help in designing data acquisition processes so that the maximum information can be extracted given a specific number of possible measurement sites. Limiting the data sources can increase the speed of the analysis. The key highlight of integrating LUR and SSA is to make processes like air quality monitoring flexible because LUR can consider limited but accessible data sources. However, we need to consider some crucial aspects. First, variables should be selected carefully and used correctly in the models. Second, the design of SSA-based optimization relies on the quality of input from the LUR for putting weight on those areas for the cost function we want to achieve. It is also helpful in reflecting on the temporal dependency of air quality at a location and the spatial correlation among other locations. Third, SSA needs inputs about probability distributions and temperature change functions which is a critical aspect of optimal location selection. Hence, by using such statistical tools, big data analysis can be effective regardless of the “bigness”. Statistics has been a major component of data analysis for centuries and will be crucial in the era of big data.

3.5 Acknowledgement

The authors gratefully acknowledge funding from the European Commission through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, <http://www.geo-c.eu/>). Jorge Mateu has also been partially funded by the grant MTM2016-78917-R from the Spanish Government of Science and Competitiveness.

Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models

Published as: Gupta, Shivam, Edzer Pebesma, Jorge Mateu, and Auriol Degbelo. "Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models." *Sustainability* (2071-1050) 10, no. 5 (2018).

Abstract A very common curb of epidemiological studies for understanding the impact of air pollution on health is the quality of exposure data available. Many epidemiological studies rely on empirical modelling techniques, such as land use regression (LUR), to evaluate ambient air exposure. Previous studies have located monitoring stations in an ad hoc fashion, favouring their placement in traffic “hot spots”, or in areas deemed subjectively to be of interest to land use and population. However, ad-hoc placement of monitoring stations may lead to uninformed decisions for long-term exposure analysis. This study introduces a systematic approach for identifying the location of air quality monitoring stations. It combines the flexibility of LUR with the ability to put weights on priority areas such as highly-populated regions, to minimise the spatial mean predictor error. Testing the approach over the study area has shown that it leads to a significant drop of the mean prediction error (99.87% without spatial weights; 99.94% with spatial weights in the study area). The results of this work can guide the selection of sites while expanding or creating air quality monitoring networks for robust LUR estimations with minimal prediction errors.

Keywords: air quality monitoring; land use regression; monitoring location optimisation; simulated annealing; spatial mean prediction error

4.1 Introduction

Nitrogen dioxide (NO_2) has been identified as one of the harmful pollutants affecting the quality of life of population (McConnell et al., 2010; Health Effects Institute Panel, 2010). Other pollutants of interest include fine particles ($PM_{2.5}$), elemental carbon (EC), and Ozone (O_3); each has been linked to respiratory diseases and vehicular emissions (Charpin and Caillaud, 2017). Describing the pathways from the generation of emission, dispersion and chemical transformation of pollutants in ambient air is very challenging because of its high variability over space and time (Mayer, 1999). To represent this intraurban variability in pollutant concentrations, various sophisticated exposure assessment methods were used in the recent past Khreis et al., 2017; Conti et al., 2017. There are also several research efforts investigating air pollution modelling approaches using machine learning and other computationally intensive methods Bougoudis et al., 2018; Bougoudis et al., 2016; Feng et al., 2015. Dispersion models that simulate pollutants' dispersion and reaction in the atmosphere are also often infeasible at higher spatial resolution throughout larger areas. These methods require measurements from the monitoring station (directly or indirectly) and data calculated by meteorological prediction models as inputs for predicting the extreme values of air pollutants. However, these approaches are not well suited for situations where data is scarce.

GIS and spatial analysis have increasingly become an essential tool for air pollution monitoring (Briggs, 2005). Interpolation of pollution data collected by regulatory air quality monitoring stations can help in regional patterns, but the air quality monitoring networks are very sparsely arranged to collect informed data at a city level. Land Use Regression (LUR) models are helpful to take into account air pollution variability within the cities. LUR models are a promising alternative to these conventional approaches as they establish the relationship between easily accessible land use characteristics and pollutant measurement. LUR models also require the pollutant measurements collected at the monitoring sites to select the independent variables, but are less data intensive than the other air pollution monitoring methods mentioned above.

The present study introduces an air quality monitoring network design (MND) optimisation method and demonstrates its applicability for the city of Münster (Germany). A LUR model was selected based on Beelen et al. (2013b) to model the NO_2 distribution in the study area. The selection of this LUR model assumes that it asserts the knowledge about the real NO_2 distribution for the study area considering predictors. The predictors of the regression model were then used as the input to the optimisation method. The primary objective for the method developed in this study is to find the combination of locations which minimise the spatial

mean prediction error over the entire study area for two contexts: (1) without using any weighted function; and (2) with a spatial population weighted function for high population density areas. There are two significant innovations for the proposed method. First, the flexibility to identify optimal locations for air quality monitoring devices without the obligation to feed in air quality monitoring station measurements. The method exploits variables as input data which do not include pollutants concentration values (or measurements from monitoring stations), and relies on predictors of an LUR model which are easily accessible for all the locations in the study area. Second, the determination of the locations of new monitoring stations, which can optimally contribute to improve air quality monitoring efforts (and help lessen data scarcity challenges).

The paper is organised as follows. Section 4.2 gives a brief overview of the previous research conducted in this area. In Section 4.3, we describe the study area and the data considered for the optimisation method. We propose the optimisation method adopted as the criterion for air quality MND in Section 4.4. Section 4.5 presents the results of the proposed method. Finally, Section 4.6 points at the limitations of the work, and Section 4.7 concludes the article.

4.2 Related Work

Understanding the effect of pollution on city residents requires a monitoring network that can provide a representative view of the experiences across the population (European Union, 2008; Raffuse et al., 2007). In an urban space, a network of monitoring stations can routinely measure pollutants concentration in space and time. However, placing only a few centrally located monitoring stations is not very helpful for assessing the spatial variability of air pollution over an urban area (Ott et al., 2008). The measurement of spatial air pollutant concentration variation was ineffective with single monitoring stations (Goldstein and Landovitz, 1977). Since the process of air pollution monitoring for capturing spatial variability is costly and time-consuming (Kanaroglou et al., 2005a), it is highly desirable to find the air quality MND which can use an optimal number of measurement devices and reduce costs (Nejadkoorki et al., 2011). In general, the objective of air pollution monitoring network concerns spatial representativeness (i.e., siting criteria, including fixed or mobile sites and numbers of sites), temporal resolution, and accuracy of measurement (Kuhlbusch et al., 2013).

An optimal air quality MND is essential for air pollution management (Mofarrah and Husain, 2010; Wu et al., 2010). The MND can be determined by its placement criteria based on study objectives. The determination of spatial and temporal

variations in the air pollutant concentration and its degradation are two widely used underline consideration for choosing the air quality MND (Benis et al., 2016). MND selection can be considered as an optimisation problem, i.e., searching for the best combination of potential locations (Elkamel et al., 2008). The process of comparing all possible combinations of locations and determining the optimal configuration is practically intensive, especially when optimisation aims involve multiple variables and multiple objectives for the vast area. An heuristic approach such as spatial simulated annealing (SSA) can be helpful in solving the optimisation problems for location selection at different scales (Wang et al., 2015b; Van Groenigen et al., 1999). Many studies have extended the optimisation criterion for various other constraints, such as protecting sensitivity receptors (e.g., population districts and historic buildings), response to emission sources nearby, and minimising costs (Sarigiannis and Saisana, 2008; Kao and Hsieh, 2006; Benis et al., 2016). Because of these increasingly complex constraints, optimisation of MND for monitoring air pollution is considered a demanding task (Wang et al., 2015b).

Although various approaches discussed previously have been put forward to make valuable progress, only a few have focused on the challenge of setting up a new monitoring network, and of revision and redistribution of an existing air quality MND considering their optimisation for robust LUR estimation. The primary goal of developing LUR is to predict air pollutant values, and this can help in deriving information about the exposure to harmful air quality at the city level. The evaluation of LUR models has been limited to the number of measurements available in a monitoring campaign (Hoek et al., 2008). In some cases (such as in our study with two), no or very few air quality monitoring measurement stations exist. Selection of new monitoring sites without air pollution data create an issue, as one aspect of the accuracy of the monitoring devices and related estimations deals with the aspect of "where" the data is collected.

Few studies indicated that the robustness of the LUR model to predict air pollution can be improved when more monitoring sites can be considered (Marshall et al., 2008). Other studies argue that the variability of monitoring sites is more important than the number of monitoring sites. This implies that a LUR model can be more robust when a large variety of land use characteristics are considered for monitoring network design (see Ryan and LeMasters, 2007). In a recent work, Wu et al. (2017) pointed at two current gaps in research of LUR development: the lack of rigorous methods to determine the number of monitoring sites, and the lack of systematic ways of finding out the distribution of these monitoring sites. Wu et al. (2017) also examined the aspect of tackling the number and distribution for monitoring sites required to develop LUR, but the selection criteria for optimal location selection was judged based on manual land use entities rules. The question of *how to systematically*

place monitoring stations of an air quality monitoring network is the main focus in this article.

4.3 Material

The method was applied to the city of Münster in North Rhine Westphalia, Germany (51.96° N, 7.63° E). Münster is one of the 42 agglomeration areas and one of Germany's biggest cities in terms of area (303 km²), as shown in Figure 4.1a. The city is divided into six administrative districts with nearly half the city's area being agricultural, resulting in a low average population density of approximately 900 inhabitants per km², but with a population density of about 10,000 inhabitants per km² in the city centre. For analysis, the whole city is divided into 599 grid cells of 1 km², as shown in Figure 4.1b.

As often happens for a mid-sized city in Europe, air quality data are collected by only two monitoring stations for the whole study area (European Union, 2008), and this is clearly insufficient for performing any meaningful geostatistical analysis. Moreover, various other air quality models such as CHIMERE, LOTOS-EUROS, and ENSEMBLE models (Colette et al., 2017) for retrieving air quality data were not very helpful, because the resolution of these models were meagre compared to the area of study. We can only represent nine cells of these standard models output for our area of interest (as shown in Figure 4.2). Since the data were not available for more than two monitoring stations in Münster, we added 14 simulated locations as existing air quality monitoring locations ($n = 14 + 2 = 16$) in the MND during the study. The simulation of these 14 locations was based on the argument of Ryan and LeMasters (2007) for considering different land use types, and a special focus on populated areas in city (see Figures 4.1 and A.1). We assume that the knowledge about selected LUR in the study can represent the actual state of the NO_2 concentration in the study area and hence predictors selected in the regression can be used to judge the performance of the proposed optimisation method.

In recent years, LUR has been a choice for various epidemiological and health studies to explain exposure (Khreis et al., 2017). In addition, this method demonstrates a better or an equivalent performance to other geostatistical methodologies such as kriging and conventional dispersion models (Hoek et al., 2008). The structure of the LUR method varies depending on the study objective and available data (i.e., number of variables, different distances and buffer sizes) due to location and size of the areas (i.e., city, state, and nation). In any case, a model typically involves information about traffic, road type & length, land use, and population. LUR uses multiple linear regression to derive the relation between air pollutant concentration and known

variables at a location or its surroundings. It is then applied to unmeasured locations to predict a spatially resolved average air pollution values. The regression model is of the form :

$$y = X\beta + \epsilon \quad (4.1)$$

where

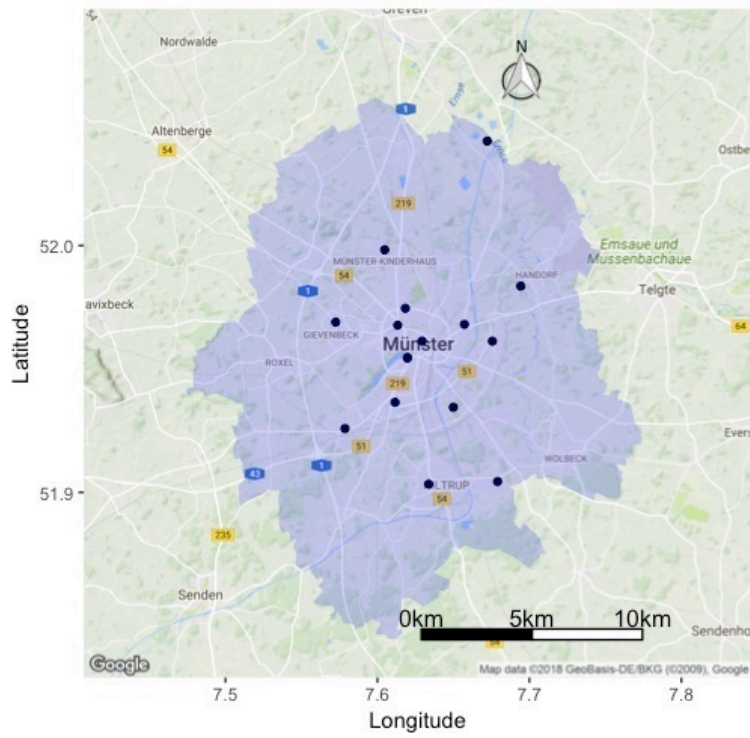
- y is an $n \times 1$ vector of air pollution concentration from monitoring sites at any particular time (in our case annual mean NO_2 concentration at monitoring stations);
- X is an $n \times k$ matrix with observations of k independent variables for the n available air pollution monitoring stations;
- β is a $k \times 1$ vector of unknown parameters that we want to estimate; and
- ϵ is an $n \times 1$ vector of errors, assumed to be independent and identically distributed.

The main strength of LUR is the empirical structure of the regression mapping and its relatively simple input/low cost (compare to dispersion modelling). A well established study for LUR model, European Study of Cohorts for Air Pollution Effects (ESCAPE), has been applied to develop local LUR models for 36 European regions for NO_x , and nitrogen dioxide (NO_2) (Beelen et al., 2013a), to explain the impact of long term exposure of air pollution on health. Methods were also hypothesised to improve the accuracy and prediction power of the LUR models using various approaches such as random forests (Brokamp et al., 2017). However, the number and distribution of monitoring sites to build LUR models were identified as one of the critical factors affecting the quality of LUR outcomes in previous studies Wu et al., 2017. Various methods were adopted in the recent past which can help in determining the number and distribution of monitoring sites (Benis et al., 2016; Wang et al., 2015b). Few previous studies also evaluated the impact of the number of monitoring sites on LUR model results Basagaña et al., 2012; Wang et al., 2012b, but the influence of the spatial distribution of monitoring sites remains poorly understood. Since monitoring stations cannot be placed everywhere, the monitoring network must be used in a way that it represents critical pollutant levels where there are no monitoring stations. A viable alternative to address the constraint of monitoring station data availability could be to consider the spread of predictor variables which are available for all the locations in study area for identifying the optimal locations which can help in improving LUR robustness.

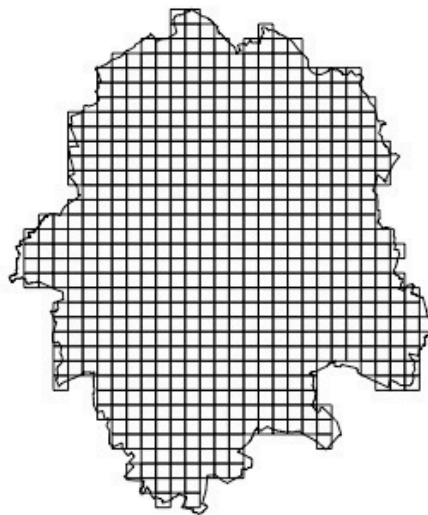
For our study, a LUR model was selected, assuming it represents annual average NO_2 for the study area, pursuing the standardised predictors selected from the ESCAPE study (Beelen et al., 2013b). GIS analyses were conducted to derive the discrete values of predictors at all potential monitoring station locations included in this study based on the selected LUR model. The predictor values were derived using the open geospatial data gathered from various sources, e.g. road network and building datasets were obtained from OpenStreetMap (OpenStreetMap contributors, 2017). These datasets were used to calculate the total length of major roads, minor roads and building counts in varying buffers sizes of 25, 50, 100, 300, 500, 1000, and 5000 m. We also calculated the distances to the nearest major or minor roads data for each monitoring location. The city council of Münster provided traffic intensity data on major roads. Minor roads traffic intensity was calculated by spreading the 25% of the total vehicles in the city based on its length. CORINE land cover data of 2012 were used to extract the land use information around the monitoring station in the buffers. The land cover classification from CORINE was regrouped into high-density residential land, low-density residential land, industry, port, urban green, semi-natural and forest areas. We calculated the surface area of each land use regrouped in each buffer considered in the study. After model selection, predictors involved (Table 4.1) were further utilised in the study for identifying the combination of optimal locations for air pollution monitoring network.

Table 4.1. LUR variables selected

Variable	Variable Description
rdcount_1000	Road count in 1000 m buffer
minrdcount_100	Minor road count in 100 m buffer
minrdcount_500	Minor road count in 500 m buffer
rdlength_100	Road length count in 100 m buffer
rdlength_5000	Road length count in 5000 m buffer
rdlength_50	Road length count in 50 m buffer
mjrdlength_300	Major road length count in 300 m buffer
dist.mjrd	Distance to major roads
minrdlength_5000	Minor road count in 5000 m



(a) Administrative boundary and monitoring station locations in the study area.



(b) Study area divided into 599 grid cells.

Figure. 4.1. Study area: City of Münster.

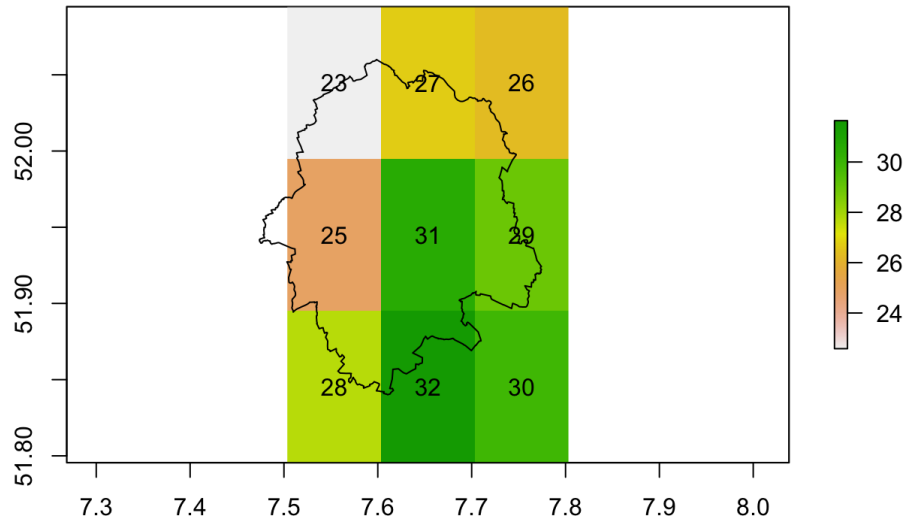


Figure 4.2. NO_2 concentration ($\mu g/m^3$) map predicted by CHIMERE model as of 20 October 2017 for Münster.

4.4 Method

Our knowledge of air pollution monitoring is largely based on limited data. The present paper takes a new look at MND using a new optimisation method which identifies the potential locations in the study area. The optimisation methods were developed to identify the new MNDs by random perturbation over potential locations, considering associated predictors within the LUR models. The random selection of the monitoring sites, and the optimisation of their spatial distribution in the study area was implemented using Spatial Simulated Annealing (SSA).

4.4.1 Spatial Simulated Annealing (SSA)

Having an objective before initiating the monitoring campaign is helpful in distributing the sensors optimally. The objective must be in agreement with the aim of the study. If the objective requires monitoring the excess level of pollutants, one approach is to locate the monitoring sites in the study area where pollutants concentration is higher. Usually, such design of the network is in agreement with the intention to monitor the maximum exposure values, and it does not represent the whole picture about the air pollution in the city. Air quality varies on a comparatively small scale, as the pollutant concentration at a particular place relies heavily on local emission sources, structures and atmospheric flow conditions (Britter and Hanna, 2003; Ott et al., 2008; Rösli et al., 2000). The location selection for MND can be optimised numerically using approaches such as geostatistical surveys, where probability sampling is not considered (De Gruijter et al., 2006). Finding the optimal locations for monitoring networks manually is practically impossible

given the extraordinary number of possible combinations, even with the coarse grid than we considered for our study. This process can be automated by using the spatial numerical search algorithm called spatial simulated annealing (SSA). The method implements optimisation by generating a sequence of new possible monitoring locations. A new monitoring location is considered by selecting random locations and shifting them in a random direction over the random distance in two dimensional space defined in the optimisation process. After each such perturbation, the mismatch between the current MND's discrete value and the observed new MND's value is quantified for combination of locations according to a predefined optimisation criterion function. In the 2000s, Van Groenigen et al. (1999) introduced a simulated annealing algorithm as a method to find the spatial distributions of monitoring points based on an optimisation criterion function. Over the years, the approach has been used for investigating optimal sample designs for criterion as variogram estimation (Lark, 2002), thinning existing sampling networks (Boer et al., 2002), optimisation of geostatistical surveys (Brus and Heuvelink, 2007), phased sampling designs for containment urban soil (Wang et al., 2014a), and Kriging with External Drift (KED) variance minimisation for rainfall prediction (Wadoux et al., 2017), to name just a few.

The SSA approach to finding an optimal monitoring design involves the minimisation of the optimisation criterion value using an objective function such as the average kriging variance across the area of interest (Van Groenigen et al., 1999). The optimal location identification problem is formalised as a single-objective optimisation function, pointing at criteria involving discrete-valued variables. For some monitoring design objectives, the optimality of a particular combination of monitoring sites may be judged based on more than one criterion, considering not only the quality of the resulting information obtained but also the cost of acquiring the data.

4.4.2 Optimisation Criterion Estimation

The optimal MND is the configuration for which the criterion chosen is minimal. The criterion we used in this study is to minimise the spatial mean prediction error for the selected LUR model. Under the assumption of independent and identically distributed (IID) errors, the (ordinary least squares) estimator $\hat{\beta}$ of β is a linear combination of the observations y :

$$\hat{\beta} = (X'X)^{-1}X'y \quad (4.2)$$

and the predicted pollutant concentration \hat{y}_o for a new location with predictor variables x_o is given by

$$\hat{y}_o = x_o\hat{\beta} \quad (4.3)$$

with x_o the $k \times 1$ vector with predictor variables at that location.

The estimation error in β is

$$\text{Var}(\hat{\beta} - \beta) = (X'X)^{-1}\sigma^2 \quad (4.4)$$

and hence the error in \hat{y}_o is

$$\text{Var}(\hat{y}_o - y_o) = \text{Var}(x_o\hat{\beta} - x_o\beta) = x_o\text{Var}(\hat{\beta} - \beta)x_o' = x_o(X'X)^{-1}x_o'\sigma^2 \quad (4.5)$$

with σ^2 the variance of ϵ .

Considering a two-dimensional study area A , we want to decrease the prediction error using n observation sites using a network design D^n . The optimisation process starts with a existing MND (can also be done using randomly selected monitoring design) $D_0 \in D^n$, consisting of observation points s_o, \dots, s_n with corresponding predictor variable vectors x_o, \dots, x_n . At the first optimisation step, the monitoring sites are transformed into a random vector with only one element different from the initial, yielding a new monitoring design D_1 . During the optimisation process, the prediction error is calculated at each node of the rasterised study area A .

The different monitoring sites in the MND are supposed to cover a diverse range of representative areas. Therefore, to optimise the MND considering all the sites in MND, we average over all potential locations s_o inside the study area A . Since σ^2 acts as a constant, we leave it out and use as criterion:

$$\frac{1}{|A|} \sum_{x_o \in A} x_o(X'X)^{-1}x_o' \quad (4.6)$$

with $|A|$ the size of the area (expressed as the number of grid cells representing the area). In this optimisation criterion, manipulating the MND leads to the modification of X .

The typical aim for air pollution monitoring is to accurately measure air quality and understand the impact of ambient air pollution on population. However, because of the limited number of possible monitoring station sites, we need to prioritise the MND optimisation in a way that it can help in collecting precise information about air quality in the residential area of the cities. With this aim, we fine tuned the above mentioned optimisation method to identify the optimal locations which can help in

decreasing the spatial mean prediction error prioritising areas where the population resides. This newly extended optimisation criterion can be written as:

$$\frac{1}{|A|} \sum_{x_o \in A} P(s_o)(x_o(X'X)^{-1}x_o') \quad (4.7)$$

where $P(s_o)$ is the population at location s_o . We established the weight for our study by focusing on the areas with more than 1500 houses per grid cells (see Figure A.1 in Appendix A).

We use as scenario a setting where (due, e.g., to budget and infrastructure constraints) the number of air quality monitors deployed n are fixed. We aim to find the optimal locations for the specific n monitors, which can help in developing the robust LUR model for cities. The optimisation method developed in this section will allow us to compare the different possible combinations for optimal MND. In this study, we only consider a static design for the monitoring network, i.e., locations of the network do not change over time. With a finite number of candidate locations N derived by discretising the whole study area A , we can get N^n combinations, and select the one which minimises the optimisation criterion. Every time a new possible MND is generated, the optimisation criterion mentioned earlier is calculated and compared with the criterion value of the previously selected monitoring design. The new location will be accepted if the MND leads to a smaller optimisation criterion overall. If the MND leads to a larger optimisation criterion, the new design might sometimes be accepted, based on the probability of acceptance defined in the parameters of the annealing algorithm, expressed as:

$$P(\text{acceptance}) = e^{\frac{e(\text{old}) - e(\text{new})}{\text{temperature}}} \quad (4.8)$$

where *temperature* is a control parameter which controls the number of remaining iterations. The temperature defined in the annealing algorithm decreases from a positive starting value to zero as the count of iterations increases. From Equation (4.8), we see that, at a given temperature, the larger the increase in criterion value, the smaller will be its probability to accept a worse network design. In addition, for lower values for temperature (at larger number of iterations), the probability of accepting a worse design becomes lower. The repetition for a new design creation continues until the total number of defined iterations has been achieved. We refer to the work of Heuvelink et al. (2010) for a more detailed explanation of the optimisation algorithm used in our study.

4.4.3 The Optimisation Procedure

The steps of the optimisation algorithm are the following:

1. A LUR model is chosen for the study area by selecting predictors that explain the air pollutant considered.
2. An initial (possibly existing) monitoring design D_0 is defined, consisting of n observations to be optimised.
3. The area of study A is discretised, the raster is defined with N raster nodes.
4. The optimisation criterion (proportional to average prediction error over A) is calculated using Equation (4.6) or Equation (4.7);
5. D_0 is modified, returning a monitoring design D_1 and the new mean prediction error is calculated
6. A new monitoring design is accepted if it reduced the optimisation criterion value or rejected as the basis for further optimisation based on Equation (4.8).
7. The optimisation continues until the proposal monitoring designs are no longer accepted, based on energy transition and iteration parameters.

The schematic overview of the proposed optimisation method described above is shown in Figure 4.3. All geospatial and statistical calculations for the study were conducted using the R packages `sp` (Pebesma and Bivand, 2015), `sf` (Pebesma, 2017). For running SSA, we used the R package `spsann` (Samuel-Rosa et al., 2017). The source code for the proposed optimisation method can be accessed from Github (Gupta, 2018a).

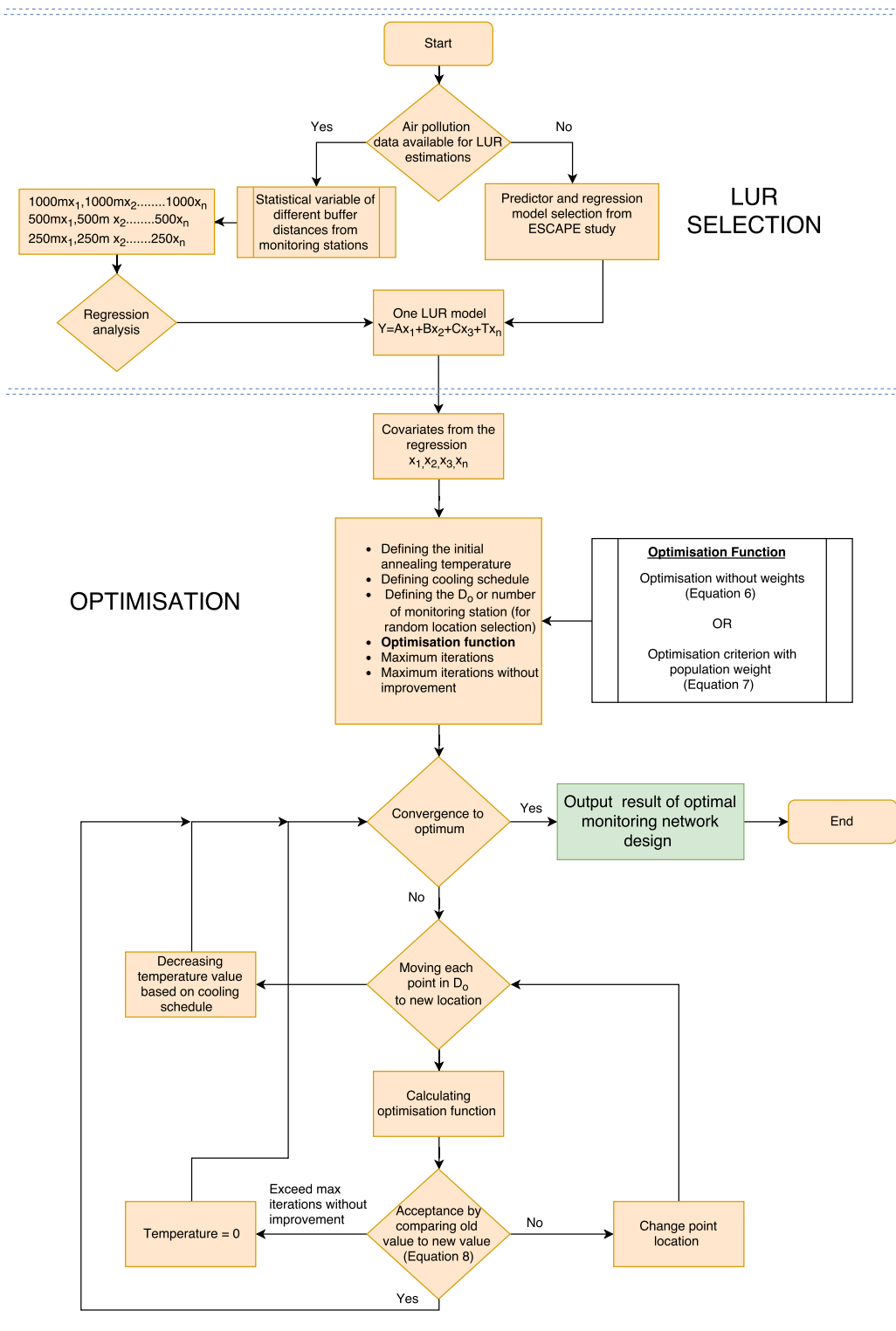


Figure. 4.3. Schematic overview of the proposed optimisation method. Since no LUR regression model was available for the study area at the moment of the analysis, the LUR model from the ESCAPE study was used in this paper.

4.5 Results

Nine (out of 58) predictor variables were used from the selected LUR model (Table 4.1). The variables were selected based on the existing knowledge about their relationship with the pollution concentration data and their inclusion in previous LUR models (Beelen et al., 2013a). We used all 599 raster cell as the potential locations for monitoring stations and the associated values of selected predictors at those locations in SSA to find the optimal MND. The inputs for running both of our optimisation algorithms require a raster grid with potential sites (599 cells) for monitoring stations, the predictor of the LUR as covariates for each raster cell, an initial monitoring network design D_0 with the location of existing monitoring stations, the temperature and the probability of acceptance parameters for annealing.

4.5.1 Optimisation without a Weighted Function for the Study Area

The optimal MND identification started with $n = 16$ monitoring network points, based on already existing monitoring network design (D_0). The method iterated to select any MND which can minimise the criterion value recognising the optimisation method (Equation 4.6), and representing a minimum spatial mean prediction error for the study area. Figure 4.4 represents the criterion values achieved at different probability parameter values while keeping other annealing parameters constant during the whole process. Figure 4.5 shows that several worsened designs were accepted in the beginning, succeeding that the mean prediction error steadily decreased. After about 3500 iterations, no substantial further reduction was achieved, indicating that the algorithm reached a nearly optimum configuration design, as was confirmed by observing the similar patterns of decline at different runs. Overall, the defined criterion was considerably dropped from 156.5 to 0.1918595 with 0.3 as probability of acceptance, which shows an improvement of about 99.87%.

It can be inferred from the results in Figure 4.6 that the final optimised monitoring locations have a larger density around the city centre than in the initial monitoring design, and this is consistent with the population-based optimisation criterion for MND defined earlier.

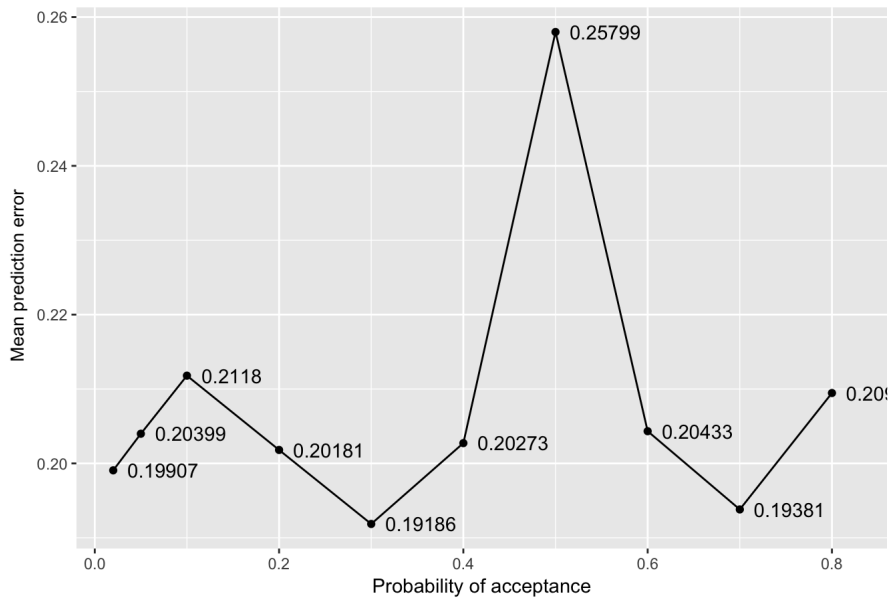


Figure 4.4. Spatial mean prediction error achieved by SSA at different probabilities of acceptance using the optimisation method without weights.

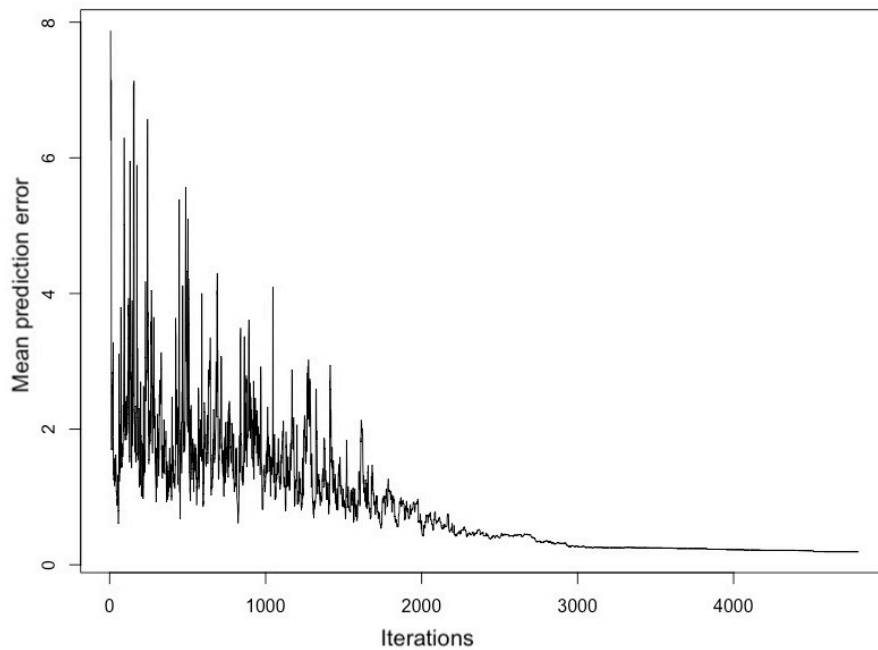
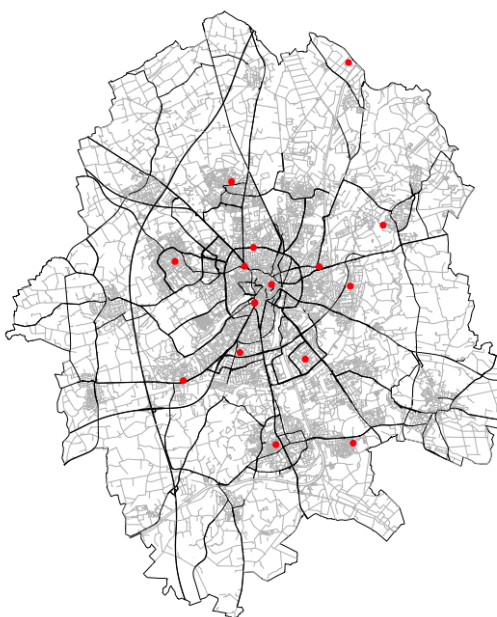
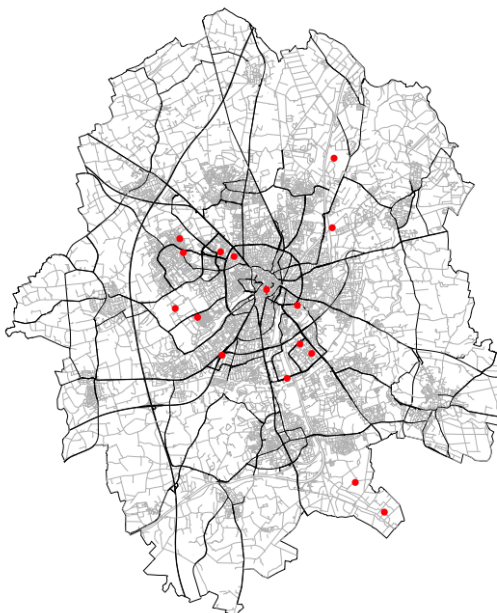


Figure 4.5. Energy transition while running optimisation in SSA using parameters of 0.3 probability of acceptance after removing five higher values.



(a) Initial monitoring network design (D_0).



(b) Optimised monitoring network design after using criterion.

Figure. 4.6. Monitoring network designs realised after using the first optimisation criterion.

4.5.2 Optimisation with a Population Weighted Function for the Study Area

After realising the optimal MND that represents the least spatial mean prediction error for the study area, we considered focusing on the areas with high population density (calculated based on the number of residential buildings around each potential location for the monitoring stations). Equation (4.7) was used as a criterion for implementing the weighted optimisation method.

Again, the configuration was considered optimal based on the least energy (spatial mean prediction error) obtained during the calculation of the criterion value at various probabilities (Figure 4.7). Figure 4.8 shows the optimised monitoring design obtained after running the algorithm. As a comparison with Figure 4.6b, an alternative design was obtained using the weighted spatial areas of highly populated housing areas (see Figure A.1). The new monitoring design has a close resemblance to the previous optimal configuration but this time with more emphasis on the area with weight (green areas in the Figure 4.8b represents populated housing areas) than on areas with no weight (or non-housing areas). The optimised MND represents a significant drop in the spatial mean prediction error, from 571,492.23 to 332.8651, a difference of about 99.94%. The optimal configuration obtained at various other probability of acceptance presented graphically can be found in Appendix A.2. The optimised monitoring networks also placed the monitor locations towards the boundary of the study area. This is a well-known effect while running SSA, expressed by Van Groenigen et al. (1999) in the literature.

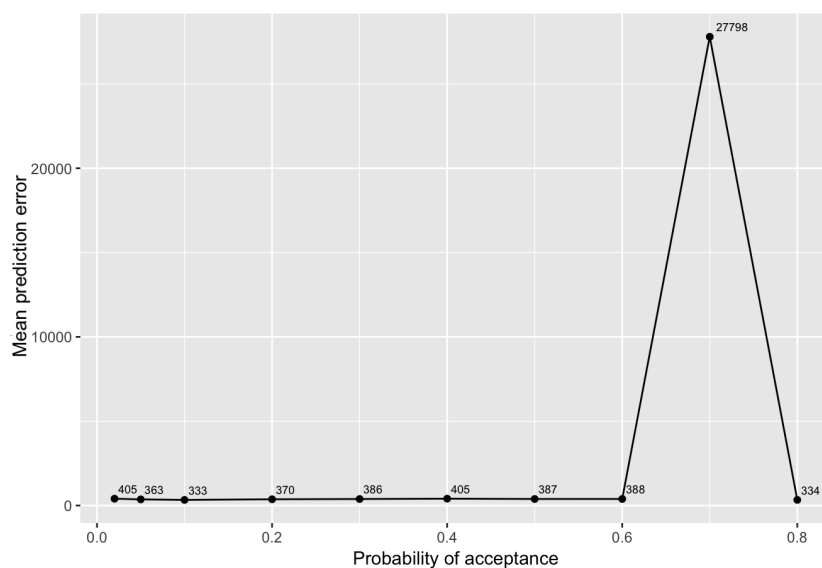
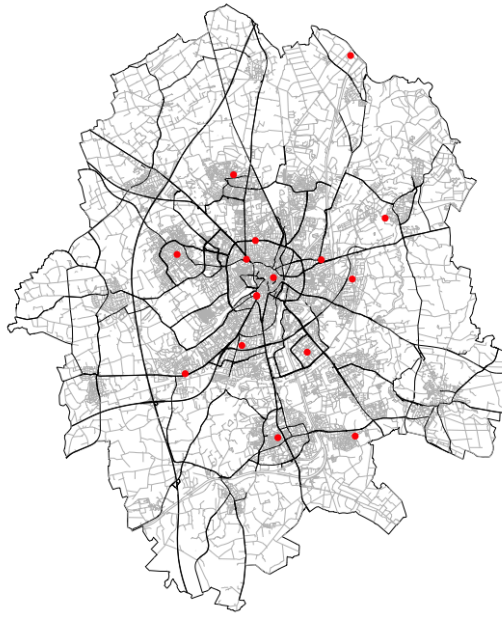
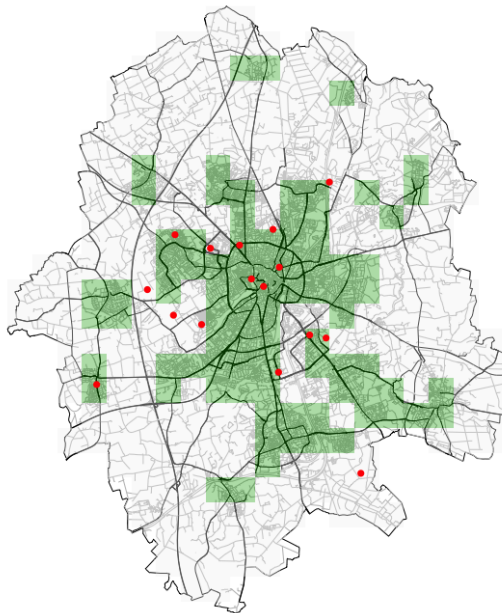


Figure 4.7. Spatial mean prediction error achieved by SSA at different probability of acceptance using optimisation method with population weighted criterion.



(a) Initial monitoring network design (D_0).



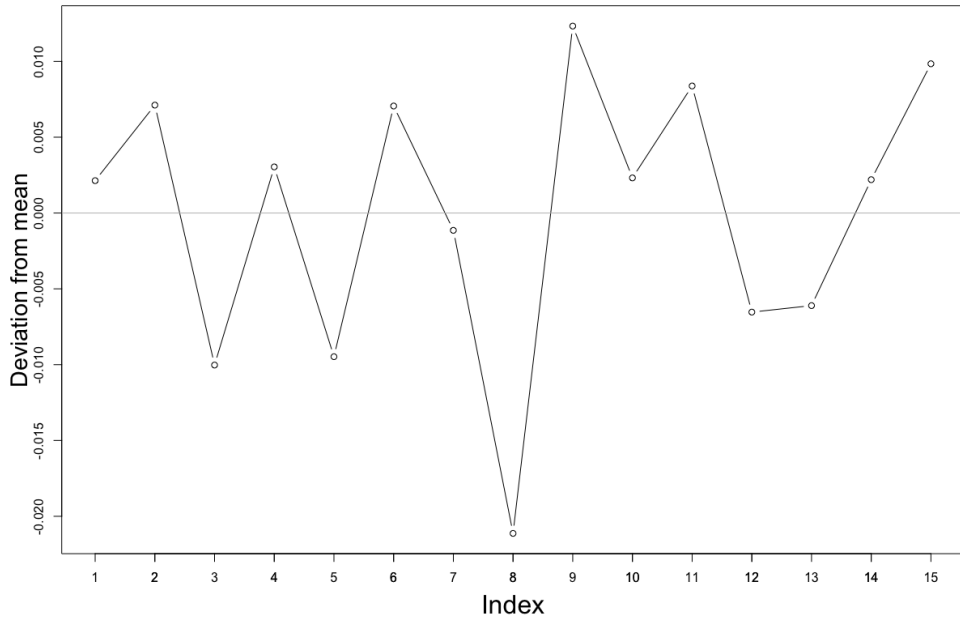
(b) Monitoring network design after population weighted optimisation.

Figure. 4.8. Monitoring network designs obtained using a population weighted optimisation criterion.

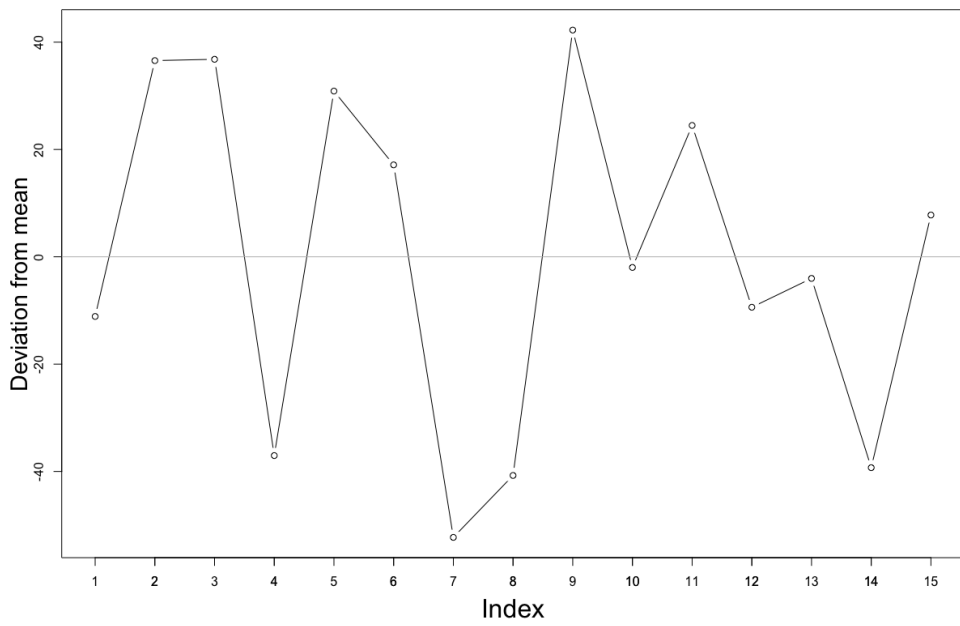
4.5.3 Sensitivity of the Optimisation Methods

The distribution of optimal monitoring sites at various locations and its criterion values can be influenced by various parameters or with different iterations. Indeed, SSA is a stochastic algorithm and does not generate a unique spatial distribution of monitoring stations at each run (i.e., algorithm execution). To investigate the variation in final criterion values for each complete process of optimisation, we ran the algorithm 15 times each for both without weighted optimisation and with population weighted area optimisation. The 15 runs for each involve exactly the same parameters and the probability of acceptance as the run with the least optimisation criterion values according to the graphs of previous runs (see Figures 4.4 and 4.7) (i.e., 0.3 probability of acceptance for non weighted and 0.1 probability of acceptance for weighted area, respectively). Figure 4.9 shows the results of the investigation for change in optimal criterion values. The resulting energy state for the optimisation criterion without weights returned values with mean 0.2058 and standard deviation of 0.0091 or 4.42%. In the case of the population-weighted optimisation run, the criterion values obtained were with the mean of 398.030 and standard deviation of 31.53 or 7.3%. These small deviations from the mean values can be considered as insignificant in comparison with the improvement of around 99% in spatial mean prediction error for both the methods.

