REPORT SERIES IN AEROSOL SCIENCE
N:o 246 (2022)

Derivation of Black Carbon Proxies in an Integrated Urban Air Quality Monitoring Network

Pak Lun Fung

Institute for Atmospheric and Earth System Research (INAR) / Physics
Faculty of Science
University of Helsinki
Helsinki, Finland

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Exactum CK112, Pietari Kalmin katu 5, on January 21st, 2022, at 12 o'clock noon.

Helsinki 2022

Author's Address: Institute for Atmospheric & Earth System Research (INAR) / Physics
P.O. Box 64
FI-00014 University of Helsinki
pak.fung@helsinki.fi

Supervisors: Professor Tareq Hussein, Ph.D.
Department of Physics
University of Jordan
Institute for Atmospheric & Earth System Research (INAR) / Physics
University of Helsinki

Professor Tuukka Petäjä, Ph.D.
Institute for Atmospheric & Earth System Research (INAR) / Physics
University of Helsinki

Docent Martha Arbayani Zaidan, Ph.D.
Institute for Atmospheric & Earth System Research (INAR) / Physics
University of Helsinki
School of Atmospheric Sciences
Nanjing University

Reviewers: Professor Jean Sciare, Ph.D.
Climate and Atmosphere Research Center
Cyprus Institute

Docent Santtu Mikkonen, Ph.D.
Department of Applied Physics
University of Eastern Finland

Opponent: Professor Stephan Weber, Ph.D.
Institute of Geoecology
Technische Universität Braunschweig

ISBN 978-952-7276-72-3 (printed version)
ISSN 0784-3496
Helsinki 2022
Unigrafia Oy

ISBN 978-952-7276-73-0 (pdf version)

Helsinki 2022
http://www.FAAR.fi

Acknowledgements

First things first, I would like to show my sincere appreciation to my three supervisors: Tareq Hussin, Tuukka Petäjä and Martha Zaidan, and especially Tuukka for serving as the custos of the defence. I would also like to thank Stephan Weber to be the opponent, and Jean Sciare and Santtu Mikkonen to be the pre-examiners. Although we did not have the chance to meet, I am still grateful to have you all for the thesis and the defence.

In my view, this doctoral thesis and defence do not only cover what I accomplished for my doctoral degree, but also reflect the scientific journey that I have gone through. Therefore, I would start thanking all the supervisors I had before my doctoral study. They have shown me what scientific research should be. Thank you Timo Vesala and Ivan Mammarella for opening the door to a new world of research during my master degree, and Liisa Kulmala for the recognition of my strength of being a researcher.

First being a summer worker under Tareq Hussein's team, he brought me to world of air quality research. I would like to thank him, together with Tuukka Petäjä and Markku Kulmala, for backing me up to work flexibly and providing me with research guidance. I would also like to thank my another supervisor Martha Zaidan for our numerous fruitful discussion on both scientific and personal level. Besides, I appreciate a lot that Pauli Paasonen brought the air quality guidance group together. The regular meetings with the group has been important to me for the sense of belonging to INAR and the identity of being an INARian.

Being a member in the HOPE project and the MegaSense programme has widened my horizon to collaborate with many great researchers from outside our institute. I have learnt a lot regarding interdisciplinary research. I would like to thank all my collaborators. Special thanks to Salla, Krista, Naser, Andrew and Samu, for the day-to-day support and working as a team in Ubikampus.

Apart from work-related gratitude, I would like to acknowledge friends and colleagues whom I valued a lot during my research journey. I might not have very close contact with every one of them but they definitely enlighten me in their own ways: Ahmed, Kira, Sini, Janne, Stephany, Aki, Toprak, etc.

Last but not least, I would like to show appreciation to my friends and family in Hong Kong. Despite the geographical separation and time difference, they have been providing continuous emotional support which helped me through countless dark and bright nights in Finland, especially in the pandemic time. Jason, Chloe, Hodar and Yukko, you are not forgotten!

Along this dream chasing journey, it can be sometimes difficult to find my own position, and, I have to admit that, from time to time, I wanted to give up and had the question asking myself why I dragged myself into this mess. Thanks to the support and inspiration, I seldom feel walking alone on my research path.

The graduation might put an end to the doctoral work, yet the progress of bridging knowledge gaps in research never stops. I will take this as a little milestone in my life. Instead of ‘taking’ from the society, it is perhaps to right time to ‘give’ back to the science community. From here, I will review what I have done and plan for the future steps!

Pak Lun Fung aka Alan
Helsinki, December 2021

Pak Lun Fung
University of Helsinki, 2022

Abstract

Air pollution is one of the biggest environmental health challenges in the world, especially in the urban regions where about 90% of the world's population lives. Black carbon (BC) has been demonstrated to play an important role in climate change, air quality and potential risk for human beings. BC has also been suggested to associate better with health effects of aerosol particles than the commonly monitored particulate matter, which does not solely originate from combustion sources. Furthermore, BC has been recommended to be included as one of the parameters in air quality index (AQI) which is communicated to citizens. However, due to financial constraints and the lack of the national legislation, BC has yet been measured in every air quality monitoring station. Therefore, some researchers developed low-cost sensors which give indicative ambient BC concentrations as an alternative. Even so, due to instrument failure or data corruption, measurements by physical sensors are not always possible and long data gaps can exist. With missing data, the data analysis of interactions between air pollutants becomes more uncertain; therefore, air quality models are needed for data gap imputation and, moreover, for sensor virtualization. To complement the current deficiency, this thesis aims to derive statistical proxies as virtual sensors to estimate BC by using the current air quality monitoring network in Helsinki metropolitan area (HMA).

To achieve this, we first characterized the ambient BC concentrations in four types of environments in HMA: traffic site (TR: 0.77–2.08 μg m^{-3}), urban background (UB: 0.51–0.53 μg m^{-3}), detached housing (DH: 0.64–0.80 μg m^{-3}) and regional background (RB: 0.27–0.28 μg m^{-3}). TR, in general, had higher BC concentrations due to the close proximity to vehicular emission but decreasing trends (–10.4 % yr^{-1}) likely thanks to the fast renewal of the city bus fleet in HMA. UB, on the other hand, had a more diverse source of BC, including biomass burning and traffic combustion. Its trend had also been decreasing, but at a smaller rate (e.g. UB1: –5.9 % yr^{-1}). We then narrowed down the dataset to a street canyon site and an urban background site for BC proxy derivation. At both sites, despite the low correlation with meteorological factors, BC correlated well with other commonly monitored air pollutant parameters by both reference instruments and low-cost sensors, such as NO$_x$ and PM$_{2.5}$. Based on this close association, we developed a statistical proxy with adaptive selection of input variables, named input-adaptive proxy (IAP). This white-box model worked better in terms of accuracy at the street canyon site (R^2 = 0.81–0.87) than the urban background site (R^2 = 0.44–0.60) because of the scarce missing gaps in data in the street canyon. When compared with other white- and black-box models, IAP is preferred because of its flexibility and architectural transparency. We further demonstrated the feasibility of sensor virtualization by using statistical proxies like IAP at both sites. We also stressed that such virtual sensors are location specific, but it might be possible to extend the models from one street canyon site to another with a calibration factor. Similarly, the proposed methodology can also be applied to estimate other air pollutant parameters with scarcity of data, such as lung deposited surface area and ultrafine particles, to complement the existing AQI.

Keywords: virtual sensor, statistical proxy, missing data, black-box, white-box

Contents

1 Introduction **11**

2 Methods **18**

2.1 Measurement sites 18

2.2 Instruments 20

2.3 Statistical proxies 21

2.3.1 White-box (WB) approach 21

2.3.2 Black-box (BB) approach 23

2.4 Trend analysis 24

2.5 Evaluation metrics 25

3 Results and discussion **27**

3.1 Characteristics of BC in HMA 27

3.2 Comparison of BC with NO_x and $PM_{2.5}$ 30

3.3 Correlation of BC with other parameters 32

3.4 Evaluation of statistical proxies involved 33

3.5 Integration of low-cost sensor (LCS) data into the statistical proxies . . 38

3.6 Feasibility and limitations of sensor virtualization 40

4 Review of papers and the author's contribution **42**

5 Conclusion and outlook **43**

References **45**

List of publications

This thesis consists of an introductory review, followed by four research articles. In the introductory part, these papers are cited according to their roman numerals.

I Luoma, K., Niemi, J. V., Aurela, M., **Fung, P. L.**, Helin, A., Hussein, T., Kangas, L., Kousa, A., Rönkkö, T., Timonen, H., Virkkula, A., & Petäjä, T. (2021). Spatiotemporal variation and trends in equivalent black carbon in the Helsinki metropolitan area in Finland, *Atmos. Chem. Phys.*, 21(2), 1173–1189. `https://doi.org/10.5194/acp-21-1173-2021`

II **Fung, P. L.**, Zaidan, M. A., Sillanpää, S., Kousa, A., Niemi, J. V., Timonen, H., Kuula, J., Saukko, E., Luoma, K., Petäjä, T., Tarkoma, S., Kulmala, M., & Hussein, T. (2020). Input-adaptive proxy for black carbon as a virtual sensor, *Sensors*, 20(1), 182. `https://doi.org/10.3390/s20010182`

III **Fung, P. L.**, Zaidan, M. A., Timonen, H., Niemi, J. V., Kousa, A., Kuula, J., Luoma, K., Tarkoma, S., Petäjä, T., Kulmala, M., & Hussein, T. (2021). Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration, *J. Aerosol Sci.*, 105694. `https://doi.org/10.1016/j.jaerosci.2020.105694`

IV Zaidan, M. A., Hossein Motlagh, N., **Fung, P. L.**, Lu, D., Timonen, H., Kuula, J., Niemi, J. V., Tarkoma, S., Petäjä, T., Kulmala, M., & Hussein, T. (2020). Intelligent Calibration and Virtual Sensing for Integrated Low-Cost Air Quality Sensors, *IEEE Sens. J.*, 20(22), 13638–13652. `https://doi.org/10.1109/JSEN.2020.3010316`

Abbreviations

$adjR^2$	Adjusted coefficient of determination
AE	Aethalometer
AQI	Air quality index
BB	Black-box (models)
BC	Black carbon
CO	Carbon monoxide
CO_2	Carbon dioxide
DH	Detached dousing (sites)
DT	Decision trees
eBC	Equivalent black carbon
EC	Elemental carbon
FMI	Finnish Meteorological Institute (Ilmatieteen laitos)
GAM	Generalized additive model
HMA	Helsinki metropolitan area
HSL	Helsinki Regional Transport Authority
HSY	Helsinki Region Environmental Services (Helsingin seudun ympäristöpalvelut-kuntayhtymä)
IAP	Input-adaptive proxy
IARC	International Agency for Research on Cancer
INECE	International Network for Environmental Compliance & Enforcement
K-S	Kolmogorov-Smirnov (test)
LASSO	Least absolute shrinkage and selection operator
LCS	Low-cost sensor
LDSA	Lung deposited surface area
LSTM	Long short-term memory
MAAP	Multi-angle absorption photometer
MAE	Mean absolute error
MLR	Multiple linear regression
NARX	Non-linear auto-regressive (models) with exogenous variables
NN	Neural network
NO_2	Nitrogen dioxide
NO_x	Nitrogen oxides
O_3	Ozone
OLS	Ordinary least squares

P	Pressure
PM	Particulate matter
PM_{10}	Particulate matter of diameter less than 10 aerodynamic micrometer
$PM_{2.5}$	Particulate matter of diameter less than 2.5 aerodynamic micrometer
PNC	Particle number concentration
r	Pearson correlation coefficient
R^2	Coefficient of determination
RB	Regional background (sites)
rBC	Refractory black carbon
RF	Random forest
RH	Relative humidity
$RMSE$	Root mean square error
SD	Standard deviation
SMEAR III	Station for Measuring Ecosystem-Atmosphere Relations III
SNN	Shallow neural network
SO_2	Sulfur dioxide
SVR	Support vector regression
Temp	Temperature
TR	Traffic (sites)
UB	Urban background (sites)
UFP	Ultrafine Particles
UHel	University of Helsinki (Helsingin Yliopisto)
VIF	Variance inflation factor
WB	White-Box (models)
WD	Wind direction
WHO	World Health Organization
WS	Wind speed

1 Introduction

Air pollution, being one of the biggest environmental health challenges in the world, has become the world's fourth-largest risk factor for premature death. Nearly one out of every ten worldwide deaths (in total of 4.5 millions) were linked to outdoor air pollution exposures in 2019 (Health Effects Institute, 2020). According to World Health Organization (WHO, 2019), about 90% of the world's population lives in urban areas where air pollution concentration exceeds safe limits at present.

Particulate matter (PM) is one of the key components determining urban air pollution. PM can be described by a combination of varying concentration (number, surface area and mass), shape and chemical composition. Atmospheric black carbon (BC) consists mostly of agglomerated sub-micron PM, and it is emitted as a by-product of incomplete combustion (Bond et al., 2013). In urban areas, the typical combustion sources are traffic and domestic wood burning (Helin et al., 2018; Rönkkö & Timonen, 2019). In the combustion process of carbon-based fuel, BC is produced in the flame, and it is then released to the atmosphere as carbon agglomerates. BC from traffic has been found at the particle size of ∼100–150 nm whereas BC from biomass combustion has been detected at ∼300 nm (Saarikoski et al., 2021). Besides, BC is capable of being transported over a long distance, and this could also contribute to the urban BC source (Järvi et al., 2008). Due to the limited atmospheric lifetime and unevenly distributed sources, atmospheric BC is characterized by large spatial and temporal variation (Bond et al., 2013). In the atmosphere, BC particles can change during the ageing process via particle growth and surface reactions (Timonen et al., 2019).

