Estimating Black Carbon Levels With Proxy Variables and Low-Cost Sensors

Xiaoli Liu, Francesco Concas, Naser Hossein Motlagh, Martha Arbayani Zaidan, *Senior Member, IEEE*, Pak Lun Fung, Samu Varjonen, Jarkko V. Niemi, Hilkka Timonen, Tareq Hussein, Tuukka Petäjä, Markku Kulmala, Petteri Nurmi, and Sasu Tarkoma, *Senior Member, IEEE*

***Abstract*—We develop a portable and affordable solution for estimating personal exposure to black carbon (BC) using low-cost sensors and machine learning. Our approach uses other pollutants and environmental variables as proxies for estimating the concentrations of BC and combines this with machine learning-based sensor calibration to improve the quality of the inputs that are used as proxies in the modeling. We extensively validate the feasibility of our approach and demonstrate its benefits with benchmarks conducted on real-world data from two different urban locations with different population densities and characteristics. Our results demonstrate that our approach can accurately estimate BC (R^2 higher than 0.9) without relying on a dedicated sensor. The results also highlight how calibration is essential for ensuring accurate modeling on low-cost sensor measurements. Our results offer a novel affordable and portable solution that can be used to estimate personal exposure to BC and, more generally, demonstrate how low-cost sensors and proxy modeling can increase the spatiotemporal scale at which information about BC level is available.**

***Index Terms*—Air quality, black carbon (BC), low-cost sensor, machine learning, proxy.**

Manuscript received 15 September 2023; accepted 22 January 2024. Date of publication 5 February 2024; date of current version 9 May 2024. This work was supported in part by the Nokia Center for Advanced Research (NCAR); in part by the Academy of Finland Projects under Grant 345008, Grant 339614, and Grant 355330; in part by the Business Finland Project under Grant 8782/31/2022; and in part by the funding from Technology Industries of Finland Centennial Foundation to Urban Air Quality 2.0 Project. *(Corresponding author: Xiaoli Liu.)*

Xiaoli Liu, Francesco Concas, Naser Hossein Motlagh, Petteri Nurmi, and Sasu Tarkoma are with the Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland (e-mail: xiaoli.liu@helsinki.fi).

Martha Arbayani Zaidan is with the Department of Computer Science and the Institute for Atmospheric and Earth System Research, University of Helsinki, 00014 Helsinki, Finland.

Pak Lun Fung, Tuukka Petäjä, and Markku Kulmala are with the Institute for Atmospheric and Earth System Research, University of Helsinki, 00014 Helsinki, Finland.

Samu Varjonen is with the Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland, and also with the Atmospheric Composition Research, Finnish Meteorological Institute, 00101 Helsinki, Finland.

Jarkko V. Niemi is with the Air Quality Unit, Helsinki Region Environmental Services Authority, 00066 Helsinki, Finland.

Hilkka Timonen is with the Atmospheric Composition Research, Finnish Meteorological Institute, 00101 Helsinki, Finland.

Tareq Hussein is with the Institute for Atmospheric and Earth System Research, University of Helsinki, 00014 Helsinki, Finland, and also with the Environmental and Atmospheric Research Laboratory, Department of Physics, School of Science, University of Jordan, Amman 11942, Jordan.

Digital Object Identifier 10.1109/JIOT.2024.3361977

I. Introduction

PORTABLE personal air quality sensors facilitating the monitoring and mitigation of personal pollution exposure are becoming increasingly affordable and available thanks to advances in miniaturization and sensor technology. As exposure to pollutants has been linked with many acute and chronic diseases [1], [2], [3], [4], [5], [6], several types of cancer [3], [7], [8] and even the occurrence and severity of COVID-19 [9], [10], [11], [12], these devices provide essential information for mitigating personal health risks and contribute toward improved health. Besides benefits to individuals, portable air quality sensors help to improve the resolution and coverage of air quality information [13], offering policymakers and other stakeholders information about current pollutant characteristics at high spatial and temporal resolution. This helps to mitigate the global cost of pollution and evaluate the effectiveness of countermeasures designed to tackle pollution. Indeed, estimates suggest that 2–5% of global GDP is spent on the treatment of diseases linked to poor air quality [14], [15], making pollution a truly global problem.

Among the different pollutants, black carbon (BC) (also known as soot) is among the worst pollutants to affect individuals. BC is linked with several chronic health conditions, including cancer and respiratory diseases [16], [17], and it contributes to global climate change by absorbing heat [18]. What makes BC particularly problematic is that the BC particles linger in the air for a long time while bonding with chemicals and other substances [17], [19]. This makes soot, besides a harmful substance in its own right, a carrier of harmful compounds, including airborne viruses.

While portable and affordable (i.e., costing \$100–\$1000) sensors for many pollutants, such as particulate matter (PM) and common aerosols (e.g., NO_2 and CO_2), are widely available [13], [20], [21], [22], [23], [24], unfortunately, this is not the case for BC level. Indeed, the dominant approach for estimating the BC level currently is to rely on professional-grade measurement technology which typically is highly expensive to operate and maintain with a single sensor typically costing over \$50 000 [25], [26]. This lack of affordable and portable sensors for BC is unfortunate as it limits the information about personal exposure to harmful pollutants and the resolution of the available information. The limited resolution of information is also challenging to policymakers as they must base their decisions on aggregate information without having a detailed view of how BC levels vary in different parts of the urban environments.

We contribute a portable and affordable solution for estimating BC levels using low-cost air quality monitoring sensors and machine learning. The key idea in our approach is to take advantage of *proxy variables* that are integrated into a single sensor unit, and that measure environmental variables and concentrations of other pollutants. The proxy variables are used as input to a machine-learning model that estimates the current BC level. Sensor utilities on portable air quality monitoring devices are affected by cross-sensitivities which results in the measurements of different pollutants being correlated. This suggests that the concentrations of one pollutant can be estimated at least with a reasonable accuracy from the concentrations of other pollutants. This is particularly useful for BC, due to its tendency to linger in the air. Naïvely modeling relationships between different pollutants, however, is not sufficient, as sensor utilities on portable devices often suffer from significant inaccuracies [27], [28]. To overcome these inaccuracies, we combine proxy variables with machine learning-based calibration which helps to improve the quality of the sensor measurements used for proxy modeling. Machine learning-based calibration has recently emerged as a powerful solution for improving the accuracy of low-cost sensors, and to provide an alternative to laboratory-based calibration. Machine learning techniques are effective in dealing with air quality and environmental data, which are often nonlinear, and machine learning algorithms can learn from large amounts of environmental data and identify complex patterns to perform low-cost sensor calibration [28], [29]. As we demonstrate in this article, the integration of machine learning-based calibration with proxy modeling is essential for achieving accurate BC-level estimation performance.

We validate our approach through extensive experiments carried out on measurements from two locations with differing urban densities and characteristics, and over a long period of time (20 months and 26.5 months of measurements). The results demonstrate that our approach can reliably estimate BC levels with an R^2 higher than 0.9 without relying on a dedicated sensor. The results also highlight how sensor calibration is essential for improving the quality of the measurements that are used as input for the proxy modeling. Taken together, our work offers a novel portable and affordable solution for estimating BC concentrations, significantly extend the scope, scale, and spatiotemporal resolution at which information about air pollutants can be captured on low-cost sensors.

1) *Feasibility:* Demonstrate the feasibility of estimating BC levels on proxy variables using various machine learning models and pollution data at two reference stations collected in a long period.
2) *Novel Approach:* We estimate BC levels using proxy variables leveraging intelligent machine learning and sensor calibration. Specifically, our solution combines data from low-cost sensor data with precalibrated proxy variables on low-cost air quality sensors for estimating BC levels.
3) *High Accuracy:* The estimating BC levels on low-cost sensors can achieve an accuracy close to those provided by expensive high-quality instruments.