Besides the variation of final criterion values per run, the sensitivity of our method to the number of monitoring stations in the optimisation method was also investigated. For this, we changed the numbers of monitoring sites in the input parameters. Figure 4.10 compares the criterion value obtained at a different monitoring station involved in the optimisation method. We again used the same optimisation method parameters with the least criterion values of 0.3 and 0.1 probability of acceptance from the results obtained in previous runs (Figures 4.4 and 4.7) for this investigation. More monitoring stations yielded better results; the spatial mean prediction error decreased with an increase in monitoring sites. However, we kept other parameters such as initial temperature, chains and temperature change for all the optimisation runs unchanged. It would be interesting to investigate the influence of the change in parameters and number of monitoring stations on the outcome of the optimisation method.

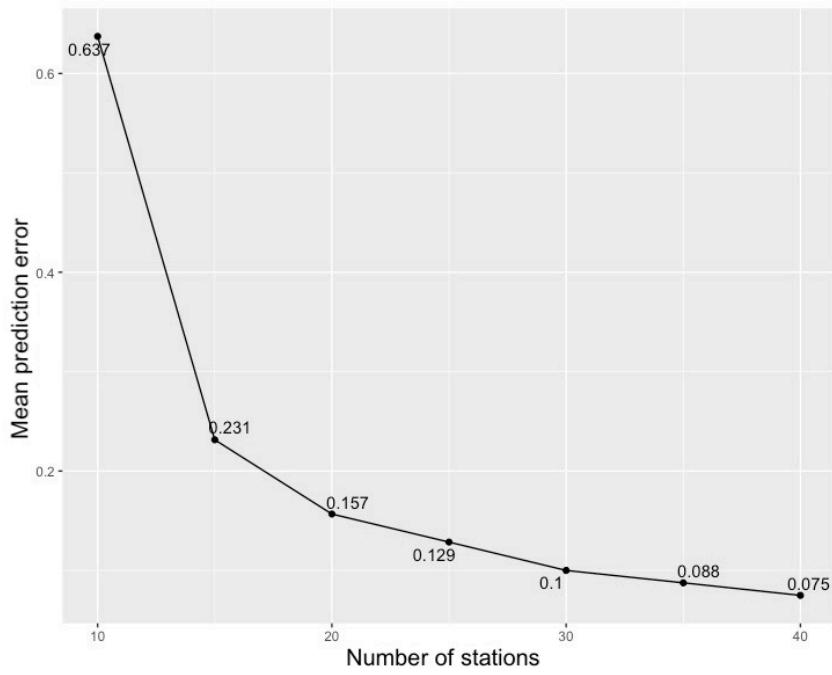


(a) For optimisation without weights.

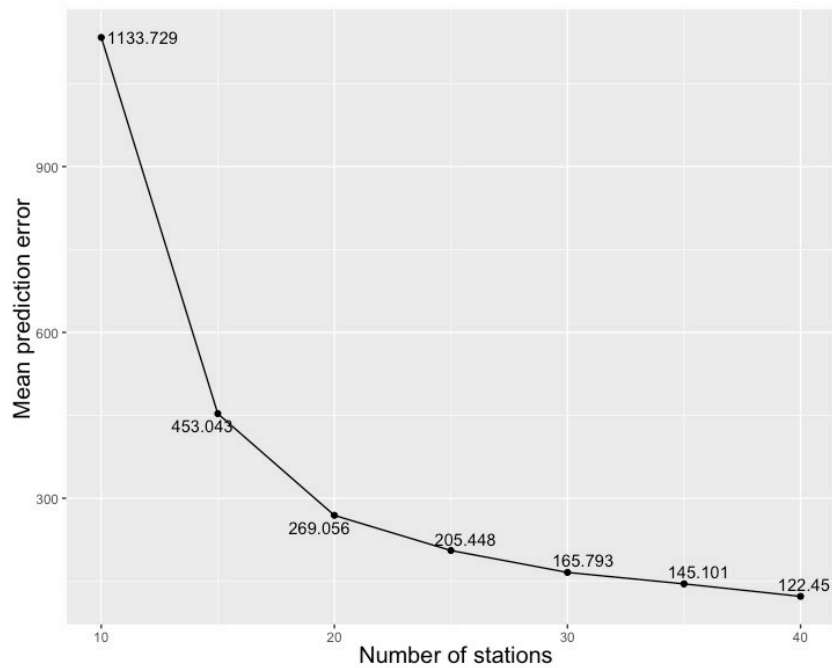


(b) For population weighted optimisation.

Figure 4.9. Deviation of the spatial mean prediction error values from mean value obtained after 15 repetitions with same parameters.



(a) Using optimisation method without weights (0.3 Probability of acceptance).



(b) Using population weighted optimisation method (0.1 probability of acceptance).

Figure. 4.10. Summary of least spatial mean prediction error values obtained for different numbers of monitoring stations.

4.5.4 Comparative Analysis

From the above results, we can infer that the combinations of locations obtained from our optimisation methods employed in the selected study area give satisfactory results towards the development of a reliable, less assumption based, less data intensive, easy to use and low-cost methods for identifying the optimal location for air quality monitoring at a city level. The proposed methods also consider the enhanced efforts to anticipate areas of more importance (such as residential areas) in cities for planning air pollution monitoring network, without using pollutants monitoring station values as inputs.

Since LUR have been a useful exposure estimation method in various epidemiological and public health studies. Location optimisation considering such methods can be useful to understand exposure in cities. At this point, it could be useful to compare the results of the proposed methods with results from previous similar research works (Sarigiannis and Saisana, 2008; Kao and Hsieh, 2006; Benis et al., 2016; Mofarrah and Husain, 2010; Wu et al., 2010; Wu et al., 2017), where other strategies were adopted to identify the optimal locations for air quality MND. The main difference with these approaches is that the proposed optimisation method used LUR predictor variables values which are available for all the candidate locations of the study area for optimal location identification, rather than the computationally intensive datasets inputs from dispersion models (which sometime are not accessible at the city level). The proposed method tries to address the fundamental input data required for air quality monitoring initiatives with the significant decrease in the assumption parameter involvement in the previous studies (such as dispersion model or emission sources and its assumptions) for location optimisation. In addition, the proposed method is more flexible and low cost compared to others, as it uses the easily accessible geospatial data for LUR estimation. The approach is also less data intensive compared to other existing approaches for optimal design of MND (Kao and Hsieh, 2006; Sarigiannis and Saisana, 2008; Wu et al., 2010). Moreover, few of the methods in the literature are suitable for application on a smaller scale, such as cities, because of spatial resolution limits of the input datasets (Benis et al., 2016; Wu et al., 2017). As illustrated in Section 4.5.2, the current method does not suffer from this drawback, and is applicable at the city level.

4.6 Discussion

In this paper, we have presented an optimisation method that can help identifying the optimal MND for robust LUR estimations. The method utilises the predictors selected in the regression model for optimal location identification. The optimisation

method shown in this study was initially developed to select the combination of locations which can represent the least spatial mean prediction error in air pollution estimation without giving weights to any specific regions in the study area. Figure 4.6 presented the outcome obtained by applying this optimisation method. Furthermore, to consider the relevance of precise air pollution information close to population, the initial optimisation method was further adapted to prioritise the specific regions of importance in the study area. The weights for the population were added to the optimisation method for identifying the optimal combination of locations that minimise the spatial mean prediction error while prioritising populated regions in the study area. SSA was used to implement the optimisation. Figure 4.8 showed the results of the modified optimisation method for the study area. Overall, the proposed approach can be of interest to air quality management authorities, researchers trying to monitor the air pollution, particularly if considering LUR estimation methods, and to the city councils to better collect air quality data in future. As discussed in Gupta et al. (2018b), it has the potential of addressing big data challenges in environmental monitoring. The method can also be of use in other domains, such as sound pollution studies, which also utilise land use regression estimations.

The standard criteria used for ambient air quality assessment are the number and locations of monitoring stations. The number of monitoring stations and locations affects the degree of detail for pollution monitoring across regions. According to the EU directive of 2008 (European Union, 2008), a minimum of one monitoring station per million inhabitants over agglomerations and additional urban areas of more than 100,000 inhabitants is required for placing air quality monitoring stations. A large number of EU member states seems to follow the strict minimum regarding number of monitoring stations, and this leads to a low resolution air quality data inventory for cities. Previous work has shown that using a limited number of monitoring stations in cities will not be sufficient for determining the patterns of air pollutants due to their complex behaviour (Wang et al., 2015b). Given that the collection of ambient air quality data is not possible at all locations in the study area, optimising the distribution of monitoring stations can help in collecting precise representative information for all the locations in the study area.

Randomly selecting monitoring station locations in the area could result in a clustered or dispersed design which may be ineffective and not representative for the real aim of the study. The efficiency of spatial monitoring design can be increased by embedding prior knowledge about the random field (Wang et al., 2012a). According to Wu et al., 2010, the design of monitoring network involves three significant considerations: (1) determining the design criteria; (2) estimating the concentration of pollutant; and (3) solving an optimisation problem. The proposed optimisation methods favour all three considerations. Firstly, it takes into account the design criterion which focuses on identifying the combination of locations for decreasing

the spatial mean prediction error in the area. Secondly, the selection of MND in the optimisation method is dependent on the linear function of predictors which estimate air pollution concentration. The optimal design will decrease the error for the estimated air pollution values based on predictors involved in the optimisation process. Figure A.2 in Appendix A.1 shows the histograms of various predictors used in the present optimisation study. It is worth mentioning here that the selection of predictors for optimisation may differ from those considered in the ESCAPE study if real data about monitoring stations is available or the number of monitoring stations involved in the initial monitoring design (D_0) changes for area-specific LUR estimations. Lastly, the proposed optimisation methods in the current study can help in solving the problem for robust LUR estimation for the study area by identifying the optimal location underlying the specific LUR model. The weight-based optimisation method also supports solving the optimisation problem based on a specific goal of the study by prioritising specific regions in the study area. The final optimal MNDs obtained from the proposed optimisation method were successful in selecting a combination of the broad range of locations (such as roads or residential areas) while also giving higher priority in the area-weighted optimisation. Thus, we complement the finding of Wu et al. (2017) regarding the advantage of mixed site MND for LUR exposure analyses.

This study optimised MND by minimising the spatial mean prediction error while using a given LUR model. To the best of our knowledge, such an approach has not been used before concerning LUR based air quality monitoring network optimisation. It is essential to find the optimal locations for the case of future monitoring campaigns plans to readjust the existing network or to develop a new MND for the city. The two significant advantages of the proposed optimisation method are: (1) flexible covariate integration, which allows the integration of other possible variables of interest for optimisation; and (2) autonomy to monitoring data, hence avoiding dependencies on monitoring data for identifying locations for MND. Overall, the flexibility offered by the proposed methods can be helpful in developing the optimal MND for the area with no or negligible amount of air pollution monitoring data. Furthermore, we would like to emphasise that this method only applies to optimisation of the MND based on an already selected LUR model. In case of unavailability of air pollution monitoring data for estimating LUR, various already existing standard LUR models (e.g., ESCAPE (Beelen et al., 2013b)) regression models can be considered for optimising the MND as we did in our study. This particular advantage can help in setting up the air pollution monitoring network from a very early stage for underlying a selected LUR model. Another significant advantage of the proposed methods can be to overcome the limitation of the LUR models concerning transferability. According to Hoek et al. (2008), transferability of LUR models depends on the similarity of the area regarding land use. On the other hand, Johnson et al. (2010) stated that LUR models are not transferable most of the time. With the help of the proposed

method, it is possible to initiate air quality monitoring considering the specific LUR of interest hence making it transferable. Based on the study focused on transferability by Allen et al. (2011), it was suggested that locally calibrated models performed better than the transferred model. The proposed method can be used as a tool for transferring the selected model and re-calibrating it locally by optimising the locations of monitoring station using predictors. Furthermore, the method provides resilience for increments and decrements of weights as per the aim for distribution of monitoring sites.

4.6.1 Limitations and Future Work

There are also limitations to the proposed method. First, the selection of a LUR model is vital for implementing the optimisation method, which means that, in the case of unavailability of monitoring station data for LUR model estimation, the regression model needs to be assumed from previous studies. In this case, the selection would be arbitrary, and may not provide a correct representation of the air pollution variability in the area of study. The availability of data and selection procedures for predictors selection in regression model also impacts the outcome. Second, the underlying assumptions of multiple regression concerning linearity between dependent and independent variables in the regression, independence and normal distribution of error term may create bias in the interpretation of the final results, which are the typical limitations for many simplistic LUR based studies. Third, the initial MND (D_0) considered in the study is comprised of 16 monitoring stations, of which 14 locations are simulated, and 2 originally had a known location. These two stations, in reality, will not be considered for relocation easily based on various objective functions. Further research should be dedicated to fine tune the existing criterion functions which can restrict perturbation of permanently located monitoring stations but only allowing optimal location identification for additional monitoring stations (Van Groenigen et al., 1999). These limitations also highlight the difficulty of having proper data about air pollution and concerning predictors. There may also be several other sources which can inherit errors, concerning the open data and simulated locations of monitoring stations used in the study. Although many optimisation methods have already been developed, the flexibility offered by our method provides room for more insights to be considered for optimisation in the future. For example, taking into account the geographical information, preliminary observations, and information on the spatial correlation could have helped in improving MND optimisation strategy (Beelen et al., 2009; Wadoux et al., 2017; Brus and Heuvelink, 2007). We have not considered using such optimisation constraints in the current study, but they could be integrated during future research.

Future work will include the testing of the method with other datasets in different cities. As to the potential of low-cost sensors to increase the spatial coverage of air pollution monitoring (which has been a matter of discussion in recent years, see Clements et al., 2017b), the proposed method can be further extended for decisions making process about where one should deploy low-cost sensors for air pollution monitoring in cities. Further data collection is required to determine precisely how air pollution monitoring network can be developed and optimised considering various LUR models for higher resolution air pollution monitoring in the study area. Further studies can also be carried out, utilising enhanced data collection procedures including systemic crowd-sourcing approaches and higher resolution remote sensing data from the various missions (such as Sentinel-4 and -5P), to enable air pollution monitoring efforts at the higher spatial scale in cities. Thus, encouraging the efficient monitoring of air pollution distribution and gathering information about the possible exposure of the population can serve as a base for improving environmental sustainability and urban health.

4.7 Conclusions

LUR models provide the opportunity to take into account within-city variability of air pollution concentration for epidemiological and public health studies. In the present study, we aimed at improving the robustness of LUR by identifying the combination of locations which can decrease the spatial mean prediction error for air quality estimation in the cities. A statistical optimisation method was developed to optimise locations in the study area. The initial version of the optimisation method focused on identifying the locations for MND which can help in representing the air pollution estimates with minimal spatial mean prediction error considering an area of interest. This version was then further modified to include the weighted function of the population to determine optimal locations which can represent the estimates with least spatial mean prediction error in high density populated spaces of the study area. The methods require all predictor variables in selected LUR to be known at all of the potential locations for calculating their significance in the optimisation process. The optimisation method does not rely on monitoring station data for monitoring site placement, thus giving independence for planning and readjustments of the optimal air quality MND for the cities with no or insignificant amount of air quality data. Furthermore, we demonstrated that, by distinguishing between weighted areas, the optimisation method could be a helpful tool in air quality monitoring and exposure studies. The proposed optimisation method is an efficient way to achieve air quality estimates with minimal prediction errors to understand air pollution variability and support the sustainable air quality control efforts for the urban spaces. One possible extension of the method could be the inclusion of Wu and co-workers' work on selecting the number of monitoring stations required and their optimal

design for robust LUR estimation. Moreover, the possibility to prioritise particular areas of interest may be considered as a useful control in the air quality control and exposure assessment related decision-making processes. Additional further work could include assessing how the method performs when provided with various quality of LUR models and data sources for a given urban area as input. In all, the proposed optimisation method can be a helpful tool in air quality MND that enables LUR estimations with fewer errors for preventing air pollution exposure and advancing urban health sustainability.

Author Contributions: Shivam Gupta and Edzer Pebesma conceived the optimisation method. Shivam Gupta developed and performed the main proportions of the study. Edzer Pebesma and Jorge Mateu provided advice regarding the optimisation method and performance. Auriol Degbelo supported in shaping the presentation of the paper.

Acknowledgments: The authors gratefully acknowledge funding from the European Commission through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, <http://www.geo-c.eu/>).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations: The following abbreviations are used in this manuscript:

MND	Monitoring Network Design
LUR	Land Use Regression
ESCAPE	European Study of Cohorts for Air Pollution Effects
OECD	Organisation for Economic Co-operation and Development
WHO	World Health Organisation
PM	Particulate Matter
EC	Elemental Carbon
O ₃	Ozone
CORINE	coordination of information on the environment

Optimisation of VGI based Air Quality Monitoring Networks for Cities

Based on: Gupta, Shivam, Edzer Pebesma, Auriol Degbelo, and Ana Cristina Costa. "Optimisation of VGI based Air Quality Monitoring Networks for Cities." *ISPRS International Journal of Geo-Information*, (Under Review).

Abstract

Air quality has had a significant impact on public health, the environment and eventually on the economy of countries for decades. Effectively mitigating air pollution in urban areas necessitates accurate air quality exposure information. Recent advancements in sensor technology and the advent of volunteered geographic information (VGI) open up new possibilities for air quality exposure assessment in cities. Yet, citizens and their sensors are put in areas deemed to be subjectively of interest (e.g. where citizens live, school of their kids or working spaces), and this leads to missed opportunities when it comes to optimal air quality exposure assessment. In addition, while the current literature on VGI has extensively discussed data quality and citizen engagement issues, few works, if any, offered techniques to fine-tune VGI contributions for an optimal air quality exposure assessment. This article presents and tests an approach to minimise land use regression prediction errors on citizen-contributed data. The approach was evaluated using a dataset (N=116 sensors) from the city of Stuttgart, Germany. The comparison between the existing network design and the combination of locations selected by the optimisation method has shown a drop in spatial mean prediction error by 52%. The ideas presented in this article are useful for the systematic deployment of VGI air quality sensors, which can aid in the creation of higher resolution and more realistic maps for air quality monitoring in cities.

Keywords: Air quality monitoring, sensor location optimisation, crowdsourcing, citizen engagement, volunteered geographic information, Land Use Regression, Spatial Simulated annealing

5.1 Introduction

Air pollution is currently a global fret, which can be linked to the extensive population growth and urbanisation, together with their aftereffect in traffic, industrialisation and energy consumption (Molina et al., 2004). Human health is closely linked to the air we breathe (Barer, 2017), as evidence from recent studies for the adverse health effects has shown (Brown and Bowman, 2013; Bauernschuster et al., 2017). A recent report of the World Health Organization (WHO) report suggests that 92% of the world's population live in places that exceed the recommended annual mean concentrations of Particulate Matter ($PM_{2.5}$) WHO, 2016. Because of the growing health effects of chronic exposure to ambient air pollution, policy makers and scientists are showing an increased interest in monitoring air pollution at a higher spatial resolution. Various recent studies from spatial epidemiology and public health have set out a specific interest in traffic based pollution (Jerrett et al., 2007; Hamra et al., 2015; Khreis et al., 2017), particularly in Stuttgart, Germany (Bauer et al., 2018). Generally, air pollution monitoring is done by environmental or government organisations using a network of fixed monitoring stations. Typical regulatory decisions are taken based on long duration temporal trends and statistics (Conti et al., 2017), while considering certain conditions related to hotspots are estimates based on real-time data, if available. Interpreting the pathways from the generation of emission, dispersion and chemical transformation of pollutants in ambient air pollution concentrations is very challenging due to its high spatiotemporal variability (Mayer, 1999). In the recent years, land use regression (LUR) is widely used in various health and epidemiological studies to estimate air pollution at a finer spatial scale in the urban areas (Nunen et al., 2017; Wolf et al., 2017; Weichenthal et al., 2016). However, due to economic reasons, the number of air quality monitoring stations in cities is usually sparse and limited, and this considerably limits an accurate assessment of the intraurban variability of air pollution.

Citizens and environmental agencies are exploring the potential of small, low-cost air quality monitoring sensors to enable detailed real-time information on air quality in the city (Jiao et al., 2016; Snyder et al., 2013; Yi et al., 2015). Several low-cost sensor deployments have been conducted in recent years extending from citizens investigating air quality in the houses and surrounding areas, to networks of sensors to measure community-level air quality, to a vast network of sensors covering the cities (Jiao et al., 2016; Shusterman et al., 2016). However, the datasets provided by low-cost sensor network are argued for less accuracy (Fang and Bate, 2017; Castell et al., 2017; Schneider et al., 2017a). Despite such limitation, the demand for sensor technology is high, driven by widespread concern about the air pollution as well as an interest in reducing the personal exposure (Clements et al., 2017b). While crowdsourcing approaches for air quality data gathering and related technologies are

escalating, research to inform the translation of low-cost air quality sensor data into real application remains limited. The term "low-cost" may be interpreted differently depending on the end users and the specific purpose of the study. For instance, U.S EPA Tier 3 instruments can be low cost (2000-3000 USD) for regulatory authorities but not for general citizens who are willing to participate in the data collection process (Watkins, 2013). Therefore, in our study, we refer to low-cost sensors as a device which cost less than 200 Euros that can be used by individuals or community for air quality monitoring.

In order to capture the spatial variability in detail, accuracy of data will profoundly be relating to "where" the data is collected. To better understand exposure in microenvironments at a right level it is crucial to take into account the spatial coverage of air pollution monitoring networks. Inappropriate location selection may lead to over or underestimation of pollution originated from various emission sources in the city. When considering low-cost sensors to gather air quality data, previous studies suggest that generally, the datasets generated with the help of citizen or community participation approaches inherit serious data gaps and the measurements collected are from irregularly spread sensors (Schneider et al., 2017a). Since the process of air pollution monitoring to capture spatial variability involves specific cost and time (Kanaroglou et al., 2005b), it is desirable to optimise the monitoring locations. Hence, to overcome the data gaps and irregular spatial spread of sensors to make data collection more efficient, there is a need for methods that can help in extending wide-spread and optimal location identification.

Participatory data approaches can be helpful to enable detailed air quality data collection, but exploiting the datasets generated from these low-cost devices requires tools and techniques for data cleaning and processing. The vast amount of dynamic, varied, detailed, and interrelated datasets from citizen participatory approaches could be enhanced by preparing the protocols and infrastructure that enables scientifically sound data collection (Bonney et al., 2014). The deployment of systematically spread low-cost sensors for urban air quality monitoring can be useful for air quality data collection. With the potential of low-cost air quality monitoring sensors to increase the spatial coverage (Elwood et al., 2012), along with its application to foster participation (Clements et al., 2017b), it is desirable to systematically identify the optimal placement of monitoring stations to make the best use of advanced sensor technology and citizen engagement efforts.

The present paper aims to develop a method which can help in systematically identifying the optimal locations of citizen sensors for air pollution monitoring. The method is tested using citizen-contributed data from the city of Stuttgart, Germany as a scenario. The primary objective of the optimisation method includes the identification of the most advantageous spread and optimal monitoring site locations

that minimises mean prediction error for Land Use Regression (LUR) estimations of air quality parameters for the study area. LUR is a method for spatial exposure assessment. It helps to model pollutant concentrations (e.g. particulate matter, nitrogen dioxide) at any location using various environmental characteristics of the surrounding area (e.g. traffic intensity, elevation, land use type). The spatial simulated annealing (SSA) algorithm was used to run the objective function for finding the optimal monitoring network design.

The main contributions of this study can be summarised as follows :

1. we formulated the optimisation method that can help in the placement of low-cost citizen sensor with an aim to minimise the mean prediction error of the air quality estimation method: Land Use Regression (LUR). Based on our understanding, this is the first formulation of the low-cost sensor placement problem in the context of systematic location identification for improving LUR estimations
2. we reflected on the properties of the objective function which takes into account the wide-spread network aspect along with the objective to decrease the mean prediction error for a given LUR model during the optimisation process.
3. we provided a case study of the city of Stuttgart to demonstrate the applicability of our approach

While existing works on analysing the quality of volunteered geographic information have mostly aimed at examining the degree to which a fact contributed by a volunteer is likely to be true (see e.g. Goodchild and Li, 2012), this work approaches the question of quality of VGI from a slightly different angle. By trying to find the spatial distribution of volunteers which can minimise the global prediction error of the air pollution monitoring network, this work intends to inform the coordination of VGI efforts for air pollution monitoring at the city level. As such the method proposed can be classified as belonging to the fourth type of VGI validation process identified in (Sieber and Haklay, 2015), namely ‘measure of fitness by way of completeness’ (not the amount of points, but the promise of detail or spatial extent). There are a couple of methods in the literature to tackle the VGI quality aspects of positional accuracy, thematic accuracy, and topological consistency, but a general lack of methods addressing other aspects such completeness, temporal accuracy, or vagueness, as a recent review by Senaratne et al. (2017) reminded. Criteria such as road density (Haklay, 2010), or errors of omission/commission (Jackson et al., 2013) were used as surrogates for completeness in previous articles, but completeness in this work is approximated using the combination of two criteria: spatial spread and minimal prediction error of the air pollution monitoring network.

The rest of the paper is organised as follows. In Section 5.2 we present a brief overview of the previously done work on the topic. Section 5.3 describes the study area and the data used in the study. A new methodology for optimisation is described in the Section 5.4. In the next Section 5.5 we present the results and discuss the objective function used in the proposed optimisation method. Section 5.6 presents the discussion regarding the developed optimisation method. We draw the conclusion in Section 5.7.

5.2 Related work

The deployment of the network of air quality monitoring stations is of vital importance for various air quality monitoring methods. Various air quality methods and their regulation exist in the literature, and the adoption of crowdsourced air quality data for filling data gaps has also been a critical discussion in recent years, especially for the vision of the smart city and citizen engagement. This section briefly discusses previous related work on the topics of crowdsourced data integration approaches and air pollution monitoring.

5.2.1 Citizen participation/ VGI

The effect of pollution on city residents requires a monitoring network that can provide a representative view of the experiences across the population while considering the wide distributions of pollution levels and socioeconomic location conditions across the monitoring sites. Low-cost sensors are the technologies that can be helpful in advancing air pollution monitoring by gathering a massive amount of spatiotemporal air quality data. Various low-cost air pollution sensors have already been successfully integrated into long-term deployments to access fine-grained air pollution information (Yi et al., 2015). In practice, these data sources can help in facilitating ongoing indications of changes in air quality, rather than absolute measurements (Gabrys and Pritchard, 2018). The applicability of low-cost sensors for future air pollution monitoring is well recognised in literature (Snyder et al., 2013; Castell et al., 2017; Clements et al., 2017b). Extending the application of these sensors by involving citizens or communities for environmental data collection has increased in the recent past (Jovašević-Stojanović et al., 2015; Clements et al., 2017b). Through volunteered geographic information (VGI) or crowdsourcing data methods, a large number of individuals may be engaged to collect data about phenomena impacting the city life. In general, (and as indicated in Lisjak et al., 2017) the involvement of citizens not only provides an opportunity to close data gaps but also brings the policy-making process closer to people. Citizens are willing to get involved in air pollution monitoring studies and get aware of the

ambient environment (Clements et al., 2017b; Budde et al., 2017). With the help of citizen participation, hundreds of low-cost sensors can be dispersed in an urban environment that can facilitate data collection simultaneously. This gathered data can promote the development of improved models that can explain the pollutant variability within the urban environment.

Education and involvement of communities for air quality monitoring is not only crucial for improving public health but also for building awareness about the sources of air pollution, exposure causes and impact of other pollutants on health. Engaging citizens may also support in deploying the network of low-cost air quality monitoring sensors that can be of significant potential for improving the spatial coverage of pollutant's variability in urban space and can foster citizen participation (Clements et al., 2017b). Despite these advantages, while utilising these alternative data sources, attention is needed to undertake valid capturing and representation of the data. The design of the air quality monitoring network is of vital importance for extracting precise and detailed spatial variability information of air pollution in the city. Most of the datasets generated with the help of citizen or community participation approaches suffer from serious data related gaps and the measurements collected are usually from irregularly spread sensors, which may represent the air quality for only a small number of areas (Schneider et al., 2017a). The irregular distribution of air quality data acts as a barrier in utilising such observations for air quality mapping applications for the cities. Another important consideration while monitoring detailed air quality in the city involves the selection of monitoring sites and the number of sensors involved in an air quality monitoring network. The number of sensors involved and their locations can affect the expected outcome of the air quality modelling approaches which utilises the data specific modelling approaches (Hoek et al., 2008; Kanaroglou et al., 2005b; Wang et al., 2012b; Basagaña et al., 2012).

The selection of monitoring site is challenging because of various parameters such as local land use type, emission source, electricity connections, installation requirements of the equipment and aim of the study. Increasing the number of sensors in the monitoring network also increases the costs and efforts for data process and information generation. Systematic location and size selection for sensor network deployment can be a useful consideration to gather the optimal amount of data with the proper spread as per the fitness of purpose. It is not practically possible to gather air quality measurements at all the locations in a city, nor is it required. Few carefully chosen locations which can fit the purpose of the study with the specified number of sensors in the network can be helpful in representing the air quality for the city in detail. The necessity of formulating the requirements for low-cost sensor network deployment for the specific purpose and at a specific location is essential. Clements et al. (2017b) recommended the identification of the research

question as one of the key consideration for planning the deployment of the citizen participation based air quality sensor networks. Their discussion suggests that the research question and pollutant(s) of interest should govern the size and locations of the air quality sensor network. For instance, if the aim of the study is to reliability measure air quality in a city over a large area (which is the primary focus of our study), sensor locations are important but also the data redundancy aspect, pollutant variation and sensor density within the network (Clements et al., 2017b). For the systematic deployment of the low-cost air pollution monitoring sensor network, we could combine the crowdsourced data location selection with scientific models and their variables, to achieve better spatial coverage.

5.2.2 Air quality monitoring methods

Geospatial tools have become a useful tool for modelling air pollution in the recent years. To represent the intraurban variability of pollutants, various exposure assessment methods were proposed (Khreis et al., 2017; Conti et al., 2017). Approaches include interpolation of fixed-site monitoring stations, dispersion modelling, remote sensing, land use regression (LUR), proximity and various other deterministic methods (Hystad et al., 2011). Each method has their inherent limitations that may restrict their application for developing detailed air pollution maps for the cities. For instance, the dispersion models that simulate pollutants dispersion and reaction in the atmosphere are often infeasible at higher spatial resolution throughout larger areas (Jerrett et al., 2005a). The interpolation of pollution data collected by regulatory air quality monitoring stations can help in regional patterns identification, but the networks are very sparsely arranged to collect informed data, hence, limiting their application for detailed air pollution modelling (Hoek et al., 2008). Over the years, LUR modelling has demonstrated better or equivalent performance to other geostatistical and dispersion methods and is therefore being considered for application in various epidemiological exposure studies (Hoek et al., 2008). However, the scarcity of sensor data may impact the outcome of LUR models. In order to address the sparsity and scarcity challenges, robust and compact systems which can be wide-spread are desired to capture the spatiotemporal variations of air pollutants (Peng et al., 2014).

Usually, the measurement of air pollution in urban space is possible with the help of a network of air quality monitoring sites. In practice, EU states are required to comply with the directive, framework and legal requirement for assessment and management of ambient air quality as described in Air Quality Directive 2008/50/EC (European Union, 2008b). The methods for monitoring air quality currently involve the use of fixed monitoring station networks in the European cities. Monitoring of air pollutants is primarily performed using analytical instrumentation, such as optical

and chemical analysers. However, installation of single monitoring stations will not help effectively in monitoring air pollution (Goldstein and Landovitz, 1977), nor will the placement of monitoring stations ad-hoc or in few centrally located areas be adequate to infer the pollutants' detailed spatial variability in a city (Ott et al., 2008). The air quality maps observed presently are very scarce as the analysers used in the observation network are complicated, bulky and expensive, together with a significant amount of resources required to maintain and calibrate them (Chong and Kumar, 2003). These constraints lead to the low number of air quality monitoring stations which are generally not adequate to capture the small-scale spatial variability of air pollutants in the urban environment. As said previously, recent advancements related to sensor technologies have resulted in relatively low-cost and small devices for measuring air quality (Borrego et al., 2016; Spinelle et al., 2017). The emergence of low-cost devices was also recognised by the policymakers and was recommended to be embodied in the air quality directives (Borrego et al., 2015; Watkins, 2013). Current air pollution monitoring networks can benefit from the use (and efficient spatial distribution) of these low-cost sensing devices in the context of volunteered geographic information.

In all, crowdsourced data/volunteered geographic information has a significant potential to improve current air pollution monitoring networks, notably through the provision of new data points to expand their spatial coverage. The spatial arrangement of sensors usually influence how well the spatial distribution of air pollutants and their impact can be captured (Wang et al., 2012b). Although various approaches have been proposed to identify the optimal locations for air pollution monitoring sensor placement (Benis et al., 2016), a limited number of studies have focused on general location selection aspect for deployment of low-cost sensors (Weissert et al., 2017b; Kim et al., 2017). To our knowledge, no previous studies have considered the application of the optimisation method approach for systematically locations selection for crowdsourcing air pollution data for robust LUR estimations for the cities.

5.3 Material

5.3.1 Study Area

The proposed method was applied for the city of Stuttgart (48.7758° N, 9.1829° E), which is the capital and largest city of the state of Baden-Württemberg, Germany. The city of Stuttgart, is also the capital of Baden-Württemberg state (pop. 11 million, 36,000 square kilometers) and the Administrative Region of Stuttgart (pop. 4 million, 11,000 km^2), is located in the centre of the very densely populated southwestern

Stuttgart Region (population 2.7 million, 3,700 km²), close to both the Black Forest and the Swabian Jura. It covers an area of 207.35 km² and lies in a bowl-shaped valley about 270m above the sea level on the back of the Neckar river. The city centre is situated in a lush valley, ringed with vineyards and forests, and the river and has a population of 62,8032 (as of 31 December 2016) (Baden-Württemberg, 2018). Air pollution is a severe concern in the city due to its topographic conditions and industrialisation. Few German newspapers also called the city of Stuttgart as “the German capital of air pollution” (Deutsche Welle, 2016b). In 2016, the city authorities issued alert, first-ever warning in Germany concerning air pollution (Deutsche Welle, 2016a). Currently, city environmental protection authorities are utilising four monitoring stations to gather data about air quality in the city (City of Stuttgart, 2018). Measurements from three out of four stations are available as open data. Figure 5.1 shows the locations of stations whose data is openly available on Umwelt Bundesamt portal (Umbelt Bundesamt, 2018). Figure 5.2 presents the study area with the crowdsourcing sensor network.

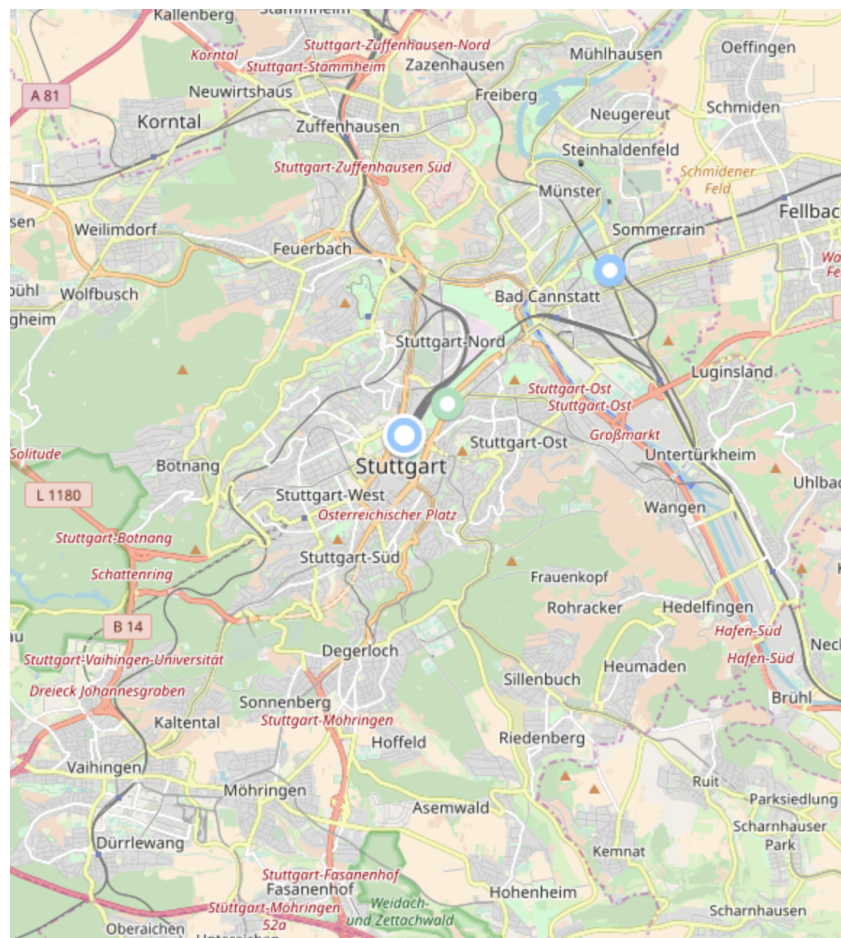


Figure. 5.1. Stations in Stuttgart (Source: Umwelt Bundesamt)

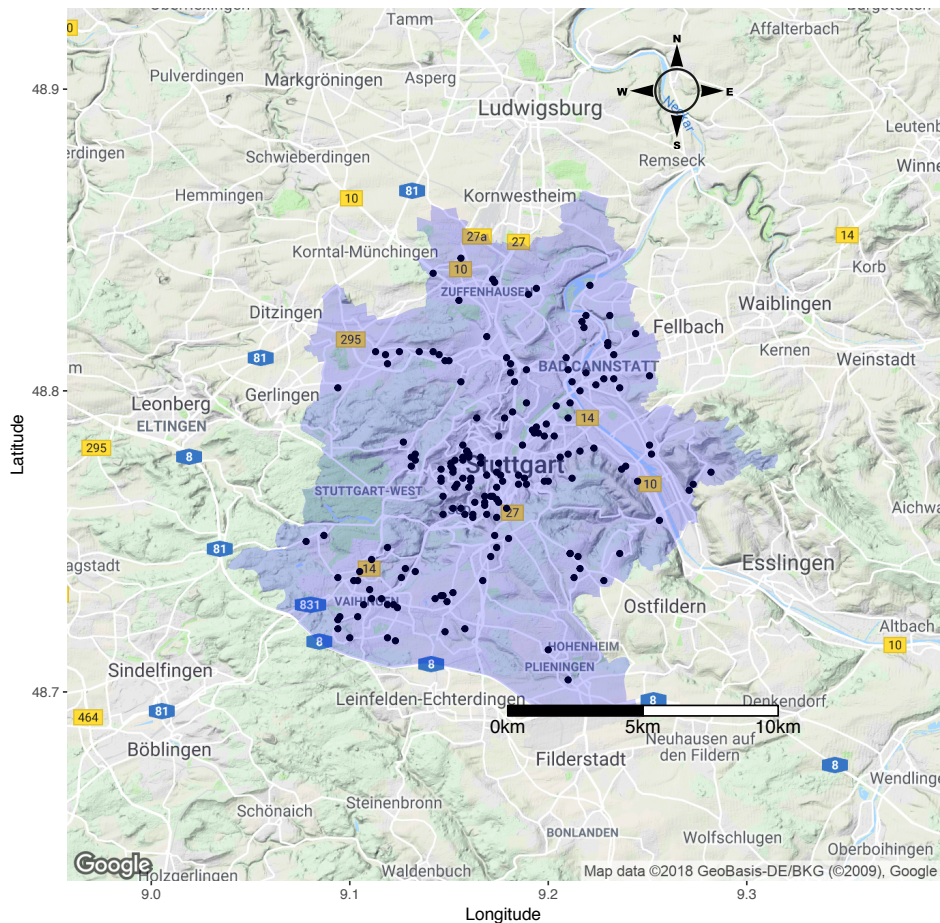


Figure. 5.2. Study area: City of Stuttgart and the existing citizen sense air quality network

5.3.2 Data

The sources of data that are used for this study can be divided into two categories, as detailed in the following section:

Citizen sense air pollution data

Four official monitoring stations alone will not be enough to fully assess the amplitude of the air pollution issue in Stuttgart, as well as the effectiveness of measures taken to mitigate it. Fortunately, the city of Stuttgart is having a dense network of citizen sensed low-cost air pollution monitoring sensor network developed by OK labs (OK Labs, 2018b). Ground measurements of PM_{10} and $PM_{2.5}$ from low-cost sensors were available as open data for various locations (OK Labs, 2018a). Initially, the data of total 594¹ sensors were downloaded for the 1st week of 2018 (1st-7th

¹The dataset does not allow for statements on the actual number of contributors.

Jan 2018). Further data cleaning for removing sensors with no values in the specified period and the pollutant of interest lead to final 117 monitoring sensors, which were used in this study. The measurements are collected using SDS011 sensors, with the measurement unit of $\mu\text{g}/\text{m}^3$ OK Labs.

Land use regression (LUR) variables open data

A LUR model needs several geographical predictors variables (e.g. land use type, road count, distance to roads, traffic, and terrain variables for specific buffers around the monitoring stations) as input. In the modelling process, the air pollutant measurements are considered as the dependent variable, and geographical predictor variables are considered as the independent variable to establish a regression model that can help in estimating the air pollution at unmeasured locations. The regression model is of the form :

$$y = X\beta + \epsilon \quad (5.1)$$

with

- y an $n \times 1$ vector of air pollution concentration from monitoring sites at any particular time (in our case annual mean NO_2 concentration at monitoring stations),
- X an $n \times k$ matrix with observations of k independent variables for the n available air pollution monitoring stations,
- β a $k \times 1$ vector of unknown parameters that we want to estimate, and
- ϵ an $n \times 1$ vector of errors, assumed to be independent and identically distributed.

The values about the predictor variables were extracted from the open data available on the internet. The buildings and road datasets were downloaded from Open Street Maps (OSM) and Geofabrik services (Geofabrik GmbH Karlsruhe, 2018; OpenStreetMap contributors, 2017); population data was downloaded using European Data Portal (Open.NRW, 2018), altitude data was downloaded from Bundesamt für Kartographie und Geodäsie open data portal (Bundesamt für Kartographie und Geodäsie, 2018), and the land use data was downloaded from CORINE Land Cover (CLC) (Copernicus, 2018).

5.4 Method

As mentioned in Section 5.1, the present paper seeks to develop an optimisation method which can take into account fitness-of-purpose objective for VGI data collection. As Clements et al. (2017b) suggested, the identification of the research question before planning the deployment of the VGI based air quality sensor network can help in useful data collection. The optimisation method presented later takes the question *what are the set of locations in the city where measurements are required for estimating air pollution with minimal LUR prediction error?* as the research question. It uses Spatial Simulated Annealing (SSA) to run the objective function.

Spatial Simulated Annealing (SSA) is a random search algorithm that explicit deals with spatial vicinity. It is the spatial version of the probabilistic techniques Simulated annealing, which was developed by Van Groenigen et al. (1999) for spatial soil sampling design optimisation. The SSA technique mimics the cooling of the metal phenomena to reach global optima, like simulated annealing. In the starting phase of the annealing process, the locations for sensors can change a lot, with low probability even at not so optimal locations. As the process cools down with time, changes in locations become smaller, and acceptance of worse designs of monitoring network becomes less likely. During the optimisation process, the algorithm takes several hundred or thousands of iterations to identify the optimal configurations. The SSA algorithm is widely used in sampling design for mapping (Heuvelink et al., 2010; Van Groenigen et al., 1999). The SSA algorithm requires an objective function, whose output value acts as ‘energy’ in the optimisation process. The optimal design identification is made based on the set of the location which represents the minimal energy of all iterations in SSA. Hence, the objective functions should be formalised as a single objective optimisation function, pointing at discrete-valued variables which calculate the energy value.

5.4.1 Optimisation objective function

The optimisation is performed based on some rules and objectives that are used as a function. The optimisation objective function is usually composed of one or many constraints or aspects which are calculated by using the explanatory variables of a given LUR model in our case. The objective function is implemented using SSA, where it estimates the objective function value (also called the energy of annealing) to identify the set of locations which fulfil the given optimisation objectives.

In our study, the first aspect of the objective function was to identify the set of locations which can decrease the spatial mean of prediction error (PE) for the study area. It can be expressed as follows :

$$PE = \frac{1}{|A|} \sum_{x_o \in A} x_o (X'X)^{-1} x_o' \quad (5.2)$$

with $|A|$ the size of the area (expressed as the number of grid cells representing the area), x_o, \dots, x_n set of predictor variables values for all potential location in A , X and X' being the matrix and transpose matrix of the predictor variables for the randomly selected monitoring design $D_o \in D^n$ in optimisation process. For the objective function, manipulation of the set of locations leads to the modification of X matrix values.

For a two-dimensional study area A , the prediction error is calculated using n observation sites using a network design D^n . The optimisation process starts using a network configuration fed in as input or by randomly selecting monitoring design $D_o \in D^n$ (if just n is defined), consisting of observation points s_o, \dots, s_n with corresponding predictor variable vectors x_o, \dots, x_n . During the optimisation process, the monitoring sites are transformed into a random vector with only one element different from the initial, yielding a new monitoring design D_1 . The optimisation process, compute the prediction error for each D_x utilising each node of the rasterised study area A , until the minimum value is achieved. For the further details about the above-mentioned objective function, we suggest referring to the work done by Gupta et al. (2018a).

Along with the requirement of the objective function to decrease the mean prediction error for the study area, the second aspect in the objective function was to enforce the wide-spread distribution of sensors in the study area. The wide-spread deployment is necessary because it can help in providing granularity to air pollution data, better informs the identification of pollution sources and supports in more conclusive studies on the effect of air pollution on the quality of life in cities (Mitchell and Dorling, 2003; Kumar et al., 2015b). The widespread deployment also helps in reducing the uncertainties associated with the modelled forecasting results. Hence, we extended the application of objective function developed by Gupta et al. (2018a) to consider the wide-spread distribution of sensors for VGI based monitoring network design.