WHO (2012) pointed out that BC might not be directly toxic, but it can act as a universal carrier of other chemical components with varying toxicity which can bring severe effects on human health, for example benzo(a)pyrene that is produced along with BC during combustion (Hellén et al., 2017). BC particles of the size range of ∼100 nm can penetrate deep into the respiratory system and all the way to the alveolar region where the particles can be transported in the blood circulation system and further into the organs. Long-term exposure to BC, even at low levels (Brunekreef et al., 2021), could cause, for example, cardiopulmonary disorders, respiratory illnesses and diseases that are not related to allergies (Janssen et al., 2011). International Agency for Research on Cancer (IARC, 2014) stated that diesel exhaust particles composed of BC are classified as 2B carcinogens, and impose harm on human health and environment.

Besides potential health effects on human beings, BC also plays an important role in climate change. Due to its black appearance, BC absorbs solar radiation in the atmosphere over a large wavelength range (Bond et al., 2013; IPCC, 2013). The settling of the BC on snow or ice sheets can lower the reflectivity, thus increasing in radiation absorption and further speeding the heating and melting of the snow (Flanner et al., 2009). This emphasizes the impact of BC emissions and induced warming in the Arctic (Klimont et al., 2017). Reductions of BC emissions can, therefore, have a cooling effect, but the additional interaction of BC with clouds is uncertain, which could lead to some counteracting warming effects (IPCC, 2021). More locally, BC could lead to poor visibility and bad air quality (Novakov et al., 2003).

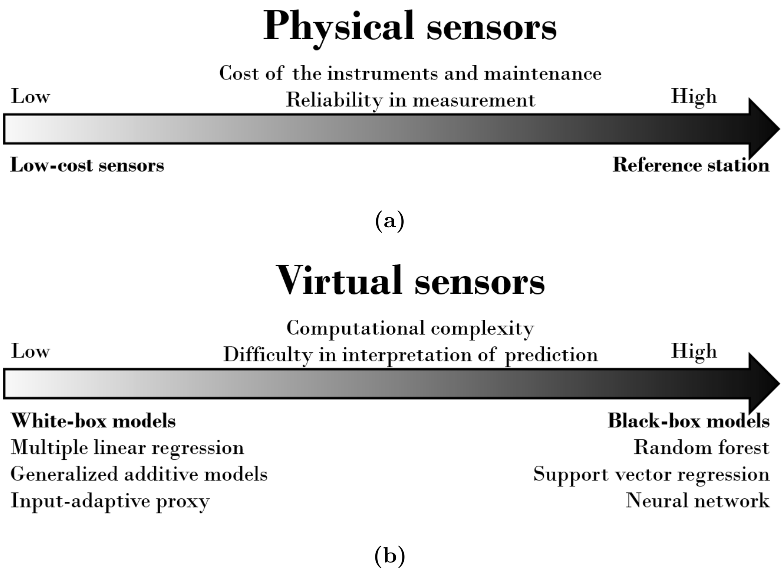
Due to its negative influences on human's health, climate change and air quality, it has been recommended to include BC alongside with the other air quality parameters, including particulate matter of diameter less than 10 and 2.5 μm (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2) and carbon monoxide (CO), for the calculation of air quality index (AQI) because BC concentration can associate better with health effects of aerosol particles than just PM (Achilleos et al., 2017), which does not solely originate from combustion sources (WHO, 2012). However, it has yet to be adopted (WHO, 2021). According to the International Network for Environmental Compliance & Enforcement (INECE, 2008), reduction of BC emissions has been considered as a cost-effective way to reduce a major cause of global warming. In order to reduce BC emissions, one of the suggestions is to implement a BC footprint similar to the footprint of carbon dioxide (CO_2) (Timonen et al., 2019). For these suggestions to be taken into action, we demand for BC measurements to be covered more extensively.

To understand the whole picture of the characteristics of urban BC and the interactions of BC with other air pollutants, a dense air quality monitoring network is important to capture their temporal and spatial variations. Currently, BC mass concentration can be estimated in at least three ways, none of which fully represent BC (Sharma et al., 2017): conversion of light absorption to give equivalent black carbon (eBC), thermal desorption of elemental carbon (EC) from weekly integrated filter samples to give EC, and measurement of incandescence from the refractory black carbon (rBC) component of individual particles using a single particle soot photometer. Yet, no uniform metrics exist for emissions, concentrations, or impacts characterization and even a precise definition of BC is missing (Timonen et al., 2019). Even using the same estimation method, the use of different correction algorithms and mass absorption cross-section values could alter the resulting BC concentrations (Luoma et al., 2021). Apart from these concerns, building a dense network based on only the reference level instruments could be expensive, attributed to not only the instruments themselves, but also the maintenance and the workforce to sustain the measurements. Currently, apart from the regulated pollutants, some other important health-concerned parameters (e.g. BC) are not necessarily measured continuously in every national station. The monitoring of these parameters is not required by the national environmental legislation in many countries (e.g. Kutzner et al., 2018).

Because of this, researchers have been looking for alternatives, and among them, low-cost sensors (LCSs) have been widely explored in the past decade for their usefulness as an additional component in the air quality monitoring system (e.g. Morawska et al., 2018). The classification of physical sensors is a spectrum depending on the cost of the instruments and the reliability in measurement (Figure 1a). In particular, LCSs, which are usually less than one tenth of the price of reference instruments (Lagerspetz et al., 2019), has been developed and deployed in monitoring networks in bulk (e.g. Caubel et al., 2019); however, the data quality of the LCSs remains a major issue that hinders the wide spread of LCS implementation in practice. Accurate field measurements from the LCSs will give a better resolution for a model in an urban setting compared to a model constructed on just laboratory results (Krecl et al., 2018). Recently, extensive research on LCSs have dedicated their studies on developing in-field sensor calibration (e.g. Concas et al., 2021) to enhance the feasibility of this alternative.

Even if an air monitoring network is well built with extensive reference instruments and LCSs, long data gaps can exist due to instrument failure or data corruption (Junger & De Leon, 2015; Zaidan et al., 2019). With missing data, the behavior of air pollution becomes more uncertain; therefore, air quality models are needed for data gap imputation and air quality prediction. It would be of great interest to develop statistical proxies-based virtual sensors to complement the current air quality monitoring. Statistical proxies are data-driven approaches, which can be broadly classified into white-box (WB) and black-box (BB) models. The classification of the two types of models is a continuum depending on the computational complexity and the ability to interpret the prediction (Figure 1b). WB models can be further classified into a few categories depending on their model structures and feature selection criteria. Simple yet apprehensible models, such as multiple linear regression (MLR, e.g. Fernández-

Guisuraga et al., 2016), generalized additive models (GAM, e.g. Järvi et al., 2009), and mix-effects models (e.g. Mikkonen et al., 2011; Fung et al., 2021a) are commonly utilized as WB models in air pollutant proxy studies. They map out the association between the explanatory variables (input variables) and the response variable (outpur variable). Apart from model structures, the criteria of selecting input variables in multi-pollutant datasets for model development have received considerable attention over the years, and a large number of feature selection methods have been proposed (Park & Klabjan, 2020) but no single feature selection method uniformly outweighs the others (Hastie et al., 2020). Examples include traditional methods like stepwise procedures (e.g. Chen et al., 2019), regularization approach (e.g. Chen et al., 2019; Šimić et al., 2020) and criterion-based procedures. The criterion-based procedures, such as best subset regression, choose the best predictor variables according to some criteria (e.g. coefficient of determination (R^2), residuals, etc). These procedures are sensitive to outliers and influential points, but involve a wider search and compare models in a preferable manner. No matter which model structures and methods of


Physical sensors

Cost of the instruments and maintenance
Reliability in measurement
Low
High
Low-cost sensors
Reference station

(a)

Virtual sensors

Computational complexity
Difficulty in interpretation of prediction
Low
High
White-box models
Multiple linear regression
Generalized additive models
Input-adaptive proxy
Black-box models
Random forest
Support vector regression
Neural network

(b)

Figure 1: A figure to explain (a) types of physical sensors in the spectrum of cost and accuracy and (b) statistical proxies as virtual sensors in the continuum depending on the computational complexity and the ability to interpret the prediction.

feature selection are used, the resulting WB models are still easy to interpret. BB models, on the other hand, refer to systems or objects which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings or underlying principles (Rudin, 2019). These include, but are not limited to, random forest (RF, e.g. Kang et al., 2018; Masih, 2019) and neural networks (NN, e.g. Cabaneros et al., 2019; Zaidan et al., 2019; Fung et al., 2021b). One could optimize the models by adjusting the hyper-parameters, for example learning rate, number of trees in RF and number nodes in NN. However, these hyper-parameters do not have any physical meanings to the output pollutant variable. BB models generally give better estimations in terms of accuracy but provide limited transparency and accountability on the results (Zaidan et al., 2019).

When the proxy is validated and proven to be reliable and accurate, the proxy in turn can be converted to a virtual sensor (Figure 2), i.e., sensor virtualization (Martin et al., 2021); in other words, converting data sources by physical sensors to virtual sensors via validated statistical proxies. The implementation of statistical proxies as virtual sensors will lower the costs by providing added value of the already measured parameters without a physical observation site upgrade, operation costs or maintenance costs (Tegen et al., 2019). Virtual sensors provide an alternative when a physical sensor cannot be

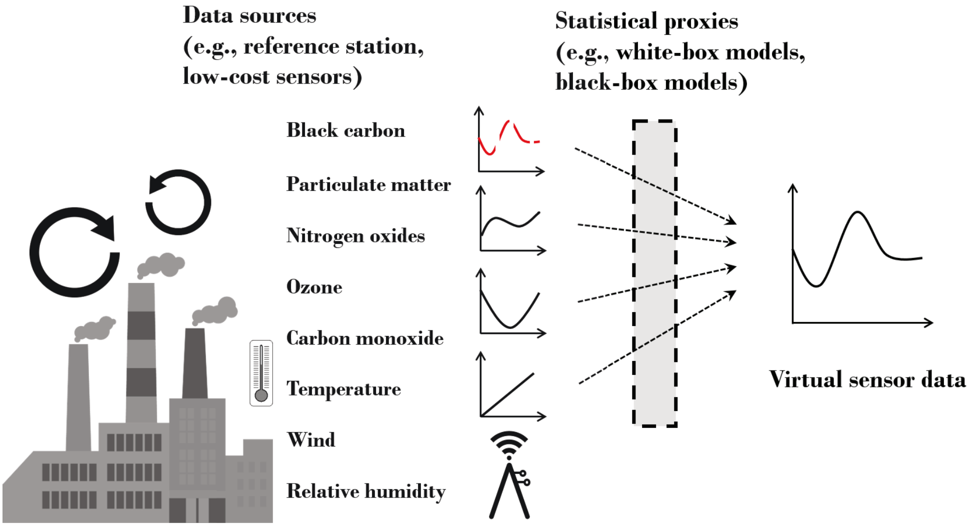


Figure 2: The process of sensor virtualization.

placed in a preferred position due to spatial conditions (e.g. lack of space for a sensor) or hostile environments (e.g. exposure to acids or extreme temperatures). The resulting delay or inaccuracy of the measurement, when installing the sensor in a less suitable spot, may be compensated by virtual sensors (Tegen et al., 2019). Furthermore, the use of virtual sensors could provisionally improve the data quality by compensating drifts of physical sensors (Albertos & Goodwin, 2002) and reducing signal noise by physical sensors (Albertos & Goodwin, 2002), which are regarded as well-known phenomenon rendering a sensor's accuracy (Baier et al., 2019). In the field of air quality monitoring, previous studies (e.g. Liu et al., 2019; Woo et al., 2016) have demonstrated the validity of virtual sensors to complement the existing physical measurements.

This thesis first presents the results of in-situ measurements of BC concentration conducted at various environments in Helsinki metropolitan area (HMA), Finland. To narrow down the investigation, two measurement sites were selected for developing statistical proxies: Mäkelänkatu measurement site (Hietikko et al., 2018) operated by Helsinki Region Environmental Services (HSY), and Station for Measuring Ecosystem-Atmosphere Relations III (SMEAR III, Järvi et al., 2009) co-operated by the University of Helsinki (UHel) and Finnish Meteorological Institute (FMI). The two sites represent the environment of street canyon and urban background, respectively. The derivation of air quality proxies in HMA has been introduced in Mølgaard et al. (2013) and Shahriyer (2020) using the reference data in SMEAR III but it did not cover the other parts of HMA. This thesis explores the developments of virtual sensors by using the measurements by different levels of physical sensors, including reference station and low-cost sensors, and implementing different statistical proxies, including WB models and BB models.