II. Related Work

A. Low-Cost Sensor Calibration

Low-cost sensors are prone to noisy measurements, environmental inferences, and interdevice differences, which can result in errors in the sensor readings. Most of the works on sensor calibration focus on developing calibration algorithms or frameworks for specific pollutants (typically either PM or one or more gaseous pollutants) [30], [31], [32]. The algorithms that have been proposed cover basically all common machine learning techniques, ranging from linear models, such as univariate linear regression [22], [33], [34], [35], [36], [37], [38], [39] and multivariate linear regression (MLR), to nonlinear models, such as support vector regression (SVR) [38], random forest regression (RFR) [27], [38], [39], [40], [41], multilayer perceptron (MLP) [21], [22], [38], [42], complex neural networks [43], deep learning [44], and hybrid models combining several techniques [41].

B. BC Estimation and Proxy Variables

BC estimation is typically carried out using expensive professional-grade measurement stations. Most of the works on modeling BC concentrations focus on a global scale and use aggregate-level estimates [45], [46], [47], [48], [49] which offer limited spatiotemporal resolution and are unable to offer insights into personal exposure. The few works to consider a finer resolution have focused on specialized microenvironments, such as transportation systems or high-density city blocks and used professional-grade measurements as inputs for estimation [50], [51], [52]. There has also been some limited work on developing mobile platforms for capturing BC concentrations, but these remain proprietary and limited in use [46].

Proxy variables are defined as variables that are not directly relevant but can be utilized to serve in place of an unobservable or immeasurable variable. Proxies have been used, e.g., to forecast pollutant concentrations [53] or to fill in missing values in observations [29]. The feasibility of using air quality measurements as proxy variables for BC estimation has been demonstrated in our earlier research. Fung et al. [54] developed a simple linear regression white-box model for estimating BC by using an input adaptive approach, which manages to search for the best combination of proxy variables for the estimation using ordinary least squares (OLSs). Contrary to linear regression, Rovira et al. [55] focused on the nonlinear properties of BC by exploring two black-box models, i.e., SVR and random forest. Zaidan et al. [56] and Fung et al. [57] compared and evaluated BC estimation using white-box and black-box models using proxy variables measured at reference stations.

Compared to previous research, instead of measuring BC directly, we use proxy variables for estimating BC concentrations. We also incorporate sensor calibration as part of the estimation process to improve performance and consider a broader range of input variables. Our work extends the

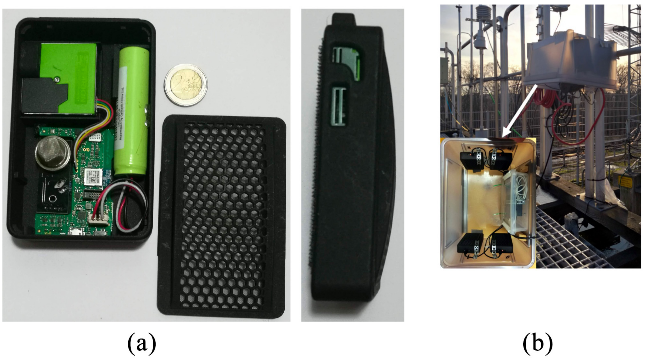


Fig. 1. Sensing systems used in our experiment. (a) Portable LCPs. (b) Installation of four LCPs on SMEAR III under a rain cover.

TABLE I
AVAILABLE INSTRUMENTS MEASURING VARIABLES USED IN BC PROXY DEVELOPMENT AND THE AVAILABLE LCPS VARIABLES

Sensing stations	Instruments	Variables measured
SMEAR III	Grimm 180/FH 62 I-R	$PM_{2.5}$, PM_{10}
	Horiba APNA 360	NO, NO_2, NOx
	MAAP Thermo Scientific 5012	BC
	Platinum resistant thermometer Pt-100	T
	Thin film polymer sensor Vaisala DPA500	RH
	Barometer Vaisala DPA500	P
	Vaisala cup anemometer	WD, WS
Mäkelänkatu	Horiba APNA-370	NO, NO_2, NOx
	Horiba APOA-370 and Thermo Model 49i	O_3
	Horiba APMA-360	CO
	MAAP Thermo Scientific 5012	BC
	Vaisala WXT 520 and Vaisala WXT536	T, P, RH, WD, WS
LCPs	Sensirion SPS30	PM_1, $PM_{2.5}$, PM_4 , PM_{10}
	MiCS-4514	CO, NO_2
	MQ-131	O_3
	BME-280	T, RH, P
	SI1133-AA00-GM	UV

literature by providing a new affordable and approach for BC-level estimating by integrating low-cost sensor calibration with proxy modeling. Our experiments demonstrate that this innovative combination of techniques helps to improve the accuracy of BC estimates significantly.

III. SENSOR MEASUREMENTS AND SYSTEM IMPLEMENTATION

The focus of our research is on providing an affordable, accurate, and portable solution for estimating BC concentrations. We accomplish this by combining modeling that uses proxy variables and intelligent sensor calibration. In this section, we describe the measurements that we use to develop and evaluate our methodology (Sections III-A and III-B) and detail a prototype implementation of our system Section III-C.

A. Reference Sensing Stations

We develop our BC proxy models using air quality data extracted from two high-quality measurement stations, namely, SMEAR III[1] and Mäkelänkatu[2] stations. The SMEAR III station is located in a suburban in the front open yard and its surface includes built, car parking, road, and vegetation areas, whereas the Mäkelänkatu station is located in a street canyon just beside Mäkelänkatu street. Using the measurements from these locations with different air pollution profiles enables developing proxies for estimating the BC concentrations that can work in different environments and thus generalizing our BC proxy model.

The high-quality reference stations are equipped with accurate professional-grade sensors measuring important pollutants and environmental factors. The important pollutants mainly include the PM and gas. Environmental factors include wind direction (WD), wind speed (WS), pressure (P), relative humidity (RH), temperature (T), and, depending on the sensor unit, other related measurements. The sensor types and corresponding measured variables from these two reference stations and low-cost sensor packages (LCPs) are presented in Table I. The reference measurements are used to develop the BC proxy model to explore the feasibility of estimating BC levels on proxy variables using machine learning methods. The two reference stations are described as follows.

SMEAR III: The reference station is operated by the Institute for Atmospheric and Earth System Research (INAR) and is located at the Kumpula Campus area at the University of Helsinki, Finland. The station is located in the front open yard and at about 150 m from a main street in the Kumpula district and it is about 4 kilometers north-east from Helsinki center in Helsinki [58]. The station is planned for research and scientific exploration and it is designed to measure the relationship between forest and atmosphere in boreal climate zone [59]. This site is categorized as a semi-urban area, a distinct surface covered with buildings, roads, and vegetation areas. The station consists of high-quality sensors mounted on a 31 m tall tower, with its base located on a rocky hill at 26 m above sea level. Its sensors can measure PM, gases, and meteorological and radiation variables.

Mäkelänkatu Station: The reference station is located at the Mäkelänkatu district in Helsinki and is operated by the Helsinki Region Environmental Services Authority (HSY). Mäkelänkatu is one of the main streets of the city that leads to the city center. The street is lined with apartment buildings and has 42 m of width. The street consists of six lanes, two tramlines, two rows of trees, and two pavements. Mäkelänkatu Street is one of the arterial roads in the city where every day different kinds of vehicles, such as cars, buses, trams, and trucks cross in it and often cause traffic congestion [60], the reason for having a high level of $PM_{2.5}$ and BC pollution. The traffic is especially high during rush hours, at 8 A.M. and 5 P.M., and it is the main source of BC in this street. This is the main reason that the reference station is placed in the vicinity of the street and it is interesting to measure air quality there. The sensing station consists of a container equipped with standard air quality measurement instruments. Most of the inlets for the measuring devices are located on the top of the container, approximately at a height of 2.8 m from ground level.