To integrate the wide-spread aspect in the optimisation objective function, we calculate the inverse mean shortest distance (IMSD) for the set of locations selected in each iteration of annealing after calculating the mean prediction error value using

Equation 5.2 in the optimisation process. The spread aspect of the objective function can be written as:

$$IMSD = \left[\frac{1}{N} \sum_i^N \min_{j \neq i} (D_{ij}) \right]^{-1} \quad (5.3)$$

where N is the number of points in the configuration considered for optimisation, and $\min_{j \neq i} (D_{ij})$ is second minimum distance between the i^{th} point to other points of configuration (as the minimum value will be 0 for each point distance to itself).

The algorithm for computing the IMSD (Equation 5.3) to enforce wide-spread distribution of points can be summarised as :

1. input of a number of points (N) with a different spatial configuration as selected in each iteration of SSA,
2. compute the distance matrix for all the points,
3. identify the second minimum value in each row of the matrix, as distance matrix will contain the first minimum value as 0,
4. compute the mean of the minimum values from each row and column of the distance matrix,
5. compute the inverse of the mean value.

After the computation of inverse mean value, the value from the mean prediction error part will be added to get a single objective function value which will be further characterised as energy state in the SSA optimisation process. Furthermore, the optimisation function also can consider the weight function to prioritise one of the two aspects (prediction error or spread function) when identifying optimal locations during the optimisation process. The weights must be equal to or larger than 0 and sum to 1. The overall equation of the objective function which identifies the set of wide-spread locations presenting minimal prediction error for the study area can be expressed as :

$$Energy = (PE \cdot W_1) + (IMSD \cdot W_2) \quad (5.4)$$

where W_1 and W_2 are the weights which can be assigned to each aspect the objective function based on the aim and fitness aspect of the VGI based air pollution monitoring initiatives. LUR prediction error and spatial spread are both critical for air pollution monitoring. The main idea behind the discussed objective function with the flexibility to consider weights is to give policymakers (e.g. coordinators of VGI initiatives) some control over prioritising their goal considering two crucial aspects of air pollution monitoring campaigns. Minimising the prediction error of the LUR implies confidence in the estimated values of air pollution at locations that were not observed. On the other side, maximising spatial spread leads, as said above, to an air quality monitoring network potentially more informative as to the identification of various unidentified pollution sources in the city.

The overall steps for the optimisation algorithm can be summarised as follows:

1. A LUR model is selected/developed (using the air pollution ground data from low-cost sensors and predictor variables). If the ground data is not available for LUR creation, already existing LUR models can be selected arbitrary or by selecting models containing specific predictor variables which are significant for the study area;
2. Initial monitoring station locations are defined as the input, consisting of N observations, which can also be feed in as a whole number ;
3. The study area A is discretised, the candidate locations are defined based on the resolution expected for the study area;
4. Random point selection in each iteration starts and calculates the constraint values using SSA;
5. The design of each previously selected configuration during the optimisation is modified until the network design is accepted based on objective function value;
6. A design will be accepted if it reduced the prediction error as well as distribute the sensor in a wide-spread fashion, depending on the weight assigned to each objective as per Equation 5.4;
7. The optimisation continues to iterate and find the set of optimal locations until the new energy value reached minimum and is not changing in further iteration based on the energy transition and other annealing parameters.

All geospatial and statistical operations for the study were carried out in the R statistical environment (R Core team, 2017), using packages *sp* (Pebesma and Bivand, 2015), *sf* (Pebesma, 2017) and *SpatialTools* (French, 2015). For running SSA, we used the R package *spsann* (Samuel-Rosa et al., 2017). The source code of the optimisation method developed in this study can be accessed from GitHub (Gupta, 2018f).

5.5 Results

The monitoring of air pollution is highly location-dependent. In order to tackle the challenges of acquiring spatially fine-grained air pollution data for cities using VGI based approach, it is crucial to pay considerable attention to *where* the air pollution data must be collected by participants. We tested the optimisation method for the city of Stuttgart where a large number of citizens are collecting air pollution data using low-cost sensors developed by OK Labs. In this section, we present the results of the tests we performed to understand the significance of the proposed optimisation method.

In our study we tested the application of the developed optimisation method for two different practical scenarios:

1. Starting a new VGI campaign
2. Finding out where to place new VGI sensors

5.5.1 Starting a new VGI campaign

Considering the advantages of new low-cost miniature sensor devices that are capable of monitoring air pollution, we first tested the application of the developed optimisation method for the aim of initiating a VGI campaign. Initiating a new campaign would mean that no crowdsourced air pollution data is available, which leads to either relying on the official monitoring station data for LUR development or start the process from scratch. Since for the study area of Stuttgart, only three monitoring stations are measuring the air pollution data (which are not enough to develop the LUR model), we are of the opinion that it would be wise to start the procedure from scratch, believing no air pollution data availability in the study area for the first test case.

To initiate the process to identify optimal location for the deployment of new sensors network, we need to follow the steps as discussed in the previous section (subsection

5.4.1). As suggested, the optimisation method requires the input of explanatory variables of a given LUR model for identifying the optimal locations. We selected the model of Austria from the ESCAPE project for $PM_{2.5}$ (Eeftens et al., 2012b), as the underlying model which can explain the $PM_{2.5}$ concentration distribution for the study area. The selected model can be presented as :

$$25.44 + 0.11 * BUILDINGS_{100} - 0.65 * SQRALT \quad (5.5)$$

The selection of the Austrian model was based on two underlying assumptions. First, the model utilises square-root of altitude (SQRALT) as one of the explanatory variables. Stuttgart is characterised by very uneven altitudes and has a valley around it. We assumed that the SQRALT could help in explaining the dynamics of air pollution. The second factor is the availability of data, the building and altitude data was easily accessible; hence we decided to use this model for testing the optimisation method for the city of Stuttgart. It is also important to point out that we have used the number (N=116) and location of existing crowdsourcing network's configuration as the initial monitoring network for the test. However, it is not mandatory to provide a configuration; the optimisation method can also select random locations as the initial configuration for a certain number of sensors if given as input. We chose to use configuration to investigate the change in optimal locations from the existing monitoring network for a selected LUR model.

Figure 5.3 represents the optimal locations identified by the optimisation method, when only considering the prediction error aspect of the objective function keeping the spread function to 0 (i.e. $W_2=0$ in Equation 5.4). As can be inferred from the figure, the resulting configuration is clustered, which can be attributed to the explanatory variables under consideration from the selected LUR model. After getting the first overview of the significant locations with prediction error aspect, we again used the optimisation method with the equal weight ($W_1=W_2=0.5$) to also consider the wide-spread location selection aspect. Appendix B shows the influence of different weight values on resulting configuration.

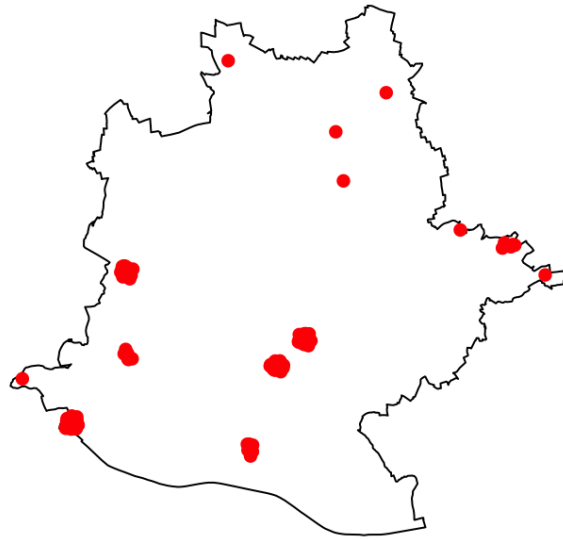


Figure. 5.3. Optimisation outcome without using the spread aspect of the objective function (N=116)

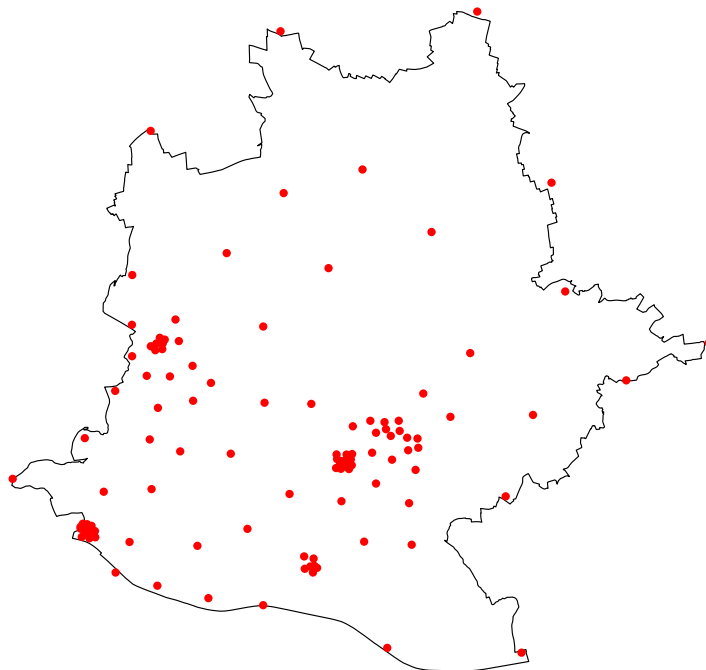


Figure. 5.4. Optimisation outcome considering the equal weight on both wide-spread as well as prediction error aspect of the objective function (N=116)

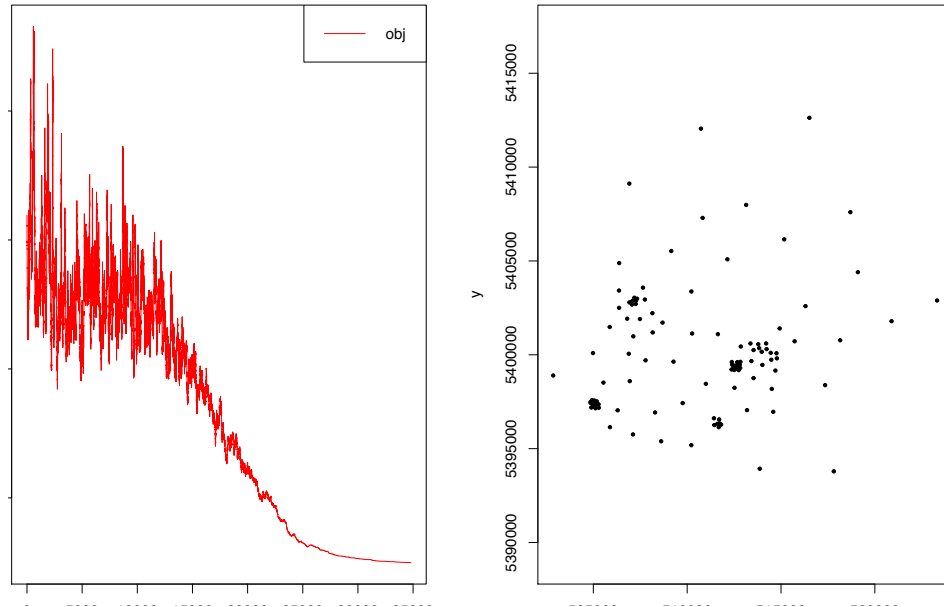


Figure. 5.5. Annealing energy transition during the optimisation with objective function laying equal weight to prediction error and wide-spread aspect (N=116)

Figure 5.4 presents the result of the optimisation with equal weight on each aspect of objective function. The outcome of this optimisation process acknowledges the wide-spread aspect. Several changes in location selection in monitoring network design can be noticed with few locations being clustered due to an equal weight of prediction error aspect. The outcome of the similar weight optimisation process decreased the prediction error from 0.018700 to 0.0089, as a percentage decrease of 52.41% in prediction error along with wide-spread configuration. As one can see, the location distributions of Figures 5.3 and 5.4 differ considerably within themselves, and from the original distribution from Figure 5.2. This visual inspection suggests two things: First, that the method works as expected, since incorporating spread as a criterion on Figure 5.4 has had the desired impact on the distribution of the network; Second, given that the difference between the distribution of locations with and without the use of the method turns out to be substantial, Figures 5.3 and 5.4 remind that randomly placing stations is not enough to take the most out of VGI air quality monitoring endeavours.

5.5.2 How many sensors should we deploy?

When initiating the air pollution monitoring campaign, one important consideration is the number of sensors desired to start the monitoring campaign. To understand the impact of the size of the monitoring network to the overall performance, we ran the optimisation method to identify the effect of the number of monitoring stations to the optimisation objective function. In the Figure 5.6 we can see the

influence of the change in monitoring network size to the prediction error value from the developed optimisation objective function with equal weight on both wide-spread as well as prediction error aspect (i.e., $W_1=W_2=0.5$). Figure 5.7 presents various optimal configuration of monitoring network obtained while running the optimisation method with the different number of monitoring station for initiating a VGI campaign.

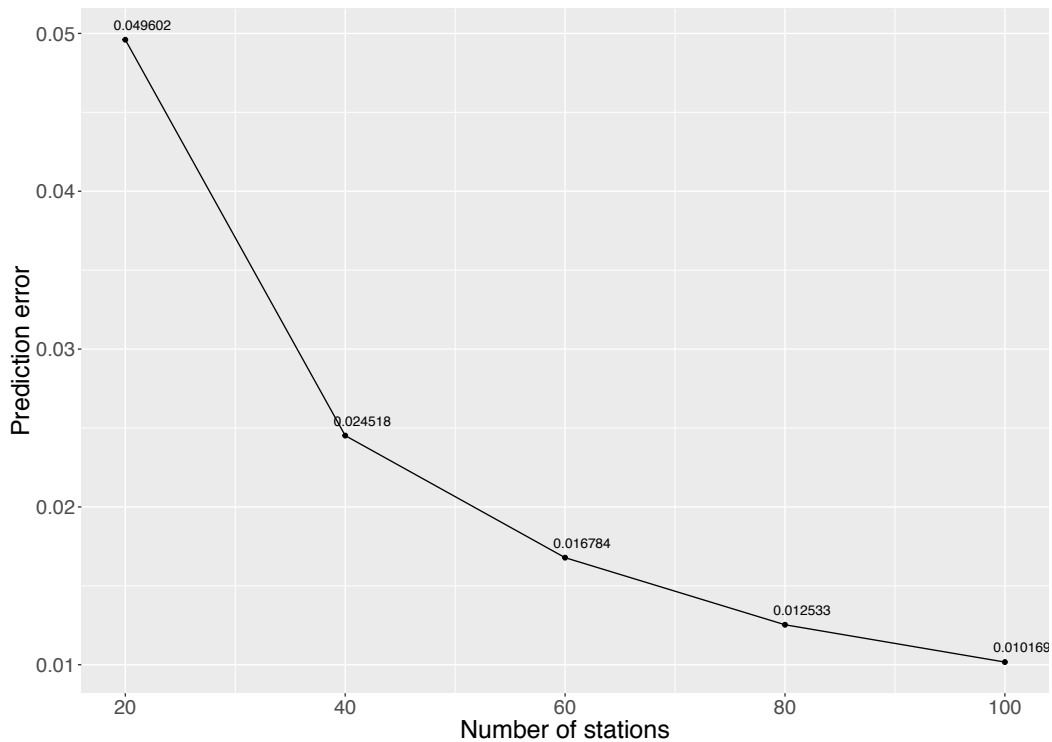
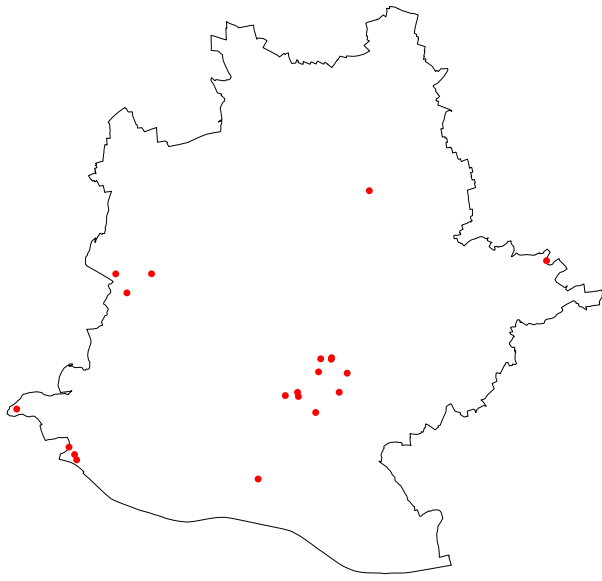
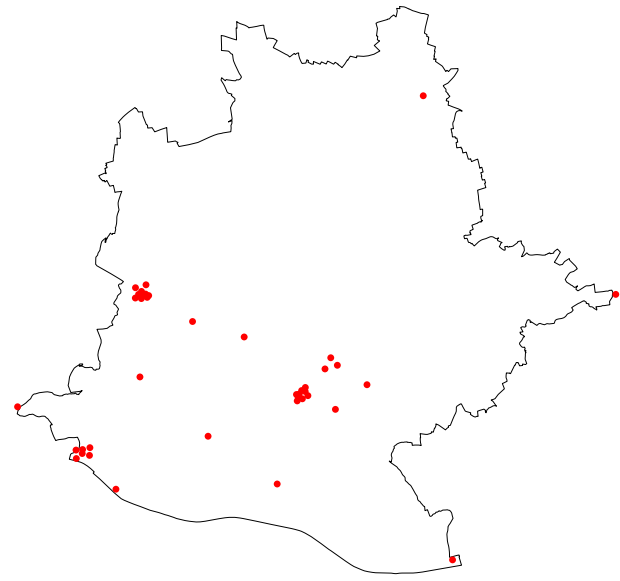


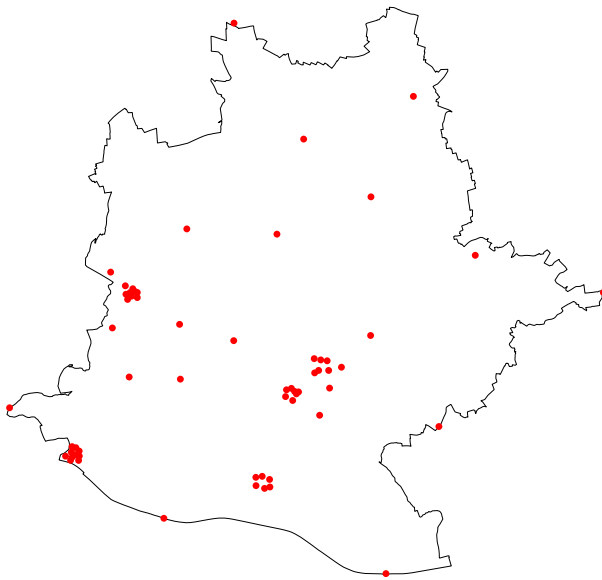
Figure. 5.6. Influence of number of monitoring sensors on the decreased prediction error aspect of objective function with equal weights on both the aspects.



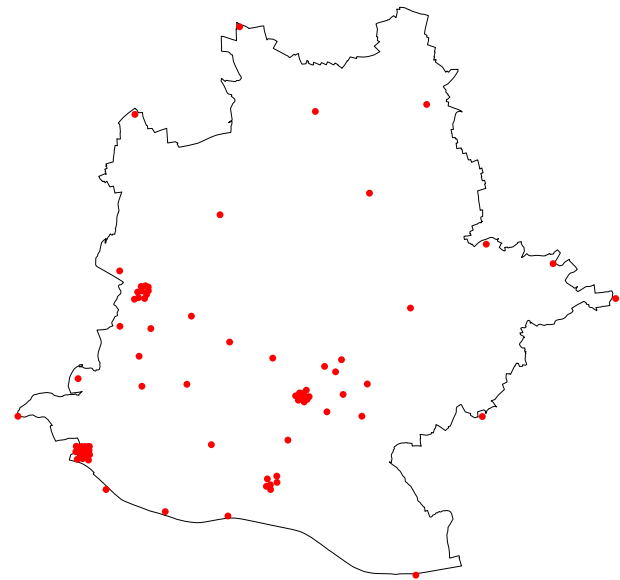
(a) With 20 sensors.



(b) With 40 sensors.

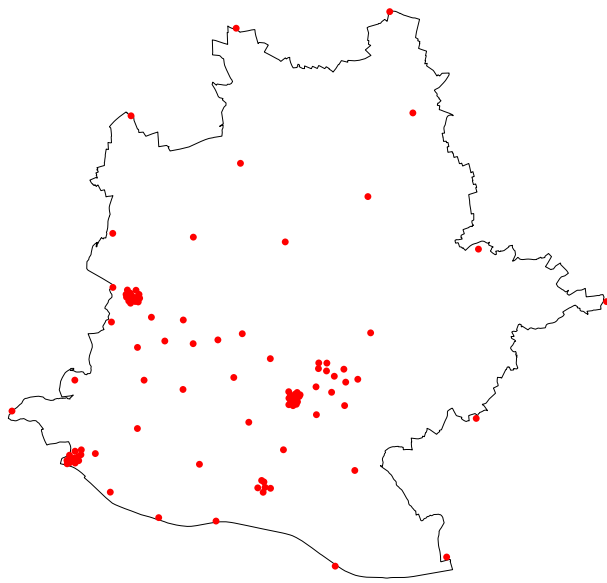


(c) With 60 sensors.



(d) With 80 sensors.

Figure. 5.7. Optimal location identified for specific number of sensors to initiate VGI campaign.



(e) With 100 sensors.

Figure. 5.7. Optimal location identified for specific number of sensors to initiate VGI campaign (Cont.).

5.5.3 Location significance

Another important factor while starting any air pollution monitoring campaign is to identify locations which are of great significance to the overall process of air pollution monitoring. Figure 5.8 presents a collective plot of all the configurations which can help in inferring the locations of significance in the study area for deploying sensors considering a given LUR model. The optimal locations obtained from various runs of the optimisation method using different numbers of monitoring stations suggests that few of the locations are indispensable. The repetitive selection of few locations in the study area highlights the significance of those locations for decreasing prediction error for a given LUR model. The results also demonstrate the potential of the optimisation method to identify locations which require significant attention and must not be neglected while initiating a new VGI campaign for air pollution data collection.

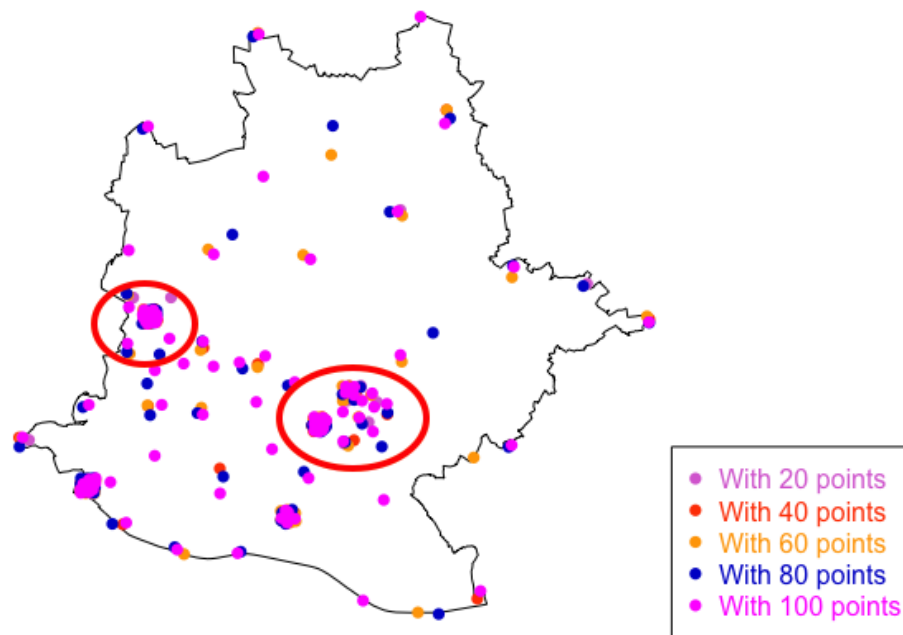


Figure. 5.8. Plot collectively representing all the configurations obtained by running objective function using different numbers of monitoring sensors which can be deployed for initiating VGI campaign to identify the location of significance.

5.5.4 Where to place new VGI sensors?

The ideas from the previous sections are useful while planning a new VGI campaign (e.g. a two-days citizen science project to gather some values about pollutant concentrations in the city), and can help VGI coordinators decide where to best channel the available resources. This section considers another scenario, namely that of extending an existing VGI network with new sensors using a systematic approach. By using the already existing VGI sensor data (i.e. the 116 sensors), we developed a LUR model. The advantage of developing a new LUR model using the air pollution data from crowdsourcing approaches is that it provided a more realistic explanation of the air pollution in the city than from an arbitrarily selected model (as we did in the previous test). The regression model developed by using the real data measured in the city can be helpful for choosing the predictor variables which can actually explain up to some extent the air pollution in the study area.

In our study, we have created the LUR model using the low-cost sensor data by following the steps suggested in the ESCAPE study (Eeftens et al., 2012b). The model uses PM_{10} concentration as the dependent variable and the following explanatory variables; square root of altitude (SQRALT), buildings in 500m buffer

(BUILDINGS_500), industries in 300m buffer (INDUSTRY_300), major roads length in buffer of thousand (MAJORROADLENGTH_1000) and low density residential land in 1000m buffer (LDRES_1000). The final model can be represented as :

$$\begin{aligned} \text{Pollutant Concentration} \sim & SQRALT + BUILDING_{500} + INDUSTRY_{300} \\ & + MAJORROADLENGTH_{1000} + LDRES_{1000} \end{aligned} \quad (5.6)$$

The quality of the model was low (R^2 of 0.1442). Nevertheless, we believe that even with low explanatory power, the developed model was more reliable to represent the air pollution in the study area than an arbitrarily selected model (Subsection 5.5.1). The explanatory variables of the developed LUR model were then used in the optimisation method for determining optimal locations.

Same as for the previous case, the optimisation method was exercised to identify optimal locations using the LUR model developed using crowdsourced data. The resulting configuration (see Appendix B) acknowledged the spread aspect for identifying locations which were widely spread as well as decreases the prediction error from 0.01870 to 0.008903, as a percentage decrease of 52.39%.

Extending the existing network

However, it is not feasible to move the existing monitoring sensors locations. Thus, we investigated the applicability of the developed optimisation method to identify the set of locations, if the VGI campaign decides to extend the existing monitoring network with 20,40,60,80 and 100 more sensors. Figure 5.9 shows the influence on the prediction error when the existing monitoring network was extended. It is apparent from the figure that with the addition of more monitoring stations systematically, the prediction error decreased. Figure 5.10 presents various configurations realised during the expansion of the existing monitoring network (in red) by adding the specified number of sensors (in green).

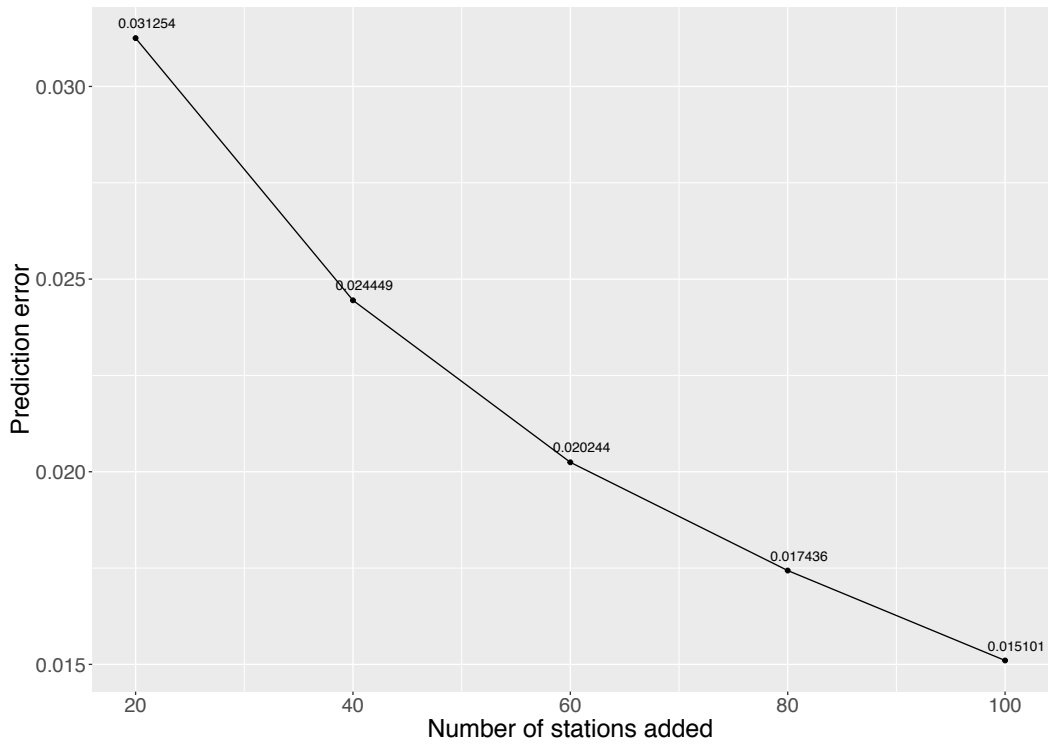
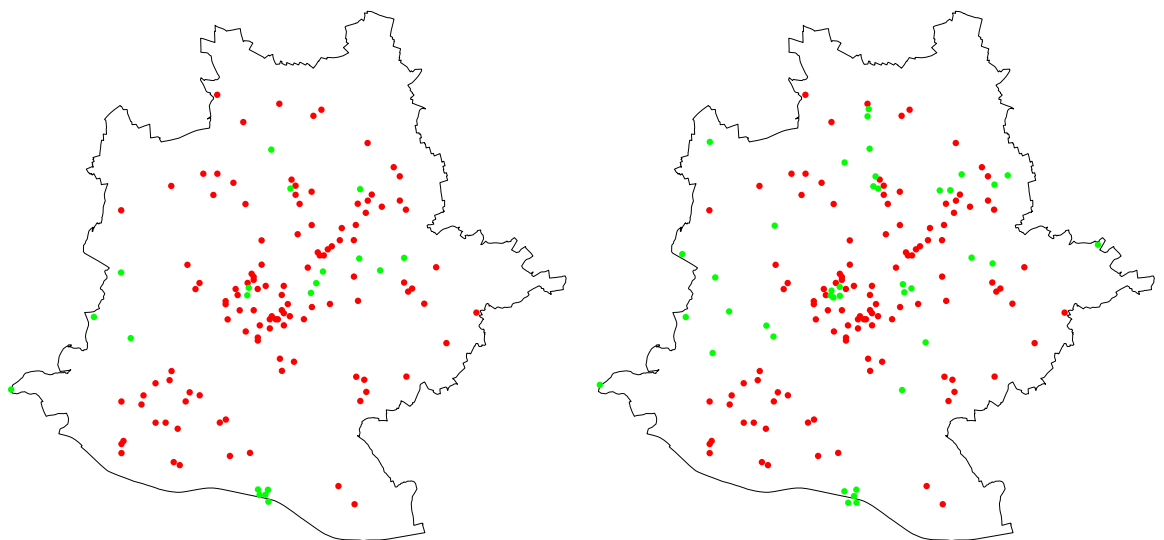


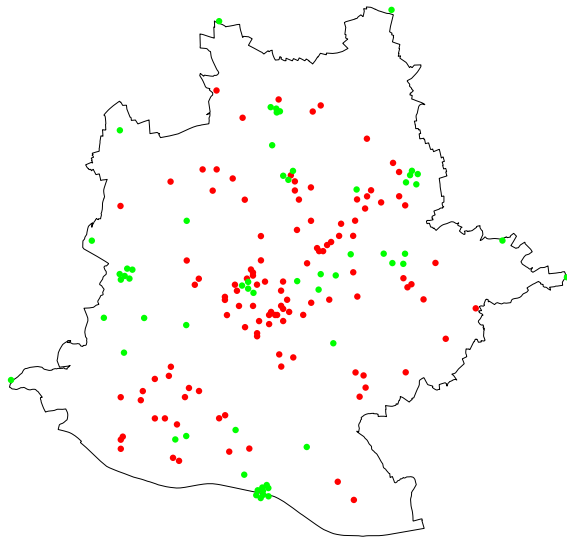
Figure. 5.9. Diagnostic study to capture the impact of extending the number of monitoring sensors into the existing VGI based monitoring network



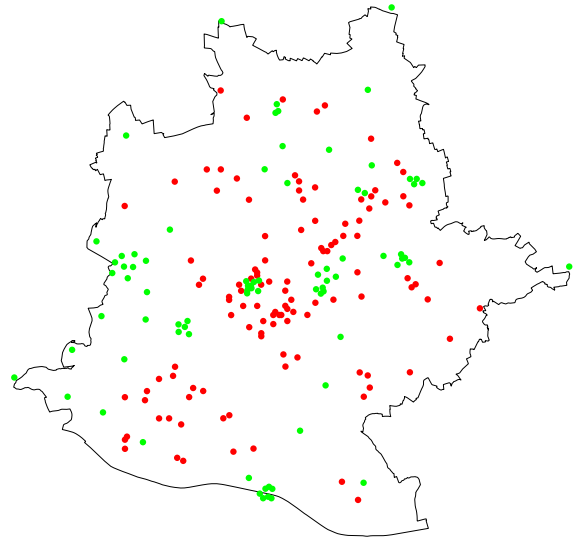
(a) With 20 sensors.

(b) With 40 sensors.

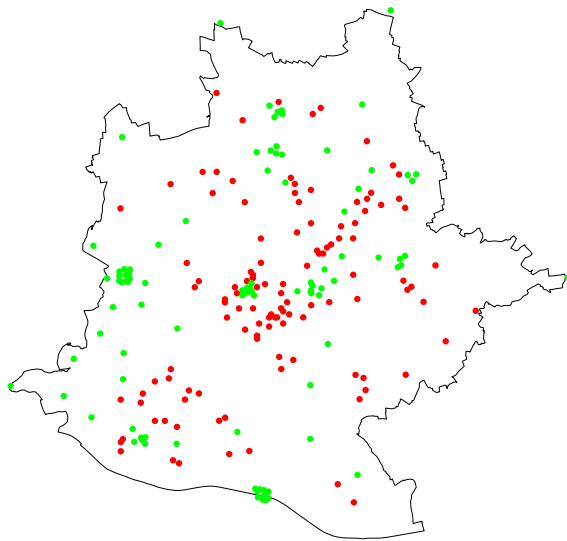
Figure. 5.10. Optimal location identified for extending the existing crowd sourcing monitoring network using proposed objective function.



(c) With 60 sensors.



(d) With 80 sensors.



(e) With 100 sensors.

Figure. 5.10. Optimal location identified for extending the existing crowd sourcing monitoring network using proposed objective function (Cont.).

5.6 Discussion

This section reflects on the significance of the study as well as its limitations, and points at future work.

5.6.1 Significance

The study demonstrates the application of optimisation method which can aid in the systematic deployment of low-cost sensors for detailed air quality monitoring while considering the scientific models like LUR. Low-cost sensors can provide data with very high spatial and temporal resolution, which is not feasible with conventional measurement approaches. The study has provided means to combine low-sensor datasets to a scientifically recognised air pollution modelling approach to facilitate a better air pollution data collection in the city. The developed optimisation method builds upon ideas suggested by Gupta et al. (2018a), to optimise air quality monitoring networks for VGI campaigns. The significant performance of the proposed optimisation method to decrease the mean prediction error by 52% along with wide-spread sensor network, demonstrate its applicability to enable systematic planning for VGI campaign for efficient air pollution monitoring.

The wide-spread VGI campaign sources can be useful for overcoming the issues connected to data quality, such as field duplication, data duplication and irregular spread of sensors as pointed out by Clements et al. (2017b) and Budde et al. (2017). The optimisation method also helps in leading the way to first define the research question (LUR in our case) to drive the data collection process. By defining the objectives before data collection, the method can also be useful for reducing the cost of deployment by limiting the number of sensor nodes required. The method can also be beneficial to identify locations which are easily accessible for sensor maintenance and calibration, for example by using population-weighted optimisation (Gupta et al., 2018a). Such extensions can assist in decreasing sensor failure and replacement costs for successful long-term deployment. If the population weights are considered, the optimisation method foster construction of LUR models with network design incorporating area close to population and roads, which can better characterise the full range of pollutant concentrations close to population (Wu et al., 2017).

Since the currently available sophisticated monitoring stations are not capable of expressing the air pollution variability at a detailed spatial scale, the wide-spread and less prediction error based low-cost monitoring network can be an alternative for gathering measurements, which can be detailed and informative. Using alternative

data sources also helps in overcoming the sparsity and scarcity challenges existing in the literature. The resilience of the developed optimisation objective function to prioritise the wide-spread and prediction-error aspect could be advantageous for developing a systematic crowdsourcing sensor network whose measurements can be used in versatile air quality modelling approaches. The spatial spread aspect of the proposed optimisation method helps in shrinking the effects caused by spatial correlation in LUR residuals (which usually exist Beelen et al. (2009)). However, by using weighted least squares (WLS) instead of ordinary least square (OLS) for Equation 5.2 or considering the kriging prediction error based optimisation as suggested by Van Groenigen et al. (1999) would have presented an analogous effect on the spatial spread of optimal configuration as the spread aspect of the proposed optimisation objective function achieved for declustering points. The method also inherits the flexibility offered by LUR and SSA, making it more implementable even in the cases where availability of data is limited. The outcome of the optimisation objective function considering the wide-spread distribution aspect can also help in distributing the points in different land use type, which can be constructive for developing robust LUR models as suggested by Wu et al. (2017).

The current state of sensor technologies with relatively large measurement uncertainties lead to concerns regarding engaging the citizens in the data collection process. Observing spikes while collecting data using VGI approaches may promote behavioural change which can help in preventing exposure to bad air quality. On the other side, this may also lead to the panic situation, possibly negating any health benefit. Nevertheless, relating the spikes to the geographical variables like road counts, traffic, and other emission sources by using LUR models may help in wide-informing citizens about their actions and local area contributions. Furthermore, the low-cost sensors data might not be monetised for proper air quality applications, but the LUR approach used in the study can act as a tool to process and visualise the data; the resulting analysis and the corresponding information generated can be easily monetised.

Overall, the optimisation method can help in defining the locations for systematic VGI campaign planning, which anticipates the wise use of the participation efforts along with reducing the error for air pollution modelling. The use of open and easily accessible data for VGI campaign systematic deployment, make this approach more implementable. Another major benefit of deploying VGI sensors is their ability to measure real-time data and provide immediate feedback that helps in improving the air pollution monitoring strategy systematically with the help of the proposed optimisation method. This also gives the opportunity to serve as a tool to help in building the capacity of participants to understand air pollution and the influence of geographical variables in the proximity, which can also explain air pollutant's variability. The wide-spread distribution aspect in the proposed optimisation method

also enables citizens to identify potential sources of air pollution otherwise unknown to regulatory authorities. Altogether the optimisation method proposed in the study helps in enabling VGI-based systematic collaboration and foster discussions that are important for understanding the applicability of low-cost sensors for detailed air quality monitoring in cities.

5.6.2 Limitation and Outlook

Along with the advantages, the proposed optimisation method also brings along some challenges and limitations. One of the critical limitations for the application of the low-cost sensor data for air pollution monitoring is the reliability of the measured data. Further challenges include short working time, inadmissible data and calibration challenge (Clements et al., 2017b). In the study, we used the data from low-cost sensors to develop a LUR model. The quality of the LUR model developed was low ($R^2=0.1442$) which concerns the quality of data produced by the low-cost devices, and the locations from where they were collected (Wu et al., 2017). However the field is in transition, future sensors may overcome quality flaws with better deployment strategies.

In addition to these limitations related to the use of low-cost sensor data, there are limitations concerning the proposed optimisation method. To begin with, the selection of LUR model is the first step to find the optimal location, which means that if we do not have a LUR model for the study area, we have to select one from the previous studies by specifying some assumption based rules for model selection. The selection of a LUR model based on some assumptions may not involve variables that are convincing enough to explain air pollution in the study area. Other limitation of the approach concerns the use of a LUR model and the underlying assumptions of multiple linear regression (e.g. linearity between dependent and independent variables, independent and normal distribution of error terms may create biases in interpreting the outcomes, which are the typical limitations for any simplistic regression-based studies). Limitations also exist for the SSA approach, as it is a stochastic method, every different run of optimisation method may yield different monitoring network designs. The process of optimisation is also very time-consuming, depending on the input parameters of annealing, variables used for computing the objective function and the study area size. While running the optimisation for the study, the process took 6-8 hours for one optimisation outcome.

As can be seen from the results, the output of the optimisation ended up being clustered. This clustering can sometimes be caused by the spatial auto-correlation of the predictor variables which lead to all points close to each other. The reliability of the LUR used for the optimisation may also contribute to the clustered results. Devising

the methods that address these limitations by taking into account robust LUR, and information on the spatial correlation and interpolation based constraints can be helpful in improving the design objectives of the study. We have not considered such factors in our study but future work could consider integrating it. Extending the developed optimisation method to consider the population distribution weights as proposed by Gupta et al., 2018a, can also be useful in identifying the locations close to living spaces. A population-based weight can be useful in two ways. Firstly, for identifying locations where the citizen lives, which can make the initiation of VGI campaign easier. Secondly, it promotes the gathering of air pollution data which is representing the real exposure of the population in the living spaces of the city. For the practical implementation of the proposed optimisation method for VGI approaches, future work can focus on integrating the optimal location identification method with citizen observatory based projects like FLAMENCO Project (2018). Integration with citizen observatory based projects can be fruitful because the optimisation method can identify the locations and citizen observatory can identify the participants at the optimal location, making the overall flow of VGI-campaign initiation easy.

As discussed in previous studies related to low-cost sensors deployment (Clements et al., 2017b; Budde et al., 2017), the field of low-cost sensors for environmental monitoring is in transition, and more work is needed to continue exploring the potential of low-cost sensors for air pollution monitoring. With the help of low-cost sensor systematic deployment initiatives by using citizen participation approaches, it is possible to bring forward a whole new system which anticipates the development of open data platforms like OK Labs (2018a). These initiatives also help in connecting other systems that utilise air quality data like health informatics, housing companies, sustainable urban planning; thereby helping in enabling the development of tools and techniques which can improve Quality of Life (QoL) in cities.

5.7 Conclusion

In this paper, we propose an optimisation method that can help in the systematic deployment of air pollution monitoring sensors for VGI approaches. The systematic deployment of the monitoring station in the city is desired to enable detailed air pollution monitoring with significant accuracy. The optimisation method takes into account two important aspects, namely, the decreasing prediction error for a given LUR model and the wide-spread distribution of locations in the study area. While identifying optimal locations, the optimisation method considers the weight for each aspect, if required to give priority to any of the aspects. The decreased prediction error aspect can help in developing a robust LUR model, and the wide-spread

distribution aspect support in making the data collection approach more versatile and informative. The applicability of the optimisation method was demonstrated using two practical cases: 1.) initiating a new VGI campaign, and 2.) placing new VGI sensors. In the first test case, the optimisation method identifies the set of locations using the explanatory variable of an already existing LUR model. This approach is used to initiate a VGI campaign for the cities where no air pollution data is available to develop the LUR model. In the second test case, a LUR model was developed using the VGI based air pollution data source. The results of the optimisation method exercise revealed a significant decrease in prediction error (by 52%) while taking into account the wide-spread distribution. The method can thus be considered as a useful tool to policymakers for systematic planning of the size and location of VGI campaigns. The availability of more accurate and open data, improved low-cost sensor for reliability and systematic deployment of sensors in VGI campaign may help in refining the performance of the proposed optimisation method for more robust results. Future work can involve integrating the optimisation method with citizen science observatories to identify participants at the optimal locations identified by the objective function, which can help in detailed air quality monitoring in cities using VGI approaches.

Acknowledgements: The authors gratefully acknowledge funding from the European Union through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, <http://www.geo-c.eu/>).

Author Contributions: The contributions of the respective authors are as follows: Shivam Gupta (SG) and Auriol Degbelo (AD) conceived and designed the study; SG performed the analysis; EP supervised the research; SG wrote the main manuscript and all other authors revised the manuscript.

Conflict of interests: The authors declare no conflict of interest.

Abbreviations:

The following abbreviations are used in this manuscript:

ESCAPE	European study of cohorts for air pollution effects
IMSD	Inverse mean shortest distance
LUR	Land Use Regression
OLS	Ordinary Least Squares
PE	Prediction Error
$PM_{2.5}$	particulate matter (PM) that have a diameter of less than 2.5 micrometers
QoL	Quality of Life
SQRALT	Square root of altitude
SSA	Spatial Simulated Annealing
USEPA	United States Environmental Protection Agency
VGI	Volunteered Geographic Information
WHO	World Health Organisation
WLS	Weighted Least square

Connecting Citizens and Housing Companies for Fine-grained Air Quality Sensing

Based on: Gupta Shivam; Degbelo Auriol; Pebesma Edzer. Connecting Citizens and Housing Companies for Fine-grained Air Quality Sensing. *GI_Forum 2018 Journal* (Under review)

Abstract

The complex nature of air quality suggests the need for fine scale air quality monitoring in cities. With 1 in 8 death worldwide associated with air pollution in 2012, communities have started partnering with academic institutions, state and federal agencies to assess local air quality and address these concerns. Participatory Sensing (PS) has recently become one popular method for collecting air quality information. It offers the prospect of collecting data at finer levels of granularity, but is subject to at least two significant challenges: data gaps (due e.g., to the lack of calibration, maintenance and replacement of sensors), and concerns of citizens with respect to their privacy protection. In this paper, we argue that including housing companies as stakeholders of participatory sensing frameworks may be beneficial in overcoming the aforementioned challenges. A survey (N=18) investigating the perception of housing companies to participate in air quality monitoring for cities, suggests that housing companies are indeed willing to participate in air quality data collection. The ideas presented in the article are pertinent to the design of more robust and privacy-aware participatory sensing frameworks.

Keywords: Air pollution, participatory sensing, privacy, volunteered geographic information, housing companies

6.1 Introduction

Air pollution has a huge impact on human health. Research has shown associations with a broad range of health endpoints such as mortality, asthma and low birth weight (Kelly and Fussell, 2015). In 2015, long-term exposure to ambient fine particulate matter air pollution ($PM_{2.5}$) was associated with 4.2 million deaths, representing 7.6% of total global deaths (Cohen et al., 2017). The rapid urbanisation accompanied with surging air pollution have now led to increasing regulatory measures for air pollution monitoring. Nevertheless, not much has helped to decrease the exposure since the large segment of the population continues to reside in areas with air quality concerns (Brauer et al., 2015). Air pollution concentration monitoring is usually taken care of by environmental or government authorities using networks of fixed monitoring stations, equipped with sophisticated instruments which are specialised for measuring various kind of pollutants such as carbon monoxide (CO), nitrogen oxides (NO_x), sulphur dioxide (SO_2), ozone (O_3) and particulate matter (PM). These sophisticated instruments are considered reliable because the governing authority ensures the standard procedures for instrument calibration, data collection and analysis. Usually, regulatory control measures are taken after a long time series data analysis of the collected data. However, regulatory air monitoring systems in general do not assess variability in air quality at a sufficiently detailed spatial scale (Jerrett et al., 2005b). This is not possible due to the cost and expertise required for utilising expensive air quality monitoring stations.