In summary, the main aims of this thesis are illustrated in Figure 3, which are to:

1. characterize the ambient black carbon concentration from the reference measurements in Helsinki metropolitan area (**Papers I–III**)
2. develop a novel statistical proxy for black carbon as a white-box virtual sensor in case of missing data with adaptive input variables (**Paper II**)
3. compare and evaluate different statistical proxies when considered as virtual sensors (**Papers II & III**)
4. explore the possibility to integrate virtual sensors into low-cost air monitoring sensors by using statistical proxies (**Paper IV**)

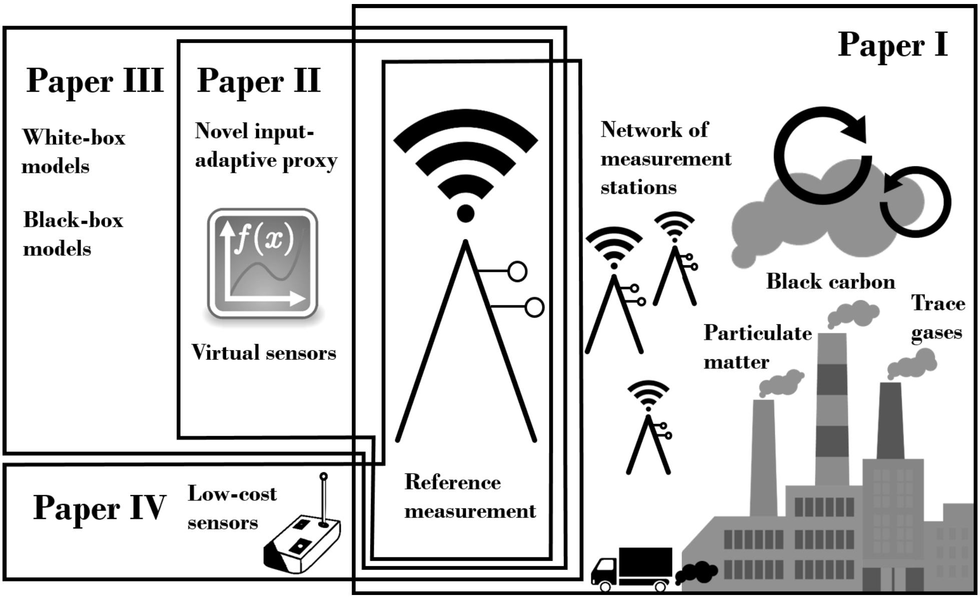


Figure 3: A figure to explain the scope of the four papers included in this thesis.

2 Methods

Paper I investigated the long-term trend of BC in HMA, Finland, thus measurements of longer period (2005–2018) were collected. Measurements of shorter period are needed for proxy development: **Papers II–III** focused on the data in two full years in 2017 and 2018 whilst **Paper IV** analyzed the data during the period of the LCS campaign from March 2018 to June 2019. The following subsections describe the measurement sites, instruments, statistical proxies and analysis involved in this thesis. A summary of the four papers is presented in the end of this section as Table 1.

2.1 Measurement sites

In **Paper I**, four types of environments were investigated in HMA: traffic sites (TR1–6, in red), urban background sites (UB1–2, in blue), detached housing sites (DH1–5, in magenta) and regional background sites (RB1–2, in green). The corresponding locations are shown in Figure 4 and the detailed description of all the sites can be found in **Paper I**. RB2 is not shown in the map because it is located about 200 km outside HMA and the location is out of scope of this thesis. Meteorological data in Pasila weather station (black diamond) was used in **Paper I**. From these measurement sites, TR2 (Figure 5a) and UB2 (Figure 5b) were selected for the development of proxies in **Papers II–IV**. They are Mäkelänkatu measurement site (Hietikko et al., 2018) operated by HSY and SMEAR III (Järvi et al., 2009) co-operated by UHel and FMI, respectively. The two sites are 500 meters away from each other, which, respectively, represent the environments of urban background and traffic sites.

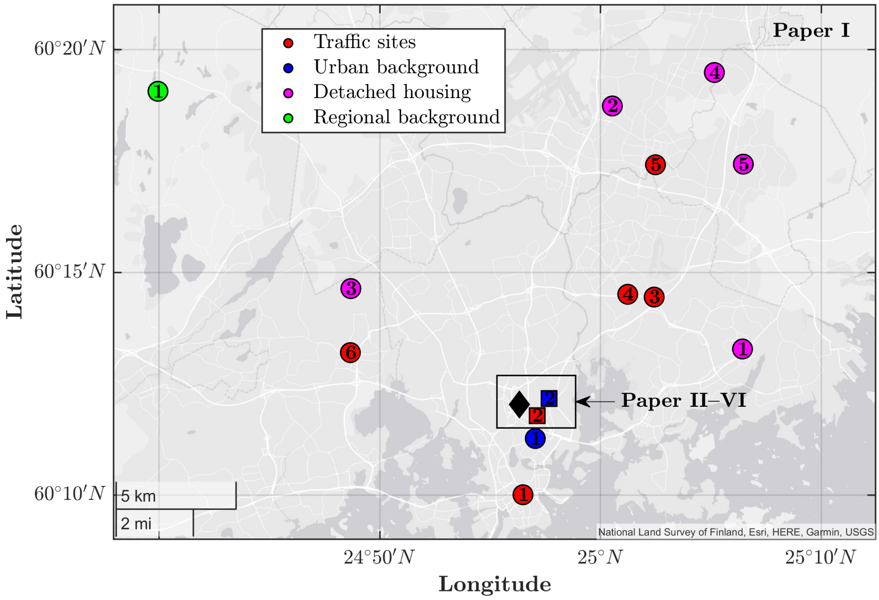
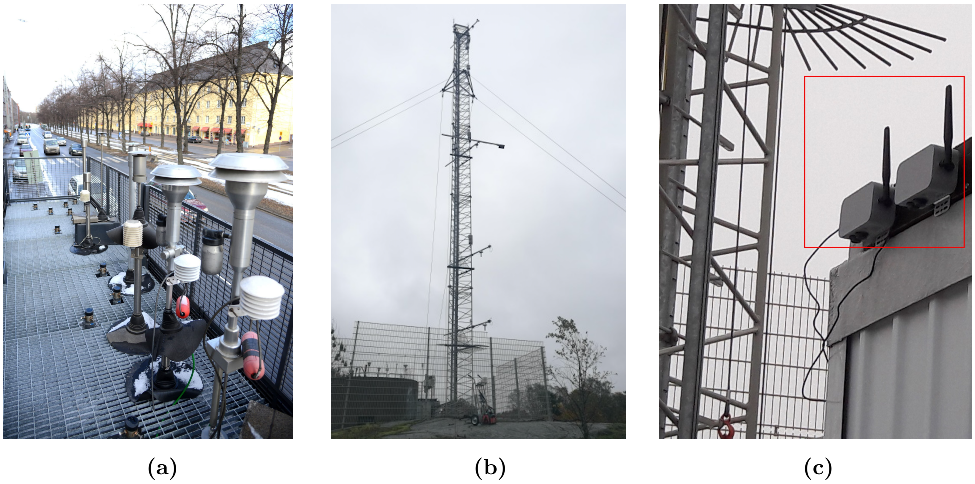


Figure 4: Location of the measurement stations in Helsinki metropolitan area. **Paper I** collected data in all the stations shown on the map (both circle and square markers). **Papers II–IV** focused on urban background UB2 site and street canyon TR2 site for statistical proxy development (square markers). Markers of red, blue, magenta and green color with corresponding site number represent the four different environments: traffic (TR) sites, urban background (UB) sites, detached housing (DH) sites and regional background (RB) sites, respectively. The black solid diamond represents the location of Pasila weather station.



(a) (b) (c)

Figure 5: The measurement tower at the urban background UB2 site is shown in (a) while the measurement container at the street canyon TR2 site is shown in (b). Data were collected from there in all the four papers. (c) shows the Clarity LCSs positioned side-by-side at the measurement tower at UB2 site. The collected data was used in the integration of LCS analysis. All the three pictures were originally presented in **Paper IV**.

2.2 Instruments

All of the measurements of BC were conducted by using a multi-angle absorption photometer (MAAP, Thermo Fisher Scientific) except at DH4, DH5 and RB2. The MAAP determines the absorption coefficient of aerosol particles at wavelength 637 nm by collecting the particles on a fiber filter and by measuring the intensity of light penetrating the filter and the intensity of light that is scattered from the filter at two different angles (Petzold & Schönlinner, 2004). The absorption coefficient was determined from these measurements by using a radiative transfer model (Petzold & Schönlinner, 2004). The BC concentrations were obtained from the absorption coefficient by using a constant mass absorption cross-section value of 6.60 m^2 g^{-1} (Petzold & Schönlinner, 2004). The other stations used Aethalometer (either AE31 or AE33, Magee Scientific), which has similar measuring principles (Luoma, 2021) but determines the absorption coefficient of aerosol particles at seven wavelengths (370, 470, 520, 590, 660, 880, and 950 nm). All BC measurements were conducted optically. The head of the sampling line was located 4 m above the ground. The concentration of BC was measured for particles smaller than 1 μm. Sample air was dried with an external dryer or by warming up the sample to 40°C at most of the stations, but at UB2, the sample air was not dried. The BC concentration was derived from the light absorption of the particles, and hence the measured BC was technically equivalent black carbon (eBC, Petzold et al., 2013), but we use 'BC' in the rest of the thesis for clarity.

For the instruments of the rest of the air pollutants measured in HMA air monitoring sites and meteorological parameters used in Pasila weather station, a detailed specification was described in **Paper I**. **Paper II** reported the meteorological measurements in TR2 and UB2 and **Paper IV** documented the specification of the Clarity LCSs (Figure 5c).

2.3 Statistical proxies

Statistical proxies of BC were developed at the street canyon TR2 site and the urban background UB2 site. They are measurement sites in HMA which measure tens of other parameters including air pollutants and meteorological conditions in a long-term basis by employing a suite of automated sensors and manual monitoring programs. The various parameters measured at the same station make statistical proxy derivation feasible.

2.3.1 White-box (WB) approach

WB models are ones of the approaches for the statistical proxy derivation in this thesis. It is easy to comprehend the influence of different input variables on the output variable and usually faster in computation (Figure 1b). This allows users to optimize the model based on the actual physical properties rather than the statistical regression loss. Due to the usually present missing data conditions, **Paper II** explored the idea of developing a novel WB model with a capability of filling up missing data, with the resulting product named input-adaptive proxy (IAP). **Paper III** evaluated and compared some other WB models, such as least absolute shrinkage and selection operator (LASSO) and decision trees (DT), with IAP. In this subsection, we explain the methods of two highlighted WB models.

The novel input-adaptive proxy (IAP) IAP, as a WB model, makes use of the model structure of multiple linear regression (MLR) with ordinary least squares (OLS) techniques, together with a variable selection method of criterion-based procedures (**Paper II**). It first examined the correlation of output variable with the input features by Pearson correlation coefficient (r). It pre-selected the most correlated input features and created sub-models with maximum three input features. The model applied an extra regularization by using Tukey's bisquare weighting function, which depends on the residuals, leverages from regression fits and the estimates of the standard deviation of the error terms, with a tuning factor 4.685 (Wang et al., 2018) as a robust fitting in case data are contaminated with outliers that often takes place in field measurements. We ran the regression and evaluated every sub-model. According to the evaluation metrics used, we ranked the sub-models based on their performance in descending order. In practice, datasets are typically incomplete; in case of missing data, some data points in the input variables in the best sub-model can possibly be missing, hence the imputation cannot be fully achieved. IAP further imputed the missing data with the second best performing sub-model, and so on, until all the voids were filled up. Some of the sub-models were subject to rejection under two conditions: (1) strong multi-collinearity among the input parameters using variance inflation factor (VIF, Kleinbaum et al., 2013) and (2) violation of the normality assumption of residuals using Kolmogorov-Smirnov (K-S) test (Steinskog et al., 2007). Based on the situation of missing data, the automated IAP searched for the best sub-model option from the ranking chart. Hence, each data point might be estimated differently depending on the data availability. Since IAP has been developed to fill up the missing data itself, no missing data imputation was required before modeling, unlike the other models we used. The full description of this model can be found in **Papers II & III**.

Least absolute shrinkage and selection operator (LASSO) As one of the example of WB models, LASSO was used and compared against IAP in **Paper III**. LASSO was first introduced by Tibshirani (1996) and later extensively used in air pollutant prediction (e.g. Van Roode et al., 2019). Like IAP, it is also a MLR but uses regularization term as variable selection method. The regularization imposes a penalty on different parameters of the model to reduce the model freedom and eliminates redundant predictor variables, yet deals poorly with multi-collinear variables. In this way, the model improves the generalization capacity. LASSO makes use of L1-norm penalty by using a geometric sequence (Tibshirani, 1996), where λ is a hyper-parameter that controls the penalty term for the strength of shrinkage. In **Paper III**, λ was optimized by obtaining the minimum mean square error in a five-folded cross-validation.

2.3.2 Black-box (BB) approach

BB models, which lie on the other end of the spectrum in Figure 1b, has been demonstrated to work better in terms of accuracy; however, they provide limited transparency and accountability regarding the outcomes (Rudin, 2019). The commonly used BB models in the field were selected. Random forest (RF), support vector regression (SVR), shallow neural network (SNN) and long short-term memory (LSTM) were used in **Paper III** while non-linear auto-regressive exogenous model (NARX) was used in **Paper IV**. The evaluation of the different WB approaches and BB approaches was performed in **Paper III**. Since most BB approaches do not accept dataset with missing data. Different interpolation techniques were performed before data analysis: linear interpolation (**Paper III**), nearest neighbor method (**Paper III**) and Akima cubic Hermite interpolation method (**Paper IV**). Below we describe some of the BB approaches used: RF, SNN and NARX.

Random forest (RF) RF is constructed on each subset by aggregating the results generated from all individual decision trees (Kang et al., 2018; Masih, 2019), as individual decision tree tends to over-fit (Yu et al., 2016). Different random subsets from the original dataset with replacement are generated. The same learning method on each sample is trained later and finally outputs of each model are simply weighted. This bootstrap-aggregated ensemble method reduces bias and error variance and improves generalization (Van Roode et al., 2019). In **Paper III**, Breiman's random forest algorithm was applied for each decision split to determine the number of input variables to be selected at random (Breiman, 1996). As the outcome was obtained by aggregating all individual trees, the importance of input features were not easily depicted as one of the characteristics of BB models. The full description of this model can be found in **Paper III**.