B. Low-Cost Sensor Packages

Fig. 1(a) presents one of the LCPs used in our study. Each LCP is built on top of the BMD-340 System on Module (SoM), which is powered with a 3500-mAh battery and enclosed in a 3D-printed case made of ESD-PETG filament. General battery life before recharging via micro USB interface

[1] https://www.atm.helsinki.fi/SMEAR/index.php/smear-iii

[2] https://www.hsy.fi/en/residents/theairyoubreathe/monitoring-stations-helsinki-metropolitan-area/Pages/makelankatu.aspx

is about 26 h. Each LCP is connected to mobile phones via Bluetooth low energy (BLE) for transmitting the measured data to the back-end (server layer). The mobile phones are connected to the server through the 4G network or Wi-Fi. Each LCP reports measurements periodically. The reported readings include T, RH, P, carbon monoxide (CO), nitrous dioxide (NO_2), ozone (O_3), PM of various masses and sizes, the amount ultraviolet (UV) light, the GPS position, and the timestamp [61].

To collect data for calibration and validation purposes, we install four LCPs near high-quality reference stations. By installing four LCPs close to each other in the same environment we aim to ensure sensors' consistency and sensor failures. Whereas sensors' consistency means the LCPs generate similar measurements while operating in the same environment [24]. In addition, by installing multiple LCPs, we plan to recover from the sensor failures whereas if a sensor fails to operate due to power drainage or other reasons other LCPs still continue measurements. As shown in Fig. 1(b), we install the LCPs in pairs facing each other under a rain cover, mounted onto SMEAR III. The LCP presented in Fig. 1(a) is a portable low-cost sensor that can be attach to citizen's bag for tracking the measurement of air pollutants. The micro-sensors installed inside LCP in Fig. 1(a) are the same as the mini-sensors inside the four LCPs installed near the high-quality reference station in Fig. 1(b). While the LCPs in Fig. 1(b) are powered by connecting them to the electricity grid. The LCPs are set up to transmit their readings every 2 min and the LCPs measurement campaign was carried sparsely between November 2019 and February 2020. In the main experiment, we demonstrate how calibration and proxy models can operate together using measurements from low-cost sensors for sensor calibration and BC estimation.

C. System Implementation

The overall methodology has been implemented following the framework shown in Fig. 2. The framework consists of three layers: 1) a sensing layer; 2) a server layer; and 3) an application layer. In the sensing layer, low-cost mini-sensors and stationary reference station sensors continuously measure the air quality and transmit the measurements to the server layer (i.e., back-end). The difference between mobile and stationary mini-sensors is that the portable mini-sensors are connected to mobile phones for transmitting the air quality data to the server (deployed on the edge or in the cloud), while the mini-sensors in reference stations directly transmit the air quality data to the server using available 4G/5G connections through a Rest application programming interface (API). The backend links the measurements with those collected from a professional grade reference station and is responsible for learning the calibration and proxy models used by our approach. Data transmission can take place through any type of wireless medium, including short-range communication system, e.g., WiFi and BLE, or long-range cellular or IoT communication protocols [62]. The data transmission interval can be considered to follow a desired update rate and transmission interval. Our current deployment uses 2-min cycle for sampling air quality measurements, and these are sent to a server hourly.

The server layer is responsible to processing the data. First, the air quality data are preprocessed, synchronized, and the quality of data is checked. Next, the processing pipeline calibrates the measurements of the low-cost sensors. The calibrated data are used both on the portable device to provide information about the air quality, and given as input to the proxy modeling to estimate BC levels. The BC proxy estimates are further transmitted to the application layer where they can be accessed by the end user. Hence, the end users can not only observe the accurate pollutants concentrations of the measured pollutants from the portable device but also access information about the personal BC exposure. The calibration models are currently deployed on portable low-cost sensors and used in two projects to compensate pollutant concentrations. Finally on the application layer, pollution hotspot maps are made according to the data from the portable devices and these can used to support further applications, such as route planning.

IV. BC Estimating With Proxy Variables

We first develop the proxy modeling approach and demonstrate its overall feasibility using data from two reference stations. Specifically, in each location, we estimate the BC levels using the measurements of other pollutants and environmental variables in the same location. The approach of testing on two locations with different characteristics is used to demonstrate the generality of approach. We then further build on this result and develop the proxy modeling and calibration for low-cost sensors, and demonstrate the feasibility and benefits of our overall approach in the subsequent sections.

A. Proxy Estimation Pipeline

The proxy modeling pipeline follows a traditional machine learning pipeline. First, the preprocessing step uses a multivariate imputation by chained equations (MICEs) imputer [63] to fill in missing values. In the experiments, the imputer is trained separately for each training set split, and the same imputer is then used both for the training and testing data. Next, the features are scaled using standardization. Similarly to the imputer, the scaler is learned separately for each training data split. Training the imputer and the scaler only on the training data and separately for each fold to prevent data leakage and keep bias to a minimum.

After preprocessing, we perform feature selection using recursive feature elimination (RFE) on the whole set of features, until only one feature is left. In our experiments, we train the models using the highest ranked feature, adding the rest of the features one by one by following the ranking, and computing the performance for each set of features. This procedure is performed separately for every training split. For estimating BC concentrations, we test machine learning models that have shown robust performance for other pollutants: MLR, SVR, decision tree regression (DTR), adaboost regression (ABR), gradient boosting regression (GBR), RFR, and MLP. Indeed, the machine learning models used in our work are based on regression models. Since we believe

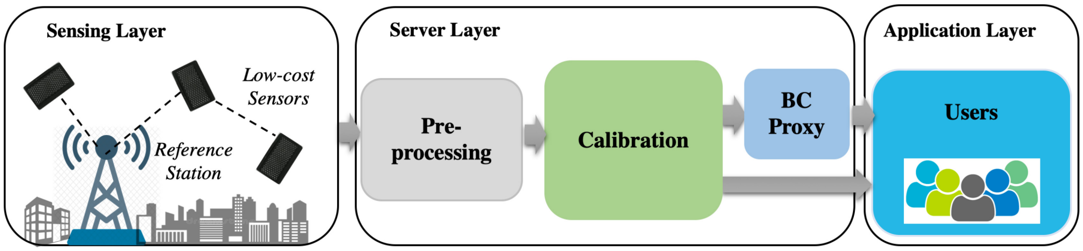


Fig. 2. Framework for the implementation of a BC proxy.

TABLE II
ERROR MEASUREMENTS USED FOR PERFORMANCE EVALUATION

RMSE	MAE	MAPE	MBE	R^2
$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert\cdot 100\%$	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)$	$1-\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

that varying machine learning architecture will not provide significant performance differences, therefore the architecture of the models is determined based on the default settings provided by the machine learning scikit-learn library [64].

B. Performance Evaluation

We evaluate the performance of the different models using tenfold cross-validation. The folds are generated by using the KFold function in the Scikit-learn library in Python and the data set is equally divided into folds according to time due to the characteristics of time-series data. We consider all common error measures, root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), and R-squared (R^2) listed in Table II, where y_i, $\hat{y}_i$, and $\bar{y}_i$ represent the target value, the predicted value, the mean of observed target values, respectively, and n is the number of samples. We use multiple measures because different measures can rank models differently since different measures focus on different aspects of performance [28]. RMSE focuses on outliers, MAE focuses on the average performance, MAPE expresses the error in proportion to the target values, MBE measures bias, and R^2 measures the correlation. We use MAE as the cost function for training the models as we are interested in a model that approximates well the target data on average without focusing on outliers. For every error measure, we obtain the overall score by averaging the scores obtained in every fold. We also plot target diagrams, which allow us to quickly visually compare the performance of different models, in line with best practice in air quality research [65], [66]. Target diagrams are plotted by using centered RMSE (CRMSE) and MBE divided by the standard deviation of the target values. CRMSE is computed similarly to RMSE, but subtracting from the predicted and target values their respective means.