Geographic information systems (GIS), deterministic models (e.g., AIRMOD, RLINE, SHEDS), and remotely sensed data have been the bases for most of the air pollution modelling efforts (Özkaynak et al., 2013). With recent progresses in GPS/GIS enabled devices like cell phones, low-cost navigation devices and measurement sensors, individual citizens can also contribute to the flow of geospatial data about health-related environmental factors (Richardson et al., 2013; Fang and Lu, 2012). These activities are broadly referred to as participatory GIS, crowdsourcing, or volunteered geographic information (Mooney et al., 2013). Various environmental monitoring technologies and communications have led to the increased availability of air pollution sensing devices which are affordable and easy to use (Jiao et al., 2016). These technologies can result in rapid evolution of air pollution monitoring approaches (Jiao et al., 2016; Fang and Lu, 2012). With the help of these affordable technologies for collecting large environmental datasets, participatory sensing has been one of the popular methods quite recently because of peoples' concerns about the negative impact of environmental factors (Commodore et al., 2017).

Participatory sensing (PS) is defined in this work after Christin et al., 2011 as people's voluntary use of devices, to contribute data for their own benefit and/or the benefit

of the community. For environmental monitoring, the data aspect is of particular importance. Our understanding of the complex relationship between air pollution and human health has improved substantially with time, but data gaps and the resultant uncertainties still limit our ability to fully assess the adverse impact of air pollution. To fill this gap and for increasing public participation, non-professionals and citizen communities have emerged recently to help in the data collection process (Clements et al., 2017b). The core of any successful PS approach entails three essential ingredients: technology, data and people. By making use of advancing computing powers, high-performance networks, storage and evolving sensors, PS can be conducive in collecting and analysing environmental data. A typical PS application involves data collected by devices (cellphone enabled/independent sensor) of volunteers, which is then forwarded to a central server for processing. The captured data is augmented with meta-data such as location, time, identification and context information, for further processing and made available to individuals or communities depending on the needs.

Scientific research can also benefit from well-prepared crowd-sourced observation campaigns. A high-density sensor network of PS has a significant potential for improving spatial monitoring of environmental variables (Schneider et al., 2017b). Thus, with the help of PS, it is possible to provide a picture of various environmental variables at spatial scale and resolution, which is not previously available (Fang and Lu, 2012). Community participation may also help to build trust and empower the participants and communities, especially when the data is community managed and owned, giving them an opportunity to be equally weighted with industry and regulatory bodies (Clements et al., 2017b). In the recent past, PS has been effectively used to monitor various environmental phenomena such as noise, air, radiation and water pollution (Hemmi and Graham, 2014; D'Hondt et al., 2013; Weissert et al., 2017a; Little et al., 2016). Despite their advantages, PS approaches have their own issues, for example human error, varying data type, reliability of low-cost sensor measurements, data quality, the stability of data sources, malicious use of deployed devices, maintenance and calibration issues, and the privacy of participants (Richardson et al., 2013).

The aim of this article is to discuss the prospect of involving housing companies as stakeholders of air quality sensing initiatives. Including them as a player could be helpful to address two issues of current participatory sensing frameworks, namely data completeness challenges and privacy protection concerns. There is a strong conceptual overlap between the terms of participatory sensing and volunteered geographic information (Mulalu, 2018; Haklay et al., 2018), and both are used interchangeably throughout the article. Section 6.2 presents participatory sensing frameworks, and elaborates on some of the current issues in PS. Section 6.3 suggests to involve housing companies as a third player in PS framework to address issues

related to the sparsity of PS nodes, maintenance of sensors, and location privacy threats for the participants. Section 6.4 presents results of a survey we conducted to assess their willingness in joining PS initiatives for air quality monitoring. Finally, Section 6.5 and 6.6 put forward the discussion and the conclusion of the work, respectively.

6.2 Background

Environmental research for cities requires fine-grained data for an accurate assessment of city phenomena. GIScience can be helpful in addressing various environmental challenges of city by using a wide range of key concepts that contribute to urban intelligence - representation, connection, coordination, measure, networks, movements, participation or even sensors, to list few of them (Batty et al., 2012). These urban sensing approaches have the potential to generate a “data commons”, that is, a data repository generated through decentralized collection, shared freely, and amenable to distributed sense-making not only for the pursuit of science but also advocacy, art, play, and politics (Cuff et al., 2008). GIScience’s approach of connecting urban citizens as the active sensors for data collection have the capacity to contribute effectively to the spatial intelligence of the cities (Roche, 2014). The major advantage of using PS frameworks for environmental monitoring is its potential to increase the spatial resolution of atmospheric measurements to identify variations below the city or regional levels, even down to street-level or below (Apte et al., 2017; Gabrys and Pritchard, 2018). As Goodchild, 2007 pointed out, the most important value of such information may be in what it tells us about local activities in various geographic locations that go unnoticed. There are numbers of citizen science programmes that actively collect source data from members of the public for environmental monitoring including air pollution (Honicky et al., 2008; Boulos et al., 2011; Costa, 1999; Elen et al., 2012). Nonetheless, there are still some important open issues concerning PS frameworks. These are briefly discussed below.

6.2.1 Data Completeness Challenges

Traditional data collection methods for air pollution levels at intra-city level are often sparse (Schneider et al., 2018). Compared to conventional procedures for collecting air pollution data using sophisticated monitoring stations, PS can be very different (Elwood et al., 2012). For instance (and unlike PS), any information with low or no administered value along with difficulties in data collection, may not be present in an official monitoring sources. Balanced spatial spreads for data sources are expected in official monitoring sources, yet this may not be the case in PS. Also, data quality aspects in PS need considerable attention. The list of components of spatial data

quality vary from author to author (see Degbelo, 2012; Devillers et al., 2010), but the focus of this section is on accuracy and completeness. Both are critical, since inaccurate or incomplete environmental data not only impacts the action taken by policy makers, but the general public perception's towards the environment also.

When monitoring air pollution or other environmental disturbances with low-cost technology, citizen-led initiatives are typically challenged about the validity or accuracy of their data (Gabrys and Pritchard, 2018). In the case of air quality monitoring, the current evaluations of low-cost sensors unveil that available particulate matter (PM) sensors exhibit reasonable performance (AQ-SPEC, 2017). Often these data sources can provide ongoing indications of changes in air quality, rather than absolute measurements (Gabrys and Pritchard, 2018). Since PS is open to the contributions of volunteers, there is room for inclusion of corrupted data. Indeed, people sometimes can act selfishly and exploit the resources for their benefit, and PS frameworks are prone to such inimical users behaviours (Mousa et al., 2015). Users, for instance, can start using outdoor air quality sensors to measure the air quality inside their houses.

With respect to completeness, the spatial density of the overall PS network is key to the inferences which can be made based on the PS data. It has been discussed in the literature (Gibson et al., 2000) that the amount of spatial detail in a dataset influences the types of patterns which can be detected during the analysis process. In the context of PS, the number of participants actively contributing data does not necessarily correlate with the amount of spatial detail in the PS dataset. For example, many people collecting data at one location (immediate surroundings) can also contribute redundant data, which can be useful for assessing data quality (Budde et al., 2017), yet put some limits on spatial coverage of the data collection process (Jaimes et al., 2012; Thepvilojanapong et al., 2010). Furthermore, PS approaches suffer sometimes from representativeness issues (also called lurkers phenomena (Lombi, 2018)) where large number of participants do not actively contribute to the campaign. Other factors which may contribute to incomplete data collection in the PS framework include the maintenance of the sensors and their calibration. For example, temperature and humidity have a large effect on gas-phase air quality sensors leading to a decreased sensitivity which requires re-calibration and cleaning up over time (Lewis et al., 2016; Masson et al., 2015). Citizens are interested in participation but may not be willing to handle repeated replacement, calibration and maintenance of the sensors for proper measurements. If not well maintained, these sensors can produce corrupt data, and bring about adverse consequences for the overall data analysis.

6.2.2 Privacy Concerns

A PS incorporates people who may have ethical concerns about their privacy. If participants use personal devices to collect sensor data (e.g. Fang and Lu, 2012), one of the key challenges in integrating them for PS discussed in the literature is “Privacy” (Kotovirta et al., 2012; Liu et al., 2018). As PS involves the creation of data including the participant’s location and time (see Christin et al., 2011), the disclosure of such data comes with location privacy threats for participants which should be mitigated. These threats deserve attention because “[o]ur precise location uniquely identifies us, more so than our names or even our genetic profile” (Duckham and Kulik, 2006). Participants’ ambivalence due to privacy concerns may slim down their interest for participating and contributing in the data collection process.

The problem of privacy is not new, and several works (Richardson et al., 2013; Cuff et al., 2008; Bowser and Wiggins, 2015; Kotovirta et al., 2012; Liu et al., 2018; Kessler and McKenzie, 2018) pointed out the need for addressing it. The technical challenges in providing privacy in PS, originate from the simultaneous presence of several mutually untrusted (and/or potentially unknown) entities, including participants, data consumers and service providers (Eugster et al., 2003). Various methods have been proposed to preserve and increase awareness about the privacy of the citizens (Agrawal and Srikant, 2000). Infrastructures which can serve to participants and data collectors using cryptographic tools were proposed (De Cristofaro and Soriente, 2013). Various methods like k-anonymity and l-diversity which blurs sensitive information, have also been considered to protect participants information in the PS system (Huang et al., 2010). However, there is still considerable ambiguity with regards to participants’ privacy when PS systems are extended based on realistic assumptions (Christin et al., 2011). Methods like anonymous reputation architectures (Androulaki et al., 2008), pseudonyms (Li and Cao, 2013) were also proposed to solve privacy and incentive related conflicts in PS. However, they are vulnerable to identity-based attacks (Niu et al., 2018). Also, the privacy-enhancing peer-to-peer reputation system from Kinatader and Pearson, 2003 was not well utilised because of lacking trust between PS participants and data collectors. Personal data may in principle be gathered, analysed and used with participant’s consent (Taylor et al., 2016), but overall, the discussion on privacy threats is still ongoing in PS (Jiang et al., 2018)

6.3 Method

The previous section has pointed out some open issues in PS framework with respect to data completeness, and privacy. In essence, PS frameworks still suffer from

challenges related to the sparsity of sensor networks, maintenance of sensors, and location privacy threats for the participants. The rest of the article discusses the prospect of involving housing companies as stakeholders of air quality sensing initiatives to address these issues. The argument is presented first (Subsection 6.3.1), followed by the design of a survey to assess the actual interest of housing companies in being involved (Subsection 6.3.2). Results of the survey are presented in Section 6.4, and complemented by a critical reflection of their implications for PS and GIScience in Section 6.5.

6.3.1 Involving housing companies

Decent housing is essential for both individual and economic growth. It impacts individuals' well being, health and inclusion in society (Hulchanski, 2002). Housing companies are the partners in the urbanisation which own residential buildings and offer maintenance (and further services) to their tenants. These residential spaces are usually with arrangements such as single family home ownership, condominiums and cooperatives are rented on tenure. where maintenance and services are managed some public or private entities. Some housing companies have become partners with various organisations, education institutions and government to develop new services (e.g., services related to mobility) for the residents (Bäumer, 2004). Product-oriented and social services are new marketing strategies.

Location is one of the main selling points for housing companies because view, safety, and facilities in vicinity are criteria which may influence a buyer's decision. Air pollution is one of the hidden element which is also attached to the location and can impact the buyer's behaviour. In the recent years, the traditional housing sale prices in certain parts of the cities are declining because of the environmental factors associated with the location (Le Boennec and Salladarré, 2017; Chen and Chen, 2017). According to Eurostat's recent statistics on the quality of life indicator "natural and living environment", 77% of EU citizens believe that the environment has an impact on their quality of life and 87% believe protection of the environment to be at least in part, the responsibility of citizens. 95% of EU citizens feel that protecting the environment is important to them personally (Eurobarometer, 2011). These statistics suggest that involving housing companies in PS initiatives can lead to a win-win situation for both. On the one hand, housing companies can tackle the above-mentioned cost decline issues, and develop environment-related services by getting involved in the PS process. In particular, offering services where the health impact caused by environmental aspects are covered can influence buyers' overall choice for some housing property locations. On the other hand, PS can also get some of the issues mentioned in Section 6.2 better addressed. The proposed approach can draw upon existing low-cost tools for air quality monitoring (see Clements et al.,

2017b for a recent review) and support the vision of future smart cities (Batty et al., 2012; Degbelo et al., 2016).

6.3.2 Would housing companies want to join participatory sensing?

There has been very few studies in PS research, if any, looking into the wishes of housing companies regarding their involvement as stakeholders. Understanding the perception of housing companies regarding their participation in PS frameworks is important, if they are to be involved in the process of improving air quality monitoring close to the citizen's living space, and help fill the gaps of the PS. The current research investigated housing companies' perception about two main indicators of quality of life, namely "health" and "natural and living environment".

The target population in this research consisted of executive and planning officials of various housing companies in Germany. The housing companies are from the network of companies who utilise various GIS applications and services for their work. The survey was administered online and by paper mails to 179 individuals of 71 housing companies established in 42 major cities of Germany. The survey ran from June 2017 to January 2018. Participation in the survey was voluntary, and participants did not receive compensation for their participation. The questionnaire consisted of 16 questions (all in German) among which multiple choice questions, likert questions and dichotomous questions. The survey questionnaire was designed to collect the perception of housing companies regarding: (1) Quality of Life (QoL) indicators in general, (2) "health" and "natural and living environment" QoL Indicators, (3) using low cost air quality sensors to collect air quality data, and (4) sharing data with public and related institutions. Of course, there are other important aspects as well to be covered, especially in relation to specific interests and public requirements. We did not include more, however, to prevent the survey from becoming complex and long in size, as the survey was aimed at higher officials of the housing companies.

In total, we received 18 responses (1 response online, and 17 via paper mails). Of these, one was incomplete and was discarded from the analysis. It is not, at this point, possible to make definite statements about non-response bias (i.e., the degree to which sampled respondents differ from the survey population as a whole, see Johnson, 2012). The reason for this is that recent statistics about the number of housing companies in Germany are scarce. To give an impression of orders of magnitudes to the reader, a report by Consilia Capital (Moss, 2011) in 2011 listed about 84 housing companies in Germany. Given that 71 companies were randomly sampled, it's likely that non-response bias has been little for the current dataset.

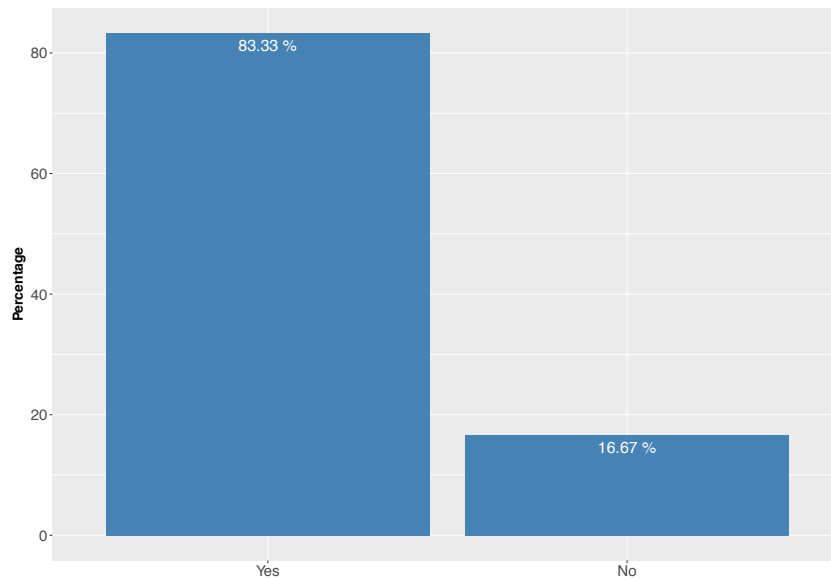


Figure. 6.1. Have you ever considered quality of life indicators for the development and planning of housing?

6.4 Results

The main results of the questionnaires are now presented. They touch on the four topics of the survey, namely (1) QoL indicators in general, (2) QoL indicators for “health” and “natural and living environment”, (3) using low cost air quality sensors to collect air quality data, and (4) sharing data with public and related institutions.

QoL indicators in general: 83% of the participants indicated that they have considered the QoL indicators suggested by Eurostat for development and planning purposes (Figure 6.1). But when asked about the information they have about their resident’s QoL, only 11% were very well informed, 61% of them were well informed and 28% of them were not informed (Figure 6.2). The majority of participants was well informed about their residents’ QoL suggests that housing companies may indeed be in a good position to contribute to further improve this QoL.

QoL indicators for “health” and “natural & living environment”: As shown in Figure 6.3, a large percentage of participants gave importance to health as a crucial indicator (41%), and in general about 71% believe it to be an important indicator. Regarding the indicator for “natural & living environment”, the participants surveyed did not see it as crucial as health but still view it as crucial (41%). The rest (59%) do not believe it to be a crucial indicator and ranked it as not so important (Figure 6.4). When asked about the importance of both indicators taken together, one can see that “Health and Natural & living environment QoL Indicators” are crucial: 17% believe it to be ‘very important’, 78% find it to be ‘important’, while only 6% does not completely agree to it (see Figure 6.5).

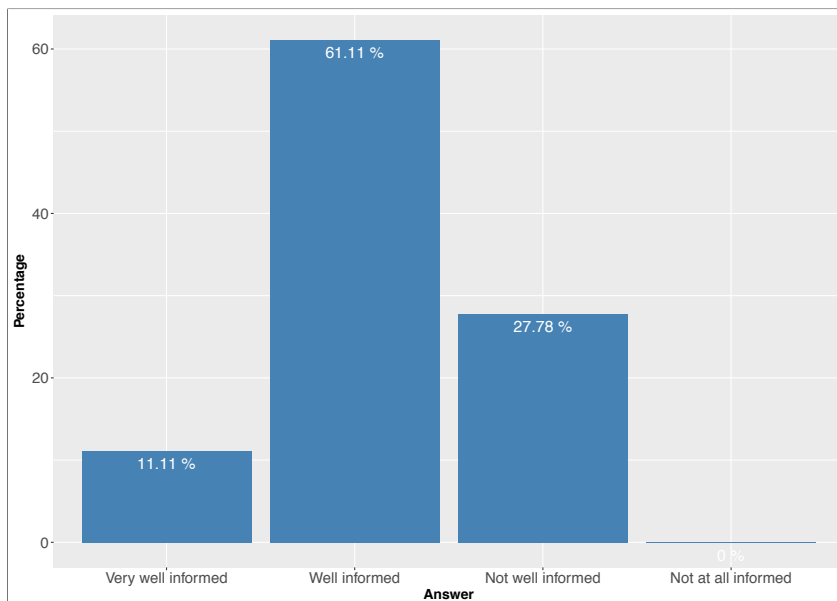


Figure. 6.2. How informed do you feel about quality of life of residents in your housing space?

Low cost air quality sensors for data: Interestingly, as indicated in Figure 6.6, a large portion (78%) of participants expressed interest in using low-cost sensors to measure air quality around housing space so that they can control it and residents can be warned to take measures.

Sharing data with public and related institutions: Only a small proportion (17%) of the participants indicated their complete interest, a large proportion (50%) of them expressed moderate interest, whereas 22% of them indicated little interest in sharing air quality monitoring data with the institutions which can help in data analysis for air quality monitoring and prediction. A very small proportion (11%) of participants were not interested in sharing the data (Figure 6.7). Moreover, when asked about sharing data with the residents directly using low-cost sensor, the majority of companies (44% said 'No' and 22% selected 'others') were reluctant. As shown in Figure 6.8, only 33% of the companies were interested in sharing the data with the residents. The negative responses for sharing data with residents may be attributed to the lack of trust due to data quality concerns existing at this point in time (Gabrys and Pritchard, 2018). Overall, the results suggest a difference - from the point of view of housing companies - between data sharing with institutions, and data sharing with residents. The exact nature of this difference, and the causes for it need further investigation (e.g., through follow-up interviews) at this point.

Limitations: The figures above give some insight into housing companies' stand-point as regards QoL indicators, as well as data collection and sharing. There are nonetheless few limitations to mention with respect to the data itself. First, there is non-response bias from which all surveys suffer (although as discussed above, this

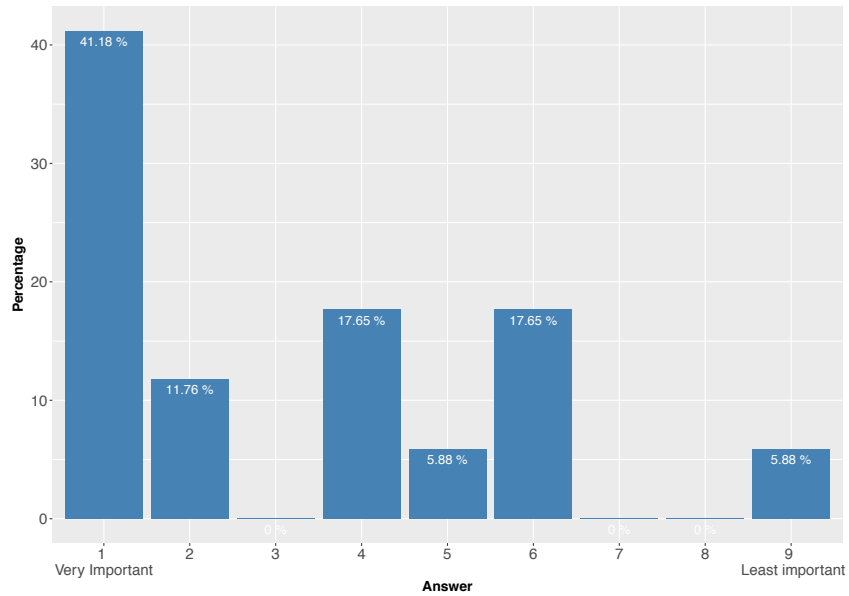


Figure. 6.3. How important is “Health” as QoL indicator for your company planning and development?

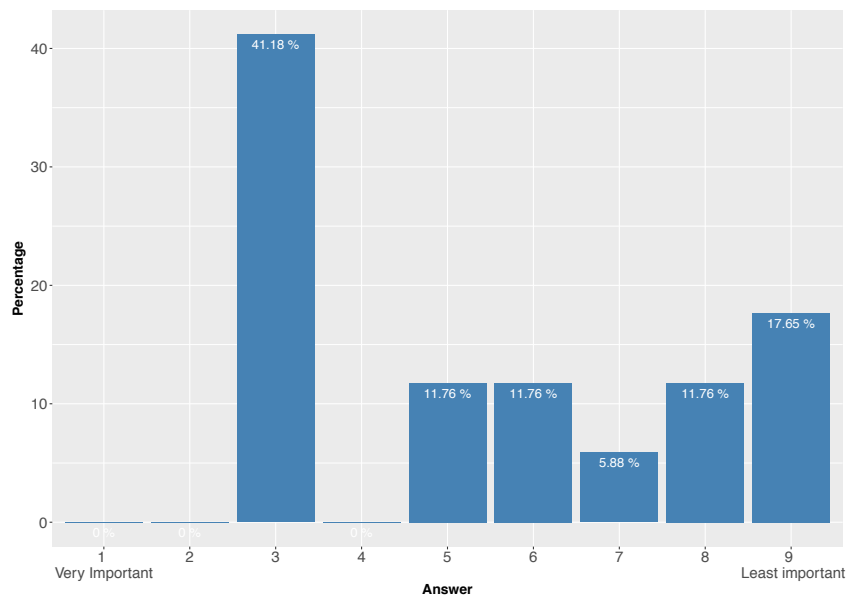


Figure. 6.4. How important is “natural & living environment” as QoL indicator for your company planning and development?

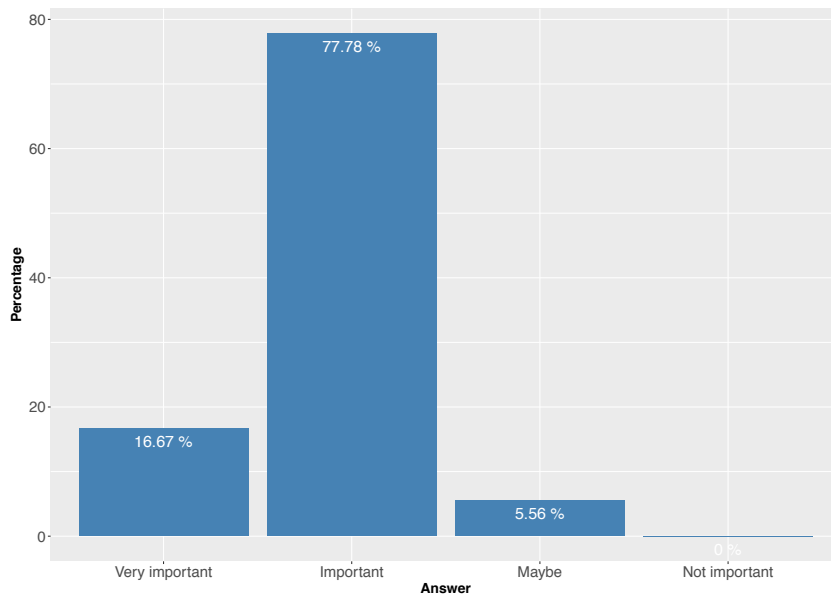


Figure. 6.5. How crucial is “health” and “natural & living environment” for housing space development plans for residents?

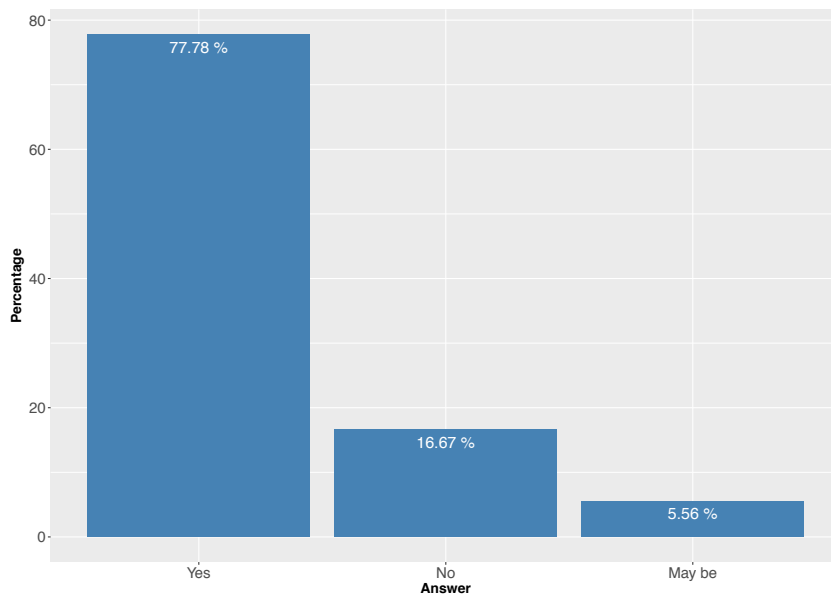


Figure. 6.6. Would you like to use low-cost sensors to measure air quality around housing space, so that you can control it and residents can take measures to breath safe?

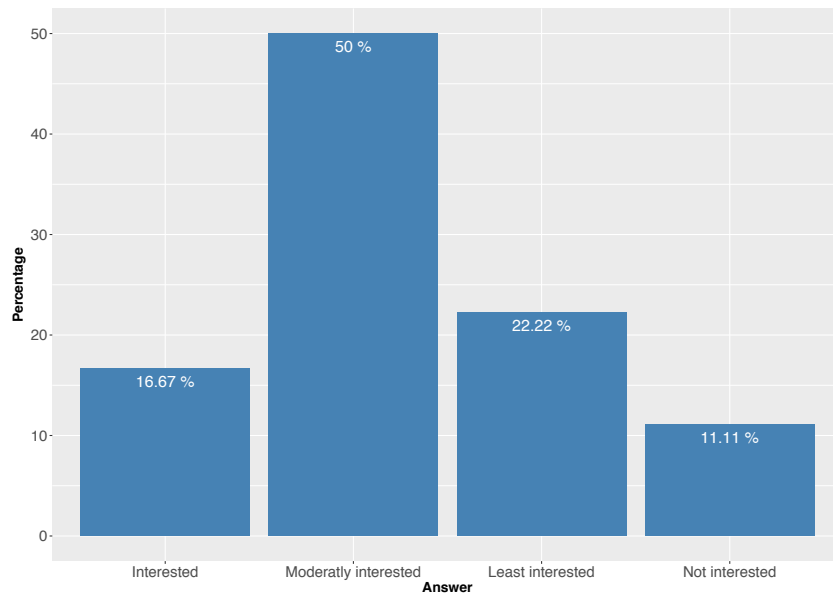


Figure. 6.7. What would be your take on sharing the air quality monitoring data with the institutions which can help in data analysis and air quality monitoring and prediction, so that residents can also get forecast of bad air quality with monitoring information?

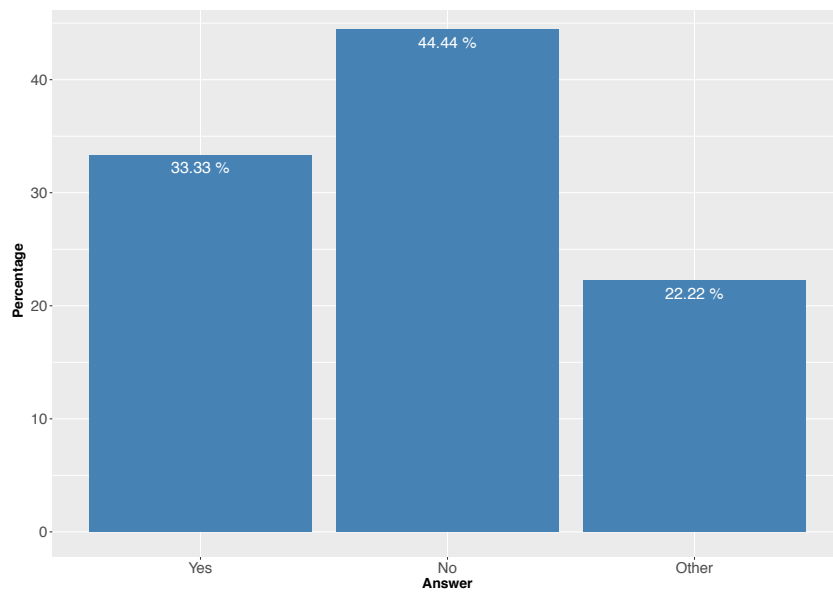


Figure. 6.8. Would you like to provide air quality information to the residents like normal other services so that they can save them self from harmful pollutant impact?

bias might be little). Second, the survey used paper mails and an online survey to examine 71 housing companies' perception as said above. However, to maintain the anonymity of the participants we have not requested the housing company's identity when responding. That is, the exact number of *distinct* companies who took part in the survey is unknown to us. Most responses obtained were from different cities (14 out of 42) and indicate different housing companies. The remaining three came from one city in Germany (i.e., Hamburg) and may have all been from the same housing company. Whether the responses all come from 15 different housing companies or 17 different, they reflect the view of 17 planning/executive members currently active in the housing business in Germany. It would be interesting to see how whether these views are shared with other planning/executive members in other countries.

6.5 Discussion

The previous section has illustrated that the housing companies surveyed not only find QoL indicators to be important, but would also be willing to use low-cost sensors to measure air quality around the housing space. This indicates that the idea of involving them as stakeholders of participatory sensing initiatives for air quality monitoring holds promise. They may play the three roles of participatory sensing applications listed in (Christin et al., 2011). They may act as *campaign administrators*, i.e., initiate participatory sensing campaigns when they invite their tenants (from time to time) to collect data for creating awareness about air quality. They may also design, implement, manage, and maintain PS infrastructures for specific houses (which is a typical role of campaign administrators). Furthermore, they may act as a *participant* when they install low-cost sensors which collect data continually about the air quality. Finally, they can also act as *end-user* when they visualise the data collected, reflect on it, and take evidence-based measures to improve the life of the residents. The next two subsections reflect on advantages and drawbacks of including them as stakeholders in PS systems.

6.5.1 Addressing data completeness challenges

As mentioned in Section 6.2, PS data suffer from issues of accuracy and completeness. Housing companies as *participants* can (in agreement with the residents) install low-cost sensors on top of buildings to collect air quality data. Involving a large number of housing companies in cities will help monitor air quality at a finer spatial granularity (addressing thereby the completeness issue). By using optimal location identification methods for air quality monitoring network spread (see Gupta et al., 2018a), PS data handling and completeness can be managed more efficiently.

Housing companies as *campaign administrators* can maintain PS infrastructures (one of the key issues at the moment), leading to more reliable nodes in the PS network and reduced chances for erroneous contributions from PS nodes. This approach also overcomes the bottlenecks of malicious data in the process caused by exploitation of the PS devices by some individuals (Budde et al., 2017). The lurker phenomenon is also here partly addressed, because the amount of data generated for air quality is no longer dependent on few individuals. Finally, housing companies as *end-users* can provide broadcasting services related to the surrounding environment, which can keep citizens updated (without being responsible for device management). Beyond addressing PS issues, the participation of housing companies in PS frameworks could (a) help them make services to their residents more attractive and updated; (b) create a profile of being ecological and innovative; (c) show the importance of proximity air quality monitoring around their property to potential customers; (d) maintain their loyalty to being resident friendly; and (e) provide data not only to their residents, but to the city at large.

There are also few drawbacks of the approach. The survey has shown, for example, that some housing companies may not be willing to share their data. This poses the question of ownership of the jointly collected air quality datasets, and it's unclear how the mediation between the different stakeholders (city council, citizens, researchers, housing companies) could be best orchestrated. In addition, successfully addressing data calibration and maintenance issues relies on the commitment of housing companies (which is likely, but may not be guaranteed). Finally, it is also possible that housing companies might act as *lurkers* in the proposed framework (but this is less likely compared to individuals level data gathering because of the spur of a competitive market on housing companies' business).

6.5.2 Addressing privacy concerns

As discussed in Section 6.2, privacy is another constraint which impacts participation in the PS approaches. The question of getting participants to contribute without identifying them has attracted lot of attention from previous work. By involving housing companies as one of the contributors for PS, we can enable shifting the sensitive information collected from the individual level to one group of people living at a certain location. With this grouping, we no longer need central data analyses and collection at the individual level. This approach has been termed "group privacy", where data is no longer gathered about one specific individual or a small group of people, but rather about large and undefined groups (Taylor et al., 2016). The involvement of housing companies can help collect data to better understand the concept of "group privacy" (Taylor et al., 2016), in the context of PS frameworks. Besides, the data collected will be inherently dis-aggregated and

therefore anonymous from inception. This, in turn, may be helpful in making data easily accessible and open for sharing. There are few drawbacks though and it is worth mentioning that there is always a trade-off between information sharing and services. There is thus the possibility of individuals missing some interesting personalised service due to the lack of more fine-grained location data. In addition, the scientific community has yet to provide effective techniques to fully prevent the identification of participants when their data is integrated with other data sources. Full anonymity of the residents, may thus not be guaranteed.

6.5.3 Further opportunities for GIScience

Beyond data completeness issues, and location privacy threat mitigation, involving housing companies in PS networks presents additional opportunities for GIScience. For instance, as pointed out in Richardson et al., 2013, spatial data holds an enormous potential for creating discovery in health research. Distributed spatial infrastructures are key to tap into this potential. A PS initiative with residents and housing companies as participants could provide valuable input for such an infrastructure. Another area where such an initiative would be beneficial is that of an open smart cities. Roche, 2014 presented four dimensions of a smart city: the intelligent city dimension (i.e., social infrastructure and civic spatial engagement practices), the digital city dimension (i.e., urban informational infrastructure), the open city dimension (i.e., governance based on the concept of open democracy) and the live city dimension (i.e., continuous adaptability to change). The approach proposed here is arguably relevant to the dimensions of digital city (feed the urban informational infrastructure), open city (enable the democratisation of environmental data collection), and live city (provide material for fine grained assessment and decision-making regarding environmental change).

6.6 Conclusion

Participatory sensing has a great potential for air quality monitoring in cities, but its success depends on the amount of participants, spread and quality of data collected. In this paper, we discussed two open issues of participatory sensing frameworks for air quality monitoring: data completeness and privacy. We proposed the inclusion of housing companies as one more stakeholder in PS framework for data collection to mitigate current data gaps & privacy issues. To understand what housing companies perceive about their inclusion, we conducted a questionnaire (N=18) administered to executive and planning staff of housing companies in Germany. The companies surveyed showed interest in using low-cost sensors for air quality monitoring and in sharing data with institutions which can analyse and process data for proper

understanding, but less inclination about sharing the data with the public. Further work is needed to establish whether people residing in the housing companies' premises would be willing to use such services at all. The view of the residents will be critical for the ultimate adoption of the proposed approach in reality.

Acknowledgements: Comments from two anonymous reviewers have helped improve the clarity of the article. We gratefully acknowledge funding from the European Union through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, <http://www.geo-c.eu/>).

A low-cost open hardware system for collecting traffic data using WiFi signal strength

Based on: Gupta, S., Hamzin, A., & Degbelo, A.,(2018). A low-cost open hardware system for collecting traffic data using WiFi signal strength. *Sensors* (Under Review).

Abstract

Road traffic and its impacts affect various aspects of wellbeing with safety, congestion and pollution being of significant concern in cities. Although there have been a large number of works done in the field of traffic data collection, there are several barriers which restrict the collection of traffic data at higher resolution in the cities. Installation and maintenance costs can act as a disincentive to use existing methods (e.g. loop detectors, video analysis) at a large scale and hence limit their deployment to only a few roads of the city. This paper presents an approach for vehicle counting using a low cost, simple and easily installable system. In the proposed system, vehicles (i.e. bicycles, cars, trucks) are counted by means of the WiFi signals. Experiments with the developed hardware in two different scenarios - low traffic (i.e. 400 objects) and heavy traffic roads (i.e. 1000 objects) - demonstrate its ability to detect cars and trucks. The system can be used to provide rough estimates of vehicle numbers for streets not covered by official traffic monitoring techniques in future smart cities.

Keywords: Traffic counter, WiFi signals, open hardware, traffic monitoring, low cost sensors, smart cities

7.1 Introduction

The rapid escalation of the population in urban spaces accompanied by increasing demands for mobility in cities (Lord and Washington, 2018), leads to substantial challenges in city planning. Increasing demands on mobility lead to growing traffic on the road, inducing suffering for citizens concerning the reduction of travel efficiency, increase in fuel consumption and health hazards from air and noise pollution caused by vehicles. In addition to being a significant source of air pollution in cities, road traffic exposes a large number of people to high daytime noise levels (Birgitta et al., 1999; Cai et al., 2017). Road traffic also leads to anthropogenic heat that together with reradiation effects from urban spaces can increase urban space temperature, resulting in urban heat islands (UHI) (Xu, 2017). Hence, it is of great importance to monitor the complex interplay of the road network and traffic conditions for better of sensing Quality of Life (QoL) in future smart cities.

Road traffic is one of the major source of air pollution in cities (Agency, 2017). It is a significant anthropogenic source of NO_x (Mertens et al., 2016), particulate matter (PM) and other harmful pollutants which impact human health (Agency, 2017). Exposure to road traffic induced air pollution can lead to various health impairment for the current as well as to the future generation. Multiple studies demonstrated the association of traffic generated air pollution to different heart-related disease in adults as well as for pregnant women (Pedersen et al., 2017; Roswall et al., 2017). Traffic data is one of the critical input variables for air pollution modelling approaches (Gulliver et al., 2018; Forehead and Huynh, 2018).

For many years, various approaches are devised to monitor air pollution at a higher resolution in the city (Forehead and Huynh, 2018). High-resolution monitoring approaches require input parameters also at a higher resolution. However, most studies used traffic models and simulation to represent traffic data because the traffic monitoring datasets are usually available for a very limited number of roads in the cities, hence limiting the possibilities to model air pollution at higher resolution in the city.

Traffic data collection was traditionally performed by using manual processes or with the application of inductive loops at certain locations (Polk et al., 1996). The inductive loops for traffic monitoring became standards in many jurisdictions and are widely utilised till date (Grote et al., 2018). Various other conventional traffic monitoring approaches include passive infrared devices, Doppler and radar microwave sensors, acoustic detectors, magnetic strips, Piezoelectric sensors, Pneumatic road tube counting devices and video vehicle detection. However, these approaches

inherit certain limitations (notably installation and maintenance costs) making them hard to deploy for detailed data collection with better spatial coverage in cities.

The vision of "smart cities" was proposed to address particular problems caused by urbanisation and to promote sustainable urban development in cities. This vision relies on the efficient application of information and communication technology (ICT) for sensing, analysing, integrating critical information which can support efficient operation and development of cities. Improved traffic control was identified in Hancke et al., 2013 as one of the possible benefits of advanced sensing in smart cities. Taking into account the emergence and rapid growth of the Internet of Things (IoT) and analytical tools, future cities may be able to enhance the execution and connectivity of urban services, reduce costs and operate on better resource management. The recent advancements in microelectronics, telecommunications and data analysis domains have led to the growing adoption of smart devices. With the application of these smart devices, it is possible to overcome detailed traffic data collection challenges. Extending the deployment of the low-cost IoT devices in a distributed model like crowdsourcing can support well-spread data collection in cities. Altogether, the recent developments in low-cost hardware and support from the crowd can help gather data, which can support transport planning and urban health risk assessment for cities.

In this paper, we present a novel system based on open-source hardware that has the potential to benefit traffic monitoring technologies because of its low-cost, privacy-preserving, ease of application and potential to large-scale deployment. Because the system is low-cost (less than \$50), it can be used to involve citizens in WiFi-based traffic data crowdsourcing projects, and this way expand traffic data collection to streets currently not covered by conventional traffic monitoring techniques. Section 7.2 briefly discusses the previous work done related to the topic. Section 7.3 describes technologies and the sensor used for the traffic monitoring and presents the algorithm used to infer the results from the proposed system. In Section 7.4, we present the results we obtained concerning the performance of the proposed system in the real world, using two different scenarios. Section 7.5 discusses the results we obtained, as well as the applications and limitations of the proposed system. Section 7.6 and 7.7 presents the future work and concludes the work.

7.2 Related work

Traffic monitoring in cities involves, among other things, estimating the number of vehicles on the road. The vehicles are tracked from point to point along the road for their information. A traffic monitoring station is used to measure traffic parameters

such as vehicle count, speed and occupancy at a specific location. The measurement at monitoring station locations is to be representative of the traffic on the road. Generally, vehicle detection and traffic surveillance involve a network of devices deployed at various roads of the cities. This section presents a brief overview of the various type of technologies and devices used for traffic surveillance with a particular focus on low-cost sensors, as it is also the primary focus of this paper. Furthermore, this section also discusses in brief privacy-related concerns while deploying traffic surveillance systems.

7.2.1 Traffic monitoring techniques

Traffic monitoring is a vital, yet challenging since to built traffic density maps traffic parameters such as vehicle count, location, speed and follow of vehicles are required. One of the essential requirement for efficient traffic systems is the reliable and real-time traffic data collecting network of devices to facilitate instantaneous decision-making. The technologies used in the devices for vehicle detection and traffic surveillance can be classified into five following categories¹:

1. Intrusive devices
2. Non-intrusive devices
3. Off-roadways devices
4. Sensor combinations devices
5. Relatively low-cost devices

Intrusive devices

Intrusive devices are installed directly into the pavement surface by creating saw-cuts or holes in the road surface, by burrowing them under the surface, or by anchoring them directly into the pavement surface. Devices such as inductive loops (IDL), magnetic detectors, micro-loop probes, pneumatic road tubes, piezoelectric and other weigh-in-motion devices are considered as intrusive devices. These devices are highly accurate for vehicle detection (> 97%) (Oh et al., 2002). However, a major drawback concerning the utilisation of the intrusive devices is the disruption of traffic caused for installation, repair and failure associated with installation in

¹See also the list of traffic sensing technologies frequently employed in traffic surveillance for data collection provided in Nellore and Hancke, 2016.

poor surfaces and use of substandard installation procedures (Mimbela and Klein, 2000). These devices are also expensive, large and consume much power which limits their implementation for better spatial coverage in cities (Balid et al., 2018). Resurfacing and repair tasks on the roads can also create the need for reinstallation of these devices. The safety of workers, those who are deploying these devices has also been a matter of concern (Balid et al., 2018).

Non-intrusive devices

Non-intrusive devices are a more reliable and cost-effective vehicle detection and surveillance devices than intrusive devices. They can be easily installed, maintained with safety with minimal disruption to traffic flow, and can provide traffic data with similar accuracy to that of inductive loop detectors (Mimbela and Klein, 2000). Non-intrusive devices include technologies such as video image processing, microwave radar, laser radar, passive infrared, ultrasonic, passive acoustic array, in which devices are mounted overhead on roadways or roadsides. These devices are capable of measuring vehicle count, presence, and passage on the road. Some devices also have the potential to provide vehicle speed, vehicle classification, and multiple-lane, multiple-detection zone coverage (Buch et al., 2011; Tang et al., 2017). However, the devices fail to perform in certain environmental condition. For instance, infrared devices can be affected by fog, and temperature change, video image processing devices detection efficiency can be hampered by weather conditions, shadows, vehicle projection into adjacent lanes, day-night transitions, vehicle/road contrasts and water salt grime or cowebs on camera lens (Mimbela and Klein, 2000). The high cost involved in the aforementioned technologies limits the large-scale integration of these devices into the traffic surveillance systems.

Off-roadways devices

These devices utilise the technologies that do not require any hardware deployment under the pavement or mounted overhead/roadside. The devices enable traffic monitoring via aircraft or satellite, as well as by probing the vehicles equipped with Automatic vehicle identification (AVI), Global Positioning System (GPS) and mobile phones (Martin et al., 2003). These technologies can help in enabling the high percentage of roads coverage. However, privacy concerns and other technology-specific limitations restrict their application (Cheung and Varaiya, 2006).

Sensor combinations devices

Due to certain limitations of individual technologies, various studies suggested the application of the off-roadway devices together with more than one technology to monitor traffic flow on the road. Applications include the combination of passive infrared with ultrasound and Doppler microwave radar, which enhanced the accuracy for vehicle detection in queues and counting them along with their height and distance discrimination (Mimbela and Klein, 2000). Nevertheless, the cost of deployment and its complexity limit the well-spread deployment of a network in the cities.

Relatively low-cost devices

The scalability and availability of traffic monitoring systems are essential for efficient and reliable, real-time traffic monitoring (Orosz et al., 2010). Devices like Magnetometer (MAG) have been found to serve the requirement (Haoui et al., 2008), but the maintenance and installation cost along with limitations with radar detectors impact the performance (Haoui et al., 2008). Low-cost, portable, and easy-to-install technologies are desired to supplement existing data sources for efficient, detailed traffic monitoring in the cities (Balid et al., 2018). The availability of new low-cost and miniaturised hardware platforms has enabled the idea of developing advanced and pervasive image-based devices, which can help in vehicle counting and traffic surveillance. Till date, several approaches have been proposed to investigate the feasibility of vehicle detection and traffic surveillance using low-cost sensors.