Shallow neural network (SNN) Artificial neural network models have been utilized in predicting air quality (e.g. Cabaneros et al., 2019; Van Roode et al., 2019; Fung et al., 2021b). The architecture of static neural networks consists of nodes which generate a signal or remain silent as activation function. The activation function in each layer determines the output value of each neuron with different weights that becomes the input values for neurons in the next hidden layer connected to it. A SNN with one hidden layer of five neurons was used in **Paper III** because Cabaneros et al. (2019) suggested that such SNN can fit any finite input-output mapping problem for non-linear relationship.

Non-linear auto-regressive exogenous model (NARX) Besides static NN, recurrent dynamic networks, such as NARX, model the dynamic of a variable (time-series) depending on its past values and on the current and past values of external driving input (exogenous inputs) (Esposito et al., 2016). NARX contains feedback connections that affect several layers of the network and this allows mapping non-linear relationship in time-series datasets (Lin et al., 1996). In **Paper IV**, the best NARX architectures were determined through grid search, and Bayesian regularization backpropagation were used for NARX parameters' estimation to ensure the model generalization and avoid over-fitting.

2.4 Trend analysis

In order to investigate the trends of the long-term air monitoring measurements in **Paper I**, we used the Mann-Kendall test and Sen's slope estimator, which are non-parametric statistical methods allowing missing data points based on Gilbert (1987). They are used widely in analyzing environmental data (e.g. Collaud Coen et al., 2020). The Mann-Kendall test quantifies whether a homogeneous long-term trend in a studied variable is statistically significant and whether it is monotonically decreasing or increasing. Sen's slope estimator estimates the magnitude of the trend. A seasonal version of Mann-Kendall test and Sen's slope estimator were applied in **Paper I** that overcame the auto-correlation problem related to the cyclic pattern of the data, for example seasonal variations, weekend effects and diurnal cycles, due to for example boundary layer dynamics or traffic rates. We used monthly median values in the trend analysis. Valid data of at least 14 days for each month were required; otherwise, the month was not taken into account in the trend analysis. Trend analysis were only applied to TR1, TR2, RB2, and UB1 in 2015–2018. Even though there were 4 years of measurements at UB2, the station was omitted from the trend analysis since the data availability at UB2 in 2016–2017 was not enough. No DH sites had data longer than one year for trend analysis.

2.5 Evaluation metrics

We calculated Pearson correlation coefficient (r, **Papers I–IV**) to investigate the linear correlation between the response variable and other explanatory variables by:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}\,, \tag{1}$$

where $\overline{x}$ is the arithmetic mean of the output variable and $\overline{y}$ is the arithmetic mean of each input variable. N is the number of valid data points in the variables of x and y. Positive r indicates positive correlation while negative r implies negative correlation. The absolute values of r ($|r|$) ranging from 0 to 1 quantify the degree of correlation from the lowest to the highest.

In order to evaluate the model performance quantitatively, coefficient of determination (R^2, **Papers III & IV**) with its variant adjusted R^2 ($adjR^2$, **Paper II**), together with mean absolute error (MAE, **Papers II–IV**) and root mean square error ($RMSE$, **Papers II–IV**), were used as diagnostic evaluation attributes, as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}\,, \tag{2}$$

$$adjR^2 = 1 - (1 - R^2) \times (\frac{N-1}{N-p-1})\,, \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|\,, \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}\,, \tag{5}$$

where y_i and $\widehat{y_i}$ are i^{th} measured and i^{th} estimated output variable by the model, respectively, and $\overline{y}$ is the expected value of the measured valuable. N is the number of complete data input to the model. Both R^2 and $adjR^2$ illustrate the linear association between two variables (the estimated output variable by proxy and the measured variable), i.e. a measure of how close the data lie to the fitted regression line. $adjR^2$ differs from R^2 in the way that $adjR^2$ considers also the degree of freedom, and adjusts the number of input terms in a model relative to the number of data points. However, neither of R^2 and $adjR^2$ consider the biases in the estimation. Therefore, we further validated the models with MAE and $RMSE$ where MAE measures the arithmetic mean of the absolute differences between the members of each pair and $RMSE$ calculates the square root of the average squared difference between the estimation and the measurement pairs. $RMSE$ is more sensitive to larger errors than MAE.

Table 1: A summary of the measurement period (column 2), measurement location (column 3), physical instruments of BC (column 4) and the statistical proxies used (column 4) in this thesis.

	Measurement period	Location	Physical sensors	Statistical proxies for BC
Paper I	2005–2018	TR1–6, UB1–2, DH1–5, RB1–2	MAAP, AE-31, AE-33	-
Paper II	2017–2018	TR2, UB2	MAAP	IAP
Paper III	2017–2018	TR2, UB2	MAAP	WB models: IAP, LASSO; BB models: RF, SVR, SNN, LSTM
Paper IV	2018–2019	TR2, UB2	MAAP	NARX

3 Results and discussion

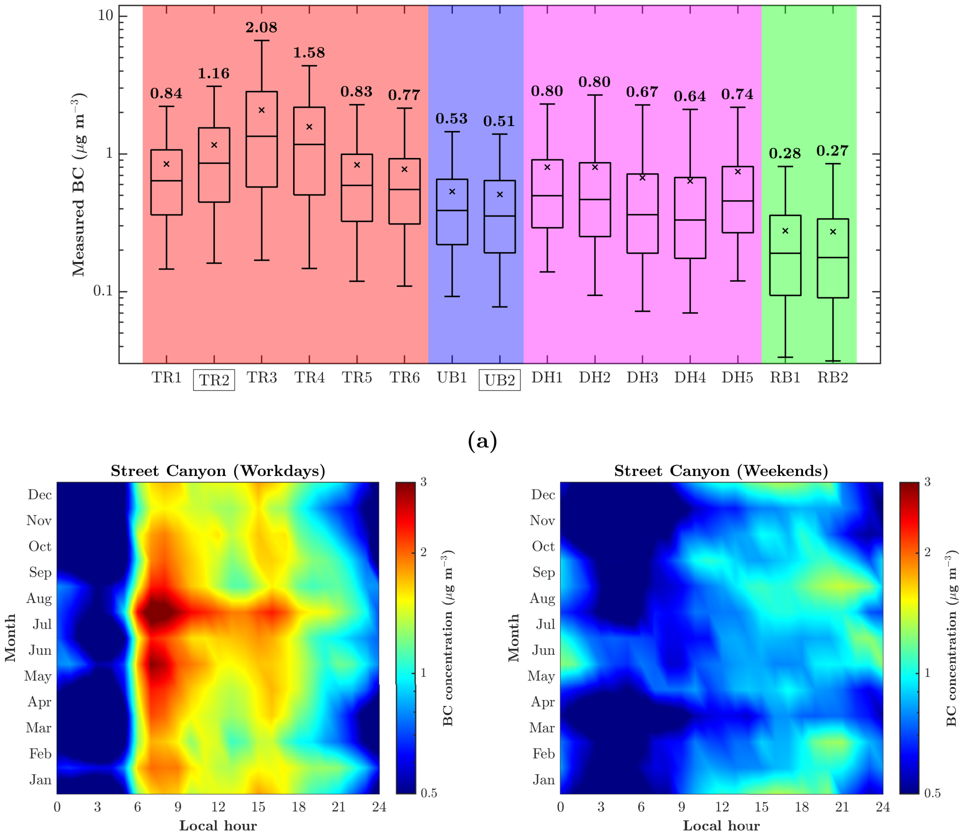
Combining the four papers in this thesis, we answer **aim 1** of the thesis by reporting the spatial and temporal ambient BC concentration from the reference measurements in HMA in Section 3.1. We also compare BC with other commonly monitored pollutant parameters, NO_x and $PM_{2.5}$, and demonstrate the correlation of BC and other parameters in Section 3.2 and Section 3.3, respectively. The demonstrated correlation could serve as a good ground for statistical proxy development and the evaluation of the statistical proxies are presented in Section 3.4, addressing **aim 2** and **3** in the thesis. We further respond to **aim 4** by exploring the possibility to integrate virtual sensors into low-cost air monitoring sensors by using statistical proxies in Section 3.5, and finally in Section 3.6, we investigate the feasibility of sensor virtualization using the studied statistical proxies.

3.1 Characteristics of BC in HMA

This subsection describes the spatial and temporal statistical diagnostics and trend analysis of the BC measurements in all measurement sites (**Papers I–III**), as well as the characterization of BC concentration at UB2 and TR2 in terms of its seasonal variation, weekend effect and diurnal cycle (**Papers II & III**).

The spatial BC concentrations were reported in **Paper I** and it is shown in Figure 6a. The highest mean BC concentrations over the year 2005–2018 were observed at the TR sites (0.77–2.08 μg m^{-3}) while DH sites had the second highest among the four environments (0.64–0.80 μg m^{-3}). The BC concentrations at TR3 and TR4 were clearly higher than the other TR sites as they both are located close to a heavily trafficked highway Ring I in HMA. At the UB sites, the mean BC concentrations were around 0.52 μg m^{-3}, which was clearly lower than at the TR and DH sites. The lowest mean BC concentrations ($\sim$0.27 μg m^{-3}) were observed at the RB sites that had no local BC sources in the vicinity.

A statistically significant (p-value$<$0.05) decreasing trend was observed for all of the stations included in the trend analysis (TR1, TR2, UB1, and RB2). The smallest absolute decrease was observed at the background stations UB1 and RB2, for which the slopes of the trends were –0.02 and –0.01 μg m^{-3} yr^{-1}, respectively. At TR1,



(a)

(b)

Figure 6: (a) A boxplot showing the statistical distribution of BC concentration measured at different locations in HMA for years 2005–2017. Shaded regions represent the characterization of the four environments: traffic sites (TR, red), urban background (UB, blue), detached housing (DH, magenta) and regional background (RB, green). The tick labels TR2 and UB2 are boxed to indicate that they were used in the proxy derivation. The horizontal line in the box and the cross represent the median and arithmetic mean at the station, respectively. The lower and upper lines of the box represent the 25^{th} and 75^{th} percentiles, respectively, while the lower and upper whiskers represent the 10^{th} and 90^{th} percentiles, respectively. (b) Contour plots showing the average monthly diurnal pattern in the street canyon TR2 site for workdays (left panel) and weekends (right panel) with x-axis the local hour in HMA, y-axis the month of a year and logarithm color scale the averaged BC concentrations.

the concentration decreased more rapidly by –0.04 μg m^{-3} yr^{-1}, and the decrease was even greater at TR2 (–0.09 μg m^{-3} yr^{-1}). In addition to the absolute trend, we also determined the relative trends by dividing the absolute slope of the trend by the overall median concentration. At TR1, UB1, and RB2, the relative trends were rather similar: –6.4 % yr^{-1}, –5.9 % yr^{-1}, and –6.3 % yr^{-1}, respectively. At TR2, the decrease was relatively steeper: –10.4 % yr^{-1}. Similar decreasing trend can be observed in other countries, for example in Germany (Kutzner et al., 2018), China (Guo et al., 2020) and the United Kingdom (Ciupek et al., 2021).

For the derivation of BC statistical proxies, we narrowed down the measurement period to 2017–2018 and all the measurements sites to TR2 and UB2, which served as good representatives of the environments of street canyon and urban background, as demonstrated in many previous studies (e.g. Järvi et al., 2009; Hietikko et al., 2018). Over the year 2017–2018, the data coverage of BC was 99% and 70% at TR2 and UB2, respectively. The low data coverage further motivates the needs for the development of statistical proxies as virtual sensors in long run. In agreement with the results of the other TR and UB stations, the BC concentration at the street canyon TR2 site (1.03±0.88 μg m^{-3}) was almost twice as high as that at the urban background UB2 site (0.47±0.46 μg m^{-3}). This is because TR2 is in proximity of a heavily-trafficked road while UB2 measured the source from residential area situated 50 meters to the north and roads with moderate traffic separated by a forest. The long distance from the source could increase dilution rate of BC, and hence lower the BC levels.

When considering the seasonal variation (Figure 6b), BC concentration in the winter appeared to be the highest (0.65±0.6 μg m^{-3}) at UB2 owing to lower mixing height (Teinilä et al., 2019) and elevated wood combustion by domestic heating (Hellén et al., 2017). At street canyon TR2 site, BC concentration in the summer appeared to be only 10% higher than the other seasons (**Paper III**). The small variation agreed with Helin et al. (2018) who observed the lack of BC seasonal variability in traffic environments. Moreover, Teinilä et al. (2019) suggested that the seasonal changing mixing height did not show correlation with the dilution of local pollution in the street canyon.

At both sites, the diurnal cycle was bimodal on weekdays due to the elevated traffic counts during working peak hours at 7–9 a.m. and at 4–6 p.m. (local hour, UTC+2 in the winter and UTC+3 in the summer) in all months as typically observed in HMA (Timonen et al., 2014). The evening peak during workdays was smaller than the morning peak because of the higher mixing height in the evening (Teinilä et al., 2019), which diluted the BC pollutants from the surface. In addition, there was a clear seasonal variation in diurnal pattern in the street canyon site. Due to low boundary layer height in summer mornings, BC concentrations were higher in summer mornings than in winter mornings (**Paper I**). During weekends, only one peak was observed in the late evening. The evening peak during weekends appears at 5–8 p.m. in the wintertime and at 8–10 p.m. in the summertime. The boosted nocturnal BC concentrations might be attributed to the increasing traffic rates along the daytime, reaching a peak approaching sunset when residents in the city return home (**Paper II**).