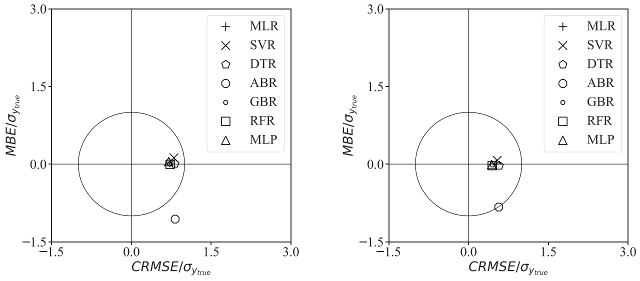
C. Comparing the Models

The results of proxy estimation are shown in Table III. Stations 1 and 2 are used to represent the SMEAR III and Mäkelänkatu reference stations, respectively, for simplicity. On the Station 2 measurements, MLP performs better than MLR, whereas for Station 1 the reverse is true. The Station 1 measurements are from a low-density urban area and from a high altitude above traffic, which results in the relationship between pollutants being simple. In contrast, the Station 2 measurements come from an area with high traffic and from a station that is closer to the street level, which results in a more complex relationship between the variables. SVR has a relatively low error, but it suffers from high bias. The performance of GBR and RFR is accurate with low bias. Overall, these two models have very similar performance across the two reference stations. The ABR by far has the worst performance. This is due to the regression models used by ABR being too simple to capture the complex relationship between variables and hence models that can simultaneously capture relationships between multiple variables are needed for BC estimation.

TABLE III
PERFORMANCE OF SELECTED MACHINE LEARNING MODELS FOR THE ESTIMATION OF BC CONCENTRATION. RMSE, MAE, AND MBE VALUES ARE EXPRESSED IN ng/m^3. RESULTS HAVE BEEN OBTAINED WITH TENFOLD CROSS-VALIDATION ON THE WHOLE AVAILABLE DATA. ALL THE AVAILABLE FEATURES ARE USED

		RMSE	MAE	MAPE	MBE	R^2
SMEAR III	MLR	271.60	167.98	70.49%	2.04	0.553
	SVR	266.14	155.75	61.39%	38.66	0.575
	DTR	404.60	222.34	80.58%	-3.74	-0.144
	ABR	593.37	500.50	289.12%	-452.21	-1.658
	GBR	271.05	158.72	63.07%	-3.63	0.581
	RFR	279.13	160.42	64.20%	-4.85	0.527
	MLP	284.33	178.59	67.91%	10.22	0.552
Mäkelänkatu	MLR	395.01	250.18	36.13%	-0.50	0.777
	SVR	398.71	241.24	30.38%	45.02	0.776
	DTR	530.81	324.67	38.18%	-28.52	0.601
	ABR	758.81	675.54	140.14%	-620.07	0.137
	GBR	355.48	219.36	28.27%	-10.75	0.820
	RFR	359.61	220.76	27.69%	-21.24	0.815
	MLP	344.45	217.09	29.41%	-6.27	0.830

The target diagrams for Stations 1 and 2 are shown in Fig. 3. In the target diagram for Station 1, every model except ABR is inside the circle and all the models inside the circle are very close to each other. In the target diagram for Station 2, the three points which represent GBR, RFR, and MLP overlap.



(a) (b)

Fig. 3. Target diagrams showing the performance of selected machine learning models for estimating BC concentration using proxy variables on high-quality data. (a) Target diagram for Station 1. (b) Target diagram for Station 2.

TABLE IV
Results of Feature Selection for Each Cross-Validation Fold, After Training Two Selected Models on Data From Mäkelänkatu. All the Performance Scores Are Calculated Using MAE, Expressed in ng/m^3

Model	MLR		RFR	
Feature set	Best	All	Best	All
Fold 1	221.93	231.42	203.21	204.27
Fold 2	266.76	266.90	228.77	228.77
Fold 3	260.12	260.16	230.74	230.74
Fold 4	197.48	207.13	190.46	190.53
Fold 5	217.24	221.94	179.12	180.77
Fold 6	244.58	259.00	222.20	227.37
Fold 7	284.73	295.80	281.41	289.66
Fold 8	238.64	297.80	211.23	213.39
Fold 9	228.14	241.47	237.04	247.08
Fold 10	215.34	220.15	190.96	190.96

RFR has the lowest MAPE, but it is the most biased one in those three models. MLP appears to be the best model, since it has the lowest RMSE and MAE, suggesting a good performance on both average values and outliers. It has also the lowest bias of the three and the highest overall correlation with the target variable. Overall, MLR is the least biased model, the best model for Station 1 seems to be GBR, and the best model for Station 2 seems to be MLP.

We also evaluated the performance of using different sets of features. The results are shown in Table IV for two selected models, MLR and RFR. Using only the best feature slightly reduces overall error, but the best-performing feature is different for different folds. Hence, in practice using all available features is sufficient as any potential improvements in performance come at the cost of generality. For this reason, in the remainder of this article, we use all input features for proxy estimation.

Overall, the results show that proxy variables can be used to estimate BC levels. The correlation in the estimates is consistently high, suggesting the proxy variables capture the overall trend in the measurements. The absolute error in the estimates is slightly higher, as can be evidenced from the MAE values. From the target diagrams, we can observe that this is due to a bias in the estimates, i.e., there is a systematic error in the estimates. This result further motivates the use of calibration as part of the pipeline as it enables eliminating the bias in the estimations—besides overcoming inaccuracy in the low-cost sensor measurements.

TABLE V
Correlations Between LCP Sensors. Each Row Indicates the Variable on Which a Correlation Is Computed, Each Column Indicates the Pair of LCP Sensors on Which the Correlation Is Computed

	LCP					
Variable	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
T	0.971	0.973	0.987	0.982	0.964	0.962
RH	0.947	0.964	0.980	0.974	0.948	0.953
P	1.000	1.000	1.000	1.000	1.000	1.000
CO	0.950	0.888	0.887	0.919	0.908	0.985
NO_2	0.996	0.993	0.994	0.994	0.994	0.996
O_3	0.046	0.360	0.104	0.704	0.304	0.126
MPM_1	0.997	0.997	0.997	0.998	0.997	0.997
$MPM_{2.5}$	0.993	0.995	0.990	0.997	0.995	0.994
MPM_4	0.988	0.993	0.985	0.993	0.992	0.988
MPM_{10}	0.990	0.994	0.987	0.989	0.990	0.984
$NPM_{0.5}$	0.998	0.998	0.998	0.997	0.997	0.997
NPM_1	0.998	0.998	0.998	0.998	0.997	0.997
$NPM_{2.5}$	0.998	0.998	0.997	0.998	0.997	0.997
NPM_4	0.998	0.998	0.997	0.998	0.997	0.997
NPM_{10}	0.998	0.998	0.997	0.998	0.997	0.997

V. BC Estimation on Low Cost Sensors

The results in the previous section demonstrated that proxy variables can be used to estimate BC concentrations with reasonable accuracy, at least if we use data from reference stations to train the models. In this section, we further demonstrate that these results generalize to low-cost sensors.

A. Estimation Pipeline

The BC estimation pipeline operates similarly to the proxy variable estimation, first using MICE to impute missing values, followed by feature scaling. As input features for calibration, we use all available features from the low-cost sensors: CO, O_3, all available PM measurements of different sizes, and weather measurements: T, RH, and P.