A method which uses continuous-wave radar was presented by Fang et al. (2007). The method uses an antenna, a microwave radio front, the analogue signal amplifier and the digital signal processor (DSP) for vehicle detection. A computer vision application enabling vehicles monitoring by using low-cost and low-complexity devices was proposed by Salvadori et al. (2015). Recently, Wifi signal based approaches were used for assessing human activity recognition (Depatla et al., 2015; Wang et al., 2013b; Hong et al., 2016; Pu et al., 2013), suggesting possible additional applications of the WiFi technology other than providing easy internet access. Approaches based on channel state information (CSI) (Won et al., 2017), link quality indicator (LQI) (Roy et al., 2011), packet loss rate (Roy et al., 2011), and received signal strength indicator were proposed for vehicle detection using radio waves (Horvat et al., 2012; Haferkamp et al., 2017). However, the methods mentioned above are not well suited for crowdsourcing applications because of their expensive specialised hardware (laptop specific WiFi cards and modules) or energy data transfer requirements. With regards to computer vision applications, low-cost devices

constrain its computational capabilities and available onboard memory, making it unfeasible for effective implementation (Salvadori et al., 2015). Approaches utilising Bluetooth low-energy beacons and smartphones were also proposed for enabling crowdsourcing of traffic data with low-cost devices (Lewandowski et al., 2018). However, dependency on smartphone and its utilisation of approach by users at the roadside seems a bit impractical. Moreover, gathering data by utilising smartphones or video analysis based system also impose the threat on the privacy of the commuters on the road.

7.2.2 Privacy and Traffic Monitoring

Traffic surveillance and vehicle detection systems are promising approaches which are cyber-physical by nature. These cyber-enabled systems face various security and privacy preserving challenges (Lu et al., 2013). If the vehicles' location privacy cannot be preserved, commuters may object to the process of being monitored in such systems. If the devices which can collect privacy sensitive data is handed over for crowdsourcing, it could be a bigger threat to the society. Hence, the privacy devices would be capable of appropriately protecting the privacy of the vehicle and the commuters on the road (Hoh et al., 2012; Yang et al., 2015). With the wider application of computer vision technologies for traffic monitoring, it is important to realise the impact of the approach on the commuters. Privacy and security is one side effect caused by the application of computer vision methods (Cote and Albu, 2017). Several approaches are proposed to solve the privacy problem in traffic monitoring approaches. Lu et al. (2008) proposed conditional privacy-preserving protocol, Lin et al. (2007) presented conditional privacy and group signature building techniques. The techniques for dealing with unlinkable pseudo-ID were also proposed by Raya and Hubaux (2007). However, these are not implemented for the latest transition of approaches to low-cost devices. Since the primary focus of the paper is to enable counting and identification of vehicles on the road using low-cost devices for crowdsourcing, it is easy to have a system which only enables the measurement of the specific parameter and discarding sensitive data collection on the origin itself.

The system proposed in this paper utilises the low-cost open hardware which utilises the WiFi signal, which is commonly available at all locations these days to traffic data. The approach has the potential to overcome various limitations existing in current systems discussed above and also preserve the privacy of the commuters.

7.3 Materials and Methods

7.3.1 System Design

In this section, we present the proposed vehicle detection and counting system utilising the received signal strength indicator (RSSI) data produced from the router and collected by the receiver, a low-cost open hardware system. The deployment plan for the proposed traffic monitoring system is illustrated in Figure 7.1

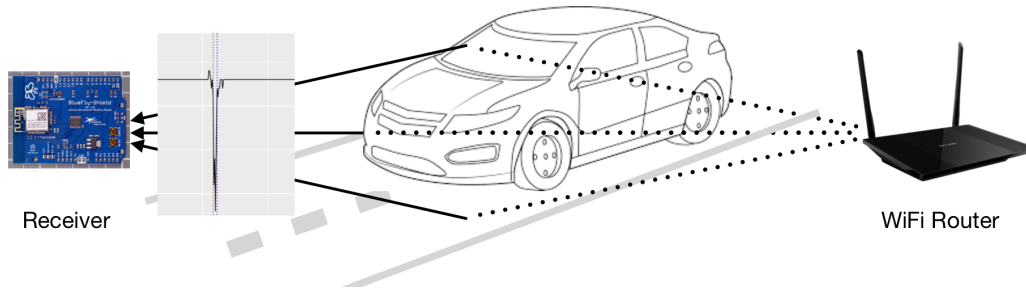


Figure. 7.1. Illustration of deployment plan for the proposed hardware system

Receiver

The receiver hardware system was developed using an Arduino Uno R3 single-board microcontroller mounted with a BlueFly-Shield (ATWINC1500) and an SD card shield V3.0 (Model: INT106D1P). The hardware system receives the WiFi signal transmitted from a TP-Link router for our study case.

Arduino Uno R3 is a simple microcontroller with simpler software structure. The Arduino works according to the modular principle with the simple procedure to add components. We mounted two components to capture and store Wi-Fi signals. The first component was the Wifi shield (ATWINC1500), which receive the Wi-Fi signal from connection created via IEEE 802.11n standard and works with the encryption type WPA2. The second component was an SD card shield V3.0 (Model: INT106D1P) that stores the Wi-Fi strength in dB and time in milliseconds. The code was written in the integrated development environment (IDE) which runs on the chip. No firmware, interpreter or operating system was involved in the process, which makes the whole procedure easier to implement and also limit the noise when receiving the signal.

Transmitter

For transmission of the WiFi signals in the study, we used the router from TP-LINK (Model: TL-WDR3600) with features like dual-band with 2.4 GHz & 5 GHz bands

and an Atheros Chip. We installed OpenWrt², an open source project based on Linux with the ability to allow specific changes in physical settings of the radio hardware such as operating frequency, transmit power and encryption. OpenWrt runs on a router only with necessary scripts. These scripts/activities can be enabled or disabled at any time. As a result, the router works in a simple version without heavy background activities. The following settings were used during our study:

- Band: 2.4GHz
- Standard: Wireless-N
- Width: 20MHz
- Encryption: mixed WPA/WPA2 PSK (CCMP)

The 2.4GHz band was used in our study instead of 5GHz because of the wide range and the compatibility with the receiver hardware system. It should be noted that introduced system structure, which includes low-cost open hardware system has not been considered in the literature. The transmitter and the receiver were installed at different (opposite) sides of the road, using WiFi signals to communicate with each other. The interference caused by the passing vehicle to the communication pathway between the two devices is measured by the receiver which is then stored in the SD card mounted present in the device. Distinct patterns of change in RSSI are observed when vehicles pass, which is captured and further utilised for analysis to count and detect the type of vehicle on the road.

7.3.2 System Implementation

For the implementation of the proposed system, we deployed the receiver and the transmitter on the roadside as shown in Figure 7.1. The proposed system was evaluated for two different scenarios: 1.) low traffic road 2.) heavy traffic road. The low traffic road in our study is the road called Heisenbergstraße, which can also be considered as the local road with fewer cars. The heavy traffic road we used as an environment in our study is Steinfurterstraße, which is one of the busiest roads in the city. Both the roads are situated in the city of Muenster, Germany. More than 500 objects for Scenario 1 and more than 2000 objects were recorded for Scenario 2 during the real-world field data collection. In order to collect the ground truth data we deployed GoPro HERO4 camera that collect video data. Figure 7.2 and 7.3 depicts the setup of the proposed system in the two different scenarios under consideration. The low traffic scenario was considered in the study to understand

²<https://openwrt.org/> (last accessed: August 21, 2018).

the accuracy of the system under limited vehicles with complexity caused by bicycles, pedestrians and other objects on the road. In contrast, the heavy traffic scenario was considered to access the performance of the proposed system with complications created by the large number and frequency of vehicles on the road with two lanes.



Figure. 7.2. Experimental setup: Scenario 1 (Low traffic road)



Figure. 7.3. Experimental setup: Scenario 2 (Heavy traffic road)

In order to evaluate the accuracy of the proposed system, it is important to prepare the ground data video stream at the same scale as the data collected from the receiver. We have also developed a web application which enables the processing of video streams and register the type, number and time stamp of the object on

the road. Figure 7.4 illustrates the developed web-application. The user of the web-application have to watch the video and select the type of object (car, truck, bike, pedestrian and so on), the system then takes the timestamp and count the total number of specific type of objects identified automatically. The input to select the type of object can be done by clicking on the web application's buttons or by using keyboard shortcuts. The outputs of the video data analysis are stored as JSON files, with each file representing a vehicle type under study with the timestamp in milliseconds, same time scale as of the dataset generated by the proposed system. During the study, the objects were mainly counted by one researcher. However, a second researcher independently redid the count on 20% of video data for Scenario 1 to confirm the number of objects counted. The Cohen's kappa coefficient was calculated to assess the agreement between the two, and the result indicates a very high overall inter-rater agreement (0.8 for bicycles and 1 for cars). The web application used can be customised as per the requirement and is available online (see Gupta, 2018h) as an open-source tool under Creative Commons licence.

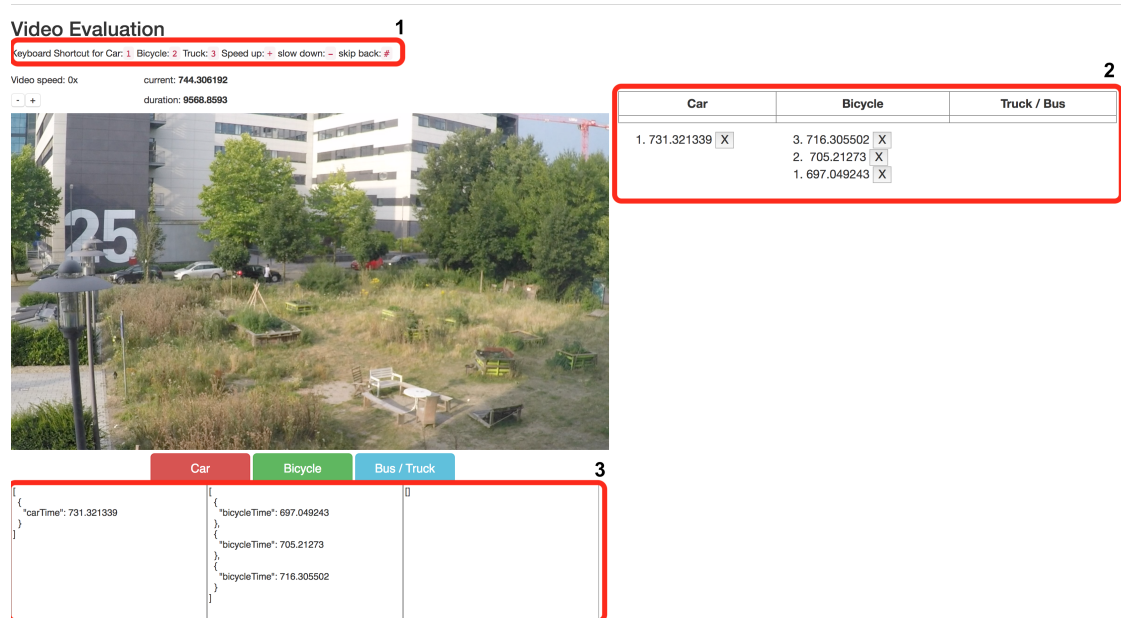


Figure. 7.4. Illustration of the web-application developed for video stream analysis

7.4 Results

This section evaluates the performance of the proposed hardware device. The RSSI changes detected by the open-hardware was stored in an SD-card. The stored data stream was then processed to evaluate the performance. The data processing involves multiple steps as discussed below:

7.4.1 Vehicle Detection Algorithm

An algorithm was used to detect the objects in the data stream collected by the device. Two parameters were used:

1. RSSI change for object detection
2. Time window for object detection

Change detection in RSSI value

Whenever an object interferes with the communication path of receiver and transmitter, the RSSI fluctuates. The pattern in fluctuation can be useful to detect the vehicle movement on the road. In order to isolate the fluctuation patterns in the data stream, during preprocessing, we removed all fluctuations of strength less than or equal to 2dB, as noise. The deletion of the 2dB signal was concluded to ignore the usual noises in the WiFi signals transmission and on the receiving end. The leftover data stream was used to detect vehicles, based on the fluctuations patterns.

In order to segregate patterns which can represent vehicle movement, the data stream was analysed. The time window needs to be determined whenever any high fluctuation is observed. In the algorithm, we identify a pattern in signal fluctuation by comparing the signal strength at each time to the preceding and subsequent time stamp of the data. During comparison, if the signal strength change is significant for the following three time stamps we start recording the initial time (say T_1) and the end time (say T_2) where the significant signal fluctuation stops. The T_1 and T_2 recorded for each fluctuation pattern provides the time window and respective signal strength that convey the presence of some object (see Figure 7.5).

After the identification of the time window and the signal strength associated, the maximum signal strength change for each time window was computed. This maximum change in signal strength in a given time window can help in identifying the type of object. However, threshold values are required to define the rule based on which the algorithm can distinguish between the different types of objects.

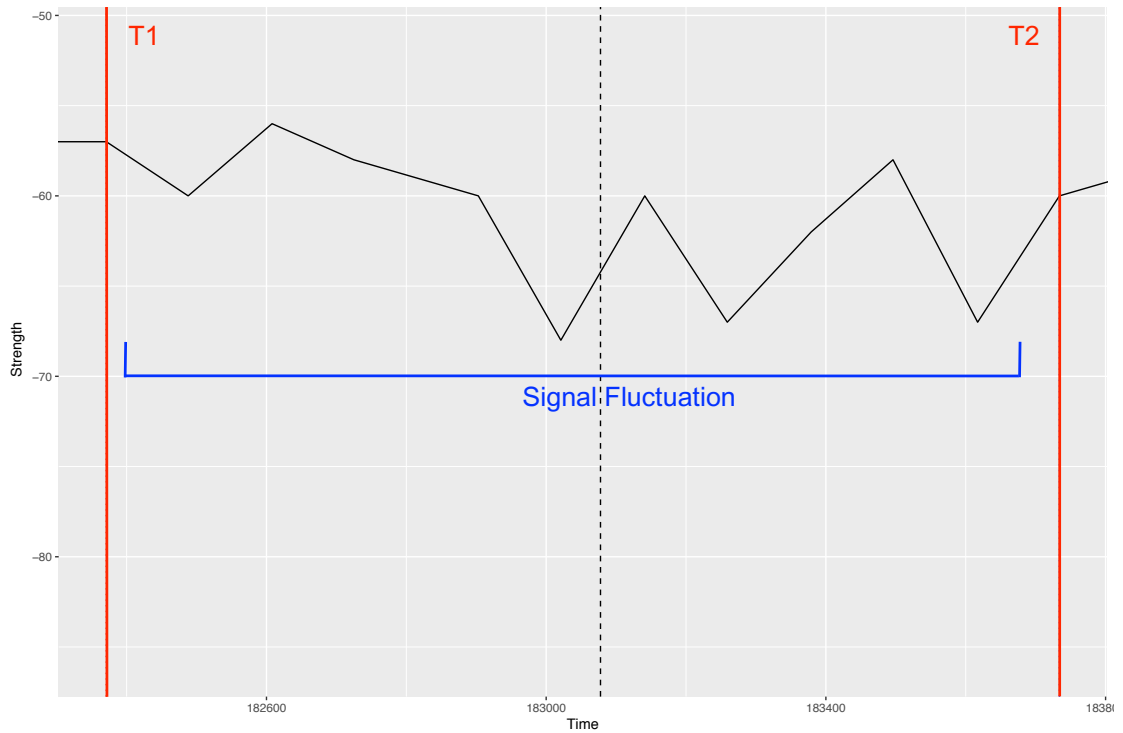


Figure 7.5. Illustration of time window and associated signal fluctuation pattern identification.

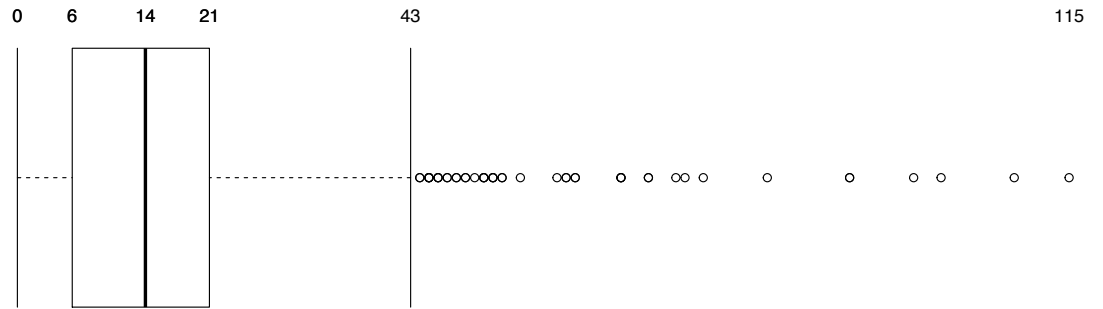
Time window for object detection

As can be inferred from the Figure 7.5, the time window is used for recording the patterns in the data stream collected. We also used the the length of time window, i.e., difference between T_2 and T_1 to identify the object in the algorithm (Equation 7.1). The intuition behind this is that different objects can produce disturbances of varied temporal length. For instance, if the fluctuation continues for a long time window, there is a possibility of having a long object between the transmitter and receiver communication path. Here also, deciding the threshold for the time window size is also important to characterise different sizes of objects.

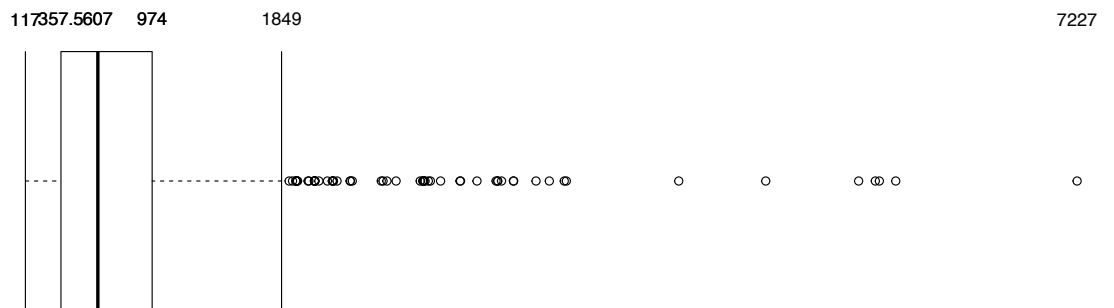
$$\text{Time window size} = T_2 - T_1 \quad (7.1)$$

7.4.2 Vehicle detection

In the study, we used the hardware device to identify the vehicle, namely bicycles, cars, and trucks. As discussed previously, defining thresholds is necessary to charac-



(a) Box plot of RSSI maximum fluctuation value parameter from the isolated patterns of data stream collected by the proposed hardware at Steinfurtstrasse.



(b) Box plot of time window size value parameter from the isolated patterns of data stream collected by the proposed hardware at Steinfurtstrasse.

Figure. 7.7. Parameters summary statistics for Steinfurterstrasse.

Based on the summary statistics (and visual inspection of the data), we defined the most plausible thresholds for vehicle identification. Threshold values were chosen so that they coincide with the values of the quartiles. Tables 7.1 and 7.2 present the threshold rules we used for vehicle detection in each of the two test scenarios under consideration.

Table. 7.1. Threshold rules for vehicle identification using Heisenbergstrasse data

Vehicle	Threshold
Cars	≥ 611
Bicycles	< 611

Table 7.2. Threshold rules for vehicle identification using Steinfurterstrasse data

Vehicle	Threshold
Trucks	>1849
Cars	>=357.5 & <=1849
Bicycles	<357.5

7.4.3 Vehicle count

After defining the threshold rules, the algorithm is capable of identifying the vehicle type. The vehicle identification finally helps in counting the number of vehicles using the particular road. We compared the accuracy of the algorithm with the ground truth data we measured after interpreting video recording into JSON files for each test case scenario. Table 7.3 and 7.4 present the comparison between the ground truth data from video and vehicle classification using the algorithm parameters we used in the study.

Scenario 1 : Heisenbergstrasse

Table 7.3. Vehicle classification using algorithm and video

Vehicle	Vehicle detected using algorithm	Vehicle detected in the video
Using time window parameter		
Cars	176	182
Bicycles	510	467
Using maximum RSSI parameter		
Cars	177	182
Bicycles	371	467

Scenario 2 : Steinfurterstrasse

Table. 7.4. Vehicle classification using algorithm and video

Vehicle	Vehicle detected using algorithm	Vehicle detected in the video
Using time window parameter		
Trucks	64	45
Cars	826	1000
Bicycles	297	66
Using maximum RSSI parameter		
Trucks	45	45
Cars	842	1000
Bicycles	31	66

7.4.4 Precision, Recall and F Measure

To understand the reliability and performance measures of the proposed device in the two different test case scenarios, we computed the precision, recall and F measure. The calculation for precision, recall and F measure was done using the following equations:

$$Precision = \frac{A}{P} \quad (7.2)$$

$$Recall = \frac{A}{V} \quad (7.3)$$

$$F\ measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7.4)$$

where, P is the total number of objects detected using the parameters of the algorithm, A is the number of P objects actually overlapping (temporally) with objects in the ground truth data, and V is the total number of objects detected in the video. Table 7.5 and 7.6 summarise the results for each object in both the test case scenarios.

For Heisenbergstrasse:

Table. 7.5. Precision, recall and F Measure using Heisenbergstrasse data

Precision	Recall	F measure
Time window change parameter		
For Cars (A = 61)		
0.3465	0.3351	0.3407
For Bicycles (A = 40)		
0.0784	0.0856	0.08188
Using maximum RSSI parameter		
For Cars (A = 64)		
0.3615	0.3516	0.3565
For Bicycles (A = 47)		
0.1266	0.1006	0.1121

For Steinfurterstrasse:

Table. 7.6. Precision, recall and F Measure using Steinfurterstrasse data

Precision	Recall	F measure
Time window change parameter		
For trucks (A = 21)		
0.3281	0.4666	0.3853
For cars (A = 425)		
0.514	0.425	0.465
For bicycles (A = 2)		
0.0067	0.0303	0.0110
Using maximum RSSI parameter		
For trucks (A = 15)		
0.3333	0.3333	0.3333
For cars (A = 451)		
0.5356	0.451	0.4896
For Bicycles (A = 0)		
0	0	0

7.5 Discussion

This study has explored the potential of using low-cost hardware for WiFi-based vehicle count. Given existing initiatives to increase free WiFi hotspots in cities (e.g. European Commission (2018a)), there is a need for techniques which take

advantage of WiFi availability for traffic monitoring purposes. The advantage of low-cost sensors over expensive, highly performant sensors is that they can be bought by a large number of citizens. As a result, they can be deployed on any road of the city, enabling data collection about traffic at places currently uncovered by official traffic monitoring techniques. It can be inferred from the results that the proposed system is capable of identifying some types of vehicles on the road with limited reliability. The classification accuracy on the low traffic road suggests the ability of the system to identify large objects like cars with higher accuracy than other objects like bikes. That is, the proposed system is useful for collecting data about cars and trucks in the residential areas, where traffic data is not generally collected.

Looking at Tables 7.5 and 7.6, one can conclude that the system should *not* be used to count bikes, as the values for precision/recall are simply too low. One can also notice that the recall values are roughly around 0.33 for cars/trucks in the tables. This suggests that counting cars/trucks using the proposed technique can be used to do *rough estimates* about the number of cars/trucks which have been on the street during a time period. The recalls being around $0.33 = 1/3$ means that if one would detect 2 cars/trucks with the code, there were roughly 6 cars/trucks on the street; if one has 1000 cars/trucks in the code, roughly 3000 cars/trucks were available in reality, and so on. Lastly, it should be noted that the numbers of vehicles identified in Tables 7.3 and 7.4 are actually quite close to the numbers of the vehicles in the videos (which suggests a good overall accuracy for the technique). However, the precision/recall values computed subsequently (Tables 7.5 and 7.6) led to much lower numbers when it comes to the overall performance. This could be due to the process of computing precision/recall values during the work. An object from the code was said to correspond to an object in the video if and only if the time T_0 of the video object was within the time window $[T1, T2]$ of the signal fluctuation (see Figure 7.5). Since the units of measurements were all in milliseconds, and the times T_0 of the video objects were manually identified, there are chances, that overlaps between the two were missed by the algorithm. Relaxing the ‘if and only if’ constraint a bit could have led to different values of precision/recall.

We believe that our system can help in overcoming the limitation imposed by other traditional existing methods. The proposed system does not require any installation within the road surface or overhead that can lead to hindrance in the normal traffic flow. This advantage of the proposed method helps in overcoming the limitations imposed by various intrusive methods as discussed by Mimbela and Klein (2000). The proposed system only performs the simple operation making the data storage and flow easy and less in volume, which helps in overcoming the barriers imposed by low-cost computer vision based traffic detection methods. Overall the proposed system addresses limitations (i.e. cost, privacy concerns, high maintenance requirements, weather and light effect) imposed by the various methods discussed in the Section

7.2. Since the system uses WiFi signals, it is tolerant to weather conditions, such as rain or thunderstorms, which is a significant advantage over various computer vision based devices, ultrasonic devices and related technologies. Therefore, it can be considered as an excellent fit to overcome the performance issues, such as the effect of weather and light condition, pointed out by Balid et al. (2018) for vision-based or radar-based low-cost devices. The system only uses RSSI of WiFi signal to access the vehicles on the road, which can be considered as the simpler version of the approach suggested by Won et al. (2017). Lastly, the technique proposed does not use any sensitive data, making it privacy-friendly during traffic data collection.

The system is also subject to some limitations, one major one being its inability to differentiate between noise and objects in a few instances. This can be because of the low power of the WiFi signals. During the study, the system represented the slow-moving pedestrian and fast-moving cars as the same object which leads to false vehicle identification or sometimes even missing the vehicle. The proposed method also has specific issues concerning the identification of a big object like trucks. The results we obtained while analysing the dataset from scenario 2 of the study suggest the limitation of the current method to identify the long objects, as the signal transmission path get used to new normal due to low time interference, missing the data in the process. The low-cost sensor also has its limitations to function at the specific temperature, or sometimes the sensor behaves unexpectedly leading to no signals or very high peaks in data. These limitations are inherent to any low-cost sensing device. Another limitation could be the fluctuating tolerance of WiFi receiver to specific signal strength, making some of the legitimate peaks of objects as noise in the dataset.

Another critical factor is the selection of thresholds during the study. The limits defined for identifying vehicles from the receiver dataset are based on the summary statistics (i.e., using quantile ranges), which is one way of making the threshold selection systematic. However, the quantile-based threshold values defined may not be ideal in all situations. Weather and other surrounding situations like the gathering of people or proximity to vehicle parking can trigger changes in values at any point during the deployment. These factors can change the observation value fluctuation to other extreme values, changing the overall quantile range of the data stream, leading to a sub-optimal threshold definition for vehicle classifications. Though threshold values are necessary for object identification, choosing the optimal threshold systematically remains a question which needs further investigation.

In all, the proposed system is capable of overcoming some limitations which exist in various traffic monitoring methods. It is also useful in addressing the privacy concerns, making the data collection process easy and fast for real-time traffic monitoring. The system tolerance to weather and luminescence related conditions

make this method more advantageous. The very low-cost of the whole infrastructure (\$50) can facilitate the traffic data collection using a large number of devices, enabling better spatial spread for detailed city-level traffic data. The easy deployment capability and operation with generally available WiFi signals make the system useful for crowdsourcing, which can encourage citizen participation and open data collection for open smart cities initiatives.

7.6 Outlook

The proposed system currently utilises the summary statistics for threshold identification for vehicle type identification and extending it to utilise machine learning based approaches may help in removing more noise in the datasets and improve the overall accuracy. Currently, the technique has been only used to detect one vehicle at a time. Extending it so that it can identify the aggregated peaks of two objects by using various statistical approaches could also be worth considering in future work. Another future work could be to improve the performance of the proposed system by developing new modules which can help in differentiating various small and big object with speed parameters. Extending the current method with approaches suggested in Roy et al., 2011; Won et al., 2017 for low-cost open hardware can also be considered for future work.

The accuracy of the proposed method is measured by using the video data which is converted to the same scale as data from the receiver by manually watching the video. This process is error-prone, as the time ascribed by the human being while counting the video object may not be the exact original time at which the object crossed the WiFi infrastructure. Future work may consider automating the current manual process with the application of vision-based analysis, such as using old phone cameras in combination with proposed hardware. The low-resolution camera from used phones or other electronic wastes with just enough resolution to differentiate between small and big objects on the road can help with the real-time feedback mechanism. Integrating the low-cost video devices data with WiFi receiver data stream can improve the overall accuracy of vehicle detection and counting.

7.7 Conclusions

In this paper, we presented a WiFi RSSI based traffic monitoring system using low-cost open hardware that is capable of providing essential functionalities for vehicle detection and counting. Real-world test results suggest that the proposed method is capable of detecting the big objects like cars in both low traffic and heavy traffic

scenarios. The proposed system is promising and can be improved significantly with more sophisticated tools and techniques for vehicle classification considering its low-cost, easiness and accessibility of technology used for its implementation. The proposed system is tolerant to weather conditions and also helps in preserving the privacy of the commuters by not using any sensitive data for vehicle identification. Since the technique is based on low-cost hardware, it has the potential to enable well spread detailed traffic data collection that can help in improving various services in the future smart cities such as transport planning, detailed air and noise pollution monitoring to improve Quality of Life (QoL).

Author Contribution: SG (Shivam gupta) and AD (Auriol Degbelo) conceived and designed the study; AH (Albert Hamzin) curated and analysed the data; Methodology developed by SG, AD and AH ; Video analysis tool developed by AH ;Validation done by AH and SG; Writing—original draft, SG; Writing—review & editing, AD and SG. Equal contribution of SG and AH.

Acknowledgement: The authors gratefully acknowledge funding from the European Union through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, <http://www.geo-c.eu/>).

Conflict of Interest: The authors declare no conflict of interest.

Abbreviations:The following abbreviations are used in this manuscript:

AVI	Automatic Vehicle Identification
CSI	Channel State Information
DSP	Digital Signal Processor
GPS	Global Positioning System
ICT	Information and Communication Technology
IoT	Internet of Things
LQI	Link Quality Indicator
MAG	Magnetometer
NO_x	Nitrogen oxides
PM	Particulate Matter
QoL	Quality of Life
RSSI	Received Signal Strength Indication
UHI	Urban Heat Islands

Synthesis

In this chapter, firstly we present a general summary of the wide variety of approaches presented in the thesis for spatial modelling of air pollution in open smart cities (Section 8.1). Following that, the results from Chapter 3 to Chapter 7, under the perspective of each research questions posed in the introduction (Section 1.4) will be summarised along with their contributions, limitations and future studies in Section 8.2. Afterwards, Section 8.3 presents the overall contributions of the thesis for Open City Toolkit (OCT). A discussion of a more general type that has not been expressed in previous chapters will be presented in Section 8.4. Finally, the outlook for the research work as a whole will be described in Section 8.5.

8.1 Summary

The purpose of this study was to investigate and develop methods to assist detailed air pollution monitoring for open smart cities. Combining the existing knowledge from literature with empirical investigations enabled us to reach the desired objectives. The literature study (Chapter 2) suggested the need for a flexible method that can mould to the constraints caused by limited data availability for enabling air pollution monitoring. With the help of the two statistical methods Land Use Regression (LUR) and Spatial Simulated Annealing (SSA), the thesis seeks to fulfil this requirement. However, air pollution monitoring at the city level is hard because of limited or no monitoring station data availability at the city level. Based on the findings from the literature study, an optimisation method was developed that can help in finding locations where the monitoring stations can be placed for initiating data collection in the city with the desired resolution. The optimisation method uses open data and identifies a set of locations which can help in developing the LUR model with a smaller spatial mean prediction error for the study area. Since it is more important to know precise ambient air quality for areas where people reside, the optimisation was further extended to find the set of optimal locations which can represent precise information about air pollutant considering the population distribution, with more focus on densely populated areas.

The air pollution data from traditional monitoring devices is limited because of their cost, maintenance and bulkiness, making them very sparsely arranged in

space and limited in numbers. This thesis investigated a systematic approach to identify optimal locations for utilising citizen participation opportunity wisely and address the sparsity and scarcity constraint of air pollution measurements, taking into account the motive of open smart cities to foster the participation of citizens, businesses and government. Furthermore, the thesis proposed the involvement of housing companies for collecting the air pollution data using low-cost sensors, which can help in overcoming certain sparsity and scarcity limitations along with various data gaps discussed in the literature concerning citizen participation. Overall, the approaches discussed in the thesis contribute to help in overcoming the limited data availability of air pollution measurements; however, the literature and the approaches we developed in this thesis also identifies a gap in the availability of detailed traffic data in the cities. Usually, traffic data is available for the limited number of roads, which limits constrain the magnification of various models for detailed air pollution monitoring within the cities. The present thesis also attempted to overcome the lack of traffic data availability by developing a low-cost open hardware based device. The hardware device uses WiFi Received signal strength indication (RSSI) for detecting and counting vehicles on the road to generate data about traffic flow on the road. The detailed summary regarding each specific research question of this thesis is discussed in the following subsections.

8.2 Summarised Results

8.2.1 *How can we use statistical methods like LUR and SSA for detailed air quality monitoring in an open smart city?*

Answering this research question was the primary focus of Chapter 3. In practice, air pollution monitoring for cities requires a significant amount data of various kinds. The significant amount of data required can be characterised as ‘big data’. Traditional statistical methods are limited in their capabilities to handle big data and challenges associated with it. In the study, we presented various big data challenges and how these challenges can impact the environmental monitoring for open smart cities, especially air pollution. In order to address the big data challenges and also to overcome barriers concerning data for detailed air pollution for open smart cities, the study discusses the application of two powerful statistical methods LUR and SSA.

The combination of the methods suggested in the Chapter 3 helps in extending the data collection process in a way that the maximum amount of data required for

detailed air pollution monitoring can be extracted from the limited number of data sources. The LUR approach provides flexibility for incorporating more theoretical knowledge about the process governing the spatial and spatiotemporal variation in air pollution. The SSA helps in identifying the finite number of locations by utilises context-aware goals, which can curb the collection of a significant amount of nonessential data. This control on the number and optimal locations for air pollution measurements making the combination of LUR and SSA method a useful process for addressing big data challenges. Furthermore, limiting the data requirements to selected variables in LUR and air pollution measurements to only optimal locations also help in overcoming the big-data challenges, which may focus considering the recent applications of new alternate data sources (e.g., IoT data) for air pollution monitoring. Even though the combination of LUR and SSA seems promising, the study also pointed out a few constraints associated with it. The incapability of the LUR method to distinguish between the impact of specific pollutants in the environment because of their symbiotic effects limits its performance. Furthermore, well-applied SSA function depends on the various essential input parameters. The calculation of appropriate parameters requires various test runs which can be time-consuming. Additionally, the SSA implementation to identify optimal locations may also require a significant amount of time, depending on the size of the area and objective under consideration.

Fundamentally, the power of statistics can help in solving various challenges of big data. The combined application of statistical methods like LUR and SSA can be a useful approach for extending air pollution monitoring at a finer scale for the open smart cities along with addressing big data challenges. However, it is also important to keep in consideration a few essential prospects. Firstly, concerning variable selection for developing legitimate LUR model. Secondly, about the design of SSA-based optimisation objective function which relies on the quality of input variables used in selected LUR model. Finally, the requirement of SSA input parameters, such as probability distributions and temperature change functions which are decisive for the quality of optimisation outcomes. The approach can also be useful for extending the monitoring of various other environmental factors impacting human health in cities, such as noise pollution.

8.2.2 How to systematically place stations for an air quality monitoring network to maximally reduce land use regression prediction errors?

Air pollution monitoring relies on the availability of monitoring station measurements. Air pollution modelling approaches like Land Use regression (LUR) uses the

monitoring station data for estimating the air pollution at the unmeasured locations. Commonly, one of the constraints for developing LUR models at a finer spatial resolution for the cities is the limited or no monitoring station data availability. To overcome the limited data availability, deploying few monitoring stations in the study area can help; however, sometimes the new sensor deployments may not be helpful for capturing the detailed spatial variability of air pollution in a city. Systematic selection of monitoring station locations for establishing the new monitoring station network or extending the existing monitoring network is desired for robust LUR (Hoek et al., 2008; Kanaroglou et al., 2005a). In Chapter 4, the question concerning systematic placement of air pollution monitoring stations was addressed by developing the optimisation method. The objective of the optimisation method was to find the set of locations which together can decrease the spatial mean of prediction error for a given LUR model. Usually, the air pollution monitoring stations are very sparsely arranged which act as a barrier for detailed city-level air pollution monitoring. Even though it is possible to access open data about the explanatory variables for LUR estimations, the limited number of sparsely arranged monitoring stations measurements restricts the computation for precise air pollution monitoring at a finer spatial resolution in cities. In order to overcome such a limitation in air pollution modelling, an optimisation method was developed which was presented in Chapter 4.

The optimisation method requires a given LUR model, and its predictor variables for finding the set of locations represent the minimal LUR prediction error for the study area. If the study area has monitoring station measurements for LUR model creation, optimisation function can use the particular LUR's covariates in SSA to find optimal locations for extending or redesigning the existing monitoring network. In the case where the measurements are not available, the study suggests using one of the ESCAPE project's LUR model for identifying locations to collect air pollution data by using the optimisation method. In any case, the essential requirement for optimisation method is the LUR model, and the predictor variables involved in it to run optimisation function.

Generally, air pollution monitoring aims to understand the impact of pollutants on the population living in the cities. In the study, the optimisation method mentioned above was further extended to find the optimal locations considering the weight of the population while identifying optimal locations. The extended optimisation approach helps in identifying the locations by prioritising the areas where the population is higher, as in, finding locations which decrease the prediction error for the populated areas of the city. The application of the optimisation method was demonstrated by the case study of the city of Münster. The result shows that the optimisation method performed well by decreasing the spatial mean of prediction error from the initial monitoring network configuration to final optimal configuration

by 99.87% and 99.94% using optimisation method without and with population weight respectively for the given LUR model. The result of the study was promising, which demonstrates the applicability of the method discussed in Chapter 3.

The optimisation method has the advantage of being flexible regarding input covariates integration and autonomy to monitoring station measurements for identifying locations to initiate air pollutant measurement collection. Which means, without the availability of air pollution measurement data we can start the process to identify locations that can help in detailed air pollution monitoring in cities. The optimisation approach also enables the transferability of the already existing LUR models, hence making the air pollution monitoring initiative more cost-effective (Hoek et al., 2008; Allen et al., 2011). Furthermore, the extended version of the optimisation method provides resilience for increasing and decreasing the weighted regions of the study area making it more flexible concerning applicability. The study also has few limitations concerning the selection of a particular LUR model in the case where there is no existing data for developing the LUR model for the study area. The selection of LUR can be arbitrary or based on specific variables of significant importance, leading to assumption based selection. The limitations also include the inherent assumptions of multiple regression concerning linearity between dependent and independent variables, independence and normal distribution of error terms which are typical for any regression studies (Beelen et al., 2013a; Beelen et al., 2009). Another limitation is regarding the utilisation of SSA, as it is a stochastic approach, which implies, every run of the optimisation with the same data and parameters may provide different outcomes (Helle and Pebesma, 2015).

Overall, the optimisation method developed was capable of identifying the optimal locations with a well-spread monitoring station locations in the study area. It is subject to the boundary effect, which is common for any SSA based optimisation (Van Groenigen and Stein, 1998). The extended version of the optimisation also demonstrated the acknowledgement of the population weight for identifying the optimal locations which decrease prediction error for the populated areas of the city. The optimisation method can be a useful tool for planning air pollution monitoring networks, however, sometimes the clustering of optimal locations may arise in optimisation. The reason for clustering can be attributed to the spatial autocorrelation in the variables used for optimisation or the annealing parameters used for SSA. It is important to investigate the legitimate parameters for SSA to prevent attaining local optima during the optimisation process. The process to identify the legitimate parameters for annealing may be tedious, as it may require various test runs. The annealing process may also be time-consuming, which usually depends on the size of the area under consideration, the number of variables used for computation and computation requirements for calculating the objective function values (also called energy in SSA).

8.2.3 *How can citizen participation curb the air pollution data sparsity constraint for air quality monitoring?*

Citizen participation is one of the significant aspects concerning the vision of smart cities (Risimati and Gumbo, 2018). Citizen participation can promote action to prevent harmful exposures or improve local air quality. Traditionally air quality was measured using the sophisticated pieces of equipment that are expensive, bulky and require high-maintenance. The recent advancements in the miniaturisation and other information and communications technologies (ICT) have brought to market many low-cost sensors. These low-cost sensors extend the opportunity to collect data at higher spatial and temporal resolution, which is hard to achieve using conventional devices. Individual citizens and environmental communities are showing interest in collecting air quality data. Such participation can help in gathering the vast amount of data that comes with large-scale deployment (Schneider et al., 2017b; Clements et al., 2017a). Citizens are ready to participate but to understand the spatial variability of air pollution, the question of *where* they need to collect the air quality data is essential. Most of the times, citizens install the sensors at their place of residence, which sometimes lead to repetitive data or false data into the data collection process, hence limiting the full application of citizen sensors for air pollution monitoring (Budde et al., 2017). In the Chapter 5, the study developed an optimisation method which can assist in the systematic deployment of low-cost air quality sensors for detailed air pollution monitoring in cities. The study presents and tests the approach to utilise the citizen participation for air pollution monitoring using LUR models. The combination of locations in the optimisation method was identified using the objective function which takes into account two constraints: 1.) well-spread citizen sense network and 2.) set of locations which represent the minimal spatial mean prediction error for a given LUR in the study area.

The developed optimisation method was tested for two different cases. Both the cases use the equal weight for the two optimisation constraint discussed above. The monitoring stations used were from the crowdsourcing network existing in the city of Stuttgart. Initially, the method was tested for the case where it was assumed that no monitoring station data exist to develop LUR model, only predictor variables based on previously existing LUR studies (such as ESCAPE (Beelen et al., 2013b)) was available from several open data sources. The study tested the optimisation method using the previously existing model of Austria (Beelen et al., 2013b), considering that the model contains altitude as one explanatory variable for $PM_{2.5}$ concentration in the study area. Stuttgart is located at an altitude ranging from 207 m to 549 m; hence we assumed that the Austrian model could be an alternative LUR for the study. The number of stations and their locations was used as they exist in the real world and has nothing to do with the LUR model. The results

obtained were clustered as well as spread-out considering the two constraints of the optimisation method. The clustering can be because of the LUR model not being the real representation of the pollution in the study area. Another explanation could be the spatial auto-correlation of the variables involved in optimisation. Moreover, the results demonstrated a distinct improvement by decreasing the prediction error in the resulting configuration compared to the original configuration we gave as input by 52.42% for a given ESCAPE Austria LUR model.

The second and final test performed was based on the LUR model developed using the citizen sensed air quality data. We used the procedure suggested by ESCAPE (see Beelen et al., 2013b) for developing a regression model. The open data from various data sources were used as explanatory variables for the study. The regression model we obtained was with a very low R^2 value of 0.1422, but this model could be a better representative for explaining the pollution for the study area than arbitrary or assumption based model selection (as we did in the previous case). The optimisation method was then tested using the developed LUR model, resulting in the identifying configurations for monitoring network that were less clustered and with better spread than the previously tested cases for assumed LUR model. The optimisation method performed considerably well by identifying the well-spread monitoring station network while also decreasing the prediction error in the optimised configuration compared to the original configuration given as input by 52.39%.

The tests of the study present promising results regarding the performance of the optimisation method considering the objective of identifying locations which are well-spread and represents minimal prediction error for the study area. The advantage of the method is its resilience to incorporate available open data for the optimal location identification to initiate or redesign existing air pollution monitoring network. The method also allows laying custom weights on the two constraints, which must be equal to or larger than 0 and sum to 1. Primarily, the application of the developed optimisation method can encourage systematic practice for VGI based opportunities by identifying locations. Limited locations but essential location selection for overall air pollution modelling process may lead to wisely use citizen participation and also limiting the flow of nonessential data into the process. The study also discusses the significance of the number of participants required for efficient air pollution monitoring for the city, making it cost-efficient. Another significant advantage of the approach is the open access availability of the data to various authorities and public for environmental issues as well as for the researchers and businesses to perform analyses and then share the services with the users. The proposed approach helps in enabling systematic VGI based data collection process that may open up channels for further communication among researchers, analysts and might help advancing science.

Along with the advantages, the method also suffers few limitations concerning the reliability of measurements and maintenance aspect of the sensors. Since the sensors were from low-cost sensors, the need for re-calibration and maintenance are usually required, which may limit the regular, uninterrupted flow of data. Also, the frequent involvement of citizen to actively commit on a day to day bases to keep up sensors running may not be practically possible. Sometime, citizens may also act selfishly and use devices for personal benefits, such as for indoor air pollution monitoring, leading to false data in the pollution monitoring process. Another major limitation is the behaviour of the citizens in the data collection process. With time the participants' interest to collect data fades, leading them to act as lurkers, which may impact the whole process of air pollution monitoring. Privacy concerns are also triggering the unwillingness to participate in the data collection process, which leads us back to the lack of data. The optimisation also suffers from the limitation as discusses in the previous study concerning LUR and SSA. The process of optimisation using the objective function we developed is tedious. For our analysis, the process took hours to days for identifying the optimal parameters and deciding best configuration for further analysis.

Altogether, the optimisation method has the potential to initiate systematic VGI based air pollution data collection. The method helps in first deciding the research goal and then helps in planning by identifying the optimal locations for air pollution data collection, making the whole process of participatory data gathering and analysis more efficient as suggested by Clements et al., 2017a. Moreover, the optimisation method generated clustered results, which can be attributed to factors concerning the variables used, annealing parameters and the LUR considered. Moreover, these factors can be improved if more open data is available to develop variables which can explain air pollution better for the study area. The approach also has the vulnerability to cluster around the extreme values producing few clusters in the optimisation method. Future research can consider developing the constraints in objective function which can limit the sampling of points on the well-distributed range of values of variables under study and not only extreme values.

8.2.4 How can housing companies act as stakeholders in participatory processes for air pollution monitoring to address data gaps?

In practice, citizen participation suffers from challenges related to data gaps (due, e.g., to the lack of calibration, maintenance and replacement of sensors), lurker phenomena from individuals and concerns of citizens about privacy. The study in Chapter 6 helps in addressing such challenges which limit air pollution data

collection by using participatory sensing approaches, which is not limited to citizens. The study discusses the importance of involving the private-public stakeholders in the air pollution data collection process. The study proposes the idea of involving housing companies as stakeholders of a participatory sensing initiative to support air pollution monitoring close to where people live. The study proposed that the housing company can play three different roles as suggested in the literature for participatory sensing framework:

1. *as campaign administrators*, i.e., initiate participatory sensing campaigns when they invite their tenants. Also, the housing companies design, implement, manage, and maintain PS infrastructures for houses spaces,
2. *as participants* when they install low-cost sensors which collect data continually about the air quality,
3. *as end-users* when they visualise the data collected, reflect on it, and take evidence-based measures to improve the QoL of the residents.

The participation of housing in air pollution monitoring initiatives can be a useful alternative for air pollution data, which also overcomes some limitations of the citizen participation approach. The limitation with regards to the unwillingness to maintain and re-calibrate sensors from time to time can be resolved by involving housing companies as campaign administrator. The misuse of the sensors by citizens can be mitigated if housing companies can act as a participant. Finally, the lurker phenomena and privacy concern limitations can be addressed, if the housing company act participant and end user by broadcasting the data to the public and based on group privacy implementation. In all, the proposed approach can overcome various limitations of the citizen participation and encouraging “good” air pollution data collection from a well maintained low-cost sensor network.