3.2 Comparison of BC with NO_x and $PM_{2.5}$

BC, NO_x and $PM_{2.5}$ are considered as one of the traffic-related pollutants in urban area; therefore, it is of great interest to compare BC with NO_x and $PM_{2.5}$ and to see whether BC can be estimated by them. Similar as BC concentration, NO_x and $PM_{2.5}$ concentrations were higher at TR2 than that at UB2. The difference in mean NO_x concentration was very distinctive, almost five times higher at TR2 (68.1±65.1 μg m^{-3}) than that at UB2 (11.8±16.6 μg m^{-3}). On the contrary, the environmental separation of the $PM_{2.5}$ concentration was not as clear as in BC and NO_x. The mean $PM_{2.5}$ concentrations at TR2 and UB2 were comparable, 7.04±4.58 μg m^{-3} and 5.37±6.86 μg m^{-3}, respectively (**Paper II**). This unclear environmental separation could be due to the more diverse source of $PM_{2.5}$ in urban area (e.g. Karagulian et al., 2015).

To see how the decrease in BC concentrations compared to the trends in other air pollutants, we conducted the trend analysis also for the $PM_{2.5}$ and NO_x data (Table 2). The only parameter for which we did not observe a statistically significant decreasing trend was $PM_{2.5}$ at TR2 (p-value>0.05). While the trend in BC concentration was –10.6% yr^{-1}, the trends in NO_x and $PM_{2.5}$ concentration at TR2 were –19.7% yr^{-1} and –7.1% yr^{-1}, respectively. The concentrations of BC and NO_x decreased relatively faster than the concentration of $PM_{2.5}$ (**Paper I**).

Since the pollutant emissions from the traffic sources generally decreased, it clearly affected the trends in BC and NO_x, and, to a lesser extent, the trend in $PM_{2.5}$. This difference in extent is because $PM_{2.5}$ is not as sensitive to changes in primary traffic-related emissions as BC and NO_x which are originated from local traffic sources. Similar results were also found in Krecl et al. (2017). While the decrease in local traffic emissions seem to be one of the probable explanations for the decreasing trends, the changes in the long-range transported pollution also influenced the trends. Statistically significant trends were observed also at RB2, indicating that the long-range transported pollution and the emissions in the regional area had also decreased.

Table 2: Statistical diagnostics and results of trend analysis of the measurements of BC, NO_x and $PM_{2.5}$ at the selected stations: street canyon TR2 site and urban background UB2 site. The statistical diagnostics (all in μg m^{-3}) were conducted in 2017–2018, including arithmetic mean (Mean), standard deviation (SD), 25th percentile (P25), median (P50) and 75th percentile (P75), which are shown in column 3–7, respectively. Data coverage (%) in these two years is presented in column 8. Column 9–10 show the absolute trend (μg m^{-3} yr^{-1}) and relative trend (% yr^{-1}), respectively, included in the trend analysis conducted in 2015–2018.

Location	Variable	Mean	SD	P25	P50	P75	Data %	Absolute trend	Relative trend
Street	BC	1.03	0.88	0.43	0.79	1.37	99	–0.09	–10.4
canyon	NO_x	68.1	65.1	23.8	47.2	92.5	99	–11.0	–19.7
TR2	$PM_{2.5}$	7.04	4.58	3.99	6.00	8.83	97	–0.46*	–7.10*
Urban	BC	0.47	0.46	0.18	0.34	0.60	70	-	-
background	NO_x	11.8	16.6	3.30	7.07	14.1	83	-	-
UB2	$PM_{2.5}$	5.37	6.86	1.96	3.69	6.94	42	-	-

* p-value>0.05, not statistically significant.

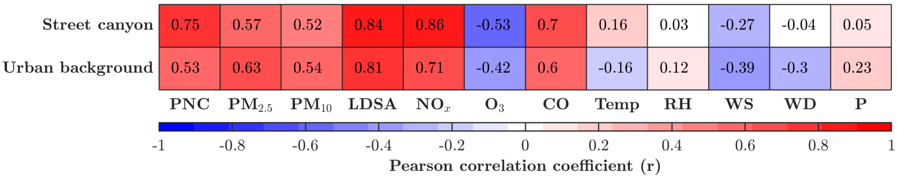
At TR2, NO_x seemed to decrease at a doubled rate compared to BC, which could be caused, for example, by the fast renewal of the city bus fleet. According to the Helsinki Regional Transport Authority (HSL), in 2015, about 17% of the HSL buses were Euro VI, and in 2018, the fraction had increased to about 50%. A study by Järvinen et al. (2019) showed that moving to Euro VI buses from enhanced environmentally friendly vehicles efficiently decreases the NO_x emissions regardless the types of fuels (Aakko-Saksa et al., 2020). Although NO_x emission control has been successful, urban roadside NO_2 concentrations have not decreased as expected. The increase in diesel cars and their potentially high primary NO_2 emissions might have been one factor in slowing down the decline of concentrations (Anttila, 2020). Some other factors like the overall change in traffic density and traffic policy could also potentially influence the concentrations of the traffic-related pollutants. Moreover, at TR2, the short time-series cause uncertainties in the trends, and, for example, the year-to-year variability caused by the meteorological conditions could lead to the apparent decrease in pollutant concentrations (**Paper I**).

3.3 Correlation of BC with other parameters

This subsection describes the correlation of BC and other air pollutants in HMA, in particular at the two highlighted measurement sites. In order to find out the most linearly correlated terms with BC, we computed their overall Pearson correlation coefficient (r). If BC has a high correlation with some of the other parameters, it provides a good ground to develop statistical proxies to estimate BC based on the highly correlated parameters.

Figure 7 illustrates the correlation of BC and other air pollutants at the street canyon TR2 and the urban background UB2 site. The hue of each box represents the degree of correlation and the color differentiate whether the correlation is positive (red) or negative (blue). The logarithm of BC was highly correlated with part of the gaseous and aerosol variables in both sites. The correlations were high at the street canyon TR2 site, in particular with lung deposited surface area (LDSA), NO_2 and particle number concentration (PNC) (r = 0.75–0.85). We also notice that the correlation on workdays was much higher than that on weekends (**Paper II**). It is due to the fact that a great portion of NO_2 and the number of particles also come from traffic as BC emissions do (Helin et al., 2018) in proximity of a heavily-trafficked road. At urban background UB2 site, the correlations of BC with aerosol and gaseous compounds were generally slightly lower (**Paper III**). No clear discrepancies between workdays and weekends were observed. However, in winter and spring, the correlation of BC with LDSA was unexpectedly high (**Paper II**). These high correlations provided a basis for proxy development based on other existing measurements.

In alignment with previous studies (e.g. Hussein et al., 2004; Teinilä et al., 2019), the linear association of BC and meteorological data remained low in both sites. Wind speed ranked top ($|r|$ = 0.27–0.39) at both TR2 and UB2 sites among the other meteorological variables, which were also suggested in previous studies (e.g. Järvi et al., 2008; Teinilä et al., 2019). At TR2, RH, P and WD had very low correlation with BC concentration ($|r| < 0.1$). On the other hand, the correlation of these meteorological variables with BC concentration appeared to be higher ($|r|$ = 0.12–0.30). No clear discrepancies between workdays and weekends were observed in the urban background environment (**Paper II**). By running trend analysis to factors like WS and Temp, it is likely that the decreasing trends in the BC concentration cannot be explained by the local meteorology (**Paper I**). Despite the low correlation of BC with meteorological data, we still included them as input variables because the data availability of meteorological data is usually higher and they are useful when the other measurements happen to be suspended.



	PNC	$PM_{2.5}$	PM_{10}	LDSA	NO_x	O_3	CO	Temp	RH	WS	WD	P
Street canyon	0.75	0.57	0.52	0.84	0.86	-0.53	0.7	0.16	0.03	-0.27	-0.04	0.05
Urban background	0.53	0.63	0.54	0.81	0.71	-0.42	0.6	-0.16	0.12	-0.39	-0.3	0.23

Figure 7: A heatmap showing the correlation of BC and other parameters measured at urban background UB2 site (upper row) and at street canyon TR2 site (lower row). The hue of each box represents the degree of correlation and the color differentiate whether the correlation is positive (red) or negative (blue). The number in each box calculates the Pearson correlation coefficient (r). Note that all the aerosol and gaseous parameters including BC are in logarithm.

3.4 Evaluation of statistical proxies involved

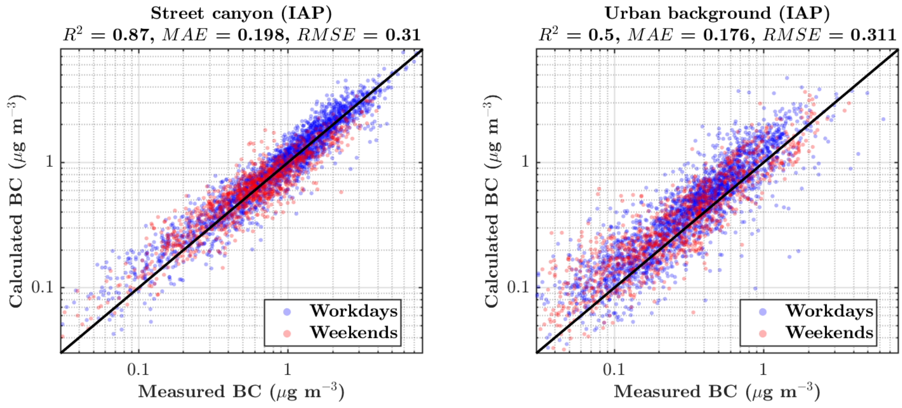
This subsection describes the performance of the statistical proxies we used for estimating BC in terms of accuracy and other aspects, such as flexibility, complexity and efficiency. We compared input-adaptive proxy (IAP, **Papers II & III**) and least absolute shrinkage and selection operator (LASSO, **Paper III**) as white-box (WB) models. In addition, random forest (RF, **Paper III**), support vector regression (SVR, **Paper III**), shallow neural network (SNN, **Paper III**), long short-term memory (LSTM, **Paper III**) and non-linear auto-regressive network with exogenous inputs (NARX, **Paper IV**) were also evaluated as black-box (BB) models.

Except for NARX, which were developed with a different timeframe, R^2 for all the WB and BB models were higher than 0.8 (Table 3, WB: R^2 = 0.81–0.87; BB: R^2 = 0.86–0.87) at the street canyon TR2 site. As demonstrated in scatter plot Figure 8a as one of the examples, the data points were more concentrated along the 1:1 line, which indicates that the models performed well in terms of accuracy. In spite of a higher R^2, the fitting for RF and SVR diverged from the 1:1 line at the lower tail (**Paper III**).

Table 3: An evaluation table of all statistical proxies used for estimating BC models at the urban background UB2 site and the street canyon TR2 site. The evaluation matrix includes R^2, MAE (μg m^{-3}) and $RMSE$ (μg m^{-3}). The statistical proxies are classified as white-box (WB) and black-box (BB) models. The measurement period is January 2017–December 2018, unless stated otherwise.

Location	Evaluation metrics	Statistical proxies							
		WB models		BB models					
		IAP	LASSO	RF	SVR	SNN	LSTM	NARX1 (LCSs)*	NARX2 (LCSs)*
Street canyon TR2	R^2	0.87	0.84	0.87	0.86	0.87	0.87	0.90	0.86
	MAE	0.199	0.200	0.190	0.197	0.191	0.181	0.141	0.202
	$RMSE$	0.310	0.342	0.303	0.321	0.308	0.300	0.289	0.342
Urban background UB2	R^2	0.50	0.60	0.55	0.41	0.49	0.64	0.81	0.79
	MAE	0.176	0.141	0.146	0.166	0.157	0.151	0.219	0.248
	$RMSE$	0.311	0.278	0.297	0.339	0.317	0.265	0.406	0.427

* The data collection period was shorter (March 2018–June 2019) because of the LCS campaign.
NARX1 has input variables of Temp, RH and calibrated $PM_{2.5}$.
NARX2 has input variables of Temp, RH, calibrated $PM_{2.5}$ and modeled CO_2.



(a)

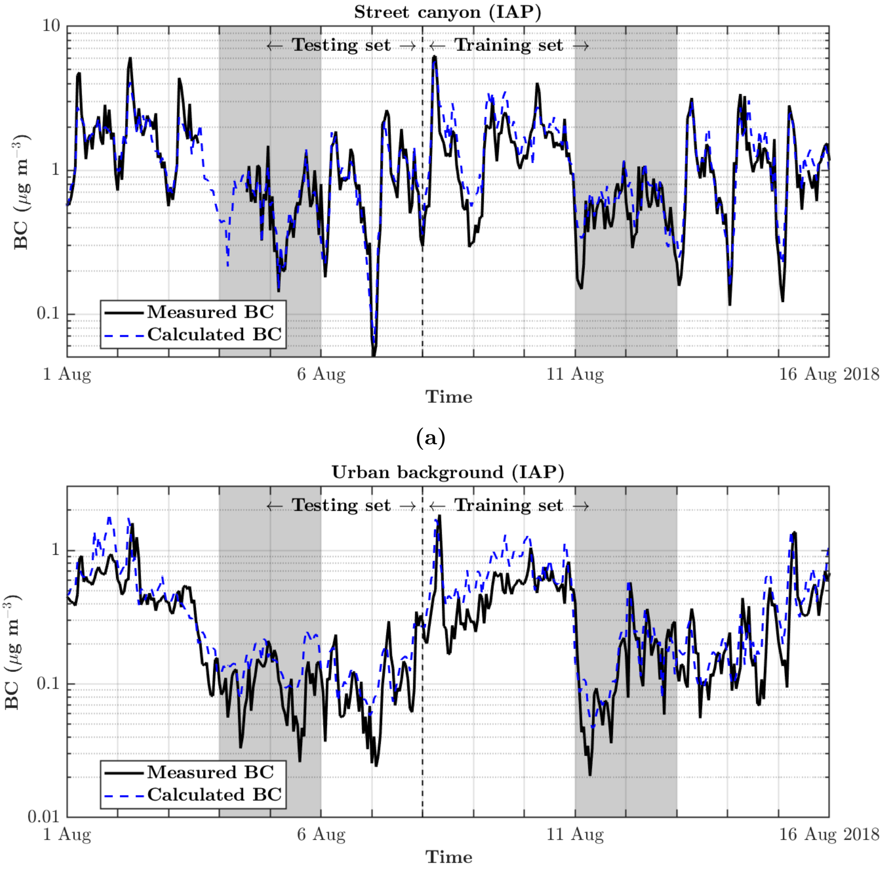
(b)

Figure 8: Scatter plots in logarithm scale of calculated BC concentration (μg m^{-3}) by using statistical proxy IAP against measured BC concentration (μg m^{-3}) by reference instrument MAAP at (a) urban background UB2 site and (b) street canyon TR2 site. Blue and red circles represent data points for workdays and weekends, respectively. The corresponding evaluation metrics R^2, MAE and $RMSE$ are shown as the sub-titles in the figure.