B. Experiments

We conduct our experiment on the SMEAR III reference station and nearby installed LCPs [Fig. 1(b)]. There are four LCP sensors marked with LCP1, LCP2, LCP3, and LCP4. The correlations between those four LCP sensors are presented in Table V. The LCPs overall have very high consistency, but there were periodic hardware failures on three of the four sensors that affected the O_3 measurements. The O_3 measurements of LCP3 are the most consistent and best match with the reference station, and hence we use this low-cost sensor as a basis in the experiment and evaluation. Note that there are also periods where some data from the reference station—or the LCPs—is not available due to maintenance, hardware failures, connectivity failures, and other factors. Smaller gaps in the data are handled using imputation whereas longer gaps have been excluded from the analysis.

The evaluation using a similar process as in the previous section, i.e., using separate training and testing splits and calculating a wide range of error measurements. We validate the calibration pipeline and the overall BC estimation pipeline using experiments with a separate train-test split

(50/50) instead of cross-validation due to having less data, for the reasons described above. For a short period (around a week), the ground truth BC measurements are missing due to a hardware failure happening during a holiday period. We impute the values for these missing days by training a proxy model using the reference station measurements only and using the output of this model on the measurements from low-cost sensors as the BC estimate for this period. Air quality measurements are heavily correlated in time [28] and hence removing these measurements would result in discontinuity that breaks the models that are trained on the data. For validating the calibration results, we can only consider periods where the calibrated sensor values of the proxy variables can be compared to the reference values. As shown in Section IV, most of the co-pollutants are only available intermittently. The use of a 50/50 split ensures there are sufficient measurements for training and testing the calibration models for all the variables considered in the evaluation. The reference stations we have used are among the leading observation stations worldwide and hence the issue of missing data unfortunately is a reality that any data modeling approach must address—which is also why we incorporate it into our evaluation. As error metrics we use the same measures as before, i.e., the same error measures as in Section IV, namely, RMSE, MAE, MAPE, MBE, and R^2, and as models we consider MLR, SVR, DTR, ABR, GBR, RFR, and MLP.

To obtain a baseline for comparison, we also evaluate the performance of a model trained on LCP3 against a model trained on SMEAR III. To make the model as similar as possible, instead of using all available features from reference station, we select features as similar as possible to the ones we select from the LCP3, namely, CO, O_3, $PM_{2.5}$, PM_{10}, T, RH, and P. To ensure consistency in the comparison of models trained on SMEAR III and LCP3, we use the same train-test split for every model. This means that the timestamp indices of the training data used to train a Station 1 model are the same as the indices in the training data used to train an LCP3 model and the same is true for the test data.

C. Model Comparison

Table VI compares the BC proxy models trained on the low-cost sensor data to those trained from the reference sensor data. The results generally are very similar and the best-performing models with the lowest bias are GBR and RFR. When trained on LCP3 data, the performance of these two models is close to the performance obtained with reference station data (i.e., SMEAR III). The MAE is slightly higher for the low-cost sensors and the correlation is smaller, due to noise in the measurements. Nevertheless, the same general trend remains and the differences in performance are not significant. Fig. 4 further demonstrates this point by comparing the estimates BC levels between the models trained on LCP3 data and reference station data. The general trend is accurately captured and both models are capable of distinguishing between harmful and nonharmful levels of BC. Indeed, the main difference between the two models comes during the highest BC levels where the lower sensitivity of the low-cost sensors may result in underestimating the overall level of BC.

TABLE VI
PERFORMANCE OF SELECTED MACHINE LEARNING MODELS FOR THE ESTIMATION OF BC CONCENTRATION USING PROXY VARIABLES FROM LOW-COST AIR QUALITY SENSORS. RMSE, MAE, AND MBE VALUES ARE EXPRESSED IN ng/m^3. RESULTS HAVE BEEN OBTAINED WITH A 50-50 VALIDATION SPLIT WITH RANDOM SAMPLING ON THE AVAILABLE DATA. ALL AVAILABLE FEATURES ARE USED

Model	RMSE	MAE	MAPE	MBE	R^2	Source
MLR	221.88	154.53	65.44%	-8.83	0.656	SMEAR
	289.72	182.82	74.97%	-3.37	0.414	LCP 3
SVR	240.77	145.76	58.78%	38.80	0.596	SMEAR
	308.44	166.88	56.37%	70.66	0.336	LCP 3
DTR	171.20	90.80	36.42%	0.94	0.795	SMEAR
	186.52	104.83	41.78%	-1.15	0.757	LCP 3
ABR	208.92	160.33	85.98%	-77.53	0.695	SMEAR
	252.85	206.99	121.07%	-103.96	0.554	LCP 3
GBR	154.04	98.09	44.13%	-4.20	0.834	SMEAR
	177.89	120.40	50.63%	1.45	0.779	LCP 3
RFR	136.12	72.08	32.11%	-2.46	0.871	SMEAR
	145.80	85.59	35.07%	1.63	0.852	LCP 3
MLP	164.18	109.99	45.00%	17.93	0.812	SMEAR
	183.55	124.69	51.46%	-7.16	0.765	LCP 3

We also explore the potential of transferring the models trained on one low-cost sensor to other low-cost devices. The sensors often contain variations across devices and in practice it is not possible to use every device to train the proxy model as this would require co-locating them next to the reference station for a sufficiently long period. In case calibration transfer is possible, then only a small set of sensors could be placed close to a reference station and the other devices could simply use the model trained from these measurements [13]. To test this, we test the models trained on LCP3 against the measurements obtained on the other low-cost sensors. As mentioned, the O_3 sensor on these devices had some hardware issues and hence O_3 was excluded from the model. The experiment is performed by using the best two models: 1) GBR and 2) RFR. Figs. 5 and 6 show the results for the GBR and RFR models. The target diagrams align for all devices and all points are inside the target. This suggests that a model trained on an LCP works well on other LCPs without need to retrain it.

VI. PROXY VARIABLES CALIBRATED ON LOW-COST SENSORS

The previous section demonstrates that BC concentrations can be estimated with reasonable accuracy from proxy variables and on low-cost sensors. As the final step, we demonstrate how calibrating the low-cost sensor measurements that are used as input for the proxy model further improves the overall performance.

A. Calibration of Proxy Variables

As calibration targets, we select variables from Station 1 corresponding to variables available from our LCPs, having a low percentage of missing values in the period of the LCP measurement campaign, namely, T, RH, P, CO, $PM_{2.5}$, and PM_{10}. We remove the samples where values are missing

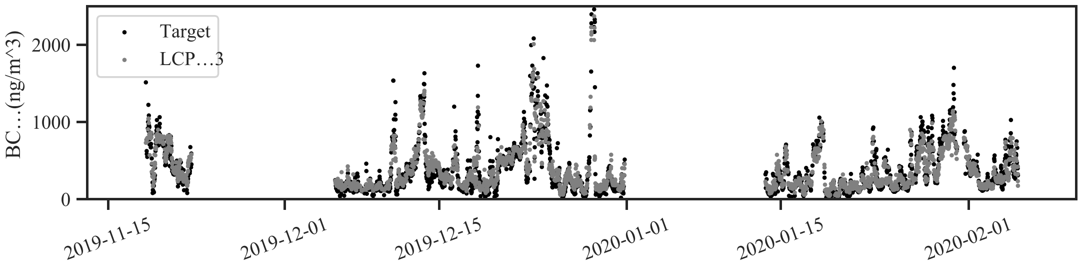


Fig. 4. Target BC values versus values predicted by a BC proxy model trained on LCP 3.