The study further investigated the perception of housing company for getting involved in a participatory sensing approach through a survey questionnaire to the executive and planning personnel of 42 major housing companies of Germany. The result of the survey reveals the majority (78%) of housing companies who responded (18 responses from 71 housing companies in Germany) were interested in being part of participatory sensing framework to aid in collecting air pollution data using low-cost sensors.

The proposed approach also sustain some limitations. Issues concerning the ownership of the collected data, housing companies commitment to calibration and maintenance and chances of housing companies as a lurker in the data collection

process may impact the whole data collection process. Also, concerning privacy, the trade-off between information sharing and services may lead to individuals missing some interesting personalised service. The full anonymity of the residents, may not be guaranteed as the scientific community is yet to provide a full-proof solution for privacy.

All in all, the involvement of housing companies can help in collecting data from well maintained, well spread and in reasonable proximity to peoples' living spaces. The housing can not only provide data but can also initiate the data collection process which can help in monitoring air pollution for the whole city with great detail. Housing companies can also provide services to the citizens by developing various new service models. As results suggested, few companies were interested in being part of the initiatives; however, the response rate we have in our survey was pretty low which was expected as we administered the survey to the executive staff of the housing companies. Future research may consider surveying the people living in the housing company premises and housing companies to get a broader picture about general public interest in air pollution-related services if housing companies collect data.

8.2.5 *How can we take advantage of WiFi networks to collect detailed traffic data in the city?*

Road traffic is amongst the major sources of air pollution. In Europe, traffic contributes to 40% of NO_x and $PM_{2.5}$ pollution. In order to monitor air pollution in cities, traffic data is one of the crucial inputs for air pollution monitoring. Especially in the case of the LUR method we discussed in the previous studies, traffic data is an essential input (Beelen et al., 2013b). Traffic data is also useful for city governance to survey existing traffic conditions which may lead to air pollution outburst at specific locations in the city. For modelling air pollution at higher resolution, we need to have high-resolution input data too. Usually, traffic data is not available at higher resolution; only a few major roads are monitored in cities considering their significance to the transportation system and local authorities. However, for modelling air pollution, detailed traffic related data can lead to better understanding of traffic-related emission impact on air pollution and its spread. To curb the limited traffic data availability in cities, in Chapter 7 we presented a method to gather traffic flow data on the road. The method utilises the open hardware to measure traffic data using WiFi, which is available in abundance nowadays. The hardware device uses the WiFi received signal strength indication (RSSI) to identify the objects on the road.

The traditional way of measuring traffic intensity on the road requires various kind of sophisticated devices ranging from high-resolution cameras to Infrared (IR) and ultrasonic devices. However, these devices are expensive, need high maintenance which can interrupt traffic, bulky and not compatible in various weather conditions making it hard to gather traffic data at a detailed level in cities. To overcome all these constraints we developed a traffic monitoring system, which is capable of using WiFi RSSI to detect the object and count it for traffic data collection. The method identifies the object based on the interference cause, which is further analysed to identify the vehicle type, such as the car, truck and bikes.

The developed hardware was tested in two different real-world scenarios: a.) under low traffic condition, b.) under heavy traffic conditions. Both scenarios were real-world situations at two different roads in the city of Münster. The hardware was installed on the opposite sides of the road with WiFi transmitter on one side and open hardware we developed as a receiver on the other side of the road. Vehicle passes from the middle of the two devices. Whenever, any object crosses between the devices, lead to an interference in the signal flow. This interference is recorded by the receiver, which is further utilised to detect the object and identify the type of object. The proposed system is capable of identifying some types of vehicles on the road with limited reliability.

Various sophisticated methods exist to analyse RSSI changes, we have used the simple amplitude and time width based approach to analyse the RSSI for object identification and detection. The hardware performs better in the low-traffic scenarios, making it useful for the residential and low traffic zones of the city. These areas are not usually monitored for traffic, but by using the proposed hardware, it is possible to monitor now. The advantage of the developed hardware and the proposed method is its ease, performance and low cost. Another significant advantage is its ability to work in extreme weather conditions, where some of the sophisticated devices suffer to provide results. The advantage also includes its ability to monitor traffic independent of any connection to mobile devices or utilising any sensitive data like video recording, hence preserving the privacy of the commuters. One of the notable drawbacks of the method was the unreliable signal fluctuations, which sometimes can lead to wrong results. There are also some limitations concerning the hardware processing power such as storage capacity, processing power; however, it can be addressed by increasing the memory on the board.

Primarily, the hardware and method are capable of identifying the objects on the road with high precision. The deployment of the developed low-cost sensors at a large scale in cities can help in collecting the traffic intensity values which can contribute to the detailed information of traffic flow in cities. The collection of traffic data can further help in detailed air pollution and noise modelling. Use of the proposed

hardware can also help in enabling smart traffic and commute systems with real-time data. The significant contribution of the developed for transport monitoring is its low-cost for enabling crowdsourcing. The real-time privacy protected crowdsourcing data can be useful, as it can act as a feedback system, which can help in enabling quick services for city transportation and related domains such as air pollution monitoring. Moreover, further work is needed to make the idea more implementable and efficient to distinguish between objects with more accuracy.

8.3 Contribution to Open City Toolkit and its Significance

Detailed air quality information can be essential for improving quality of life (QoL) in cities. This thesis focused on solving various issues on the road towards detailed air pollution monitoring using open data and open source tools for the open smart city. To implement the approaches discussed in the thesis, we utilised open data and open source tools modules like QGIS, GRASS and the R statistical environment. The scripts were developed to evaluate and utilise the available open data for city-level LUR model creation, optimal location identification for sensor placement in cities which can foster further data collection by involving government, private-public stakeholders and citizens, as discussed in Chapters 4, 5, 6 of the thesis. These R scripts together act as tools in the Open City Toolkit (OCT) which is a deliverable under the scope of the EU H2020 project GEO-C. Furthermore, the thesis also tried addressing the lack of detailed traffic data for the city by developing the open hardware which is also a contribution to the OCT for enabling open data collection and exploitation.

In all, the contributions of this thesis for the OCT can be summarised as follows :

1. R scripts for developing LUR model using open data (Chapter 3)
 - a) GitHub repository : OCT-LUR-Scripts (Gupta, 2018d).
 - b) Content: Source code, data, results and detailed guide for utilising the contribution.
2. R scripts for optimal location identification for robust LUR (Chapter 4)
 - a) GitHub repository : AQ-MND-optimisation (Gupta, 2018a)

- b) Content: Published paper, source code, data, results and detailed guide for utilising the contribution.
3. R scripts to download and utilise citizen sensed air pollution open data and optimisation method (Chapter 5)
 - a) GitHub repository : VGI-AQM-Optimisation (Gupta, 2018f)
 - b) Content: Source code for downloading crowdsourced air pollution data, the source code for analysis, data used in analysis, results and a manual for utilising the contribution.
 4. Open hardware tool for road traffic data collection (Chapter 7)
 - a) GitHub repository : WiFi-Hardware-setup (Gupta, 2018g)
 - b) Content: Detailed specification of the open hardware developed for traffic monitoring, the source code to set up, and data analysis module.
 5. Manual video data extraction tool for ground truth dataset generation (Chapter 7)
 - a) GitHub repository : WiFi-traffic-videotool (Gupta, 2018i)
 - b) Content: The source code of the web-application and instructions for using the video tool.
 6. Datasets used and created during the project (Chapter 4,5, 7)
 - a) GitHub repository : ESR08-OCT-Datasets (Gupta, 2018b)
 - b) Content: The dataset used in during the study and their links to access from Zenodo.

Upcoming contribution

1. A comic version of the whole thesis illustrating the significance of each chapter and its importance for enabling open smart cities.
 - a) GitHub repository : ESR08-Thesis-comic (Gupta, 2018c)

- b) Content: The repository will contain the comic representing the contribution of the work done by ESR08 concerning Open smart cities.

2. Shiny app for developing LUR using open data

- a) GitHub repository : OpenLUR-Shinyapp (Gupta, 2018e)
- b) Content: The repository will contain the shiny app for developing LUR model using open data for Open smart cities.

8.4 General Discussion

The recent proliferation of the visions of smart, resilient, healthy, green or sustainable cities around the globe is the part of strategic countermeasures by various institutions to the growing cities as the nexus of human and societal developments. The advantage of these bottom-up approaches is their promise to reduce risk and barriers to enable less affluent cities and communities to undertake the visions mentioned above. Urbanisation, along with urban growth brings about more environmental severe problems, such as deforestation, air, water and pollution. The European Union (EU) is fostering efforts in devising strategies for achieving urban developments in a smart way for its metropolitan areas by investing in ICT research and innovation and for developing policies that can stimulate improvement in the QoL of citizens and be making cities sustainable and liveable (Caragliu et al., 2011). It is expected that more extensive developments and practical use of digital technologies can provide citizens with better QoL and wellbeing (Europe Commission, 2015). For the realisations of the vision of smart cities and their impact on the general public, it is believed that the services provided by either local authorities or public-private stakeholders should be initiated in a collaborative, sustainable and creative way, thus making the most of any opportunity and potential for quality of life improvement. Considering this line of thought, this thesis attempted to address few limitations in detailed air quality monitoring for the open smart city, hence can facilitate fast and precise information about air quality to citizens for improving their quality of life and well being in cities.

Air pollution has been a topic of debate for ages, and still, we are not able to control it well. Air pollution is impacting the epicentre of our civilisation. Cities, extensively and causing global health and economic impact, leading to decreased quality of life. There is also inequality in air pollution distribution and exposure in cities, which may lead to environmental inequalities causing social and economic challenges in cities. Relying on distinct authorities and policymakers for addressing these global

issues is not practically the solution, as we can see from the history. Residents and businesses of the city can also be part of the solution as much as they are part of the problem. Collective efforts are required to address such issues of our generation. While the thesis explored the ways to enable detailed air pollution monitoring in cities utilising open and participatory approaches, Section 2.4 referred to various other modelling techniques used for air quality monitoring. The thesis proposed methods which are flexible and able to help in utilising open available data and tools to model air pollution and to optimally place monitoring devices for data collection. The case studies presented in the thesis seek to reflect on the practical adoption of the proposed approaches for overcoming some of the barriers to detailed air pollution monitoring.

As recent research suggests, air pollution is one of the great killer of our age (Landrigan, 2017) and especially in urban areas, where combustion and traffic-related emissions are impacting the QoL profoundly. We decided to dedicate the research under the general topic of “ Sensing Quality of Life ” for addressing this significant health burden. Although the thesis has not implemented the outcome of the research on the ground, the approaches developed during the research work suggest promising statistical significance making it worth for consideration in the planning and implementation phase. The convergence of smart cities and open data initiatives is fast unfolding designating the outcome of the thesis more applicable. The overall outcome of the thesis contributes by providing various approaches which can promote monitoring air quality at a detailed level in the city using Land Use Regression (LUR) method as the bases for modelling air pollution. However, the concepts in this thesis take into account the limited reliability of the results because of various assumption going under the LUR models and the stochastic method SSA such as the linear relation between the pollutant concentration and explanatory variables, independence and normal distribution of error term, which may not represent the reality.

Furthermore, it is important to be aware of a few considerations before using the results from the proposed approaches. The outcomes are generated using open access data sets, which means the inherent errors in data sources may influence the outcome. This can also be seen the other way as the use of the input data can partly explain the model errors. The approach used in the study provides the flexibility in such a way that LUR modelling can be initiated from whatever possible data sources we have in hand. Nowadays at least the geographical data from the Open Street Map (OSM) is available along with various open remote sensing data hubs (Copernicus, 2018), which can be the useful source for generating variables required for LUR development as per projects like ESCAPE (Eeftens et al., 2012a; Beelen et al., 2013b). Moreover, there has been an increasing discussion in the literature about opening up of city data, which will make the LUR modelling approach more useful as

city-level data can be used as input. The approach is also flexible in the sense that it allows incorporation of whatever data is collected from participatory approaches and a significant amount of freely available remote sensing data, augmenting the work of (Schneider et al., 2017b; Clements et al., 2017a; Kumar et al., 2015b; Wulder and Coops, 2014; Mayfield et al., 2017)

In the case of absence (or limited availability) of air pollution measurement data for LUR modelling in the city, the thesis suggested optimisation method which can be used by environmental authorities for systematically placing new monitoring devices in the city or by environmental activist and businesses to install low-cost sensors as part of the participatory approach. The optimisation methods allow considering variables which can be significant for predicting air quality in the city using the previously developed LUR model as the starting step for initiating the air pollution data collection. However, it is important to note that arbitrary selection of model from the existing LUR models in literature may not be providing the appropriate information to infer the locations for placing sensors. Hence, it is essential to understand the significance of the selected model and the variables to the optimisation. As the variables used in optimisation are selected from LUR, the flexibility of the LUR continues in the optimisation. Another important aspect of the optimisation method is the variation in the final configuration output because SSA is a stochastic method, even with the utilisation of same data and parameters the final configuration obtained vary in each iteration. Also, the consideration for boundary effects of the SSA is important while accessing the relevance of the optimisation outcomes (Van Groenigen et al., 1999).

The types of monitoring sites used in developing LUR models has been somewhat arbitrary (Wu et al., 2017) or ad-hoc (Beelen et al., 2013a). The systematic way of exploring the geographical distribution of the air pollution monitoring stations is essential, the present thesis help in systematically characterising the locations which are in the selected LUR models. Also, few critical regulatory and policy considerations need to be taken into account for identifying locations. Existing local, regional or national policies for monitoring are essential and need to be weighted while creating an objective function for optimisation (Muller and Ruud, 2018). In the thesis, the approaches discussed takes into account only LUR variable consideration, and future work may incorporate regulatory consideration to making the system more suitable for the application. No particular configuration of monitoring stations can simultaneously provide precise information about all the pollutants because of the LUR models' shortcoming to make a distinction between the influence of the different air pollutants, which also influences the outcomes of the optimisation method. Specific LUR and optimisation methods will address specific pollutant monitoring need. However, having widespread multiple sensors across the city

can be a useful approach of deployment in generalising the approach for various pollutant concentration measurement. In the [LUR CROWDSOURING SPREAD]

The traffic data is hard to acquire, and authorities are not so open to sharing the useful data easily. Even though the data will be shared at some point, the amount of data is limited to certain major roads of the city. If the air pollution information needs to be estimated in detail, it is essential to have information about traffic in detail, not only temporally but also spatially. The hardware device and method devised in the thesis is low-cost, which can make its utilisation for data gathering at various locations in the city. Even citizens can collect the data and share it in an open domain for further application. With the EU's plans to install more WiFi networks in public spaces (European Commission, 2018b), it is advantageous to have approaches which can extend the usability to such initiatives along with already existing public WiFi. The method presented in the thesis can be a useful tool to re-utilise the abundance of WiFi networks for addressing various significant concerns of the city, such as traffic mobility, air pollution, sound pollution. As the approach does not have dependencies on the commuters to connect to the network or register the device, the applicability becomes much easier. Furthermore, the method for calculating traffic only analyses the change in the WiFi signal strength, making it privacy proof and so the data can easily be shared fostering real-time usability.

8.5 Outlook

In this thesis, various approaches to help in the detailed monitoring of air pollution in the open smart city have been presented. Although many approaches exist, the presented approaches leave room for many more advancements and applications in the future. Most importantly, the practical implementation of the citizen participation sensors for air pollution data based on optimisation method outcomes must be conducted to understand the practicality of the method. This will be useful especially in the case studies to understand the applicability of low-cost sensors for air pollution monitoring under systematically planned data collection initiatives. It will be important that future investigations take into account the effect of geographical information, preliminary observations, and information on the spatial correlation to help improve optimisation strategies (Beelen et al., 2009; Brus and Heuvelink, 2007; Wadoux et al., 2017). These tasks were left out during the work and could be focused in future.

Further areas for investigation worth mentioning include:

1. Exploring the ways of extending the approach for other adverse environmental factors monitoring using open data,
2. Incorporating regulatory constraints in optimisation objectives,
3. Focusing on spatiotemporal optimisation methods for real-time monitoring,
4. Shedding light on the applicability of the method for supporting various health, environment and public participation related urban sustainable development goals for cities,
5. Working on systems which can enable data gathering from all stakeholders,
6. Usability of approach for detail climate impact assessment in cities,
7. Extending approaches which can help in geomedicine and geohealth informatics.

Extending the approach for other adverse environmental factors monitoring

The Internet of Things (IoT) is rapidly gaining attention as a key enabler of the smarter cities and the future. The WHO and its partners are using IoT for helping to improve the health and well-being of people in cities. Participatory sensing is also utilising the power of IoT solutions for data collection. Encouraging the IoT based participatory sensing applications systematically by utilising the approach discussed in Chapter 5 can further help in high-resolution monitoring of various environmental factors impacting human health in cities, such as urban heat, humidity, pollutants. With the help of advancing open remote sensing tools like OpenEO, 2017, the detailed monitoring of environmental factors can be further extended with timely outcomes.

Incorporating regulatory requirements in optimisation objectives As discussed in Muller and Ruud, 2018, there are various regulatory constraints which govern the design of the monitoring station networks. It is important to take into account the measures put forward by various regulatory bodies (e.g., European Union, 2008a; Raffuse et al., 2007) while finding optimal locations for practical implementations of the services involving regulatory bodies for equal participation. For example, extending the method from Chapter 4 with adding to objective function rules concerning regulatory requirements can help in identifying locations which may honour requirement for decreasing prediction error and regulatory principles for robust LUR estimations.

Spatio-temporal optimisation methods for real-time monitoring The present work focuses on the approaches which can help in a spatial context for air pollution monitoring. Future work may involve developing the optimisation method which can be tailored for multiple pollutant measurements. It is also essential to know the temporal dimension of the environmental variables, because of their temporal dynamics. Developing the optimisation methods which can help in identifying locations in real time for spatiotemporal monitoring can be another extension. The spatiotemporal monitoring optimisation approach can be considered as the location identification approach which can help in real time identifications of the set of sensors that should register data for a specific pollutant at a specific time. However, such an approach can be used when the well-spread air pollution monitoring network with a significant amount of sensors exist. Collecting data with space and time consideration can be useful for not only capturing specific spatiotemporal dynamics but also restricting the flow of unnecessary data in the process, which can improve the speed of data analysis and information generation.

Working on the unified systems which can enable data gathering from all stakeholders The need to have a unified system for all stakeholders is essential to support the vision of open and transparent governance. Extending the platform like Open City Toolkit (OCT) (Degbelo et al., 2016) for real implementation in cities can help in enabling data sharing, participatory data gathering and collective decision making. Hence, can contribute to anticipating equal participation and responsibility sharing for combating various problems in cities. For instance, extending the tools proposed in this thesis for OCT can help in bringing together citizens, housing companies and governmental organisations. This unification can encourage dialogues between different parties to discuss there roles and contributions for enabling detailed air quality monitoring for the city.

Shedding light on extensions to support urban sustainable development goals

Since the underlining aim of the smart cities is to enable sustainable urban growth, it would be interesting to see, how the approaches presented in the thesis can be extended for supporting certain goals of United Nations Sustainability Goals (UNDP, 2016), especially for:

1. Goal 3: Good health and well being, by further developing the methods proposed in the thesis, especially from chapter 3, 4, 5 to help in monitoring and communicating citizens about health risks and healthy mobility for cities.
2. Goal 10: Reducing environmental and social inequality, by fostering participation and sharing the detailed information about the environmental factors to the people in cities. The detailed information can be generated using the

outcomes of the thesis and further extending it for environmental factors. Information sharing and inducing participation are the ways promoted to reduce social inequalities.

3. Goal 13: Climate action, by broadening the approaches proposed in this thesis for detailed information creation concerning environmental factors which can help in explaining the effects of climate change at the micro level in the city.
4. Goal 17: Partnerships for the goals, by utilising the benefit of methods in the thesis for developing the unified platforms, open data application and stimulating participation of various stakeholders of the city for monitoring and discussions.

Figure 8.1 presents a mind map highlighting the general applications of the outcome of each chapter for the UN sustainability goals to develop sustainable cities.

Working towards approaches which can contribute to geomedicine and geo-health informatics

Detailed information about the environmental factors and well participatory framework can be useful in accessing individuals exposure to adverse conditions in space and time (Jenerette et al., 2016). Various environmental factors can contribute to specific health risk scenarios such as communicable disease outbreak, thermal discomfort, cardiovascular distress (Salata et al., 2017). Detailed information can help with environmental factors can enhance the capabilities of the health care system by taking mitigation measures, providing personalised treatments and helping people to improve their quality of life in cities.

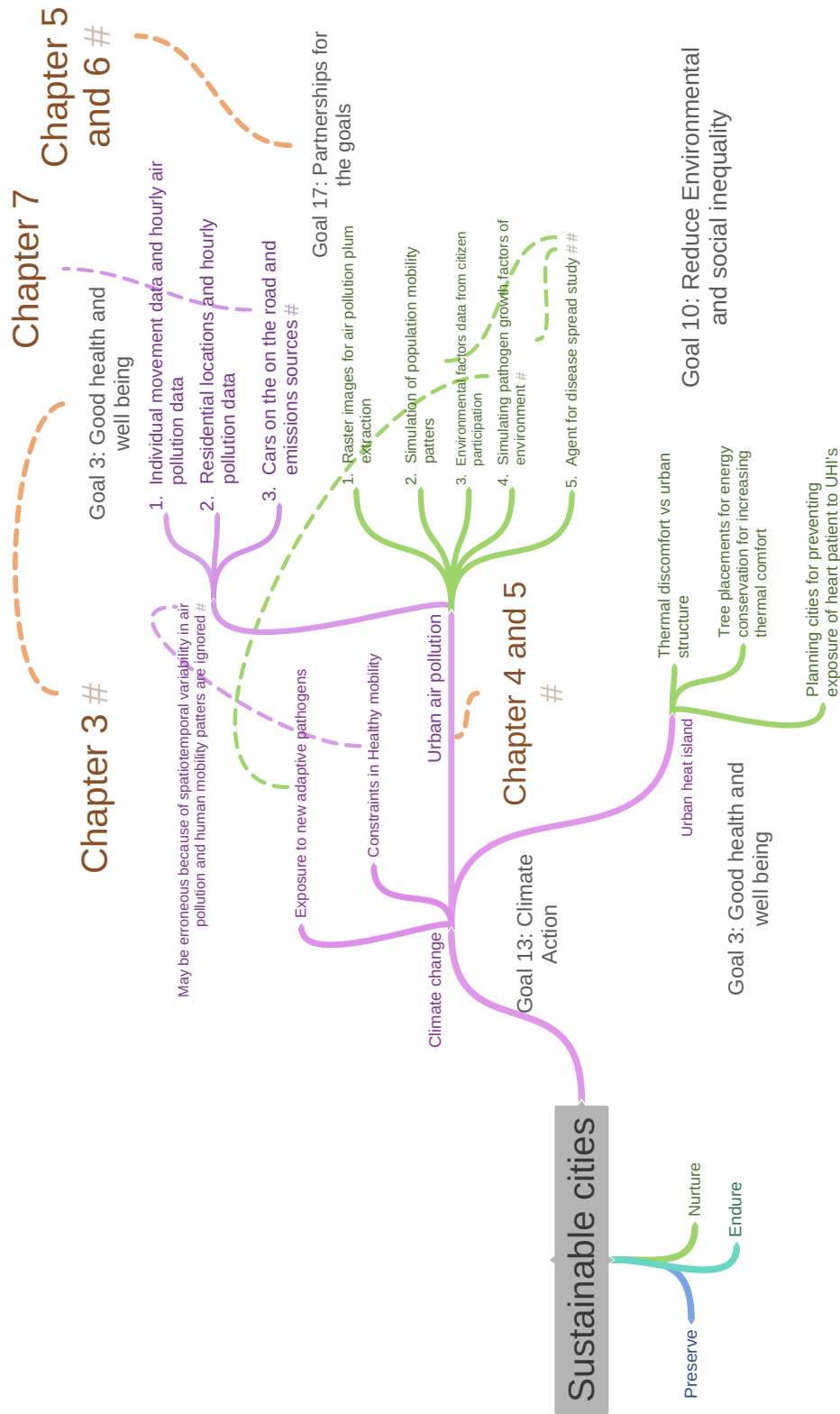


Figure. 8.1. Mind map shredding light on the possible extensions of the outcomes of this thesis for developing sustainable cities

Conclusions

Open and participatory approaches have great potential for addressing various concerns in the public domain because of their abilities to involve multiple stakeholders to overcome specific challenges which are hard to be taken care of by individuals. Cities are a multifaceted entity, and one single entity cannot quickly solve all the emerging concerns. Involving general public and private-public stakeholders will not only help to gather data but also can contribute to bringing forward new ideas and solutions. In the thesis, we explored different approaches that can help in utilising open data, citizen participation and businesses for air pollution monitoring in cities. Our results demonstrate that with the application of flexible and robust statistical methods like Land Use Regression (LUR) and Spatial Simulated Annealing (SSA), open data can be beneficial for detailed air pollution monitoring and also counter big data challenges. While the flexibility of LUR models allows the application of open data, limitations exist regarding air pollution measurement data in the city, making the whole approach for detailed air pollution monitoring coarser.

In order to address the limitations caused by lack of monitoring station measurements for monitoring air quality, the thesis developed an approach for identifying optimal locations, thereby enabling data collection. The optimisation method developed achieves the objective of identifying the set of locations in the city which can decrease the prediction error for the air pollution estimations using a given LUR. The approach was also extended to identify the set of locations which consider the population weights, to model air quality with less prediction error close to populated areas of the city. By using LUR and SSA, we were able to provide a systematic workflow that can guide additional data collection in cities. The approach has a potential to foster the systematic, collaborative collection of air pollution data involving citizens, businesses or governmental organisations for detailed air pollution monitoring approach more producible and transparent.

In practice, the monitoring stations used by governmental agencies are bulky and expensive, making their deployment limited to only a few locations in the city (usually one or two). In order to address the limited data availability constraints, the thesis also demonstrates the applicability of the optimisation method to enable systematic data collection using citizen sensed air pollution data. The citizen sense approach utilises low-cost sensors, enabling them to be deployed in bulk. Even though the

reliability of low-cost sensor measurements is still the matter of discussion, the relative changes in measurements can help in identifying the surrounding geographical variables which can explain air pollution locally, making it acceptable for LUR modelling. However, some time citizen participation may lead to false data along with several data gaps. For overcoming the data gaps in participatory frameworks, the thesis also proposed the approach for involving housing companies in the air pollution data collection process using low-cost sensors. The inclusion of housing companies in the process helps in overcoming some of the factors which can feed in false data in the air pollution monitoring, therefore encouraging private-public partnership in overcoming barriers to detailed air pollution monitoring for open smart cities.

For monitoring air pollution, traffic data is of high relevance, especially for LUR models. The thesis has also attempted to address the restrictions caused by lack of finer scale traffic data by devising a method which uses WiFi signals. The method developed utilises a low-cost sensor device which exploits WiFi signals to gather traffic data. WiFi signals are existing in abundance in cities, utilising such technologies for data collection make this approach more sustainable and implementable. The advantage of the method involves its low-cost, ability to work in adverse weather conditions and privacy-preserving abilities, which is not the case for traditional methods used for traffic monitoring.

Bibliography

- Adam-Poupart, Ariane, Allan Brand, Michel Fournier, Michael Jerrett, and Audrey Smargiassi (2014). „Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy–LUR approaches“. In: *Environmental health perspectives* 122.9, p. 970 (cit. on p. 22).
- Adams, Matthew D and Pavlos S Kanaroglou (2016). „A criticality index for air pollution monitors“. In: *Atmospheric pollution research* 7.3, pp. 482–487 (cit. on p. 21).
- Agency, European Environmental (2017). *Air pollution sources*. <https://www.eea.europa.eu/themes/air/air-pollution-sources> (cit. on p. 114).
- Agrawal, Rakesh and Ramakrishnan Srikant (2000). „Privacy-preserving data mining“. In: *ACM Sigmod Record*. Vol. 29. 2. ACM, pp. 439–450 (cit. on p. 100).
- Ahvenniemi, Hannele, Aapo Huovila, Isabel Pinto-Seppä, and Miimu Airaksinen (2017). „What are the differences between sustainable and smart cities?“ In: *Cities* 60, pp. 234–245 (cit. on p. 14).
- Akita, Yasuyuki, Jose M Baldasano, Rob Beelen, et al. (2014). „Large scale air pollution estimation method combining land use regression and chemical transport modeling in a geostatistical framework“. In: *Environmental science & technology* 48.8, pp. 4452–4459 (cit. on p. 23).
- Allen, Ryan W, Ofer Amram, Amanda J Wheeler, and Michael Brauer (2011). „The transferability of NO and NO₂ land use regression models between cities and pollutants“. In: *Atmospheric Environment* 45.2, pp. 369–378 (cit. on pp. 20, 60, 139).
- Amoako, J, M Lodh, and T Risbey (2005). *Health Impacts of Transport Emissions in Australia: Economic Costs* (cit. on p. 20).
- Androulaki, Elli, Seung Geol Choi, Steven M Bellovin, and Tal Malkin (2008). „Reputation systems for anonymous networks“. In: *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, pp. 202–218 (cit. on p. 100).
- Angelidou, Margarita (2014). „Smart city policies: A spatial approach“. In: *Cities* 41, S3–S11 (cit. on p. 13).
- Apte, Joshua S, Kyle P Messier, Shahzad Gani, et al. (2017). „High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data“. In: *Environmental Science & Technology* (cit. on p. 98).
- AQ-SPEC (2017). *South Coast Air Quality Management District Air Quality Sensor Performance Evaluation Center (AQ-SPEC)*. (Cit. on p. 99).

- Ariely, Dan, Emir Kamenica, and Dražen Prelec (2008). „Man’s search for meaning: The case of Legos“. In: *Journal of Economic Behavior & Organization* 67.3-4, pp. 671–677 (cit. on p. 1).
- Baden-Württemberg, Statistisches Landesamt (2018). *Bevölkerung und Erwerbstätigkeit*. Statistisches Landesamt Baden-Württemberg (cit. on p. 71).
- Baklanov, Alexander (2012). „Megacities: urban environment, air pollution, climate change and human health interactions“. In: *National Security and Human Health Implications of Climate Change*. Springer, pp. 103–114 (cit. on p. 2).
- Balid, Walid, Hasan Tafish, and Hazem H Refai (2018). „Intelligent Vehicle Counting and Classification Sensor for Real-Time Traffic Surveillance“. In: *IEEE Transactions on Intelligent Transportation Systems* 19.6, pp. 1784–1794 (cit. on pp. 5, 117, 118, 132).
- Ballas, Dimitris (2013). „What makes a ‘happy city’?“ In: *Cities* 32, S39–S50 (cit. on p. 15).
- Barca, Emanuele, Giuseppe Passarella, Michele Vurro, and Alberto Morea (2015). „MSANOS: data-driven, multi-approach software for optimal redesign of environmental monitoring networks“. In: *Water resources management* 29.2, pp. 619–644 (cit. on p. 26).
- Barcaccia, Barbara (2013). „Definitions and domains of health-related quality of life“. In: *Outcomes Assessment in End-Stage Kidney Disease: Measurements and Applications in Clinical Practice*, Bentham Science Publishers, pp. 12–24 (cit. on pp. 1, 28).
- Barer, Morris (2017). *Why are some people healthy and others not?* Routledge (cit. on p. 64).
- Basagaña, Xavier, Marcela Rivera, Inmaculada Aguilera, et al. (2012). „Effect of the number of measurement sites on land use regression models in estimating local air pollution“. In: *Atmospheric environment* 54, pp. 634–642 (cit. on pp. 40, 68).
- Batty, Michael, Kay W Axhausen, Fosca Giannotti, et al. (2012). „Smart cities of the future“. In: *The European Physical Journal Special Topics* 214.1, pp. 481–518 (cit. on pp. 98, 102).
- Bauer, Katharina, Thijs Bosker, Kim N Dirks, and Paul Behrens (2018). „The impact of seating location on black carbon exposure in public transit buses: Implications for vulnerable groups“. In: *Transportation Research Part D: Transport and Environment* 62, pp. 577–583 (cit. on p. 64).
- Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017). „When labor disputes bring cities to a standstill: The impact of public transit strikes on traffic, accidents, air pollution, and health“. In: *American Economic Journal: Economic Policy* 9.1, pp. 1–37 (cit. on p. 64).
- Bäumer, Doris (2004). „Come together–Involving housing companies in mobility management actions“. In: *Paper and presentation at ECOMM* (cit. on p. 101).
- Beckerman, Bernardo S, Michael Jerrett, Murray Finkelstein, et al. (2012). „The association between chronic exposure to traffic-related air pollution and ischemic heart disease“. In: *Journal of Toxicology and Environmental Health, Part A* 75.7, pp. 402–411 (cit. on p. 22).
- Beckerman, Bernardo S, Michael Jerrett, Marc Serre, et al. (2013a). „A hybrid approach to estimating national scale spatiotemporal variability of PM_{2.5} in the contiguous United States“. In: *Environmental science & technology* 47.13, pp. 7233–7241 (cit. on pp. 22, 23).
- Beckerman, Bernardo S, Michael Jerrett, Randall V Martin, et al. (2013b). „Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California“. In: *Atmospheric environment* 77, pp. 172–177 (cit. on p. 22).

- Beelen, Rob, Gerard Hoek, Edzer Pebesma, et al. (2009). „Mapping of background air pollution at a fine spatial scale across the European Union“. In: *Science of the Total Environment* 407.6, pp. 1852–1867 (cit. on pp. 23, 60, 90, 139, 151).
- Beelen, Rob, Gerard Hoek, Danielle Vienneau, et al. (2013a). „Development of NO₂ and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project“. In: *Atmospheric Environment* 72, pp. 10–23 (cit. on pp. 22, 40, 49, 139, 150).
- Beelen, Rob, Gerard Hoek, Danielle Vienneau, et al. (2013b). „Development of NO₂ and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project“. In: *Atmospheric Environment* 72, pp. 10–23 (cit. on pp. 36, 41, 59, 140, 141, 144, 149).
- Behzadian, Kourosh, Zoran Kapelan, Dragan Savic, and Abdollah Ardeshir (2009). „Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks“. In: *Environmental Modelling & Software* 24.4, pp. 530–541 (cit. on p. 25).
- Benis, Khaled Zoroufchi, Esmaeil Fatehifar, Sirous Shafiei, Fatemeh Keivani Nahr, and Yaser Purfarhadi (2016). „Design of a sensitive air quality monitoring network using an integrated optimization approach“. In: *Stochastic environmental research and risk assessment* 30.3, pp. 779–793 (cit. on pp. 38, 40, 57, 70).
- Birgitta, Berglund, Lindvall Thomas, and HS Dietrich (1999). „Guidelines for community noise“. In: *The WHO Expert Task Force Meeting on Guidelines for Community Noise*, pp. 26–30 (cit. on p. 114).
- Boer, E P J, A L M Dekkers, and A Stein (2002). „Optimization of a monitoring network for sulfur dioxide“. In: *Journal of environmental quality* 31.1, pp. 121–128 (cit. on p. 44).
- Bonney, Rick, Jennifer L Shirk, Tina B Phillips, et al. (2014). „Next steps for citizen science“. In: *Science* 343.6178, pp. 1436–1437 (cit. on p. 65).
- Borrego, C, AM Costa, J Ginja, et al. (2016). „Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise“. In: *Atmospheric Environment* 147, pp. 246–263 (cit. on pp. 8, 70).
- Borrego, Carlos, Miguel Coutinho, A Margardia Costa, et al. (2015). „Challenges for a new air quality directive: the role of monitoring and modelling techniques“. In: *Urban Climate* 14, pp. 328–341 (cit. on p. 70).
- Bougoudis, Ilias, Konstantinos Demertzis, and Lazaros Iliadis (2016). „Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning“. In: *Integrated Computer-Aided Engineering* 23.2, pp. 115–127 (cit. on p. 36).
- Bougoudis, Ilias, Konstantinos Demertzis, Lazaros Iliadis, Vardis-Dimitris Anezakis, and Antonios Papaleonidas (2018). „FuSSFFra, a fuzzy semi-supervised forecasting framework: the case of the air pollution in Athens“. In: *Neural Computing and Applications* 29.7, pp. 375–388 (cit. on p. 36).
- Boulos, Maged N Kamel, Bernd Resch, David N Crowley, et al. (2011). „Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples“. In: *International journal of health geographics* 10.1, p. 67 (cit. on p. 98).

- Bowatte, Gayan, Caroline J Lodge, Luke D Knibbs, et al. (2017). „Traffic-related air pollution exposure is associated with allergic sensitization, asthma, and poor lung function in middle age“. In: *Journal of Allergy and Clinical Immunology* 139.1, pp. 122–129 (cit. on p. 16).
- Bowser, Anne and Andrea Wiggins (2015). „Privacy in participatory research: advancing policy to support human computation“. In: *Human Computation* 2.1, pp. 19–44 (cit. on p. 100).
- Brauer, Michael, Gerard Hoek, Patricia van Vliet, et al. (2003). „Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems“. In: *Epidemiology*, pp. 228–239 (cit. on p. 21).
- Brauer, Michael, Greg Freedman, Joseph Frostad, et al. (2015). „Ambient air pollution exposure estimation for the global burden of disease 2013“. In: *Environmental science & technology* 50.1, pp. 79–88 (cit. on p. 96).
- Briggs, David (2005). „The role of GIS: coping with space (and time) in air pollution exposure assessment“. In: *Journal of Toxicology and Environmental Health, Part A* 68.13-14, pp. 1243–1261 (cit. on p. 36).
- Britter, RE and SR Hanna (2003). „Flow and dispersion in urban areas“. In: *Annual Review of Fluid Mechanics* 35.1, pp. 469–496 (cit. on pp. 4, 18, 43).
- Brokamp, Cole, Roman Jandarov, MB Rao, Grace LeMasters, and Patrick Ryan (2017). „Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches“. In: *Atmospheric Environment* 151, pp. 1–11 (cit. on p. 40).
- Brown, J, C Bowman, et al. (2013). „Integrated Science Assessment for Ozone and Related Photochemical Oxidants“. In: *Washington, DC: US Environmental Protection Agency* (cit. on p. 64).
- Brus, Dick J and Gerard BM Heuvelink (2007). „Optimization of sample patterns for universal kriging of environmental variables“. In: *Geoderma* 138.1, pp. 86–95 (cit. on pp. 44, 60, 151).
- Buch, Norbert, Sergio A Velastin, and James Orwell (2011). „A review of computer vision techniques for the analysis of urban traffic“. In: *IEEE Transactions on Intelligent Transportation Systems* 12.3, pp. 920–939 (cit. on p. 117).
- Budde, Matthias, Andrea Schankin, Julien Hoffmann, et al. (2017). „Participatory Sensing or Participatory Nonsense?: Mitigating the Effect of Human Error on Data Quality in Citizen Science“. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3, p. 39 (cit. on pp. 68, 89, 92, 99, 109, 140).
- Bundesamt für Kartographie und Geodäsie (2018). *Open Data - Freie Daten und Dienste des BKG*. Bundesamt für Kartographie und Geodäsie (cit. on p. 73).
- Cai, Yutong, Anna L Hansell, Marta Blangiardo, et al. (2017). „Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts“. In: *European heart journal* 38.29, pp. 2290–2296 (cit. on p. 114).
- Caragliu, Andrea, Chiara Del Bo, and Peter Nijkamp (2011). „Smart cities in Europe“. In: *Journal of urban technology* 18.2, pp. 65–82 (cit. on p. 148).

- Carvalho, Helotonio (2016). „The air we breathe: differentials in global air quality monitoring“. In: *The Lancet Respiratory Medicine* 4.8, pp. 603–605 (cit. on p. 18).
- Castell, Nuria, Franck R Dauge, Philipp Schneider, et al. (2017). „Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?“ In: *Environment international* 99, pp. 293–302 (cit. on pp. 64, 67).
- Charpin, Denis and Denis M Caillaud (2017). „Air pollution and the nose chronic respiratory disorders“. In: *The Nose and Sinuses in Respiratory Disorders: ERS Monograph* 76, p. 162 (cit. on p. 36).
- Chen, Dengke and Shiyi Chen (2017). „Particulate air pollution and real estate valuation: Evidence from 286 Chinese prefecture-level cities over 2004–2013“. In: *Energy Policy* (cit. on p. 101).
- Chen, Li, Zhipeng Bai, Shaofei Kong, et al. (2010). „A land use regression for predicting NO₂ and PM₁₀ concentrations in different seasons in Tianjin region, China“. In: *Journal of Environmental Sciences* 22.9, pp. 1364–1373 (cit. on p. 22).
- Cheung, Sing Yiu and Pravin Pratap Varaiya (2006). „Traffic surveillance by wireless sensor networks“. PhD thesis. University of California, Berkeley (cit. on p. 117).
- Chong, Chee-Yee and Srikanta P Kumar (2003). „Sensor networks: evolution, opportunities, and challenges“. In: *Proceedings of the IEEE* 91.8, pp. 1247–1256 (cit. on p. 70).
- Christin, Delphine, Andreas Reinhardt, Salil S Kanhere, and Matthias Hollick (2011). „A survey on privacy in mobile participatory sensing applications“. In: *Journal of Systems and Software* 84.11, pp. 1928–1946 (cit. on pp. 96, 100, 108).
- City of Stuttgart (2018). *Measuring points*. Office for Environmental Protection, Section of Urban Climatology (cit. on p. 71).
- Clements, Andrea L, William G Griswold, Jill E Johnston, et al. (2017a). „Low-Cost Air Quality Monitoring Tools: From Research to Practice (A Workshop Summary)“. In: *Sensors* 17.11, p. 2478 (cit. on pp. 5, 140, 142, 150).
- (2017b). „Low-Cost Air Quality Monitoring Tools: From Research to Practice (A Workshop Summary)“. In: *Sensors* 17.11, p. 2478 (cit. on pp. 61, 64, 65, 67–69, 74, 89, 91, 92, 97, 101).
- Cohen, Aaron J, Michael Brauer, Richard Burnett, et al. (2017). „Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015“. In: *The Lancet* 389.10082, pp. 1907–1918 (cit. on pp. 16, 96).
- Colette, Augustin, Camilla Andersson, Astrid Manders, et al. (2017). „EURODELTA-Trends, a multi-model experiment of air quality hindcast in Europe over 1990–2010“. In: *Geoscientific Model Development* 10.9, p. 3255 (cit. on p. 39).
- Commodore, Adwoa, Sacoby Wilson, Omar Muhammad, Erik Svendsen, and John Pearce (2017). „Community-based participatory research for the study of air pollution: a review of motivations, approaches, and outcomes“. In: *Environmental Monitoring and Assessment* 189.8 (cit. on p. 96).

- Conti, Gea Oliveri, Behzad Heibati, Itai Kloog, Maria Fiore, and Margherita Ferrante (2017). „A review of AirQ Models and their applications for forecasting the air pollution health outcomes“. In: *Environmental Science and Pollution Research* 24.7, pp. 6426–6445 (cit. on pp. 36, 64, 69).
- Copernicus (2018). *CORINE Land Cover*. European capacity for Earth Observation (cit. on p. 73).
- Cote, Melissa and Alexandra Branzan Albu (2017). „Teaching Computer Vision and Its Societal Effects: A Look at Privacy and Security Issues From the Students' Perspective“. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, pp. 1378–1386 (cit. on p. 119).
- Cuff, Dana, Mark Hansen, and Jerry Kang (2008). „Urban sensing: out of the woods“. In: *Communications of the ACM* 51.3, pp. 24–33 (cit. on pp. 98, 100).
- Dadvand, Payam, Bart Ostro, Francesc Figueras, et al. (2014). „Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees“. In: *Epidemiology* 25.4, pp. 518–525 (cit. on p. 20).
- Darçın, Murat (2014). „Association between air quality and quality of life“. In: *Environmental Science and Pollution Research* 21.3, pp. 1954–1959 (cit. on p. 28).
- De Cristofaro, Emiliano and Claudio Soriente (2013). „Participatory privacy: Enabling privacy in participatory sensing“. In: *IEEE network* 27.1, pp. 32–36 (cit. on p. 100).
- De Gruijter, Jaap, Dick J Brus, Marc FP Bierkens, and Martin Knotters (2006). *Sampling for natural resource monitoring*. Springer Science & Business Media (cit. on p. 43).
- De Nazelle, Audrey, Edmund Seto, David Donaire-Gonzalez, et al. (2013). „Improving estimates of air pollution exposure through ubiquitous sensing technologies“. In: *Environmental Pollution* 176, pp. 92–99 (cit. on p. 2).
- Degbelo, A (2012). „An ontology design pattern for spatial data quality characterization in the semantic sensor web“. In: *The 5th international workshop on Semantic Sensor Networks*. Ed. by C Henson, K Taylor, and O Corcho. Boston, Massachusetts, USA: CEUR-WS.org, pp. 103–108 (cit. on p. 99).
- Degbelo, Auriol, Carlos Granell, Sergio Trilles, et al. (2016). „Opening up smart cities: citizen-centric challenges and opportunities from GIScience“. In: *ISPRS International Journal of Geo-Information* 5.2, p. 16 (cit. on pp. 14, 28, 102, 153).
- Delmelle, Eric M (2014). „Spatial sampling“. In: *Handbook of Regional Science*. Springer, pp. 1385–1399 (cit. on p. 32).
- Depatla, Saandeep, Arjun Muralidharan, and Yasamin Mostofi (2015). „Occupancy estimation using only WiFi power measurements“. In: *IEEE Journal on Selected Areas in Communications* 33.7, pp. 1381–1393 (cit. on p. 118).
- Deutsche Welle (2016a). *Germany's Stuttgart asks residents to leave car at home amid high air pollution*. Deutsche Welle (cit. on p. 71).
- (2016b). *Stuttgart: Germany's 'Beijing' for air pollution?* Deutsche Welle (cit. on p. 71).
- Devillers, R, A Stein, Y Bédard, et al. (2010). „Thirty years of research on spatial data quality: achievements, failures, and opportunities“. In: *Transactions in GIS* 14.4, pp. 387–400 (cit. on p. 99).