The model performance at the urban background UB2 site was generally worse (Table 3, WB: $R^2 = 0.44$–0.60; BB: $R^2 = 0.41$–0.64) and showed more marked variation among all WB and BB models than at the street canyon TR2 site. LSTM performed best ($R^2 = 0.64$) while SVR had the lowest R^2 ($R^2 = 0.41$). LASSO, RF, SNN and SVR tended to underestimate the extreme values and the fitting slope was less than 1 (**Paper III**). In addition, IAP appeared to overestimate all data points generally for both workdays and weekends (Figure 8b).

In general, WB and BB models performed similarly in terms of accuracy which is in alignment with previous studies (e.g. Zaidan et al., 2019). When considering individual models, LSTM performed quite well in all cases because it treated also data from the previous time-step to cope with the time dependency. SNN and LSTM models had resembling architecture where SNN had only one hidden layer and did not consider short-term memory. SNN turned out to perform generally worse than LSTM. IAP and SVR had fair fitting accuracy, but IAP is still recommended because it managed to fill in missing data by other input variables without extra interpolation. This can be shown in the BC time-series at street canyon TR2 site in Figure 9a where missing gap was found on 4 August 2018 while IAP can still give estimation on that day.

The difference in the overall performance at the urban background UB2 site and the street canyon TR2 site indicated that the models are location specific because of their different pollutant sources with location specific dynamics. The street canyon has been suggested to have a relatively constant BC source (Helin et al., 2018), so the overall regression performance at the street canyon site was better than that at the urban background where the temporal variation was large attributed to the various BC sources, such as local traffic and residential wood combustion (Saarikoski et al., 2021) as well as long-range transported pollutants (Järvi et al., 2008). The model performance also depended on the missing data patterns, such as completeness of the dataset and the length of the data gap, suggested by Junninen et al. (2004). A sensitivity analysis study of missing data on these statistical proxies can be done in the future so as to determine the threshold of up to which level (missing data percentage) the statistical proxies can fill in observational gaps with good performance while remaining still robust.



(a)

(b)

Figure 9: Time-series of the measured BC concentration (μg m^{-3}) by reference instrument MAAP (black solid line) and the calculated BC concentration (μg m^{-3}) by using statistical proxy IAP (blue dashed line) at (a) the street canyon TR2 site and (b) the urban background UB2 site, in a selected period 1–16 August 2018. The vertical dotted lines separate the testing set and the training set. The shaded regions represent the weekends in the selected period.

Apart from the model accuracy presented in Table 3, **Paper III** evaluated the models by other aspects, such as flexibility, complexity and efficiency. The architecture of the WB models is transparent, which allows users to understand and modify, if necessary, the influencing variables (Rybarczyk & Zalakeviciute, 2018). Among all WB models, IAP takes maximum three input predictors in an adaptive way, while the number of input predictors for LASSO depends on lambda. For BB models, the inner components or logic are inaccessible. In case of anomalies, it is difficult to inspect and locate the problems from the model structure (Rybarczyk & Zalakeviciute, 2018).

Some BB models, for example SNN and LSTM, have many hyper-parameters, which require extensive efforts for model optimization (Esposito et al., 2016) and fail to show any physical meanings to the output variable. In terms of efficiency, LASSO and RF required the least computational resources among all the models we tested in this study. SVR and SNN took 30–50% longer time for the modeling process. Since IAP had to search for the best combination of input predictors and fill in missing data, its computational time could be five times higher than the most efficient ones. LSTM took the longest computation time, still under 10 min (**Paper III**).

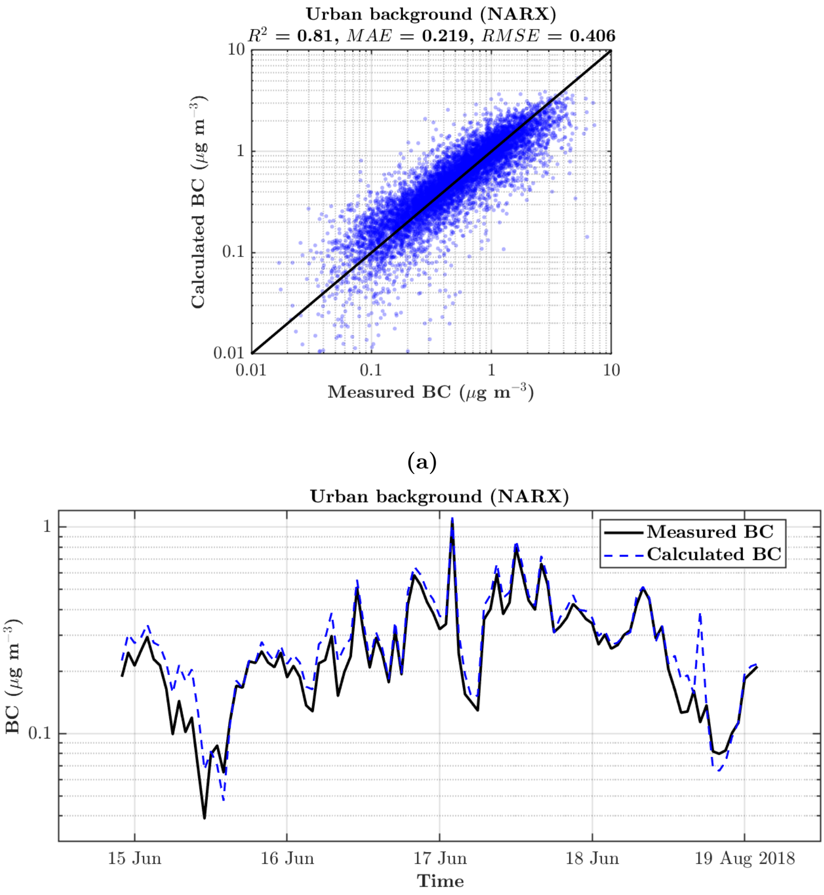
Combining all the evaluation criteria, we deduced that IAP is the best model despite the slightly higher computational time. It had relatively high accuracy at both sites. The model structure is transparent and the inner components can be inspected easily. The flexibility to select input predictors is also one of the merits when using this model. The second best model is LASSO since it performed evenly in both locations and could, therefore, be considered the most generalized model. Its fast modeling process and transparent model structure also enable it to stand out from the rest of the models.

3.5 Integration of low-cost sensor (LCS) data into the statistical proxies

This subsection investigates the feasibility of the integration of LCS data (**Paper IV**). Before moving on to the actual statistical proxy derivation, raw LCS data have to be calibrated against reference stations. In **Paper IV**, the LCS data were collected through Clarity LCSs during a campaign in March 2018–June 2019 at TR2 and UB2 site. Since weather parameters measured by LCSs had linear relationship with those measured by the reference instruments, temperature and relative humidity were calibrated using dynamic linear models. However, for other aerosol parameters, such as $PM_{2.5}$, due to the non-linear relationship, it required calibration using models that are more complex (Concas et al., 2021). Through performing sensitivity analysis, NARX was found to be the best calibration model. This non-linear model was then generalized by training it with datasets from the street canyon TR2 site and the urban background UB2 site simultaneously. After calibrating the LCSs with reference station, two BC statistical proxies using NARX structure were developed. One used temperature, relative humidity and calibrated $PM_{2.5}$ as input variables. The other used the same input variables plus CO_2 that was measured and estimated using the same LCSs.

The first BC model showed high R^2 and low MAE for both sites (Table 3, TR2: $R^2 = 0.81$, $MAE = 0.219$ μg m^{-3}; UB2: $R^2 = 0.90$, $MAE = 0.141$ μg m^{-3}). The second model showed worse performance even though it contained one more input variable because it propagated the modeling errors of LCS calibration and virtual sensors (Table 3, TR2: $R^2 = 0.79$, $MAE = 0.248$ μg m^{-3}; UB2: $R^2 = 0.86$, $MAE = 0.202$ μg m^{-3}).

Figure 10a shows a scatter plot between the measured BC by reference instruments and calculated BC by the proposed model at the urban background UB2 site. The results for the BC model were in agreement with the 1:1 reference line. Figure 10b illustrates time-series plots for BC by the reference instrument and the proposed model at the urban background UB2 site in the period of 15–19 June 2018. This further demonstrates that the statistical proxy for estimating BC tracked well the diurnal cycle of the measurement of BC concentrations obtained from reference instruments. The performance of this non-linear BC model integrated with LCS data is satisfactory in terms of accuracy.



(a)

(b)

Figure 10: (a) A scatter plot and (b) a time-series of the measured BC (black solid line) and the calculated BC with the use of LCSs using NARX model (blue dashed line) at the urban background UB2 in the period of 15–19 June 2018.

3.6 Feasibility and limitations of sensor virtualization

This subsection discusses the feasibility and limitations of converting the signals from physical sensors to virtual sensors through statistical proxies, i.e. sensor virtualization (Figure 2). In the previous section 3.4, we deduced that IAP and LASSO are the best two statistical proxies (**Paper III**); as a result, only these two proxies are considered in this section. Provided that adequate data is used to train the model in advance, virtual measurements can be given continuously with fairly high accuracy at both the urban background and street canyon sites. In case of physical sensor failure of some variables, models that take all the input variables as in the training process might not work. IAP and LASSO, nevertheless, still manage to estimate BC by fewer input variables. In extreme situations where only meteorological parameters are available, IAP is still able to work thanks to its input adaptability. Although air pollutant virtual sensors have been proven to work satisfactorily, and may assimilate a wide variety of information, it is still advisable to maintain original observational data that represents actual conditions for validation purposes (Hagler et al., 2018).

However, there are limitations in using BC models as virtual sensors (**Paper IV**). In this thesis, the two measurement sites can be characterized as street canyon and urban background. Each characterization only represents places with similar air pollutant sources; therefore, the statistical proxies trained by data measured in one environment might not work in another. Despite this limitation, it might be possible to extend the proxies from one site to another with a calibration factor. Multiple models created on measurements from different types of environments might be necessary (Jha et al., 2021). If there is enough data available, with the use of reference instruments, LCSs, and even portable sensors with citizens' involvement (e.g. Rebeiro-Hargrave et al., 2021; Kortoçi et al., 2022), from different locations not limited to street canyon and urban background sites only, a network of database can be established. Within the network, selection of input variables can be interchanged. Although virtual sensors could be a good alternative when physical sensors cannot be placed in a preferred location, extra data post-processing has to be carried out (Jha et al., 2021). This is because the calibration models and the estimations from the virtual sensors might drift due to environmental changes or physical sensors degradation (Martin et al., 2021). The statistical proxies might not be accurate anymore in case of occasional events, such as wildfire, extreme weather, etc. Therefore, the application of virtual sensors requires a long-term drift monitoring and online-adaptive models can be used to adjust the models accordingly and regularly to maintain the added value provided by the implementation of statistical proxies as virtual sensors (**Paper IV**).

Once the sensor virtualization process is validated, the continuous measurements can be utilized to update the current AQI (Monteiro et al., 2017). While there is no universally accepted method, most organizations including WHO consider PM_{10}, $PM_{2.5}$, NO_2, O_3, SO_2 and CO as parameters for AQI calculation (WHO, 2021). From the scientific point of view, these parameters are insufficient to show association with health risk of aerosol particles, especially for cardiovascular effects (Geng et al., 2013). Although WHO (2012) has recommended the inclusion of BC as one of the components in AQI alongside with the other air quality parameters, this has not been taken into action due to the unavailability of continuous BC measurements. This is partly attributed to the lack of national environmental legislation (Kutzner et al., 2018), instrument failure or data corruption (Junger & De Leon, 2015; Zaidan et al., 2019). Apart from BC, other aerosols parameters with data unavailability, such as LDSA and ultrafine particles (UFP), can also be estimated using virtual sensors with a similar methodology (Fung et al., 2021a).

4 Review of papers and the author's contribution

Paper I presents the BC, NO_x, and $PM_{2.5}$ concentrations at various environments in southern Finland and especially in the HMA. Depending on the varying local anthropogenic activities that were traffic and residential wood combustion, BC concentration also varied both spatially and temporally. In terms of long-term trends, the BC, NO_x, and $PM_{2.5}$ concentrations were statistically decreasing, and the decreasing trends were most probably due to a decrease in the local traffic emissions. For this article, I contributed by commenting and editing the manuscript. Through commenting, I also assisted in data interpretation from which has then been the basis of this Ph.D. topic.