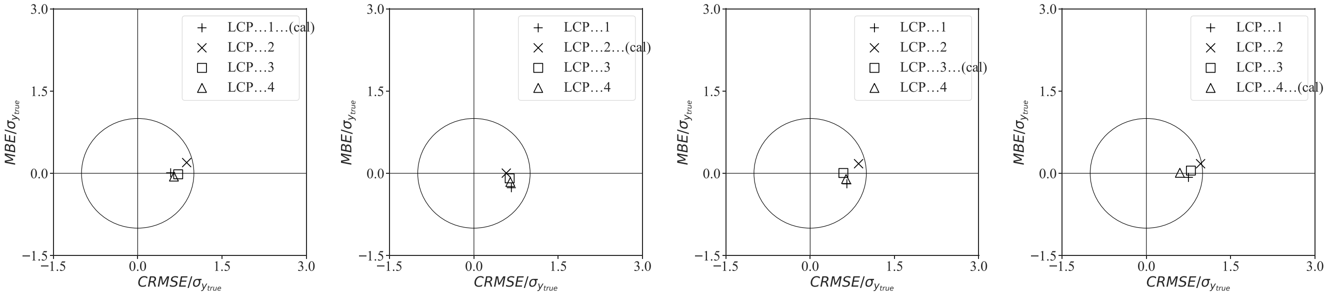


Fig. 5. Target diagrams showing the performance of a GBR model, trained on one LCP and tested on every LCP.

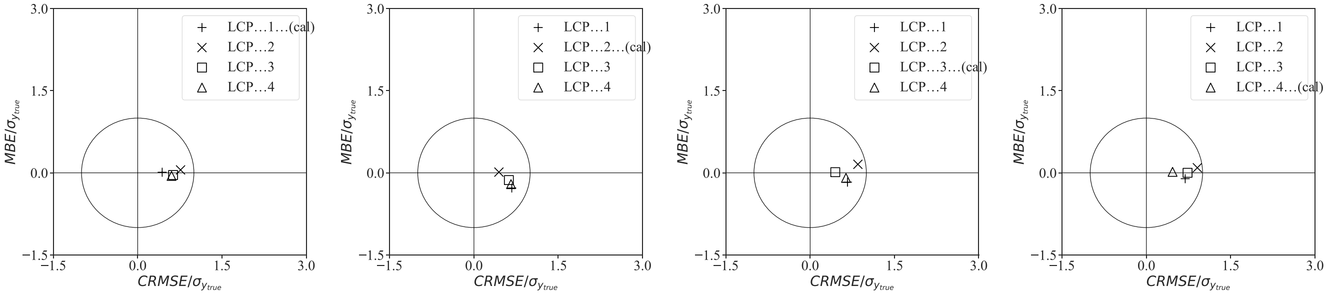


Fig. 6. Target diagrams showing the performance of an RFR model, trained on one LCP and tested on every LCP.

from these target variables, as imputing them would result in data leakage and bias. We perform variable scaling as in the previous parts of our study. We test proxy calibration using GBR and RFR since these have consistently been the best-performing models. For simplicity, we performed the experiments of this part of our study only on a single low-cost sensor as the results of the previous section demonstrated that models transfer across low-cost sensors.

In Table VII, we show the results obtained by calibrating the low-cost sensor measurements to predict pollutants to be used as proxy variables. For T we do not provide MAPE and MAE, as temperature is measured using an interval scale (°C), on which proportions do not make sense. As shown in Table VII, RFR has a better performance than GBR for every variable, therefore we decide to use RFR to perform the rest of the experiments in this part of our study. The errors are low across the board and the correlation is very high, suggesting that the calibration performs well on the measurements.

TABLE VII
CALIBRATION OF PROXY VARIABLES ON LCP 3. MAPE AND MBE ARE MISSING FROM T AS IT IS MEASURED USING AN INTERVAL SCALE (°C), ON WHICH RATIOS AND PERCENTAGES DO NOT MAKE SENSE

		RMSE	MAE	MAPE	MBE	R^2
GBR	T	0.52	0.38	—	—	0.963
	RH	3.74	2.82	3.26%	-0.17	0.818
	P	0.15	0.11	0.01%	-0.00	1.000
	CO	13.43	8.87	4.91%	0.02	0.759
	PM_{25}	1.48	0.95	143.77%	-0.02	0.773
	PM_{10}	3.29	2.06	48.13%	0.06	0.778
RFR	T	0.41	0.26	—	—	0.977
	RH	2.70	1.80	2.09%	-0.04	0.905
	P	0.11	0.07	0.01%	-0.00	1.000
	CO	11.71	6.34	3.48%	-0.17	0.817
	PM_{25}	1.49	0.86	120.56%	-0.03	0.771
	PM_{10}	2.91	1.61	31.91%	0.08	0.826

B. Using Calibrated Proxy Variables to Improve BC Estimation

Using low-cost sensor measurements as proxies to estimate BC level is beneficial as low-cost sensors help to increase the spatial and temporal resolution of information due to higher deployment density. To test the performance of estimating BC concentrations without and with calibrated proxy variables, we test multiple combinations of features and calibrated proxy variables using an RFR model.

TABLE VIII
ESTIMATION OF LCP 3 ON BC WITH AN RFR MODEL, USING COMBINATIONS OF LCP FEATURES AND CALIBRATED PROXY VARIABLES. ONLY RESULTS OF NOTEWORTHY COMBINATIONS ARE SHOWN

LCP	Calibrated proxy features				Performance				
features	T, RH, P	CO	$PM_{2.5}$	PM_{10}	RMSE	MAE	MAPE	MBE	R^2
X					122.65	76.88	31.22%	-10.40	0.896
X	X				123.29	78.38	31.65%	-4.34	0.895
X		X			127.41	79.88	32.07%	-11.14	0.888
X	X	X			125.57	78.92	31.83%	-11.22	0.891
X			X		121.62	78.95	31.82%	-9.59	0.898
X	X		X		116.66	76.41	30.87%	-7.87	0.906
X				X	117.12	77.92	30.70%	-8.44	0.906
X	X			X	116.58	76.31	30.06%	-5.73	0.906
X			X	X	117.72	77.87	30.94%	-9.46	0.905
X	X		X	X	116.13	76.55	29.97%	-6.29	0.907
X		X	X	X	123.24	77.36	30.96%	-10.06	0.895
X	X	X	X	X	122.24	76.63	31.09%	-10.00	0.897
		X	X	X	173.62	119.55	53.72%	-10.74	0.792
	X	X	X	X	149.23	97.47	42.33%	-13.29	0.847
			X	X	270.05	187.00	82.09%	-31.93	0.498
	X		X	X	149.23	97.47	42.33%	-13.29	0.847

In Table VIII, we show the results of selected combinations. As we can see, the best results are obtained with a combination of the regular variables plus calibrated proxy features T, RH, P, and PM_{10}. We obtain the second best results by adding calibrated $PM_{2.5}$. PM_{10} is chosen over $PM_{2.5}$ plausibly because the BC measured underwent an aging process at a high rate, which increases its coating thickness and hence results in a larger diameter [67]. We can also notice that the proxy variable CO does not improve the results, but it even worsens them compared to using LCP features only. This could be because CO emissions from vehicular traffic have decreased to a background level in Helsinki due to three-way catalysts in vehicles. Values close to background level are not beneficial in predicting BC in this study [68]. Using calibrated proxy features only is worse than using a combination of LCP features and calibrated proxy features together.

VII. Discussion and Roadmap

First, in Section IV, we have shown that BC concentrations can be reliably estimated from proxy variables and we have identified the best-performing machine learning for this task, which are GBR, RFR, and MLP. Second, in Section V, we have shown that estimating BC concentration using measurements from low-cost air quality sensors as proxy variables is also feasible, and similarly to the results of the first perspective, the best models are GBR and RFR, with RFR the best overall. We have also compared the results obtained with low-cost air quality sensors to results obtained from Station 1 with the same data and seen that a model built on data from low-cost air quality sensors has a performance close to the same model built on high-quality data. This result is particularly significant as it suggests portable devices carried by citizens could supplement professional-grade stations and often detailed insights into the BC concentrations in urban environments. We have also shown that a model trained on one low-cost sensor is transferrable to other sensors without the need to retrain it. Third, in Section VI, we have shown that prior calibration of low-cost air quality sensors and adding the calibrated variables to the main model for estimating BC concentration further helps to improve performance.