- D'Hondt, Ellie, Matthias Stevens, and An Jacobs (2013). „Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring“. In: *Pervasive and Mobile Computing* 9.5, pp. 681–694 (cit. on p. 97).
- Di Sabatino, Silvana, Riccardo Buccolieri, and Prashant Kumar (2018). „Spatial Distribution of Air Pollutants in Cities“. In: *Clinical Handbook of Air Pollution-Related Diseases*. Springer, pp. 75–95 (cit. on p. 18).
- Diener, Ed and Eunkook Suh (1997). „Measuring quality of life: Economic, social, and subjective indicators“. In: *Social indicators research* 40.1-2, pp. 189–216 (cit. on p. 15).
- Dirgawati, Mila, Rosanne Barnes, Amanda J Wheeler, et al. (2015). „Development of land use regression models for predicting exposure to NO₂ and NO_x in metropolitan Perth, Western Australia“. In: *Environmental Modelling & Software* 74, pp. 258–267 (cit. on p. 22).
- Dons, Evi, Martine Van Poppel, Bruno Kochan, Geert Wets, and Luc Int Panis (2013). „Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon“. In: *Atmospheric environment* 74, pp. 237–246 (cit. on p. 22).
- Duckham, M and L Kulik (2006). „Location privacy and location-aware computing“. In: *Dynamic and Mobile GIS: Investigating Changes in Space and Time*. Ed. by R Billen, E Joao, and D Forrest. CRC Press. Chap. 3, pp. 35–51 (cit. on p. 100).
- Dye, TS, CP MacDonald, and CB Anderson (1999). *Guideline for developing an ozone forecasting program*. Tech. rep. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC (United States) (cit. on p. 4).
- Eeftens, Marloes, Rob Beelen, Kees de Hoogh, et al. (2012a). „Development of land use regression models for PM_{2.5}, PM_{2.5} absorbance, PM₁₀ and PM_{coarse} in 20 European study areas; results of the ESCAPE project“. In: *Environmental science & technology* 46.20, pp. 11195–11205 (cit. on pp. 5, 149).
- Eeftens, Marloes, Ming-Yi Tsai, Christophe Ampe, et al. (2012b). „Spatial variation of PM_{2.5}, PM₁₀, PM_{2.5} absorbance and PM_{coarse} concentrations between and within 20 European study areas and the relationship with NO₂—Results of the ESCAPE project“. In: *Atmospheric Environment* 62, pp. 303–317 (cit. on pp. 79, 85).
- Elen, Bart, Jan Peters, Martine Van Poppel, et al. (2012). „The aeroflex: a bicycle for mobile air quality measurements“. In: *Sensors* 13.1, pp. 221–240 (cit. on p. 98).
- Elkamel, A, E Fatehifar, M Taheri, MS Al-Rashidi, and A Lohi (2008). „A heuristic optimization approach for Air Quality Monitoring Network design with the simultaneous consideration of multiple pollutants“. In: *Journal of environmental management* 88.3, pp. 507–516 (cit. on p. 38).
- Elwood, Sarah, Michael F Goodchild, and Daniel Z Sui (2012). „Researching volunteered geographic information: Spatial data, geographic research, and new social practice“. In: *Annals of the association of American geographers* 102.3, pp. 571–590 (cit. on pp. 65, 98).
- Eugster, Patrick Th, Pascal A Felber, Rachid Guerraoui, and Anne-Marie Kermarrec (2003). „The many faces of publish/subscribe“. In: *ACM computing surveys (CSUR)* 35.2, pp. 114–131 (cit. on p. 100).

- Eurobarometer, Special (2011). *365: Attitudes of European citizens towards the environment* (cit. on p. 101).
- Europe Commission (2015). *Digital Agenda for Europe: a Europe 2020 initiative* (cit. on p. 148).
- European Commission (2018a). *Free wireless internet hotspots in public spaces*. https://ec.europa.eu/commission/news/free-wireless-internet-hotspots-public-spaces-2018-mar-20_en (cit. on p. 130).
- European Parliament Councils (2008). *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe* (cit. on p. 4).
- European Union (2008a). „Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe“. In: *Official Journal of the European Union* (cit. on pp. 5, 152).
- (2008b). „Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe“. In: *Official Journal of the European Union* (cit. on p. 69).
- European Union, Council of (2008). „DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe“. In: *Official Journal of the European Union*, pp. L 152/1 –L 152/43 (cit. on pp. 37, 39, 58).
- Eusuf, Muhammad Abu, Mohammad A Mohit, MMR Sami Eusuf, and Mansor Ibrahim (2014). „Impact of outdoor environment to the quality of life“. In: *Procedia-Social and Behavioral Sciences* 153, pp. 639–654 (cit. on pp. 1, 15, 28).
- Fang, Jianxin, Huadong Meng, Hao Zhang, and Xiqin Wang (2007). „A low-cost vehicle detection and classification system based on unmodulated continuous-wave radar“. In: *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*. IEEE, pp. 715–720 (cit. on p. 118).
- Fang, Tianfang Bernie and Yongmei Lu (2012). „Personal real-time air pollution exposure assessment methods promoted by information technological advances“. In: *Annals of GIS* 18.4, pp. 279–288 (cit. on pp. 96, 97, 100).
- Fang, Xinwei and Iain Bate (2017). „Issues of using wireless sensor network to monitor urban air quality“. In: *Proceedings of the First ACM International Workshop on the Engineering of Reliable, Robust, and Secure Embedded Wireless Sensing Systems*. ACM, pp. 32–39 (cit. on p. 64).
- Feneri, AM, D Vagiona, and N Karanikolas (2013). „Measuring quality of life (QoL) in urban environment: an integrated approach“. In: *Cest2013, Athens, Greece* (cit. on pp. 1, 15).
- Feng, Xiao, Qi Li, Yajie Zhu, et al. (2015). „Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation“. In: *Atmospheric Environment* 107, pp. 118–128 (cit. on p. 36).
- FLAMENCO Project (2018). *Citizen Observatory – Home of the Flamenco project*. <http://citizen-observatory.be> (cit. on p. 92).

- Foley, Aoife, Beatrice M Smyth, Tomislav Pukšec, Natasa Markovska, and Neven Duić (2017). *A review of developments in technologies and research that have had a direct measurable impact on sustainability considering the Paris agreement on climate change* (cit. on p. 1).
- Forehead, H and N Huynh (2018). „Review of modelling air pollution from traffic at street-level-The state of the science“. In: *Environmental Pollution* 241, pp. 775–786 (cit. on p. 114).
- French, Joshua (2015). *SpatialTools: Tools for Spatial Data Analysis*. R package version 1.0.2 (cit. on p. 78).
- Fucic, A, V Guszak, and A Mantovani (2017). „Transplacental exposure to environmental carcinogens: Association with childhood cancer risks and the role of modulating factors“. In: *Reproductive Toxicology* 72, pp. 182–190 (cit. on p. 18).
- Fussl, Agnes, Werner G Müller, and Juan Rodríguez-Díaz (2012). „Exploratory Designs for Assessing Spatial Dependence“. In: *Spatio-Temporal Design: Advances in Efficient Data Acquisition*, pp. 170–206 (cit. on p. 25).
- Gabrys, Jennifer, Helen Pritchard, et al. (2018). „Just Good Enough Data and Environmental Sensing: Moving Beyond Regulatory Benchmarks toward Citizen Action“. In: *International Journal of Spatial Data Infrastructures Research* 13 (cit. on pp. 67, 98, 99, 104).
- Gauderman, W James, Robert Urman, Edward Avol, et al. (2015). „Association of improved air quality with lung development in children“. In: *New England Journal of Medicine* 372.10, pp. 905–913 (cit. on p. 18).
- Gent, Janneane F, Elizabeth W Triche, Theodore R Holford, et al. (2003). „Association of low-level ozone and fine particles with respiratory symptoms in children with asthma“. In: *Jama* 290.14, pp. 1859–1867 (cit. on p. 18).
- Geofabrik GmbH Karlsruhe (2018). *Downloads*. Geofabrik GmbH Karlsruhe (cit. on p. 73).
- Gharibvand, Lida, David Shavlik, Mark Ghamsary, et al. (2017). „The association between ambient fine particulate air pollution and lung cancer incidence: results from the AHSMOG-2 study“. In: *Environmental health perspectives* 125.3, p. 378 (cit. on p. 16).
- Gibson, C C, E Ostrom, and T K Ahn (2000). „The concept of scale and the human dimensions of global change: a survey“. In: *Ecological Economics* 32.2, pp. 217–239 (cit. on p. 99).
- Gilbert, Nicolas L, Mark S Goldberg, Bernardo Beckerman, Jeffrey R Brook, and Michael Jerrett (2005). „Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model“. In: *Journal of the Air & Waste Management Association* 55.8, pp. 1059–1063 (cit. on p. 22).
- Goldizen, Fiona C, Peter D Sly, and Luke D Knibbs (2016). „Respiratory effects of air pollution on children“. In: *Pediatric Pulmonology* 51.1, pp. 94–108 (cit. on p. 18).
- Goldstein, Inge F and Leon Landovitz (1977). „Analysis of air pollution patterns in New York City. Can one station represent the large metropolitan area?“ In: *Atmospheric Environment* (1967) 11.1, pp. 47–52 (cit. on pp. 37, 70).
- Goodchild, Michael F (2007). „Citizens as sensors: the world of volunteered geography“. In: *GeoJournal* 69.4, pp. 211–221 (cit. on p. 98).
- (2016). „GIS in the Era of Big Data“. In: *Cybergeo: European Journal of Geography* (cit. on p. 28).

- Goodchild, Michael F. and Linna Li (2012). „Assuring the quality of volunteered geographic information“. In: *Spatial Statistics* 1, pp. 110–120 (cit. on p. 66).
- Graaf, Shenja Van der and Carina Veeckman (2014). „Designing for participatory governance: assessing capabilities and toolkits in public service delivery“. In: *info* 16.6, pp. 74–88 (cit. on p. 14).
- Grigg, Jonathan (2018). „Air Pollution and Respiratory Infection–An Emerging and Troubling Association“. In: *American journal of respiratory and critical care medicine* ja (cit. on p. 16).
- Grote, Matt, Ian Williams, John Preston, and Simon Kemp (2018). „A practical model for predicting road traffic carbon dioxide emissions using Inductive Loop Detector data“. In: *Transportation Research Part D: Transport and Environment* 63, pp. 809–825 (cit. on p. 114).
- Gulliver, John, Kees de Hoogh, Daniela Fecht, Danielle Vienneau, and David Briggs (2011). „Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution“. In: *Atmospheric environment* 45.39, pp. 7072–7080 (cit. on pp. 22, 31).
- Gulliver, John, David Morley, Chrissi Dunster, et al. (2018). „Land use regression models for the oxidative potential of fine particles (PM 2.5) in five European areas“. In: *Environmental research* 160, pp. 247–255 (cit. on p. 114).
- Gupta, Shivam (2018a). *AQ-MND Optimisation*. <https://github.com/geohealthshivam/AQ-MND-optimisation> (cit. on pp. 47, 146).
- (2018b). *ESR08-OCT-Datasets*. <https://github.com/geohealthshivam/ESR08-OCT-Datasets> (cit. on p. 147).
- (2018c). *ESR08-Thesis-comic*. <https://github.com/geohealthshivam/ESR08-Thesis-comic> (cit. on p. 147).
- (2018d). *OCT-LUR*. <https://github.com/geohealthshivam/OCT-LUR> (cit. on p. 146).
- (2018e). *OpenLUR-Shinyapp*. <https://github.com/geohealthshivam/OpenLUR-Shinyapp> (cit. on p. 148).
- (2018f). *VGI-AQM-Optimisation*. <https://github.com/geohealthshivam/VGI-AQM-Optimisation> (cit. on pp. 78, 147).
- (2018g). *WiFi-Hardware-setup*. <https://github.com/geohealthshivam/WiFi-Hardware-setup> (cit. on p. 147).
- (2018h). *WiFi-traffic-videotool*. <https://github.com/geohealthshivam/WiFi-traffic-videotool> (cit. on p. 123).
- (2018i). *WiFi-traffic-videotool*. <https://github.com/geohealthshivam/WiFi-traffic-videotool> (cit. on p. 147).
- Gupta, Shivam, Edzer Pebesma, Jorge Mateu, and Auriol Degbelo (2018a). „Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models“. In: *Sustainability* 10.5, pp. 1–27 (cit. on pp. 75, 89, 92, 108).
- Gupta, Shivam, Jorge Mateu, Auriol Degbelo, and Edzer Pebesma (2018b). „Quality of life, big data and the power of statistics“. In: *Statistics & Probability Letters*. In press (cit. on p. 58).

- Gurjar, BR, A Jain, A Sharma, et al. (2010). „Human health risks in megacities due to air pollution“. In: *Atmospheric Environment* 44.36, pp. 4606–4613 (cit. on p. 3).
- Habermann, Mateus, Monica Billger, and Marie Haeger-Eugensson (2015). „Land use regression as method to model air pollution. Previous results for Gothenburg/Sweden“. In: *Procedia Engineering* 115, pp. 21–28 (cit. on p. 24).
- Hadley, Michael B, Rajesh Vedanthan, and Valentin Fuster (2018). „Air pollution and cardiovascular disease: a window of opportunity“. In: *Nature Reviews Cardiology* 15.4, p. 193 (cit. on p. 16).
- Haferkamp, Marcus, Manar Al-Askary, Dennis Dorn, et al. (2017). „Radio-based traffic flow detection and vehicle classification for future smart cities“. In: *Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th*. IEEE, pp. 1–5 (cit. on p. 118).
- Haklay, M (2010). „How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets“. In: *Environment and Planning B: Planning and Design* 37.4, pp. 682–703 (cit. on p. 66).
- Haklay, Mordechai Muki, Suvodeep Mazumdar, and Jessica Wardlaw (2018). „Citizen science for observing and understanding the Earth“. In: *Earth Observation Open Science and Innovation*. Springer, pp. 69–88 (cit. on p. 97).
- Hamra, Ghassan B, Francine Laden, Aaron J Cohen, et al. (2015). „Lung cancer and exposure to nitrogen dioxide and traffic: a systematic review and meta-analysis“. In: *Environmental health perspectives* 123.11, p. 1107 (cit. on p. 64).
- Hancke, Gerhard P, Bruno de Carvalho e Silva, and Gerhard P Hancke Jr. (2013). „The role of advanced sensing in smart cities“. In: *Sensors* 13.1, p. 393 (cit. on p. 115).
- Hao, Yufang and Shaodong Xie (2018). „Optimal redistribution of an urban air quality monitoring network using atmospheric dispersion model and genetic algorithm“. In: *Atmospheric Environment* (cit. on pp. 5, 22).
- Haoui, Amine, Robert Kavalier, and Pravin Varaiya (2008). „Wireless magnetic sensors for traffic surveillance“. In: *Transportation Research Part C: Emerging Technologies* 16.3, pp. 294–306 (cit. on p. 118).
- Health Effects Institute Panel, on the Health Effects of Traffic-Related Air Pollution (2010). *Traffic-related air pollution: a critical review of the literature on emissions, exposure, and health effects*. 17. Health Effects Institute (cit. on p. 36).
- Heilig, Gerhard K (2012). „World urbanization prospects: the 2011 revision“. In: *United Nations, Department of Economic and Social Affairs (DESA), Population Division, Population Estimates and Projections Section, New York*, p. 14 (cit. on p. 2).
- Helle, KB and E Pebesma (2015). „Optimising sampling designs for the maximum coverage problem of plume detection“. In: *Spatial Statistics* 13, pp. 21–44 (cit. on pp. 25, 139).
- Hemmi, Akiko and Ian Graham (2014). „Hacker science versus closed science: Building environmental monitoring infrastructure“. In: *Information, Communication & Society* 17.7, pp. 830–842 (cit. on p. 97).
- Heuvelink, Gerard B M, Z Jiang, Sytze De Bruin, and Chris J W Twenhöfel (2010). „Optimization of mobile radioactivity monitoring networks“. In: *International Journal of Geographical Information Science* 24.3, pp. 365–382 (cit. on pp. 46, 74).

- Heuvelink, Gerard BM, Daniel A Griffith, Tomislav Hengl, and Stephanie J Melles (2012). „Sampling design optimization for space-time kriging“. In: *John Wiley, Oxford* 10, pp. 207–230 (cit. on p. 26).
- Hoek, Gerard, Rob Beelen, Kees De Hoogh, et al. (2008). „A review of land-use regression models to assess spatial variation of outdoor air pollution“. In: *Atmospheric environment* 42.33, pp. 7561–7578 (cit. on pp. 21, 23, 31, 38, 39, 59, 68, 69, 138, 139).
- Hoffmann, Barbara, Susanne Moebus, Andreas Stang, et al. (2006). „Residence close to high traffic and prevalence of coronary heart disease“. In: *European Heart Journal* 27.22, pp. 2696–2702 (cit. on p. 20).
- Hoh, Baik, Toch Iwuchukwu, Quinn Jacobson, et al. (2012). „Enhancing Privacy and Accuracy in Probe Vehicle-Based Traffic Monitoring via Virtual Trip Lines.“ In: *IEEE Trans. Mob. Comput.* 11.5, pp. 849–864 (cit. on p. 119).
- Holan, Scott H and Christopher K Wikle (2012). „Semiparametric Dynamic Design of Monitoring Networks for Non-Gaussian Spatio-Temporal Data“. In: *Spatio-Temporal Design: Advances in Efficient Data Acquisition*, pp. 269–284 (cit. on p. 25).
- Hong, Feng, Xiang Wang, Yanni Yang, et al. (2016). „WFID: passive device-free human identification using WiFi signal“. In: *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, pp. 47–56 (cit. on p. 118).
- Honicky, Richard, Eric A Brewer, Eric Paulos, and Richard White (2008). „N-smarts: networked suite of mobile atmospheric real-time sensors“. In: *Proceedings of the second ACM SIGCOMM workshop on Networked systems for developing regions*. ACM, pp. 25–30 (cit. on p. 98).
- Horvat, Goran, Damir Šoštarić, and Drago Žagar (2012). „Using radio irregularity for vehicle detection in adaptive roadway lighting“. In: *MIPRO, 2012 Proceedings of the 35th International Convention*. IEEE, pp. 748–753 (cit. on p. 118).
- Hu, Mao-Gui and Jin-Feng Wang (2011). „A spatial sampling optimization package using MSN theory“. In: *Environmental Modelling & Software* 26.4, pp. 546–548 (cit. on p. 25).
- Huang, Kuan Lun, Salil S Kanhere, and Wen Hu (2010). „Preserving privacy in participatory sensing systems“. In: *Computer Communications* 33.11, pp. 1266–1280 (cit. on p. 100).
- Hulchanski, J David (2002). *Housing policy for tomorrow's cities*. Canadian Policy Research Networks Ottawa (cit. on p. 101).
- Hystad, Perry, Eleanor Setton, Alejandro Cervantes, et al. (2011). „Creating national air pollution models for population exposure assessment in Canada“. In: *Environmental health perspectives* 119.8, p. 1123 (cit. on pp. 20, 69).
- Jackson, Steven, William Mullen, Peggy Agouris, et al. (2013). „Assessing completeness and spatial error of features in volunteered geographic information“. In: *ISPRS International Journal of Geo-Information* 2.2, pp. 507–530 (cit. on p. 66).
- Jaimes, Luis G, Idalides Vergara-Laurens, and Miguel A Labrador (2012). „A location-based incentive mechanism for participatory sensing systems with budget constraints“. In: *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*. IEEE, pp. 103–108 (cit. on p. 99).

- Janssen, Stijn, Gerwin Dumont, Frans Fierens, et al. (2012). „Land use to characterize spatial representativeness of air quality monitoring stations and its relevance for model validation“. In: *Atmospheric Environment* 59, pp. 492–500 (cit. on p. 25).
- Jenerette, G Darrel, Sharon L Harlan, Alexander Buyantuev, et al. (2016). „Micro-scale urban surface temperatures are related to land-cover features and residential heat related health impacts in Phoenix, AZ USA“. In: *Landscape Ecology* 31.4, pp. 745–760 (cit. on p. 154).
- Jerrett, Michael, Altaf Arain, Pavlos Kanaroglou, et al. (2005a). „A review and evaluation of intraurban air pollution exposure models“. In: *Journal of Exposure Science and Environmental Epidemiology* 15.2, p. 185 (cit. on pp. 20, 23, 30, 69).
- (2005b). „A review and evaluation of intraurban air pollution exposure models.“ In: *Journal of exposure analysis and environmental epidemiology* 15.2, pp. 185–204 (cit. on p. 96).
- Jerrett, Michael, M A Arain, P Kanaroglou, et al. (2007). „Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada“. In: *Journal of Toxicology and Environmental Health, Part A* 70.3-4, pp. 200–212 (cit. on p. 64).
- Jiang, Qijun, Arnold K. Bregt, and Lammert Kooistra (2018). „Formal and informal environmental sensing data and integration potential: Perceptions of citizens and experts“. In: *Science of the Total Environment* 619-620. December 2017, pp. 1133–1142 (cit. on p. 100).
- Jiao, Wan, Gayle Hagler, Ronald Williams, et al. (2016). „Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States“. In: *Atmospheric Measurement Techniques* 9.11, p. 5281 (cit. on pp. 64, 96).
- Johnson, Markey, V Isakov, JS Touma, S Mukerjee, and H Özkaynak (2010). „Evaluation of land-use regression models used to predict air quality concentrations in an urban area“. In: *Atmospheric Environment* 44.30, pp. 3660–3668 (cit. on pp. 23, 59).
- Johnson, Timothy P. (2012). „Response rates and nonresponse errors in surveys“. In: *JAMA* 307.17, p. 1805 (cit. on p. 102).
- Jovašević-Stojanović, Milena, Alena Bartonova, Dušan Topalović, et al. (2015). „On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter“. In: *Environmental pollution* 206, pp. 696–704 (cit. on p. 67).
- Kanaroglou, Pavlos S, Michael Jerrett, Jason Morrison, et al. (2005a). „Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach“. In: *Atmospheric Environment* 39.13, pp. 2399–2409 (cit. on pp. 5, 37, 138).
- Kanaroglou, Pavlos S., Michael Jerrett, Jason Morrison, et al. (2005b). „Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach“. In: *Atmospheric Environment* 39.13, pp. 2399–2409 (cit. on pp. 65, 68).
- Kao, Jehng-Jung and Ming-Ru Hsieh (2006). „Utilizing multiobjective analysis to determine an air quality monitoring network in an industrial district“. In: *Atmospheric Environment* 40.6, pp. 1092–1103 (cit. on pp. 38, 57).

- Kelly, Frank J and Julia C Fussell (2015). „Air pollution and public health: emerging hazards and improved understanding of risk“. In: *Environmental geochemistry and health* 37.4, pp. 631–649 (cit. on p. 96).
- Kerckhoffs, Jules, Gerard Hoek, Jelle Vlaanderen, et al. (2017). „Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring“. In: *Environmental research* 159, pp. 500–508 (cit. on p. 22).
- Kessler, Carsten and Grant McKenzie (2018). „A geoprivacy manifesto“. In: *Transactions in GIS* 22.1, pp. 3–19 (cit. on p. 100).
- Khreis, Haneen, Charlotte Kelly, James Tate, et al. (2017). „Exposure to traffic-related air pollution and risk of development of childhood asthma: A systematic review and meta-analysis“. In: *Environment international* 100, pp. 1–31 (cit. on pp. 36, 39, 64, 69).
- Kim, Jinsol, Alexis A Shusterman, Kaitlyn J Lieschke, Catherine Newman, and Ronald C Cohen (2017). „The berkeley atmospheric CO₂ observation network: Field calibration and evaluation of low-cost air quality sensors“. In: *Atmos. Meas. Tech. Discuss* (cit. on p. 70).
- Kinateder, Michael and Siani Pearson (2003). „A privacy-enhanced peer-to-peer reputation system“. In: *International Conference on Electronic Commerce and Web Technologies*. Springer, pp. 206–215 (cit. on p. 100).
- Koh, Jin Ming, Marcus Sak, Hwee-Xian Tan, et al. (2015). „Efficient data retrieval for large-scale smart city applications through applied Bayesian inference“. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*. IEEE, pp. 1–6 (cit. on pp. 28, 33).
- Korek, Michal, Christer Johansson, Nina Svensson, et al. (2017). „Can dispersion modeling of air pollution be improved by land-use regression? An example from Stockholm, Sweden“. In: *Journal of Exposure Science and Environmental Epidemiology* 27.6, p. 575 (cit. on p. 22).
- Kotovirta, Ville, Timo Toivanen, Renne Tergujeff, and Markku Huttunen (2012). „Participatory Sensing in Environmental Monitoring—Experiences“. In: *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on*. IEEE, pp. 155–162 (cit. on p. 100).
- Kuhlbusch, Thomas AJ, Ulrich Quass, Gary Fuller, et al. (2013). „Air pollution monitoring strategies and technologies for urban areas“. In: *Urban air quality in Europe*. Springer, pp. 277–296 (cit. on p. 37).
- Kumar, Pramod, Amir-Ali Feiz, Sarvesh Kumar Singh, Pierre Ngae, and Grégory Turbelin (2015a). „Reconstruction of an atmospheric tracer source in an urban-like environment“. In: *Journal of Geophysical Research: Atmospheres* 120.24, pp. 12589–12604 (cit. on p. 25).
- Kumar, Prashant, Lidia Morawska, Claudio Martani, et al. (2015b). „The rise of low-cost sensing for managing air pollution in cities“. In: *Environment international* 75, pp. 199–205 (cit. on pp. 3, 4, 24, 75, 150).
- Kumral, Mustafa and Umit Ozer (2013). „Planning additional drilling campaign using two-space genetic algorithm: A game theoretical approach“. In: *Computers & geosciences* 52, pp. 117–125 (cit. on p. 25).
- Lagzi, Istvan, Robert Meszaros, Gyorgyi Gelybo, and Adam Leelossy (2014). „Atmospheric chemistry“. In: (cit. on p. 21).

- Landrigan, Philip J (2017). „Air pollution and health“. In: *The Lancet Public Health* 2.1, e4–e5 (cit. on p. 149).
- Lark, R M (2002). „Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood“. In: *Geoderma* 105.1-2, pp. 49–80 (cit. on p. 44).
- Le Boennec, Rémy and Frédéric Salladarré (2017). „The impact of air pollution and noise on the real estate market. The case of the 2013 European Green Capital: Nantes, France“. In: *Ecological Economics* 138, pp. 82–89 (cit. on p. 101).
- Lewandowski, Marcin, Bartłomiej Płaczek, Marcin Bernas, and Piotr Szymała (2018). „Road Traffic Monitoring System Based on Mobile Devices and Bluetooth Low Energy Beacons“. In: *Wireless Communications and Mobile Computing* 2018 (cit. on p. 119).
- Lewis, Alastair C, James D Lee, Peter M Edwards, et al. (2016). „Evaluating the performance of low cost chemical sensors for air pollution research“. In: *Faraday discussions* 189, pp. 85–103 (cit. on p. 99).
- Li, Qinghua and Guohong Cao (2013). „Providing privacy-aware incentives for mobile sensing“. In: *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*. IEEE, pp. 76–84 (cit. on p. 100).
- Lin, Xiaodong, Xiaoting Sun, Pin-Han Ho, and Xuemin Shen (2007). „GSIS: A secure and privacy-preserving protocol for vehicular communications“. In: *IEEE Transactions on vehicular technology* 56.6, pp. 3442–3456 (cit. on p. 119).
- Lipfert, Frederick W (2017). „A critical review of the ESCAPE project for estimating long-term health effects of air pollution“. In: *Environment international* 99, pp. 87–96 (cit. on p. 24).
- Lisjak, Josip, Sven Schade, and Alexander Kotsev (2017). „Closing data gaps with citizen science? Findings from the Danube region“. In: *ISPRS International Journal of Geo-Information* 6.9, p. 277 (cit. on p. 67).
- Little, Kathleen E, Masaki Hayashi, and Steve Liang (2016). „Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach“. In: *Groundwater* 54.3, pp. 317–324 (cit. on p. 97).
- Liu, Rui, Junbin Liang, Wenyu Gao, and Ruiyun Yu (2018). „Privacy-based recommendation mechanism in mobile participatory sensing systems using crowdsourced users' preferences“. In: *Future Generation Computer Systems* 80, pp. 76–88 (cit. on p. 100).
- Lombi, Linda (2018). „The Contribution of Digital Sociology to the Investigation of Air Pollution“. In: *Clinical Handbook of Air Pollution-Related Diseases*. Springer, pp. 621–636 (cit. on p. 99).
- Lord, Dominique and Simon Washington (2018). „Introduction“. In: *Safe Mobility: Challenges, Methodology and Solutions*. Emerald Publishing Limited, pp. 1–10 (cit. on p. 114).
- Lu, Rongxing, Xiaodong Lin, Haojin Zhu, P-H Ho, and Xuemin Shen (2008). „ECPP: Efficient conditional privacy preservation protocol for secure vehicular communications“. In: *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE. IEEE, pp. 1229–1237 (cit. on p. 119).
- Lu, Rongxing, Xiaodong Lin, Zhiguo Shi, and Xuemin Sherman Shen (2013). „A lightweight conditional privacy-preservation protocol for vehicular traffic-monitoring systems“. In: *IEEE intelligent systems* 28.3, pp. 62–65 (cit. on p. 119).

- Ma, Denglong, Wei Tan, Zaoxiao Zhang, and Jun Hu (2017). „Parameter identification for continuous point emission source based on Tikhonov regularization method coupled with particle swarm optimization algorithm“. In: *Journal of hazardous materials* 325, pp. 239–250 (cit. on p. 25).
- Manins, PC, Chair of Committee, et al. (2001). „Air Quality Forecasting for Australia’s Major Cities–Final Report“. In: *Project Management Committee: CSIRO Atmospheric Research, Aspendale, Australia: <http://www.dar.csiro.au/info/aaqfs>* (cit. on p. 4).
- Marans, Robert W and Robert J Stimson (2011). *Investigating quality of urban life: Theory, methods, and empirical research*. Vol. 45. Springer Science & Business Media (cit. on p. 15).
- Marchant, BP, AB McBratney, RM Lark, and B Minasny (2013). „Optimized multi-phase sampling for soil remediation surveys“. In: *Spatial Statistics* 4, pp. 1–13 (cit. on p. 26).
- Marshall, Julian D, Elizabeth Nethery, and Michael Brauer (2008). „Within-urban variability in ambient air pollution: comparison of estimation methods“. In: *Atmospheric Environment* 42.6, pp. 1359–1369 (cit. on pp. 31, 38).
- Martin, Peter T, Yuqi Feng, Xiaodong Wang, et al. (2003). *Detector technology evaluation*. Tech. rep. Citeseer (cit. on p. 117).
- Masson, Nicholas, Ricardo Piedrahita, and Michael Hannigan (2015). „Quantification method for electrolytic sensors in long-term monitoring of ambient air quality“. In: *Sensors* 15.10, pp. 27283–27302 (cit. on p. 99).
- Mateu, Jorge and Werner G Müller (2012). *Spatio-temporal design: advances in efficient data acquisition*. John Wiley & Sons (cit. on p. 25).
- Mayer, Helmut (1999). „Air pollution in cities“. In: *Atmospheric environment* 33.24-25, pp. 4029–4037 (cit. on pp. 36, 64).
- Mayfield, Helen, Carl Smith, Marcus Gallagher, and Marc Hockings (2017). „Use of freely available datasets and machine learning methods in predicting deforestation“. In: *Environmental Modelling & Software* 87, pp. 17–28 (cit. on p. 150).
- McConnell, Rob, Talat Islam, Ketan Shankardass, et al. (2010). „Childhood incident asthma and traffic-related air pollution at home and school“. In: *Environmental health perspectives* 118.7, p. 1021 (cit. on p. 36).
- Mead, Mohammed Iqbal, OAM Popoola, GB Stewart, et al. (2013). „The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks“. In: *Atmospheric Environment* 70, pp. 186–203 (cit. on p. 19).
- Mensink, C, A Colles, L Janssen, and J Cornelis (2003). „Integrated air quality modelling for the assessment of air quality in streets against the council directives“. In: *Atmospheric Environment* 37.37, pp. 5177–5184 (cit. on p. 21).
- Mertens, Mariano, Astrid Kerkweg, Volker Grewe, and Patrick Jöckel (2016). „Impact of road traffic emissions on tropospheric ozone in Europe for present day and future scenarios“. In: *EGU General Assembly Conference Abstracts*. Vol. 18, p. 10100 (cit. on p. 114).
- Michanowicz, Drew R, Jessie LC Shmool, Leah Cambal, et al. (2016). „A hybrid land use regression/line-source dispersion model for predicting intra-urban NO₂“. In: *Transportation Research Part D: Transport and Environment* 43, pp. 181–191 (cit. on p. 23).

- Mimbela, Luz Elena Y and Lawrence A Klein (2000). „Summary of vehicle detection and surveillance technologies used in intelligent transportation systems“. In: (cit. on pp. 117, 118, 131).
- Mitchell, Gordon and Danny Dorling (2003). „An environmental justice analysis of British air quality“. In: *Environment and planning A* 35.5, pp. 909–929 (cit. on p. 75).
- Mofarrah, Abdullah and Tahir Husain (2010). „A holistic approach for optimal design of air quality monitoring network expansion in an urban area“. In: *Atmospheric Environment* 44.3, pp. 432–440 (cit. on pp. 37, 57).
- Molina, Luisa T, Mario J Molina, Robert S Slott, et al. (2004). „Air quality in selected megacities“. In: *Journal of the Air & Waste Management Association* 54.12, pp. 1–73 (cit. on p. 64).
- Mooney, Peter, Pdraig Corcoran, and Blazej Ciepluch (2013). „The potential for using volunteered geographic information in pervasive health computing applications“. In: *Journal of Ambient Intelligence and Humanized Computing* 4.6, pp. 731–745 (cit. on p. 96).
- Moss, Alex (2011). *The German listed real estate sector – disproportional representation*. http://consiliacapital.com/sitmanager/uploads/ck_files/files/German%20listed%20real%20estate%20sector%20-%20disproportinoal%20representation.pdf. [Online; accessed 31-January-2018] (cit. on p. 102).
- Mousa, Hayam, Sonia Ben Mokhtar, Omar Hasan, et al. (2015). „Trust management and reputation systems in mobile participatory sensing applications: A survey“. In: *Computer Networks* 90, pp. 49–73 (cit. on p. 99).
- Mulalu, Mulalu I (2018). „Participatory Geographic Information Systems Within a Crowdsourcing Environment, With Special Reference to Volunteered Geographic Information“. In: *Handbook of Research on Geospatial Science and Technologies*. IGI Global, pp. 392–419 (cit. on p. 97).
- Muller, Nicholas Z and Paul A Ruud (2018). „What Forces Dictate the Design of Pollution Monitoring Networks?“ In: *Environmental Modeling & Assessment* 23.1, pp. 1–14 (cit. on pp. 24, 150, 152).
- Namiot, Dmitry and Manfred Sneps-Snepe (2012). „Context-aware data discovery“. In: *Intelligence in Next Generation Networks (ICIN), 2012 16th International Conference on*. IEEE, pp. 134–141 (cit. on p. 33).
- Nejadkoorki, Farhad, Ken Nicholson, and Kamal Hadad (2011). „The design of long-term air quality monitoring networks in urban areas using a spatiotemporal approach“. In: *Environmental monitoring and assessment* 172.1-4, pp. 215–223 (cit. on p. 37).
- Nellore, Kapileswar and Gerhard Hancke (2016). „A survey on urban traffic management system using wireless sensor networks“. In: *Sensors* 16.2, p. 157 (cit. on p. 116).
- Nieuwenhuijsen, Mark J, David Donaire-Gonzalez, Ioar Rivas, et al. (2015). „Variability in and agreement between modeled and personal continuously measured black carbon levels using novel smartphone and sensor technologies“. In: *Environmental science & technology* 49.5, pp. 2977–2982 (cit. on p. 4).
- Niu, Xiaoguang, Jiawei Wang, Qiongzan Ye, and Yihao Zhang (2018). „A Privacy-Preserving Incentive Mechanism for Participatory Sensing Systems“. In: *Security and Communication Networks* 2018 (cit. on p. 100).

- Nunen, Erik van, Roel Vermeulen, Ming-Yi Tsai, et al. (2017). „Land use regression models for ultrafine particles in six European areas“. In: *Environmental science & technology* 51.6, pp. 3336–3345 (cit. on p. 64).
- Oftedal, Bente, Bert Brunekreef, Wenche Nystad, et al. (2008). „Residential outdoor air pollution and lung function in schoolchildren“. In: *Epidemiology*, pp. 129–137 (cit. on p. 21).
- Oh, Seri, Stephen Ritchie, and Cheol Oh (2002). „Real-time traffic measurement from single loop inductive signatures“. In: *Transportation Research Record: Journal of the Transportation Research Board* 1804, pp. 98–106 (cit. on p. 116).
- Oiamo, Tor H, Markey Johnson, Kathy Tang, and Isaac N Luginaah (2015). „Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment“. In: *Science of the Total Environment* 529, pp. 149–157 (cit. on p. 23).
- Ojo, Adegboyega, Edward Curry, and Fatemeh Ahmadi Zeleti (2015). „A tale of open data innovations in five smart cities“. In: *System Sciences (HICSS), 2015 48th Hawaii International Conference on. IEEE*, pp. 2326–2335 (cit. on p. 14).
- Ojo, Adegboyega, Zamira Dzhusupova, and Edward Curry (2016). „Exploring the nature of the smart cities research landscape“. In: *Smarter as the New Urban Agenda*. Springer, pp. 23–47 (cit. on pp. 1, 13).
- OK Labs (2018a). *Data Archive*. OK labs (cit. on pp. 72, 92).
- (2018b). *MEASURE AIR QUALITY YOURSELF NEARLY FINISHED WITH YOUR HELP*. OK Labs (cit. on pp. 72, 78).
- (2018c). *Measurement accuracy*. OK labs (cit. on p. 73).
- Open.NRW (2018). *NRW: Zensusatlas 2011 – Bundesweite*. European Data Portal (cit. on p. 73).
- OpenStreetMap contributors (2017). *Planet dump retrieved from <https://planet.osm.org>. \url{<https://www.openstreetmap.org>}* (cit. on pp. 41, 73).
- Orosz, Gábor, R Eddie Wilson, and Gábor Stépan (2010). *Traffic jams: dynamics and control* (cit. on p. 118).
- Oser, Jennifer (2017). „Assessing how participators combine acts in their “political tool kits”: A person-centered measurement approach for analyzing citizen participation“. In: *Social indicators research* 133.1, pp. 235–258 (cit. on p. 14).
- Ott, Darrin K, Naresh Kumar, and Thomas M Peters (2008). „Passive sampling to capture spatial variability in PM_{10–2.5}“. In: *Atmospheric Environment* 42.4, pp. 746–756 (cit. on pp. 37, 43, 70).
- Özkaynak, Halûk, Lisa K Baxter, Kathie L Dionisio, and Janet Burke (2013). „Air pollution exposure prediction approaches used in air pollution epidemiology studies“. In: *Journal of Exposure Science and Environmental Epidemiology* 23.6, pp. 566–572 (cit. on p. 96).
- Pebesma, Edzer (2017). *sf: Simple Features for R*. R package version 0.5-5 (cit. on pp. 47, 78).
- Pebesma, Edzer and Roger S Bivand (2015). „Classes and Methods for Spatial Data: the sp Package“. In: *R news* 5.2, pp. 9–13 (cit. on pp. 47, 78).

- Pedersen, Marie, Thorhallur I Halldorsson, Sjurdur F Olsen, et al. (2017). „Impact of road traffic pollution on pre-eclampsia and pregnancy-induced hypertensive disorders“. In: *Epidemiology (Cambridge, Mass.)* 28.1, p. 99 (cit. on p. 114).
- Peng, JF, M Hu, ZB Wang, et al. (2014). „Submicron aerosols at thirteen diversified sites in China: size distribution, new particle formation and corresponding contribution to cloud condensation nuclei production“. In: *Atmospheric Chemistry and Physics* 14.18, pp. 10249–10265 (cit. on p. 69).
- Pilz, Jürgen, Hannes Kazianka, and Gunter Spöck (2012). „Some advances in Bayesian spatial prediction and sampling design“. In: *Spatial Statistics* 1, pp. 65–81 (cit. on p. 25).
- Polk, AE, JM Kranig, and ED Minge (1996). „Field test of non-intrusive traffic detection technologies“. In: *Intelligent Transportation: Realizing the Benefits. Proceedings of the 1996 Annual Meeting of ITS America. ITS America* (cit. on p. 114).
- Poplawski, Karla, Timothy Gould, Eleanor Setton, et al. (2009). „Intercity transferability of land use regression models for estimating ambient concentrations of nitrogen dioxide“. In: *Journal of exposure science and environmental epidemiology* 19.1, p. 107 (cit. on p. 22).
- Prüss-Üstün, Annette and Carlos Corvalán (2006). „Preventing disease through healthy environments“. In: *Towards an estimate of the environmental burden of disease. Geneva: World Health Organization* (cit. on p. 2).
- Pu, Qifan, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel (2013). „Whole-home gesture recognition using wireless signals“. In: *Proceedings of the 19th annual international conference on Mobile computing & networking. ACM*, pp. 27–38 (cit. on p. 118).
- Pudykiewicz, JA and AS Koziol (2001). „The application of Eulerian models for air quality prediction and the evaluation of emission control strategies in Canada“. In: *International journal of environment and pollution* 16.1-6, pp. 425–438 (cit. on p. 4).
- R Core team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016* (cit. on p. 78).
- Raffuse, SM, DC Sullivan, MC McCarthy, BM Penfold, HR Hafner, et al. (2007). „Ambient air monitoring network assessment guidance, analytical techniques for technical assessments of ambient air monitoring networks“. In: *Retrieved July 20, p. 2007* (cit. on pp. 37, 152).
- Rahman, Md Mahmudur, Bijan Yeganeh, Sam Clifford, Luke D Knibbs, and Lidia Morawska (2017). „Development of a land use regression model for daily NO₂ and NO_x concentrations in the Brisbane metropolitan area, Australia“. In: *Environmental Modelling & Software* 95, pp. 168–179 (cit. on p. 22).
- Raya, Maxim and Jean-Pierre Hubaux (2007). „Securing vehicular ad hoc networks“. In: *Journal of computer security* 15.1, pp. 39–68 (cit. on p. 119).
- Richardson, Douglas B, Nora D Volkow, Mei-Po Kwan, et al. (2013). „Spatial turn in health research“. In: *Science* 339.6126, pp. 1390–1392 (cit. on pp. 96, 97, 100, 110).
- Righini, Gaia, Andrea Cappelletti, Alessandra Ciucci, et al. (2014). „GIS based assessment of the spatial representativeness of air quality monitoring stations using pollutant emissions data“. In: *Atmospheric environment* 97, pp. 121–129 (cit. on p. 25).

- Risimati, Brightnes and Trynos Gumbo (2018). „Exploring the Applicability of Location Based Services to Determine the State Routes Transport Networks Integratedness in the City of Johannesburg“. In: *REAL CORP 2018–EXPANDING CITIES–DIMINISHING SPACE. Are “Smart Cities” the solution or part of the problem of continuous urbanisation around the globe? Proceedings of 23rd International Conference on Urban Planning, Regional Development and Information*. CORP–Competence Center of Urban and Regional Planning, pp. 225–234 (cit. on p. 140).
- Roche, Stéphane (2014). „Geographic Information Science I: Why does a smart city need to be spatially enabled?“. In: *Progress in Human Geography* 38.5, pp. 703–711 (cit. on pp. 98, 110).
- Röösli, Martin, Charlotte Braun-Fährlander, Nino Künzli, et al. (2000). „Spatial variability of different fractions of particulate matter within an urban environment and between urban and rural sites“. In: *Journal of the Air & Waste Management Association* 50.7, pp. 1115–1124 (cit. on p. 43).
- Ross, Zev, Paul B English, Rusty Scalf, et al. (2006). „Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses“. In: *Journal of Exposure Science and Environmental Epidemiology* 16.2, p. 106 (cit. on pp. 22, 23).
- Roswall, Nina, Ole Raaschou-Nielsen, Matthias Ketzel, et al. (2017). „Long-term residential road traffic noise and NO2 exposure in relation to risk of incident myocardial infarction–A Danish cohort study“. In: *Environmental research* 156, pp. 80–86 (cit. on p. 114).
- Roudier, Pierre, DE Beaudette, and AE Hewitt (2012). „A conditioned Latin hypercube sampling algorithm incorporating operational constraints“. In: *Digital Soil Assessments and Beyond; CRC Press: Sydney, NSW, Australia*, pp. 227–231 (cit. on p. 32).
- Roy, Swaroop, Rijurekha Sen, Swanand Kulkarni, et al. (2011). „Wireless across road: RF based road traffic congestion detection“. In: *Communication Systems and Networks (COMSNETS), 2011 Third International Conference on*. IEEE, pp. 1–6 (cit. on pp. 118, 133).
- Royuela, V, R Moreno, and E Vayá (2007). „Is the influence of quality of life on urban growth non-stationary in space“. In: *A case study of Barcelona. Institut de Recerca en Economia Aplicada* (cit. on p. 2).
- Ryan, Patrick H and Grace K LeMasters (2007). „A review of land-use regression models for characterizing intraurban air pollution exposure“. In: *Inhalation toxicology* 19.sup1, pp. 127–133 (cit. on pp. 5, 20, 22, 24, 38, 39).
- Sadalla, Edward, Subhrajit Guhathakurta, and Susan Ledlow (2005). „Environment and quality of life: A conceptual analysis and review of empirical literature“. In: *The US-Mexican border environment: Dynamics of human-environment interactions*, pp. 29–79 (cit. on pp. 2, 18).
- Salata, Ferdinando, Iacopo Golasi, Davide Petitti, et al. (2017). „Relating microclimate, human thermal comfort and health during heat waves: An analysis of heat island mitigation strategies through a case study in an urban outdoor environment“. In: *Sustainable cities and society* 30, pp. 79–96 (cit. on p. 154).
- Salvadori, Claudio, Matteo Petracca, Stefano Bocchino, Riccardo Pelliccia, and Paolo Pagano (2015). „A low-cost vehicle counter for next-generation ITS“. In: *Journal of Real-Time Image Processing* 10.4, pp. 741–757 (cit. on pp. 118, 119).

- Samuel-Rosa, Alessandro, Lucia Helena Cunha dos Anjos, Gustavo de Mattos Vasques, et al. (2017). „Package âpsann“. In: (cit. on pp. 47, 78).
- Santiago, Jose Luis, Fernando Martín, and Alberto Martilli (2013). „A computational fluid dynamic modelling approach to assess the representativeness of urban monitoring stations“. In: *Science of the total environment* 454, pp. 61–72 (cit. on p. 25).
- Sarigiannis, Dimosthenis A and Michaela Saisana (2008). „Multi-objective optimization of air quality monitoring“. In: *Environmental monitoring and assessment* 136.1-3, pp. 87–99 (cit. on pp. 38, 57).
- Sayegh, Arwa, James E Tate, and Karl Ropkins (2016). „Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees“. In: *Atmospheric Environment* 127, pp. 163–175 (cit. on p. 2).
- Schneider, Philipp, Nuria Castell, Matthias Vogt, et al. (2017a). „Mapping urban air quality in near real-time using observations from low-cost sensors and model information“. In: *Environment international* 106, pp. 234–247 (cit. on pp. 64, 65, 68).
- Schneider, Philipp, Nuria Castell, Matthias Vogt, et al. (2017b). „Mapping urban air quality in near real-time using observations from low-cost sensors and model information“. In: *Environment International* 106.May, pp. 234–247 (cit. on pp. 97, 140, 150).
- Schneider, Philipp, Nuria Castell, Franck R Dauge, et al. (2018). „A Network of Low-Cost Air Quality Sensors and Its Use for Mapping Urban Air Quality“. In: *Mobile Information Systems Leveraging Volunteered Geographic Information for Earth Observation*. Springer, pp. 93–110 (cit. on p. 98).
- Scott, E Marian (2017). „Within-urban variability in ambient air pollution: comparison of estimation methods“. In: *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data to appear* (cit. on p. 30).
- Senaratne, Hansi, Amin Mobasher, Ahmed Loai Ali, Cristina Capineri, and Mordechai (Muki) Haklay (2017). „A review of volunteered geographic information quality assessment methods“. In: *International Journal of Geographical Information Science* 31.1, pp. 139–167 (cit. on p. 66).
- Shaddick, Gavin, Matthew L Thomas, Amelia Green, et al. (2018). „Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution“. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.1, pp. 231–253 (cit. on p. 2).
- Shi, Xiaoqin, Chuanfeng Zhao, Jonathan H Jiang, et al. (2018). „Spatial Representativeness of PM2.5 Concentrations Obtained Using Observations From Network Stations“. In: *Journal of Geophysical Research: Atmospheres* 123.6, pp. 3145–3158 (cit. on p. 25).
- Shusterman, Alexis A, Virginia E Teige, Alexander J Turner, et al. (2016). „The Berkeley Atmospheric CO2 Observation Network: initial evaluation“. In: *Atmospheric Chemistry and Physics* 16.21, pp. 13449–13463 (cit. on p. 64).
- Sieber, Renée E and Mordechai Haklay (2015). „The epistemology(s) of volunteered geographic information: a critique“. In: *Geo: Geography and Environment* 2.2, pp. 122–136 (cit. on p. 66).