Paper II introduces a novel IAP to estimate air pollutant variables in case of missing data. BC estimation is the case study in the paper. By checking the correlation of BC with other existing measurements in the same site, the model manages to select the more favorable input variables for the regression. The results show the novel method could give generally accurate and continuous BC estimation at the two selected urban environments: urban background and street canyon. For this article, I performed the majority of the data analysis and writing.

Paper III compares the IAP with other WB and BB models, namely LASSO, DT, RF, SVR, SNN and LSTM. In general, the BB models demonstrate better performance in terms of accuracy, but the white-box models are better choice due to the higher transparency in model architecture. Among all tested models in the paper, IAP and LASSO are recommended due to its flexibility and efficiency, respectively, as virtual sensors. For this article, I performed the majority of the data analysis and writing.

Paper IV presents a novel method of integrating LCSs embedded with the features of intelligent calibration and virtual sensors. The paper also demonstrates how to calibrate LCSs in the field and how to extend the operation of LCSs to monitor additional air quality indicators (e.g. BC) that are not measured directly by these LCSs. I was responsible for curating data from the reference station. Moreover, I assisted in data interpretation and visualization, not to mention gave comments to the article.

5 Conclusion and outlook

Black carbon (BC) is known to have a strong influence in climate change, air quality and potential risk for human beings. Therefore, BC has been recommended to be included as one of the parameters for air quality index (AQI) calculation, along with the current parameters, by the World Health Organization (WHO). However, due to the lack of national environmental legislation of BC monitoring, BC is not measured in every national measurement stations. Besides the measurement scarcity, relying on BC measurements are not always possible due to instrument failure or data corruption, leading to long data gaps. For example, the missing data of BC in an urban background site in Helsinki metropolitan area (HMA) in 2017–2018 could reach 30%. With the missing data, the analysis of air pollution becomes more uncertain; therefore, air quality models are needed for data gap imputation and air quality prediction.

In this thesis, we explored the BC monitoring network in HMA. The ambient BC concentrations show various characteristics in different environments. Generally, BC concentrations were the highest in traffic sites (TR) due to vehicular combustion, followed by detached housing sites (DH) and urban background sites (UB) depending on the time of the day, the day of the week and the month of the year. The sources of the ambient BC in these environments were typically coupled with the mixture of household wood combustion and vehicle combustion. Regional background (RB) had the lowest concentration because there were no local BC sources in the vicinity. All the environments with long enough datasets showed a decreasing trend in BC concentrations. The study also showed that BC had a high correlation with some other commonly monitored parameters, for example NO_x and $PM_{2.5}$. These high correlation values served good grounds for statistical proxy derivation. BC was also highly correlated with lung deposited surface area (LDSA), particle number concentration (PNC) and CO and moderately correlated with PM_{10} and O_3 at the two selected sites, the urban background UB2 site and the street canyon TR2 site, linearly. At both sites, BC showed low correlation with meteorological parameters. This well responds to **aim 1**: to characterize the ambient BC concentration from the reference measurements in HMA.

At the two selected sites, we developed a novel input-adaptive method to estimate BC concentrations based on the reference station data as a statistical proxy for sensor virtualization (**aim 2**). This proposed input-adaptive proxy (IAP) was demonstrated to perform well in terms of accuracy and flexibility. As a white-box model, the model architecture of IAP is transparent; therefore, it is possible to spot the mistakes and then fine-tune the model in case of errors based on the actual physical properties of the air pollutant. It was evidenced to outperform the other models with all things considered (**aim 3**). We also explored the feasibility to integrate virtual sensors into low-cost air monitoring sensors by using statistical proxies (**aim 4**). A non-linear model structure was validated and used for the calibration of LCSs and the development of BC statistical proxy. The results showed a satisfactory performance in terms of accuracy.

The virtual sensors using the statistical proxies, both with reference data and LCS data, were demonstrated to overcome the following three weaknesses of purely physical sensors: price issues, spatial conditions and data quality deficiency. However, there are limitations in using BC models as virtual sensors. In this thesis, the two measurement sites can be characterized as urban background and street canyon. Each characterization represents places with similar air pollutant sources. The corresponding BC models are also location-specific, but it might be possible to extend the models from one street canyon site to another with a calibration model. Similar ideas could also be applied to urban background sites. Once we gather enough training data over a substantial of time from different locations, not limited to street canyon and urban background sites investigated in this thesis, a network of database can be established. Within the network, selection of predictors can be interchanged. Although virtual sensors are considered to be an alternative to overcome the three weaknesses abovementioned, additional data post-processing is often required to keep the data quality up to standard. A long-term drift monitoring and online-adaptive models are also required to adjust the proxies accordingly. These would be our next steps on top of the thesis.

In this thesis, we focused on the BC statistical proxies that can provide continuous modeling data with good accuracy as virtual sensors. Similarly, other aerosols parameters with data unavailability, such as LDSA and ultrafine particle (UFP), which are considered having negative impact on human health but lack air quality guidelines, can also be estimated with the proposed general methodology in the future.

References

Aakko-Saksa, P., Koponen, P., Roslund, P., Laurikko, J., Nylund, N.-O., Karjalainen, P., Rönkkö, T., & Timonen, H. (2020). Comprehensive emission characterisation of exhaust from alternative fuelled cars. *Atmos. Environ.*, 236, 117643. https://doi.org/10.1016/j.atmosenv.2020.117643

Achilleos, S., Kioumourtzoglou, M.-A., Wu, C.-D., Schwartz, J. D., Koutrakis, P., & Papatheodorou, S. (2017). Acute effects of fine particulate matter constituents on mortality: A systematic review and meta-regression analysis. *Environ. Int.*, 109, 89–100. https://doi.org/10.1016/j.envint.2017.09.010

Albertos, P. & Goodwin, G. (2002). Virtual sensors for control applications. *Annu. Rev. Control.*, 26(1), 101–112. https://doi.org/10.1016/S1367-5788(02)80018-9

Anttila, P. (2020). *Air quality trends in Finland, 1994–2018*. Finnish Meteorological Institute. https://doi.org/10.35614/isbn.9789523361027. Ph.D. Thesis

Baier, L., Kühl, N., & Satzger, G. (2019). How to cope with change? Preserving validity of predictive services over time. *The 52nd Hawaii International Conference on System Sciences (HICSS)*, 1085–1094. https://doi.org/10.5445/IR/1000085769

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., et al. (2013). Bounding the role of black carbon in the climate system: A scientific assessment. *J. Geophys. Res. Atmos.*, 118(11), 5380–5552. https://doi.org/10.1002/jgrd.50171

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2), 123–140. https://doi.org/10.1007/BF00058655

Brunekreef, B., Strak, M., Chen, J., Andersen, Z., Atkinson, R., Bauwelinck, M., Bellander, T., Boutron, M.-C., Brandt, J., Carey, I., Cesaroni, G., Forastiere, F., Fecht, D., Gulliver, J., Hertel, O., Hoffmann, B., de Hoogh, K., Houthuijs, D., Hvidtfeldt, U., Janssen, N., Jørgensen, J., Katsouyanni, K., Ketzel, M., Klompmaker, J., Krog, N. H., Liu, S., Ljungman, P., Mehta, A., Nagel, G., Oftedal, B., Pershagen, G., Peters, A., Raaschou-Nielsen, O., Renzi, M., Rodopoulou, S., Samoli, E., Schwarze, P., Sigsgaard, T., Stafoggia, M., Vienneau, D., Weinmayr, G., Wolf, K., & Hoek, G. (2021). Mortality and morbidity effects of long-term exposure to low-level PM2.5, BC, NO2, and O3: An analysis of European cohorts in the ELAPSE Project. Report, Health Effects Institute.

Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.*, 119, 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014

Caubel, J. J., Cados, T. E., Preble, C. V., & Kirchstetter, T. W. (2019). A distributed network of 100 black carbon sensors for 100 days of air quality monitoring in West Oakland, California. *Environ. Sci. Technol.*, 53(13), 7564–7573. https://doi.org/10.1021/acs.est.9b00282

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., Janssen, N. A., Martin, R. V., Samoli, E., Schwartz, P. E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., & Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environ. Int.*, 130, 104934. https://doi.org/10.1016/j.envint.2019.104934

Ciupek, K., Butterfield, D., Quincey, P., Sweeney, B., Lilley, A., Bradshaw, C., Fuller, G., Green, D., & Font, F. (2021). 2019 Annual Report for the UK Black Carbon Network. Report, National Physical Laboratory. https://doi.org/10.47120/npl.ENV38

Collaud Coen, M., Andrews, E., Bigi, A., Martucci, G., Romanens, G., Vogt, F. P. A., & Vuilleumier, L. (2020). Effects of the prewhitening method, the time granularity, and the time segmentation on the Mann–Kendall trend detection and the associated Sen's slope. *Atmos. Meas. Tech.*, 13, 6945–6964. https://doi.org/10.5194/amt-13-6945-2020

Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P., & Tarkoma, S. (2021). Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Trans. Sens. Netw.*, 17, 1–44. https://doi.org/10.1145/3446005

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L., & Popoola, O. (2016). Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sens. Actuators B Chem.*, 231, 701–713. https://doi.org/10.1016/j.snb.2016.03.038

Fernández-Guisuraga, J. M., Castro, A., Alves, C., Calvo, A., Alonso-Blanco, E., Blanco-Alegre, C., Rocha, A., & Fraile, R. (2016). Nitrogen oxides and ozone in Portugal: trends and ozone estimation in an urban and a rural site. *Environ. Sci. Pollut. Res.*, 23(17), 17171–17182. https://doi.org/10.1007/s11356-016-6888-6

Flanner, M. G., Zender, C. S., Hess, P. G., Mahowald, N. M., Painter, T. H., Ramanathan, V., & Rasch, P. (2009). Springtime warming and reduced snow cover from carbonaceous particles. *Atmos. Chem. Phys.*, 9(7), 2481–2497. https://doi.org/10.5194/acp-9-2481-2009

Fung, P. L., Zaidan, M. A., Niemi, J. V., Saukko, E., Timonen, H., Kousa, A., Kuula, J., Rönkkö, T., Karppinen, A., Tarkoma, S., Kulmala, M., Petäjä, T., & Hussein, T. (2021a). Input-adaptive linear mixed-effects model for estimating alveolar Lung Deposited Surface Area (LDSA) using multipollutant datasets. *Atmos. Chem. Phys. Discuss. [preprint]*, 2021, 1–33. https://doi.org/10.5194/acp-2021-427

Fung, P. L., Zaidan, M. A., Surakhi, O., Tarkoma, S., Petäjä, T., & Hussein, T. (2021b). Data imputation in in situ-measured particle size distributions by means of neural networks. *Atmos. Meas. Tech.*, 14, 5535–5554. https://doi.org/10.5194/amt-14-5535-2021

Geng, F., Hua, J., Mu, Z., Peng, L., Xu, X., Chen, R., & Kan, H. (2013). Differentiating the associations of black carbon and fine particle with daily mortality in a Chinese city. *Environ. Res.*, 120, 27–32. https://doi.org/10.1016/j.envres.2012.08.007

Gilbert, R. (1987). *Statistical methods for environmental pollution monitoring.* John Wiley & Sons.

Guo, B., Wang, Y., Zhang, X., Che, H., Ming, J., & Yi, Z. (2020). Long-term variation of black carbon aerosol in China based on revised aethalometer monitoring data. *Atmosphere*, 11(7), 684. https://doi.org/10.3390/atmos11070684

Hagler, G. S., Williams, R., Papapostolou, V., & Polidori, A. (2018). Air quality sensors and data adjustment algorithms: When is it no longer a measurement? *Environ. Sci. Technol.*, 52(10), 5530–5531. https://doi.org/10.1021/acs.est.8b01826

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Stat. Sci.*, 35(4), 579–592. https://doi.org/10.1214/19-STS733

Health Effects Institute (2020). State of Global Air 2020: A special report on global exposure to air pollution and its health impacts. Report, Health Effects Institute. https://www.stateofglobalair.org/sites/default/files/documents/2020-10/soga-2020-report-10-26_0.pdf

Helin, A., Niemi, J. V., Virkkula, A., Pirjola, L., Teinilä, K., Backman, J., Aurela, M., Saarikoski, S., Rönkkö, T., Asmi, E., & Timonen, H. (2018). Characteristics and source apportionment of black carbon in the Helsinki metropolitan area, Finland. *Atmos. Environ.*, 190, 87–98. https://doi.org/10.1016/j.atmosenv.2018.07.022

Hellén, H., Kangas, L., Kousa, A., Vestenius, M., Teinilä, K., Karppinen, A., Kukkonen, J., & Niemi, J. V. (2017). Evaluation of the impact of wood combustion on benzo[a]pyrene (BaP) concentrations; ambient measurements and dispersion modeling in Helsinki, Finland. *Atmos. Chem. Phys.*, 17(5), 3475–3487. https://doi.org/10.5194/acp-17-3475-2017

Hietikko, R., Kuuluvainen, H., Harrison, R., Portin, H., Timonen, H., Niemi, J., & Rönkkö, T. (2018). Diurnal variation of nanocluster aerosol concentrations and emission factors in a street canyon. *Atmos. Environ.*, 189, 98–106. https://doi.org/10.1016/j.atmosenv.2018.06.031

Hussein, T., Karppinen, A., Kukkonen, J., Härkönen, J., Aalto, P., Hämeri, K., Kerminen, V., & Kulmala, M. (2004). Meteorological dependence of size-fractionated number concentrations of urban aerosol particles. *Atmos. Environ.*, 40(8), 1427–1440. https://doi.org/10.1016/j.atmosenv.2005.10.061

IARC (2014). Diesel and gasoline engine exhausts and some nitroarenes. Report, IARC.