Naturally, our study also presents some limitations. First, as results in Section IV indicate, low-cost sensing components are prone to failures. In our case, O_3 sensors from three low-cost sensors failed and they needed to be removed from the data. In actual deployments, it would be essential to have mechanisms that can automatically validate the measurements and to detect such failures—at least in terms of without needing to manually inspect the values. Second, in terms of analysis, there were also some limitations. We could not perform imputation on Station 1 because it would lead to bias in the calibration of the LCPs and we could not calibrate O_3 because too many values were missing from the target value. Nevertheless, the results were consistent across all combinations that were tested. Indeed, for all imputed combinations and for those cases where O_3 values were available, the results were in line with the results of other variables and data sources, suggesting that the results are robust. Another limitation is the somewhat short measurement campaign for the low-cost sensors. Ideally, a measurement campaign should last at least a year, so that measurements can span across every season and the sensor can be tested in every condition it can encounter outdoors. However, we only had access to sparse data spanning three months during winter for this study. Ensuring sufficient retention for sensor use is a critical issue for low-cost sensors and any measurement campaigns are likely to suffer from the same issue of sparsity and limited data as our measurements. Thus, the limitations in our data reflect the characteristics of real-world data sets.

VIII. Conclusion

We contributed a novel affordable and portable solution for estimating BC concentrations using low-cost air quality monitoring devices and machine learning techniques. Our approach builds on an innovative approach that uses other pollutants and environmental variables as proxies that are used to estimate overall BC concentration. As low-cost sensors tend to suffer from noisy measurements and inaccuracies, we further incorporate sensor calibration to improve the quality of the measurements that are used as inputs for proxy modeling to enable robust and accurate modeling on low-cost sensors. We conducted experiments using a combination of ground-truth measurements from a high-quality measurement station and low-cost sensor measurements from two locations with different urban characteristics. Our results showed that a model trained on low-cost data from sensors for measuring PM of various sizes, CO, NO_2, O_3, and weather variables (T, RH, and P), approximates well the true concentration of BC, almost as accurately as a similar model trained on high-quality data from an atmospheric station. The best-performing machine learning models are GBR and RFR. The results also show that the performance of BC estimates can be improved by adding calibrated proxy variables as features, i.e., the output of models that calibrate low-cost air quality sensors to predict pollutants that correlate with BC. Overall, our research offers a new way to estimate BC using low-cost air quality sensors. This allows the monitoring of BC more densely than using conventional methods, which in turn allows better-estimating health risks faced by individuals, the generation of high-resolution pollution maps, and providing detailed information to support policy making.

References

[1] R. D. Brook et al., "Particulate matter air pollution and cardiovascular disease," *Circulation*, vol. 121, no. 21, pp. 2331–2378, 2010.

[2] M. Goldberg, "A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases," *Rev. Environ. Health*, vol. 23, no. 4, pp. 243–298, 2011.

[3] Z. J. Andersen et al., "Chronic obstructive pulmonary disease and long-term exposure to traffic-related air pollution," *Amer. J. Respir. Crit. Care Med.*, vol. 183, no. 4, pp. 455–461, 2011.

[4] Z. J. Andersen et al., "Stroke and long-term exposure to outdoor air pollution from nitrogen dioxide," *Stroke*, vol. 43, no. 2, pp. 320–325, 2012.

[5] Z. J. Andersen et al., "Long-term exposure to air pollution and asthma hospitalisations in older adults: A cohort study," *Thorax*, vol. 67, no. 1, pp. 6–11, 2012.

[6] U. Gehring et al., "Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life," *Amer. J. Respir. Crit. Care Med.*, vol. 181, no. 6, pp. 596–603, 2010.

[7] O. Raaschou-Nielsen et al., "Air pollution from traffic and cancer incidence: A Danish cohort study," *Environ. Health*, vol. 10, no. 1, p. 67, Jul 2011.

[8] E. Saber and G. Heydari, "Flow patterns and deposition fraction of particles in the range of 0.1–10 μm at trachea and the first third generations under different breathing conditions," *Comput. Biol. Med.*, vol. 42, no. 5, pp. 631–638, 2012.

[9] L. Martelletti and P. Martelletti, "Air pollution and the novel Covid-19 disease: A putative disease risk factor," *SN Compr. Clin. Med.*, vol. 2, no. 4, pp. 383–387, 2020.

[10] L. Setti et al., "The potential role of particulate matter in the spreading of Covid-19 in Northern Italy: First evidence-based research hypotheses," MedRxiv Preprint, 2020. [Online]. Available: https://doi.org/10.1101/2020.04.11.20061713

[11] X. Wu, R. C. Nethery, M. B. Sabath, D. Braun, and F. Dominici, "Exposure to air pollution and Covid-19 mortality in the United States: A nationwide cross-sectional study," medRxiv Preprint, 2020.

[12] E. Conticini, B. Frediani, and D. Caro, "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?" *Environ. Pollut.*, vol. 261, Jun. 2020, Art. no. 114465.

[13] N. H. Motlagh et al., "Toward massive scale air quality monitoring," *IEEE Commun. Mag.*, vol. 58, no. 2, pp. 54–59, Feb. 2020.

[14] *The Economic Consequence of Outdoor Air Pollution*, Org. Econ. Co-Oper. Develop., Paris, France, 2016.

[15] *Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease*, World Health Org., Geneva, Switzerland, 2016.

[16] T. A. J. Kuhlbusch, "Black carbon and the carbon cycle," *Science*, vol. 280, no. 5371, pp. 1903–1904, 1998.

[17] N. A. Janssen et al., *Health Effects of Black Carbon*, World Health Org., Geneva, Switzerland, 2012.

[18] V. Ramanathan and G. Carmichael, "Global and regional climate changes due to black carbon," *Nat. Geosci.*, vol. 1, no. 4, pp. 221–227, 2008.

[19] S. Verma, S. Ghosh, O. Boucher, R. Wang, and L. Menut, "Black carbon health impacts in the indo-gangetic plain: Exposures, risks, and mitigation," *Sci. Adv.*, vol. 8, no. 31, 2022, Art. no. eabo4093.

[20] M. A. Zaidan et al., "Dense air quality sensor networks: Validation, analysis, and benefits," *IEEE Sensors J.*, vol. 22, no. 23, pp. 23507–23520, Dec. 2022.

[21] Y. Gao et al., "Mosaic: A low-cost mobile sensing system for urban air quality monitoring," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[22] B. Maag, Z. Zhou, and L. Thiele, "W-Air: Enabling personal air pollution monitoring on wearables," *Proc. ACM Interact. Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 24, Mar. 2018.

[23] N. H. Motlagh et al., "Indoor air quality monitoring using infrastructure-based motion detectors," in *Proc. IEEE 17th Int. Conf. Ind. Informat. (INDIN)*, vol. 1, Jul. 2019, pp. 902–907.

[24] N. H. Motlagh et al., "Air quality sensing process using low-cost sensors: Validation by indoor-outdoor measurements," in *Proc. 15th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2020, pp. 223–228.

[25] E. Lagerspetz et al., "MegaSense: Feasibility of low-cost sensors for pollution hot-spot detection," in *Proc. IEEE 17th Int. Conf. Ind. Informat. (INDIN)*, vol. 1, Jul. 2019, pp. 1083–1090.

[26] M. A. Zaidan et al., "Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies," *Appl. Sci.*, vol. 9, no. 20, p. 4475, 2019.