- Singh, Nongthombam Premananda and Sharad Gokhale (2015). „A method to estimate spatiotemporal air quality in an urban traffic corridor“. In: *Science of the Total Environment* 538, pp. 458–467 (cit. on p. 21).
- Snyder, Emily G, Timothy H Watkins, Paul A Solomon, et al. (2013). *The changing paradigm of air pollution monitoring* (cit. on pp. 19, 64, 67).
- Spinelle, Laurent, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola (2017). „Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO₂“. In: *Sensors and Actuators B: Chemical* 238, pp. 706–715 (cit. on pp. 8, 70).
- Su, Judith (2018). *Portable and sensitive air pollution monitoring* (cit. on p. 4).
- Szatmári, Gábor, Péter László, Katalin Takács, et al. (2018). „Optimization of second-phase sampling for multivariate soil mapping purposes: Case study from a wine region, Hungary“. In: *Geoderma* (cit. on p. 26).
- Tang, Yong, Congzhe Zhang, Renshu Gu, Peng Li, and Bin Yang (2017). „Vehicle detection and recognition for intelligent traffic surveillance system“. In: *Multimedia tools and applications* 76.4, pp. 5817–5832 (cit. on p. 117).
- Taylor, Linnet, Luciano Floridi, and Bart van der Sloot (2016). *Group privacy: New challenges of data technologies*. Vol. 126. Springer (cit. on pp. 100, 109).
- Thepvilojanapong, Niwat, Shin'ichi Konomi, Yoshito Tobe, et al. (2010). „Opportunistic collaboration in participatory sensing environments“. In: *Proceedings of the fifth ACM international workshop on Mobility in the evolving internet architecture*. ACM, pp. 39–44 (cit. on p. 99).
- Thouron, L, C Seigneur, Y Kim, et al. (2018). „Intercomparison of three modeling approaches for traffic-related road dust resuspension using two experimental data“. In: *Transportation Research Part D: Transport and Environment* 58, pp. 108–121 (cit. on p. 5).
- Truong, Phuong N, Gerard BM Heuvelink, and John Paul Gosling (2013). „Web-based tool for expert elicitation of the variogram“. In: *Computers & Geosciences* 51, pp. 390–399 (cit. on p. 32).
- Umbelt Bundesamt (2018). *Current concentrations of air pollutants in Germany*. Umbelt Bundesamt (cit. on p. 71).
- Van Groenigen, Jan Willem, W Siderius, and A Stein (1999). „Constrained optimisation of soil sampling for minimisation of the kriging variance“. In: *Geoderma* 87.3, pp. 239–259 (cit. on pp. 38, 44, 52, 60, 74, 90, 150).
- Van Groenigen, JW and A Stein (1998). „Constrained optimization of spatial sampling using continuous simulated annealing“. In: *Journal of Environmental Quality* 27.5, pp. 1078–1086 (cit. on pp. 25, 139).
- Vardoulakis, Sotiris, Bernard EA Fisher, Koulis Pericleous, and Norbert Gonzalez-Flesca (2003). „Modelling air quality in street canyons: a review“. In: *Atmospheric environment* 37.2, pp. 155–182 (cit. on p. 21).
- Viitanen, Jenni and Richard Kingston (2014). „Smart cities and green growth: outsourcing democratic and environmental resilience to the global technology sector“. In: *Environment and Planning A* 46.4, pp. 803–819 (cit. on p. 14).

- Villanueva, Félix Jesús, Cesar Aguirre, David Villa, Maria José Santofimia, and Juan Carlos López (2014). „Smart City data stream visualization using Glyphs“. In: *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on*. IEEE, pp. 399–403 (cit. on p. 33).
- Wadoux, Alexandre M.J-C., Dick J. Brus, Miguel A. Rico-Ramirez, and Gerard B.M. Heuvelink (2017). „Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model“. In: *Advances in Water Resources* 107, pp. 126–138 (cit. on pp. 44, 60, 151).
- Wang, Bing, Bingzhen Chen, and Jinsong Zhao (2015a). „The real-time estimation of hazardous gas dispersion by the integration of gas detectors, neural network and gas dispersion models“. In: *Journal of hazardous materials* 300, pp. 433–442 (cit. on p. 25).
- Wang, Jianghao, Yong Ge, Gerard B M Heuvelink, and Chenghu Zhou (2014a). „Spatial sampling design for estimating regional GPP with spatial heterogeneities“. In: *IEEE Geoscience and Remote Sensing Letters* 11.2, pp. 539–543 (cit. on p. 44).
- Wang, Jin-Feng, A Stein, Bin-Bo Gao, and Yong Ge (2012a). „A review of spatial sampling“. In: *Spatial Statistics* 2, pp. 1–14 (cit. on p. 58).
- Wang, Kai, Hong Zhao, Yanan Ding, et al. (2015b). „Optimization of air pollutant monitoring stations with constraints using genetic algorithm“. In: *Journal of High Speed Networks* 21.2, pp. 141–153 (cit. on pp. 38, 40, 58).
- Wang, Meng, Rob Beelen, Marloes Eeftens, et al. (2012b). „Systematic evaluation of land use regression models for NO₂“. In: *Environmental science & technology* 46.8, pp. 4481–4489 (cit. on pp. 40, 68, 70).
- Wang, Meng, Rob Beelen, Tom Bellander, et al. (2014b). „Performance of multi-city land use regression models for nitrogen dioxide and fine particles“. In: *Environmental health perspectives* 122.8, p. 843 (cit. on pp. 22, 24).
- Wang, Rongrong, Sarah B Henderson, Hind Sbihi, Ryan W Allen, and Michael Brauer (2013a). „Temporal stability of land use regression models for traffic-related air pollution“. In: *Atmospheric Environment* 64, pp. 312–319 (cit. on p. 22).
- Wang, Yan, Jie Yang, Hongbo Liu, et al. (2013b). „Measuring human queues using WiFi signals“. In: *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, pp. 235–238 (cit. on p. 118).
- Watkins, T (2013). „DRAFT roadmap for next generation air monitoring“. In: *Environmental Protection Agency* (cit. on pp. 65, 70).
- Weichenthal, Scott, Keith Van Ryswyk, Alon Goldstein, et al. (2016). „A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach“. In: *Environmental research* 146, pp. 65–72 (cit. on p. 64).
- Weissert, L F, J A Salmond, Georgia Miskell, et al. (2017a). „Use of a dense monitoring network of low-cost instruments to observe local changes in the diurnal ozone cycles as marine air passes over a geographically isolated urban centre“. In: *Science of the Total Environment* 575, pp. 67–78 (cit. on p. 97).

- Weissert, LF, JA Salmond, Georgia Miskell, et al. (2017b). „Use of a dense monitoring network of low-cost instruments to observe local changes in the diurnal ozone cycles as marine air passes over a geographically isolated urban centre“. In: *Science of The Total Environment* 575, pp. 67–78 (cit. on p. 70).
- WHO (2016). *WHO releases country estimates on air pollution exposure and health impact*. World Health Organisation (cit. on p. 64).
- Wijs, Lisanne de, Patrick Witte, and Stan Geertman (2016). „How smart is smart? Theoretical and empirical considerations on implementing smart city objectives—a case study of Dutch railway station areas“. In: *Innovation: The European Journal of Social Science Research* 29.4, pp. 424–441 (cit. on p. 14).
- Wolf, Kathrin, Josef Cyrus, Tatiana Harciníková, et al. (2017). „Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany“. In: *Science of the Total Environment* 579, pp. 1531–1540 (cit. on p. 64).
- Won, Myounggyu, Shaohu Zhang, and Sang H Son (2017). „WiTraffic: low-cost and non-intrusive traffic monitoring system using WiFi“. In: *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, pp. 1–9 (cit. on pp. 118, 132, 133).
- Wu, Chih-Da, Yu-Ting Zeng, and Shih-Chun Candice Lung (2018). „A hybrid kriging/land-use regression model to assess PM_{2.5} spatial-temporal variability“. In: *Science of The Total Environment* 645, pp. 1456–1464 (cit. on p. 23).
- Wu, Hao, Stefan Reis, Chun Lin, and Mathew R Heal (2017). „Effect of monitoring network design on land use regression models for estimating residential NO₂ concentration“. In: *Atmospheric Environment* 149, pp. 24–33 (cit. on pp. 38, 40, 57, 59, 89–91, 150).
- Wu, Lin, Marc Bocquet, and Matthieu Chevallier (2010). „Optimal reduction of the ozone monitoring network over France“. In: *Atmospheric environment* 44.25, pp. 3071–3083 (cit. on pp. 37, 57, 58).
- Wulder, Michael A and Nicholas C Coops (2014). „Make Earth observations open access: freely available satellite imagery will improve science and environmental-monitoring products“. In: *Nature* 513.7516, pp. 30–32 (cit. on p. 150).
- Xu, M (2017). „Quantify Effects of Traffic, Grasslands and Water Bodies on Urban Heat Islands: Kent Vale Case Study“. In: *J Environ Bio Res* 1.1, p. 1 (cit. on p. 114).
- Yang, Chen, Bo Qin, Xiuwen Zhou, et al. (2015). „Privacy-Preserving Traffic Monitoring in Vehicular Ad Hoc Networks“. In: *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*. IEEE, pp. 22–24 (cit. on p. 119).
- Yi, Wei Ying, Kin Ming Lo, Terrence Mak, et al. (2015). „A survey of wireless sensor network based air pollution monitoring systems“. In: *Sensors* 15.12, pp. 31392–31427 (cit. on pp. 64, 67).
- Yin, ChuanTao, Zhang Xiong, Hui Chen, et al. (2015). „A literature survey on smart cities“. In: *Science China Information Sciences* 58.10, pp. 1–18 (cit. on p. 14).
- Yu, Haofei, Armistead Russell, James Mulholland, et al. (2018). „Cross-comparison and evaluation of air pollution field estimation methods“. In: *Atmospheric Environment* 179, pp. 49–60 (cit. on p. 5).

- Zhang, Kai and Stuart Batterman (2013). „Air pollution and health risks due to vehicle traffic“. In: *Science of the total Environment* 450, pp. 307–316 (cit. on p. 5).
- Zhang, Xiaole, Wolfgang Raskob, Claudia Landman, Dmytro Trybushnyi, and Yu Li (2017). „Sequential multi-nuclide emission rate estimation method based on gamma dose rate measurement for nuclear emergency management“. In: *Journal of hazardous materials* 325, pp. 288–300 (cit. on p. 25).
- Zubizarreta, Iker, Alessandro Seravalli, and Saioa Arrizabalaga (2015). „Smart city concept: What it is and what it should be“. In: *Journal of Urban Planning and Development* 142.1, p. 04015005 (cit. on p. 14).

Web pages

- BBC (2015). *China pollution: First ever red alert in effect in Beijing*. URL: <https://www.bbc.com/news/world-asia-china-35026363> (cit. on p. 2).
- Copernicus (2018). *Copernicus Open Access Hub*. URL: <https://scihub.copernicus.eu/dhus/#/home> (visited on May 30, 2018) (cit. on p. 149).
- Costa Jamie Schulte, Brooke Singer Beatriz da (1999). *AIR :: Area's Immediate Reading*. URL: <http://www.pm-air.net/credits.php> (visited on June 10, 2018) (cit. on p. 98).
- European Commission (2018b). *Free wireless internet hotspots in public spaces*. URL: https://ec.europa.eu/commission/news/free-wireless-internet-hotspots-public-spaces-2018-mar-20_en (visited on Apr. 30, 2018) (cit. on p. 151).
- European forum on eco-innovation (2018). *Eco-innovation for air quality*. URL: http://ec.europa.eu/environment/ecoinnovation2018/1st_forum/ (visited on June 22, 2018) (cit. on p. 16).
- European Respiratory Society (2017). *European Environment Agency report warns air pollution is still a major health risk in Europe*. URL: <https://www.ersnet.org/the-society/news/european-environment-agency-report-warns-air-pollution-is-still-a-major-health-risk-in-europe> (visited on Nov. 22, 2017) (cit. on p. 3).
- Eurostat (2015). *Quality of life indicators*. URL: http://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators (visited on Apr. 30, 2016) (cit. on p. 3).
- Federal Register (OFR), Office of the and the Government Publishing Office.OFR (2018). *Electronic Code of Federal Regulations*. URL: https://www.ecfr.gov/cgi-bin/text-idx?SID=e5c581756381a54c3c1976615f55b8b3&mc=true&tpl=/ecfrbrowse/Title40/40cfr58_main_02.tpl (visited on May 30, 2018) (cit. on p. 4).
- Independent (2017). *Indian medics declare 'health emergency' in Delhi as smog blankets city*. URL: <https://www.independent.co.uk/news/world/asia/india-delhi-smog-pollution-air-quality-doctors-medical-association-arvind-kejriwal-a8042931.html> (visited on Nov. 22, 2017) (cit. on p. 2).
- OpenEO (2017). *openEO - A Common, Open Source Interface between Earth Observation Data Infrastructures and Front-End Applications*. URL: <http://openeo.org/about/> (visited on May 10, 2018) (cit. on p. 152).

- UNDESA (2018). *68% of the world population projected to live in urban areas by 2050, says UN*. URL: <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html> (visited on Apr. 30, 2018) (cit. on p. 15).
- UNDP (2016). *Sustainable Development Goals*. URL: <http://www.undp.org/content/undp/en/home/sustainable-development-goals.html> (visited on Apr. 10, 2018) (cit. on p. 153).
- WHO, World Health Organisation (2014). *Children's environmental health*. URL: <http://www.who.int/ceh/risks/cehair/en/> (visited on June 5, 2018) (cit. on p. 16).
- (2015). *Global Health Observatory (GHO) data*. URL: http://www.who.int/gho/mortality_burden_disease/causes_death/en/ (visited on Sept. 30, 2016) (cit. on p. 2).
 - (2016). *Mortality and burden of disease from ambient air pollution*. URL: http://www.who.int/gho/phe/outdoor_air_pollution/burden_text/en/ (visited on June 30, 2018) (cit. on p. 16).
 - (2018a). *9 out of 10 people worldwide breathe polluted air, but more countries are taking action*. URL: <https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action> (visited on June 30, 2018) (cit. on p. 16).
 - (2018b). *The top 10 causes of death*. URL: <http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> (visited on June 5, 2018) (cit. on p. 16).
 - (2018c). *WHO Global Ambient Air Quality Database (update 2018)*. URL: <http://www.who.int/airpollution/data/cities/en/> (visited on June 30, 2018) (cit. on p. 16).

Supplementary figures from Chapter 4

This section contain the map of the study area, to represent the spread of data populated housing areas.

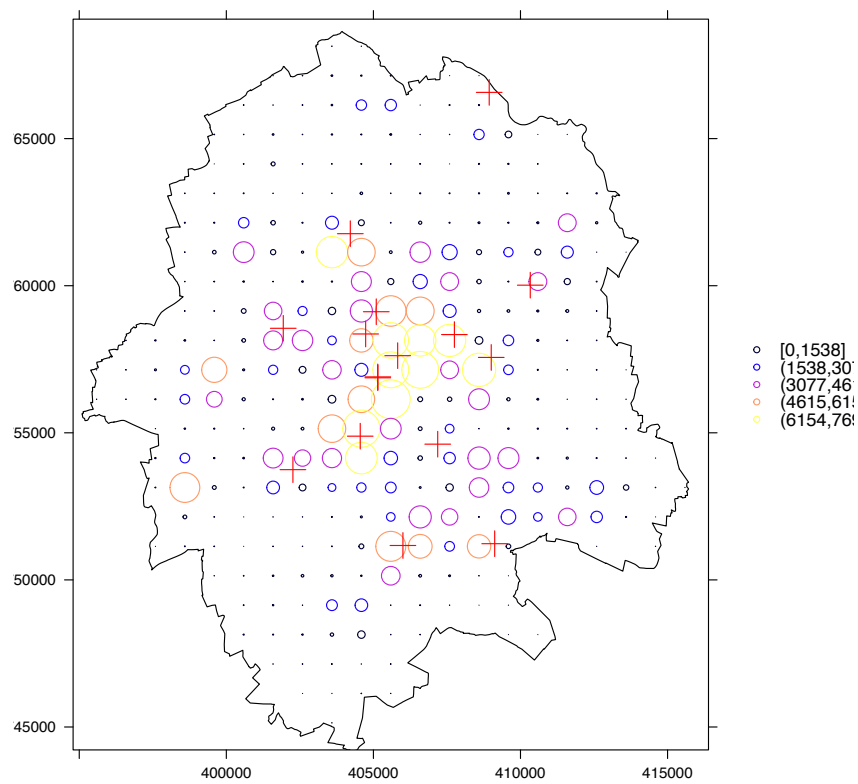


Figure. A.1. Populated housing area map with initial monitoring station locations (red plus signs) for study area

A.1

The following image shows the histogram of all the predictor variables used in the study as described in Table 4.1.

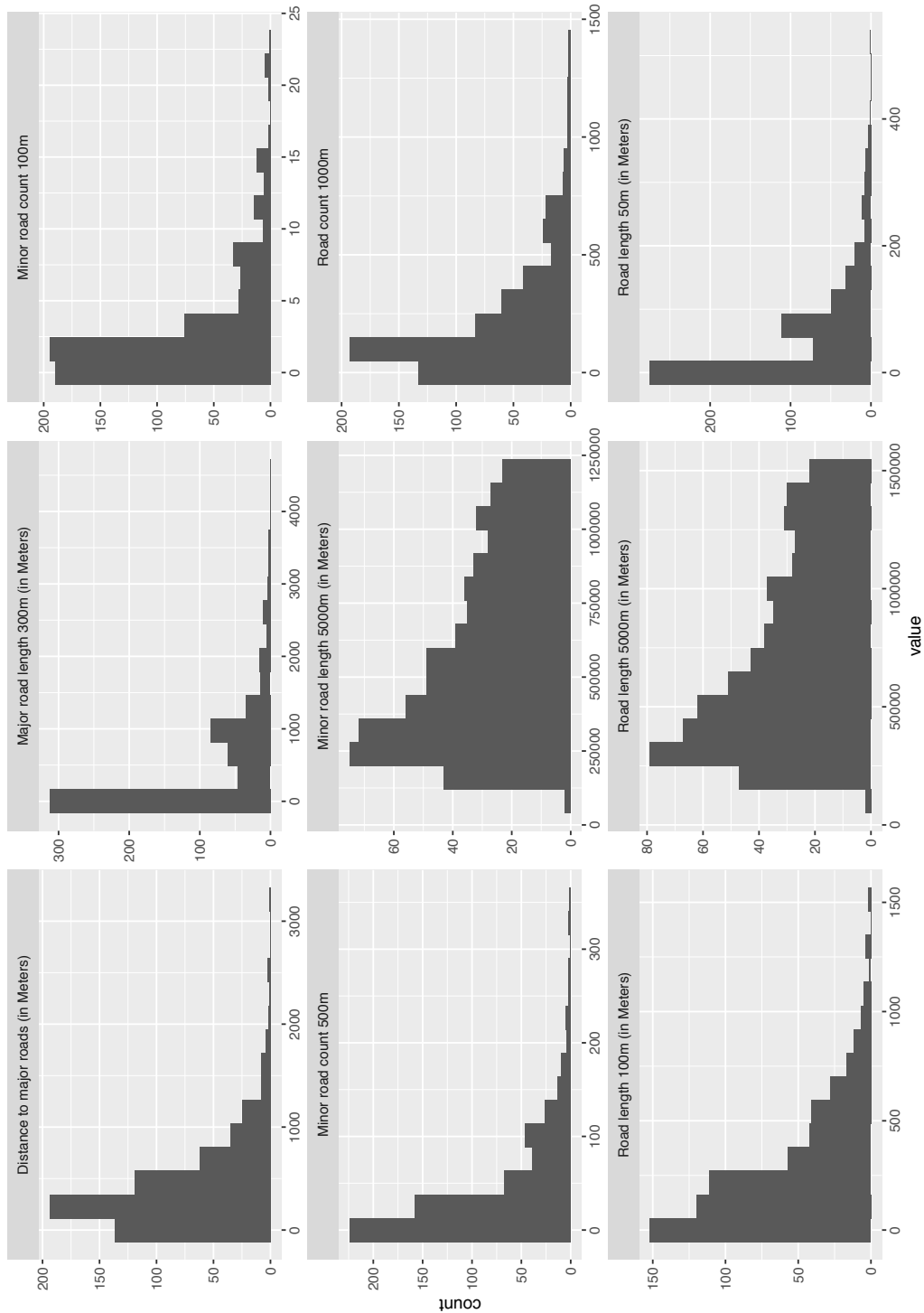
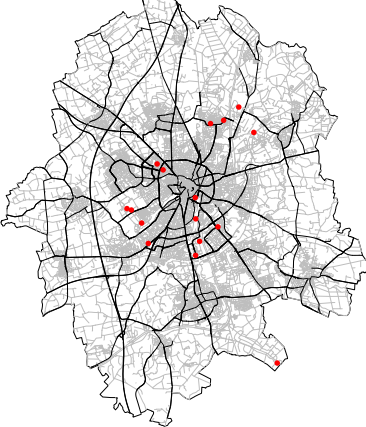
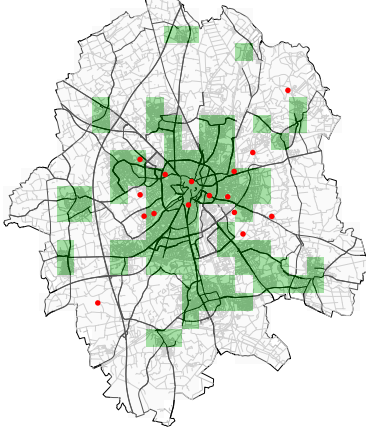
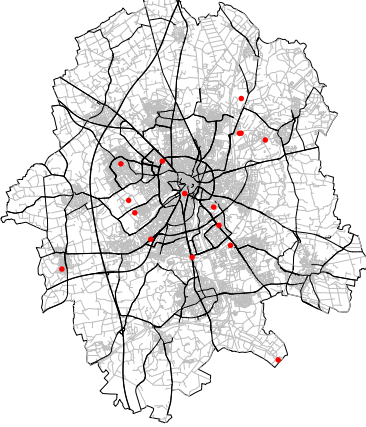
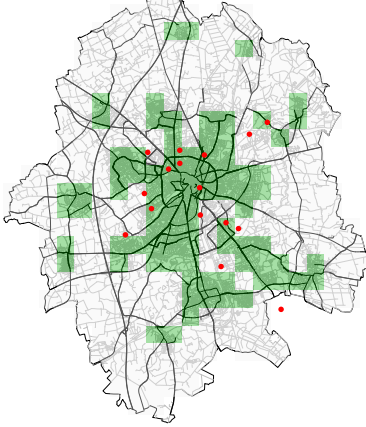
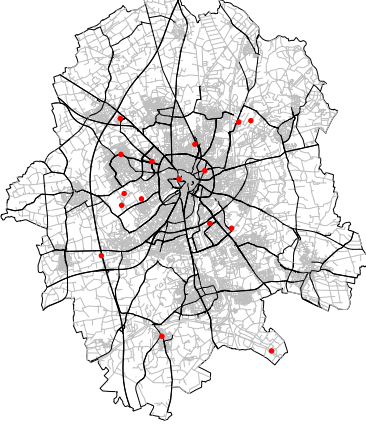
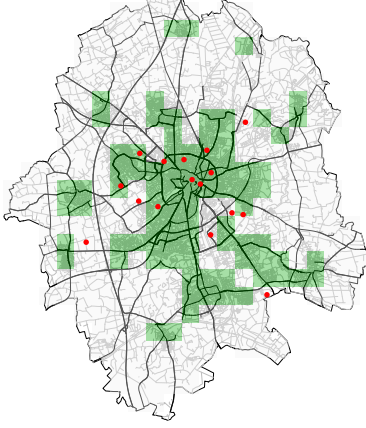
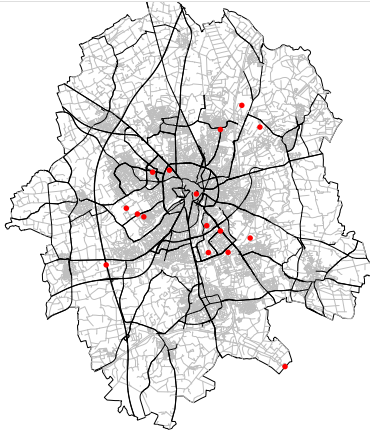


Figure. A.2. Histogram of predictor variables for LUR used in the study

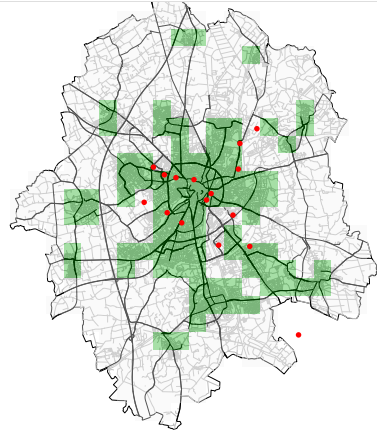
A.2

In this section we show all the configurations we realised for the study area by using optimisation method for non weighted and weighted optimisation.

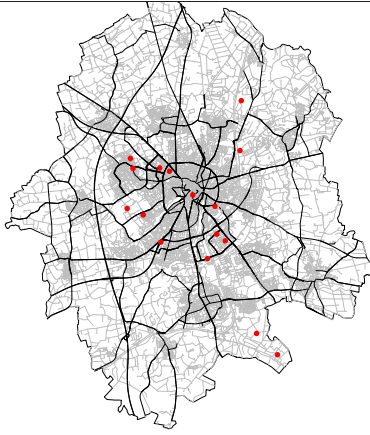
Various other configurations realised during the study for different probability of acceptance	
Without weight optimisation	Population weighted optimisation
	
Spatial mean prediction Error: 0.1990	Spatial mean prediction Error: 404.57
	
Spatial mean prediction Error: 0.2039	Spatial mean prediction Error: 363.34
	
Spatial mean prediction Error: 0.21179	Spatial mean prediction Error: 332.86



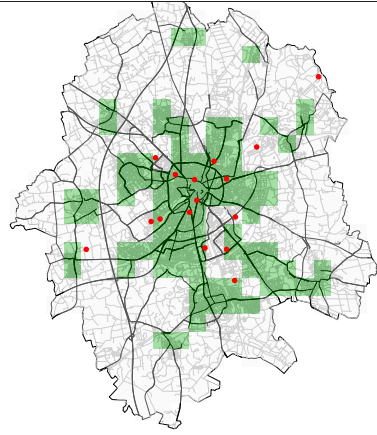
Spatial mean prediction Error: 0.2018



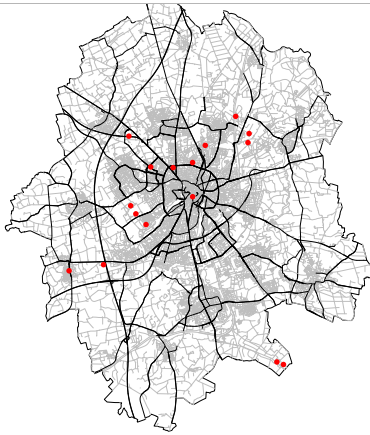
Spatial mean prediction Error: 370.25



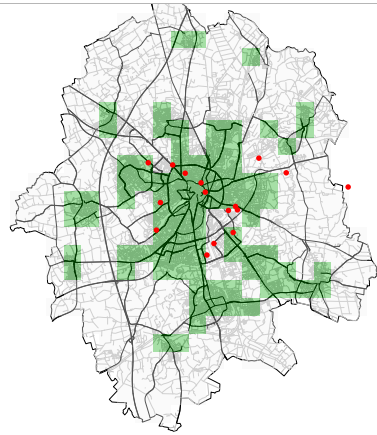
Spatial mean prediction Error: 0.1918



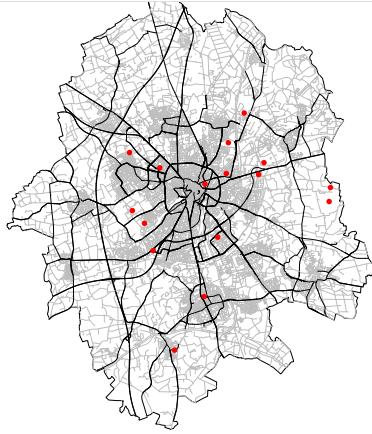
Spatial mean prediction Error: 385.52



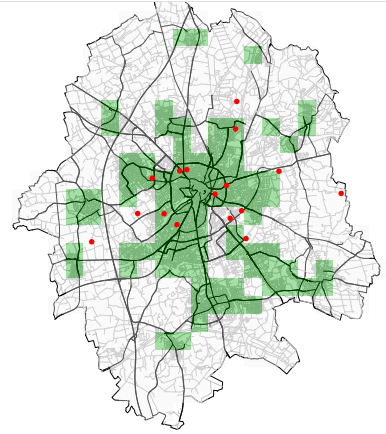
Spatial mean prediction Error: 0.2027



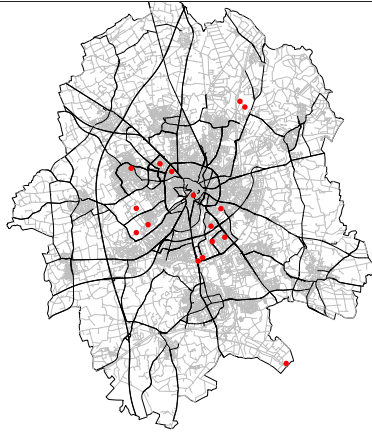
Spatial mean prediction Error: 405.39



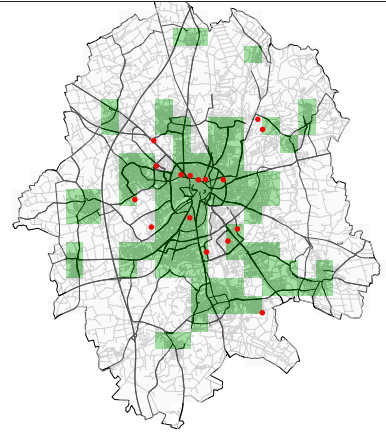
Spatial mean prediction Error: 0.2579



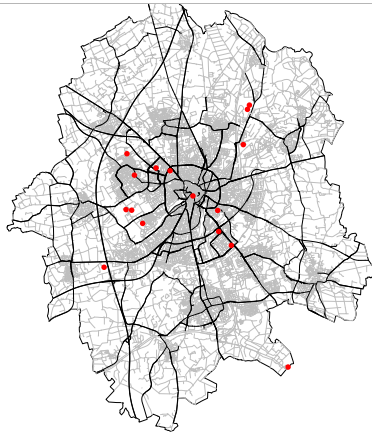
Spatial mean prediction Error: 387.15



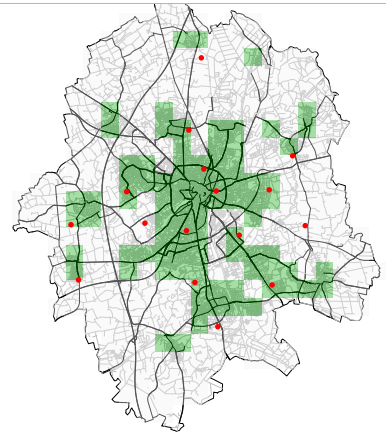
Spatial mean prediction Error: 0.2043



Spatial mean prediction Error: 387.57



Spatial mean prediction Error: 0.19381



Spatial mean prediction Error: 27798.28

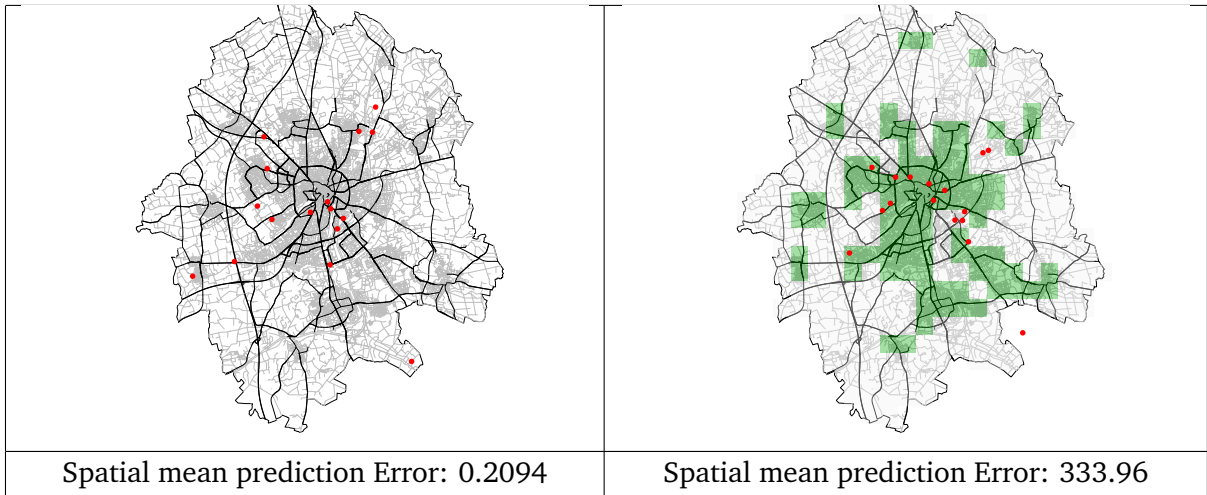
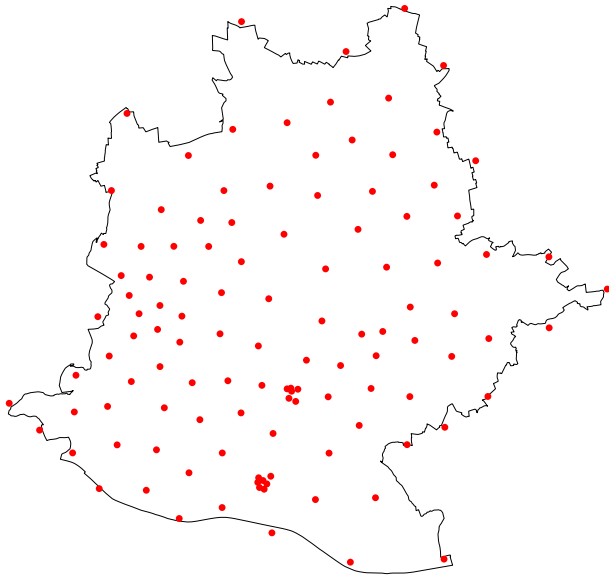
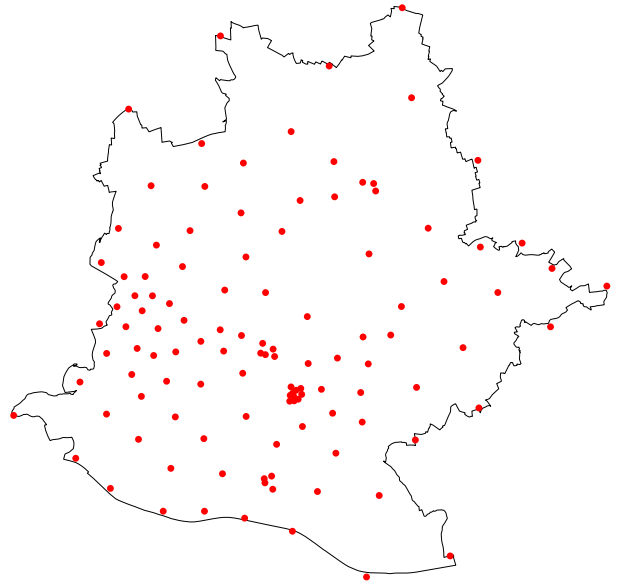


Table. A.1. All the configuration realised for optimisation at different probability of acceptance

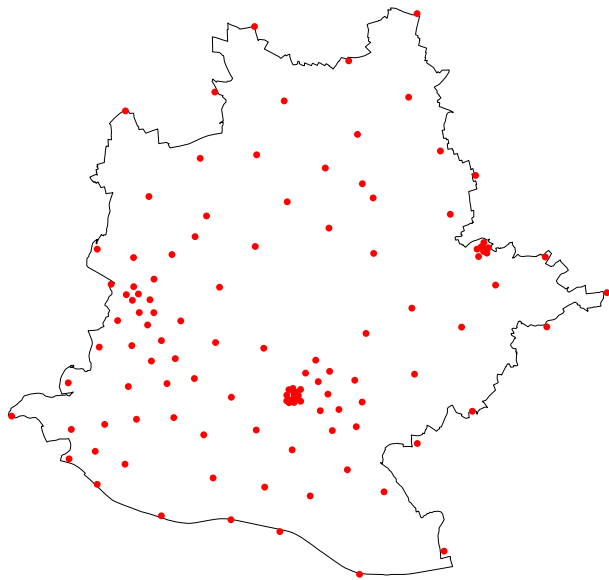
Supplementary figures from
Chapter 5



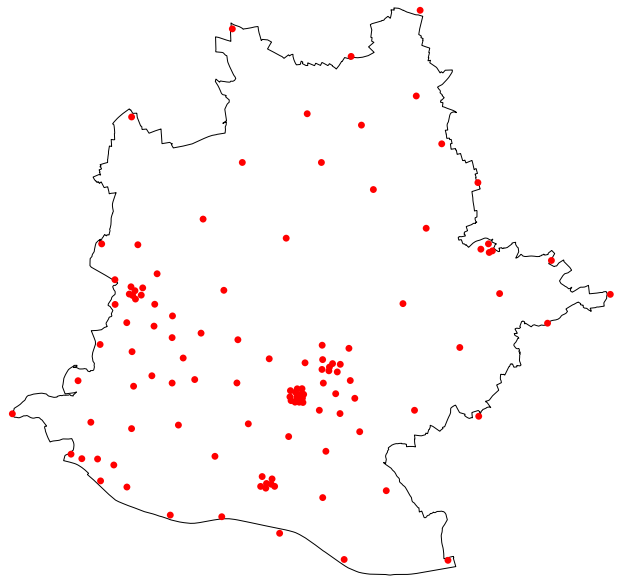
(a) $W1=0.10$ and $W2=0.90$.



(b) $W1=0.20$ and $W2=0.80$.

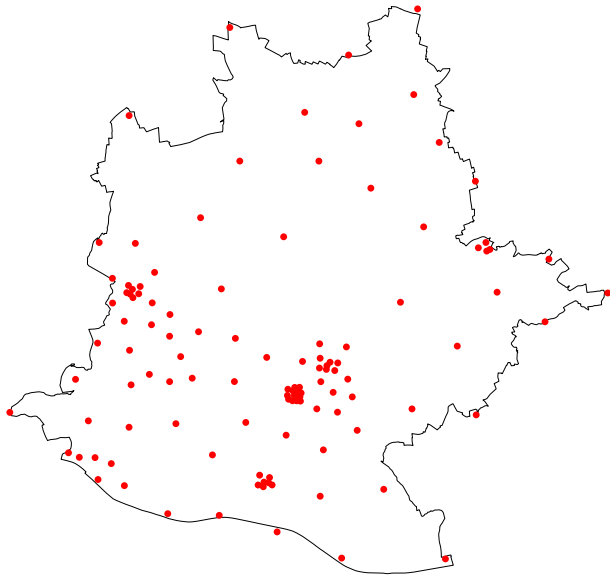


(c) $W1=0.30$ and $W2=0.70$.

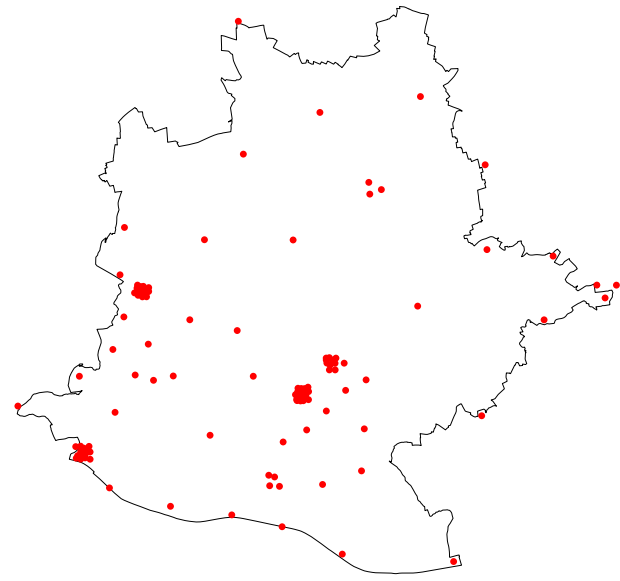


(d) $W1=0.40$ and $W2=0.60$.

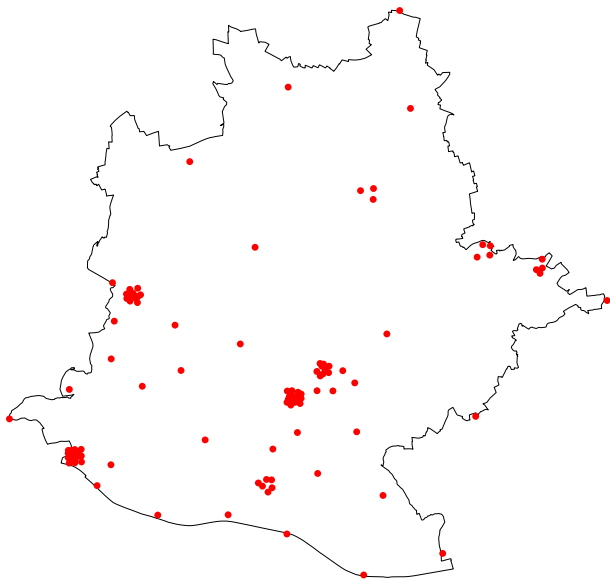
Figure. B.1. Configuration with various weights on prediction error ($W1$) and wide-spread($W2$) aspect of developed objective function.



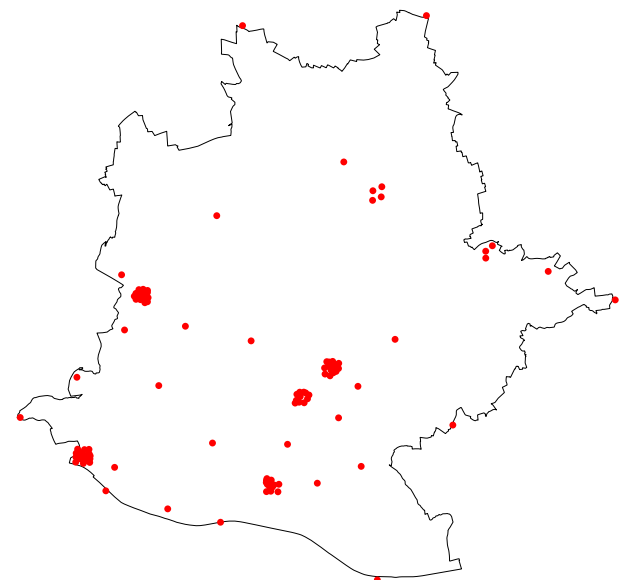
(e) $W1=0.60$ and $W2=0.40$.



(f) $W1=0.70$ and $W2=0.30$.



(g) $W1=0.80$ and $W2=0.20$.



(h) $W1=0.90$ and $W2=0.10$.

Figure. B.1. Configuration with various weights on prediction error ($W1$) and wide-spread($W2$) aspect of developed objective function.

B.1

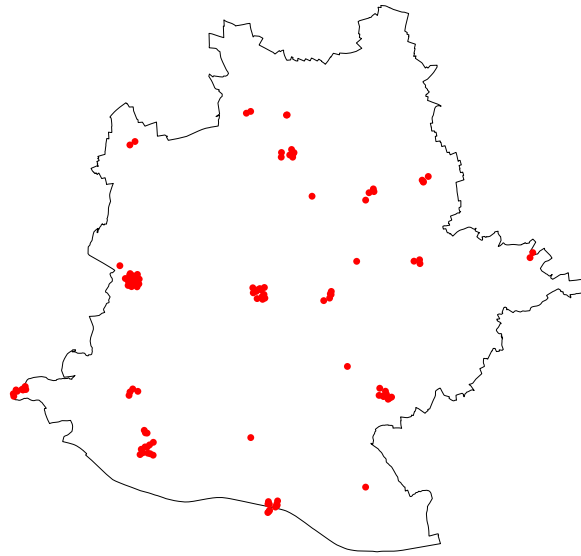


Figure. B.2. Optimal locations considering prediction error constrain for Stuttgart LUR model developed using low-cost sensor network data.

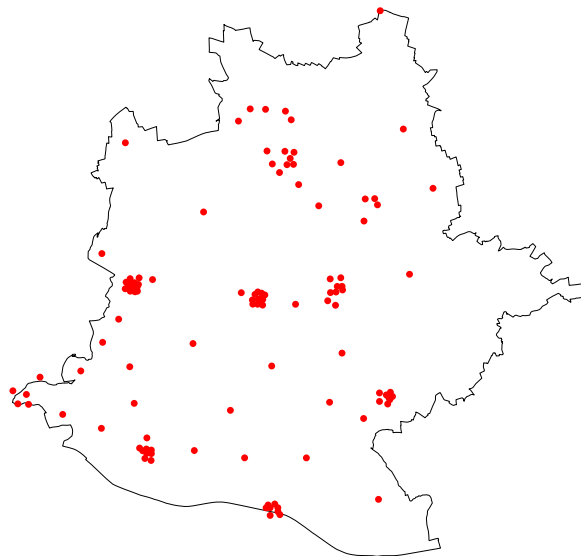


Figure. B.3. Optimal locations considering prediction error and wide-spread aspect in the objective function for Stuttgart LUR model developed using low-cost sensor network data.

List of Abbreviations

AQ	Air Quality
AVI	Automatic Vehicle Identification
CORINE	coordination of information on the environment
CSI	Channel State Information
DSP	Digital Signal Processor
EC	Elemental Carbon
ESCAPE	European study of cohorts for air pollution effects
GIS	Geographical information systems
GPS	Global Positioning System
ICT	Information and Communication Technology
IMSD	Inverse mean shortest distance
IoT	Internet of Things
LQI	Link Quality Indicator
LUR	Land Use Regression
MAG	Magnetometer
MND	Monitoring Network Design
NO_x	Nitrogen oxides
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
O_3	Ozone
PE	Prediction Error
PM	Particulate Matter
$PM_{2.5}$	Particulate Matter (PM) that have a diameter of less than 2.5 micrometers
QoL	Quality of Life
RSSI	Received Signal Strength Indication
SSA	Spatial Simulated Annealing
SQRALT	Square root of altitude
UHI	Urban Heat Islands
USEPA	United States Environmental Protection Agency
VGI	Volunteered Geographic Information
WHO	World Health Organisation
WLS	Weighted Least square