INECE (2008). Jump-starting climate protection: INECE targets compliance with laws controlling black carbon. Report, INECE.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel* *on Climate Change.* Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. In Press.

Janssen, N. A., Hoek, G., Simic-Lawson, M., Fischer, P., Van Bree, L., Ten Brink, H., Keuken, M., Atkinson, R. W., Anderson, H. R., Brunekreef, B., & Cassee, F. R. (2011). Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM10 and PM2.5. *Environ. Health Perspect.*, 119(12), 1691–1699. https://doi.org/10.1289/ehp.1003369

Jha, S. K., Kumar, M., Arora, V., Tripathi, S. N., Motghare, V. M., Shingare, A. A., Rajput, K. A., & Kamble, S. (2021). Domain adaptation-based deep calibration of low-cost pm2.5 sensors. *IEEE Sens. J.*, 21(22), 25941–25949. https://doi.org/10.1109/JSEN.2021.3118454

Junger, W. & De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmos. Environ.*, 102, 96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmos. Environ.*, 38(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Järvi, L., Hannuniemi, H., Hussein, T., Junninen, H., Aalto, P., Hillamo, R., Mäkelä, T., Keronen, P., Siivola, E., Vesala, T., & Kulmala, M. (2009). The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface–atmosphere interactions in Helsinki, Finland. *Boreal Environ. Res.*, 14, 86–109.

Järvi, L., Junninen, H., Karppinen, A., Hillamo, R., Virkkula, A., Mäkelä, T., Pakkanen, T., & Kulmala, M. (2008). Temporal variations in black carbon concentrations with different time scales in Helsinki during 1996–2005. *Atmos. Chem. Phys.*, 8(4), 1017–1027. https://doi.org/10.5194/acp-8-1017-2008

Järvinen, A., Timonen, H., Karjalainen, P., Bloss, M., Simonen, P., Saarikoski, S., Kuuluvainen, H., Kalliokoski, J., Dal Maso, M., Niemi, J. V., & Keskinen, J. (2019). Particle emissions of Euro VI, EEV and retrofitted EEV city buses in real traffic.

Environ. Pollut., 250, 708–716. https://doi.org/10.1016/j.envpol.2019.04.033

Kang, G., Gao, J., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci.*, 9(1), 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

Karagulian, F., Belis, C. A., Dora, C. F. C., Prüss-Ustün, A. M., Bonjour, S., Adair-Rohani, H., & Amann, M. (2015). Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmos. Environ.*, 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087

Kleinbaum, D., Kupper, L., Nizam, A., & Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Klimont, Z., Kupiainen, K., Heyes, C., Purohit, P., Cofala, J., Rafaj, P., Borken-Kleefeld, J., & Schöpp, W. (2017). Global anthropogenic emissions of particulate matter including black carbon. *Atmos. Chem. Phys.*, 17(14), 8681–8723. https://doi.org/10.5194/acp-17-8681-2017

Kortoçi, P., Hossein Motlagh, N., Zaidan, M. A., Fung, P. L., Varjonen, S., Rebeiro-Hargrave, A., Niemi, J. V., Nurmi, P., Hussein, T., Petäjä, T., Kulmala, M., & Tarkoma, S. (2022). Air pollution exposure monitoring using portable low-cost air quality sensors. *Smart Health*, 23, 100241. https://doi.org/10.1016/j.smhl.2021.100241

Krecl, P., Johansson, C., Targino, A. C., Ström, J., & Burman, L. (2017). Trends in black carbon and size-resolved particle number concentrations and vehicle emission factors under real-world conditions. *Atmos. Environ.*, 165, 155–168. https://doi.org/10.1016/j.atmosenv.2017.06.036

Krecl, P., Targino, A. C., Landi, T. P., & Ketzel, M. (2018). Determination of black carbon, PM2.5, particle number and NOx emission factors from roadside measurements and their implications for emission inventory development. *Atmos. Environ.*, 186, 229–240. https://doi.org/10.1016/j.atmosenv.2018.05.042

Kutzner, R. D., von Schneidemesser, E., Kuik, F., Quedenau, J., Weatherhead, E. C., & Schmale, J. (2018). Long-term monitoring of black carbon across Germany. *Atmos. Environ.*, 185, 41–52. https://doi.org/10.1016/j.atmosenv.2018.04.039

Lagerspetz, E., Motlagh Hossein, N., Zaidan, M. A., Fung, P. L., Mineraud, J., Varjonen, S., Siekkinen, M., Nurmi, P., Matsumi, Y., Tarkoma, S., & Hussein, T. (2019). Megasense: Feasibility of low-cost sensors for pollution hot-spot detection. *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, 1083–1090. https://doi.org/10.1109/INDIN41052.2019.8971963

Lin, T., Horne, B. G., Tino, P., & Giles, C. L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Trans. Neural Netw.*, 7(6), 1329–1338. https://doi.org/10.1109/72.548162

Liu, L., Kuo, S. M., & Zhou, M. (2019). Virtual sensing techniques and their applications. *2009 International Conference on Networking, Sensing and Control*, 31–36. https://doi.org/10.1109/ICNSC.2009.4919241

Luoma, K. (2021). *Aerosol optical properties, black carbon and their spatio-temporal variation*. University of Helsinki. http://urn.fi/URN:ISBN:978-952-7276-56-3. Ph.D. Thesis

Luoma, K., Virkkula, A., Aalto, P., Lehtipalo, K., Petäjä, T., & Kulmala, M. (2021). Effects of different correction algorithms on absorption coefficient – a comparison of three optical absorption photometers at a boreal forest site. *Atmos. Meas. Tech.*, 14, 6419–6441. https://doi.org/10.5194/amt-14-6419-2021

Martin, D., Kühl, N., & Satzger, G. (2021). Virtual sensors. *Bus. Inf. Syst. Eng.*, 63(3), 315–323. https://doi.org/10.1007/s12599-021-00689-w

Masih, A. (2019). Application of ensemble learning techniques to model the atmospheric concentration of SO2. *Glob. J. Environ. Sci. Manag.*, 5(3), 309–318. https://doi.org/10.22034/GJESM.2019.03.04

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petäjä, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin III, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M., & Laaksonen, A. (2011). A statistical proxy for sulphuric acid concentration. *Atmos. Chem. Phys.*, 11(11), 11319–11334. https://doi.org/10.5194/acp-11-11319-2011

Mølgaard, B., Birmili, W., Clifford, S., Massling, A., Eleftheriadis, K., Norman, M., Vratolis, S., Wehner, B., Corander, J., Hämeri, K., & Hussein, T. (2013). Evaluation of a statistical forecast model for size-fractionated urban particle number concentrations using data from five European cities. *J. Aerosol Sci.*, 66, 96–110. https://doi.org/10.1016/j.jaerosci.2013.08.012

Monteiro, A., Vieira, M., Gama, C., & Miranda, A. (2017). Towards an improved air quality index. *Air Qual. Atmos. Health*, 10, 447–455. https://doi.org/10.1007/s11869-016-0435-y

Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G. S., Jayaratnea, R., Kumar, P., Lau, A. K., Louie, P. K., Mazaheri, M., Ning, Z., Motta, N., Mullins, B., Rahman, M. M., Ristovski, Z., Shafiei, M., Tjondronegoro, D., Westerdahl, D., & Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environ. Int.*, 116, 286–299. https://doi.org/10.1016/j.envint.2018.04.018

Novakov, T., Ramanathan, V., Hansen, J., Kirchstetter, T., Sato, M., Sinton, J., & Sathaye, J. (2003). Large historical changes of fossil-fuel black carbon aerosols. *Geophys. Res. Lett.*, 30(6), 1324. https://doi.org/10.1029/2002GL016345

Park, Y. & Klabjan, D. (2020). Subset selection for multiple linear regression via optimization. *J. Glob. Optim.*, 77, 543–574. https://doi.org/10.1007/s10898-020-00876-1

Petzold, A., Ogren, J. A., Fiebig, M., Laj, P., Li, S.-M., Baltensperger, U., Holzer-Popp, T., Kinne, S., Pappalardo, G., Sugimoto, N., Wehrli, C., Wiedensohler, A., & Zhang, X.-Y. (2013). Recommendations for reporting “black carbon” measurements. *Atmos. Chem. Phys.*, 13(16), 8365–8379. https://doi.org/10.5194/acp-13-8365-2013

Petzold, A. & Schönlinner, M. (2004). Multi-angle absorption photometry—a new method for the measurement of aerosol light absorption and atmospheric black carbon. *J. Aerosol Sci.*, 35(4), 421–441. https://doi.org/10.1016/j.jaerosci.2003.09.005

Rebeiro-Hargrave, A., Fung, P. L., Varjonen, S., Huertas, A., Sillanpää, S., Luoma, K., Hussein, T., Petäjä, T., Timonen, H., Limo, J., Nousiainen, V., & Tarkoma, S. (2021). City wide participatory sensing of air quality. *Front. Environ. Sci.*, 9, 587. https://doi.org/10.3389/fenvs.2021.773778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rybarczyk, Y. & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, 8(12), 2570. https://doi.org/10.3390/app8122570

Rönkkö, T. & Timonen, H. (2019). Overview of sources and characteristics of nanoparticles in urban traffic-influenced areas. *J. Alzheimer's Dis.*, 72(1), 15–28. https://doi.org/10.3233/JAD-190170

Saarikoski, S., Niemi, J. V., Aurela, M., Pirjola, L., Kousa, A., Rönkkö, T., & Timonen, H. (2021). Sources of black carbon at residential and traffic environments obtained by two source apportionment methods. *Atmos. Chem. Phys.*, 21, 14851–14869. https://doi.org/10.5194/acp-21-14851-2021

Shahriyer, A. H. (2020). *Derivation of simple proxies for aerosol related parameters at SMEAR III site*. University of Helsinki. http://urn.fi/URN:NBN:fi:hulib-202009244141. M.Sc. Thesis

Sharma, S., Leaitch, W. R., Huang, L., Veber, D., Kolonjari, F., Zhang, W., Hanna, S. J., Bertram, A. K., & Ogren, J. A. (2017). An evaluation of three methods for measuring black carbon in Alert, Canada. *Atmos. Chem. Phys.*, 17(24), 15225–15243. https://doi.org/10.5194/acp-17-15225-2017

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Mon. Weather Rev.*, 135(3), 1151–1157. https://doi.org/10.1175/MWR3326.1

Tegen, A., Davidsson, P., Mihailescu, R.-C., & Persson, J. A. (2019). Collaborative sensing with interactive learning using dynamic intelligent virtual sensors. *Sensors*, 19(3). https://doi.org/10.3390/s19030477

Teinilä, K., Aurela, M., Niemi, J., Kousa, A., Petäjä, T., Järvi, L., Hillamo, R., Kangas, L., Saarikoski, S., & H., T. (2019). Concentration variation of gaseous and particulate pollutants in the Helsinki city centre-observations from a two-year campaign from 2013–2015. *Boreal Environ. Res.*, 24, 115–136.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Timonen, H., Aurela, M., Carbone, S., Saarnio, K., Frey, A., Saarikoski, S., Teinilä, K., Kulmala, M., & Hillamo, R. (2014). Seasonal and diurnal changes in inorganic ions, carbonaceous matter and mass in ambient aerosol particles in an urban, background area. *Boreal Environ. Res.*, 19, 71–86.

Timonen, H., Karjalainen, P., Aalto, P., Saarikoski, S., Myllari, F., Karvosenoja, N., Jalava, P., Asmi, E., Aakko-Saksa, P., Saukkonen, N., Laine, T., Saarnio, K., Niemelä, N., Enroth, J., Väkevä, M., Oyola, P., Pagels, J., Ntziachristos, L., Cordero, R., Kuittinen, N., Niemi, J. V., & Rönkkö, T. (2019). Adaptation of black carbon footprint concept would accelerate mitigation of global warming. *Environ. Sci. Technol.*, 53(21), 12153–12155. https://doi.org/10.1021/acs.est.9b05586

Van Roode, S., Ruiz-Aguilar, J., González-Enrique, J., & Turias, I. (2019). An artificial neural network ensemble approach to generate air pollution maps. *Environ. Monit. Assess.*, 191, 727. https://doi.org/10.1007/s10661-019-7901-6

Wang, N., Wang, Y., Hu, S., Hu, Z., Xu, J., Tang, H., & Jin, G. (2018). Robust regression with data-dependent regularization parameters and autoregressive temporal correlations. *Environ. Model. Assess.*, 23(6), 779–786. https://doi.org/10.1007/s10666-018-9605-7

WHO (2012). Health effects of black carbon. Report, World Health Organization. https://www.euro.who.int/__data/assets/pdf_file/0004/162535/e96541.pdf

WHO (2019). World health statistics 2019: Monitoring health for the SDGs, sustainable development goals. Report, World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/324835/9789241565707-eng.pdf

WHO (2021). Who global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Report, World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/345329/9789240034228-eng.pdf

Woo, J. H., An, S. M., Hong, K., Kim, J. J., Lim, S. B., Kim, H. S., & Eum, J. H. (2016). Integration of CFD-based virtual sensors to a ubiquitous sensor network to support micro-scale air quality management. *J. Environ. Inform.*, 27(2). https://doi.org/10.3808/jei.201500314

Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors*, 16(1), 86. https://doi.org/10.3390/s16010086

Zaidan, M. A., Wraith, D., Boor, B. E., & Hussein, T. (2019). Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models. *Applied Sciences*, 9(22), 4976. https://doi.org/10.3390/app9224976

Šimić, I., Lovrić, M., Godec, R., Kröll, M., & Bešlić, I. (2020). Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon. *Environ. Pollut.*, 263, 114587. https://doi.org/10.1016/j.envpol.2020.114587