[27] C. Borrego et al., "Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise—Part II," *Atmos. Environ.*, vol. 193, pp. 127–142, Nov. 2018.

[28] F. Concas et al., "Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis," *ACM Trans. Sens. Netw.* , vol. 17, no. 2, pp. 1–44, 2021.

[29] M. A. Zaidan et al., "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13638–13652, Nov. 2020.

[30] B. Maag, Z. Zhou, O. Saukh, and L. Thiele, "Scan: Multi-hop calibration for mobile sensor arrays," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–21, 2017.

[31] M. A. Zaidan et al., "Intelligent air pollution sensors calibration for extreme events and drifts monitoring," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1366–1379, Feb. 2023.

[32] Y. Cheng, X. He, Z. Zhou, and L. Thiele, "ICT: In-field calibration transfer for air quality sensor deployments," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 1, pp. 1–19, 2019.

[33] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," in *Proc. 2nd Int. Workshop Mobile Sens.*, 2012, pp. 1–5.

[34] C. Lin, J. Gillespie, M. Schuder, W. Duberstein, I. Beverland, and M. Heal, "Evaluation and calibration of Aeroqual series 500 portable gas sensors for accurate measurement of ambient ozone and nitrogen dioxide," *Atmos. Environ.*, vol. 100, pp. 111–116, Jan. 2015.

[35] O. Saukh, D. Hasenfratz, and L. Thiele, "Reducing multi-hop calibration errors in large-scale mobile sensor networks," in *Proc. 14th Int. Conf. Inf. Process. Sens. Netw.*, 2015, pp. 274–285.

[36] B. Maag, O. Saukh, D. Hasenfratz, and L. Thiele, "Pre-deployment testing, augmentation and calibration of cross-sensitive sensors," in *Proc. Int. Conf. Embedded Wireless Syst. Netw.*, 2016, pp. 169–180.

[37] H. Liu, H. Wu, H. Lee, Y. Ho, and L. Chen, "A system calibration model for mobile PM2.5 sensing using low-cost sensors," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jun. 2017, pp. 611–618.

[38] J. M. Cordero, R. Borge, and A. Narros, "Using statistical methods to carry out in field calibrations of low cost air quality sensors," *Sens. Actuat. B, Chem.*, vol. 267, pp. 245–254, Aug. 2018.

[39] N. Zimmerman et al., "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring," *Atmos. Meas. Techn.*, vol. 11, no. 1, pp. 291–313, 2018.

[40] C. Borrego et al., "Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise," *Atmos. Environ.*, vol. 147, pp. 246–263, Dec. 2016.

[41] Y. Lin, W. Dong, and Y. Chen, "Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–18, 2018.

[42] Y. Cheng et al., "Aircloud: A cloud-based air-quality monitoring system for everyone," in *Proc. 12th ACM Conf. Embedded Netw. Sens. Syst.*, 2014, pp. 251–265.

[43] E. Esposito, S. D. Vito, M. Salvato, V. Bright, R. Jones, and O. Popoola, "Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems," *Sens. Actuat. B, Chem.*, vol. 231, pp. 701–713, Aug. 2016.

[44] L. Chen, Y. Ding, D. Lyu, X. Liu, and H. Long, "Deep multi-task learning based urban air quality index modelling," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 1, pp. 1–17, 2019.

[45] W. F. Cooke and J. J. Wilson, "A global black carbon aerosol model," *J. Geophys. Res. Atmos.*, vol. 101, no. D14, pp. 19395–19409, 1996.

[46] M. Liu, X. Peng, Z. Meng, T. Zhou, L. Long, and Q. She, "Spatial characteristics and determinants of in-traffic black carbon in Shanghai, China: Combination of mobile monitoring and land use regression model," *Sci. Total Environ.*, vol. 658, pp. 51–61, Mar. 2019.

[47] J. Yang, S. Kang, Z. Ji, and D. Chen, "Modeling the origin of anthropogenic black carbon and its climatic effect over the tibetan plateau and surrounding regions," *J. Geophys. Res. Atmos.*, vol. 123, no. 2, pp. 671–692, 2018.

[48] G. Curci et al., "Modelling black carbon absorption of solar radiation: Combining external and internal mixing assumptions," *Atmos. Chem. Phys.*, vol. 19, no. 1, p. 181, 2019.
[49] M. Kahnert and F. Kanngießer, "Modelling optical properties of atmospheric black carbon aerosols," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 244, Mar. 2020, Art. no. 106849.
[50] I. Rivas et al., "Determinants of black carbon, particle mass and number concentrations in London transport microenvironments," *Atmos. Environ.*, vol. 161, pp. 247–262, Jul. 2017.
[51] Y. A. Awad, P. Koutrakis, B. A. Coull, and J. Schwartz, "A spatio-temporal prediction model based on support vector machine regression: Ambient black carbon in three new England states," *Environ. Res.*, vol. 159, pp. 427–434, Nov. 2017.
[52] M. Lee et al., "Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong," *Sci. Total Environ.*, vol. 592, pp. 306–315, Aug. 2017.
[53] K. Gu, J. Qiao, and W. Lin, "Recurrent air quality predictor based on meteorology-and pollution-related factors," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3946–3955, Sep. 2018.
[54] P. L. Fung et al., "Input-adaptive proxy for black carbon as a virtual sensor," *Sensors*, vol. 20, no. 1, p. 182, 2020.
[55] J. Rovira et al., "Non-linear models for black carbon exposure modelling using air pollution datasets," *Environ. Res.*, vol. 212, Sep. 2022, Art. no. 113269.
[56] M. A. Zaidan, D. Wraith, B. E. Boor, and T. Hussein, "Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models," *Appl. Sci.*, vol. 9, no. 22, p. 4976, 2019.
[57] P. L. Fung et al., "Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration," *J. Aerosol Sci.*, vol. 152, Feb. 2021, Art. no. 105694.
[58] L. Järvi et al., "The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface-atmosphere interactions in Helsinki, Finland," *Boreal Environ. Res.*, vol. 14, pp. 86–109, Apr. 2009.
[59] M. Kulmala, "Build a global earth observatory," *Nature*, vol. 553, no. 7686, pp. 21–23, 2018.
[60] R. Hietikko et al., "Diurnal variation of nanocluster aerosol concentrations and emission factors in a street canyon," *Atmos. Environ.*, vol. 189, pp. 98–106, Sep. 2018.
[61] A. Rebeiro-Hargrave, N. H. Motlagh, S. Varjonen, E. Lagerspetz, P. Nurmi, and S. Tarkoma, "MegaSense: Cyber-physical system for real-time urban air quality monitoring," in *Proc. 15th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, 2020, pp. 1–6.
[62] X. Su et al., "Intelligent and scalable air quality monitoring with 5G edge," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 35–44, Mar./Apr. 2021.
[63] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Stat. Med.*, vol. 30, no. 4, pp. 377–399, 2011.
[64] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
[65] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide," *Sens. Actuat. B, Chem.*, vol. 215, pp. 249–257, Aug. 2015.
[66] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2," *Sens. Actuat. B, Chem.*, vol. 238, pp. 706–715, Jan. 2017.
[67] Y. Wang et al., "Constraining aging processes of black carbon in the community atmosphere model using environmental chamber measurements," *J. Adv. Model. Earth Syst.*, vol. 10, no. 10, pp. 2514–2526, 2018.
[68] S. Sillanpää et al., "Long-term air quality trends of regulated pollutants in the Helsinki metropolitan area from 1994–2019 and its implications to the air quality index," *Boreal Environ. Res.*, vol. 27, nos. 1–6, pp. 1–31, 2022.

Xiaoli Liu received the Ph.D. degree in mathematics and statistics from the University of Helsinki, Helsinki, Finland, in 2017.

She is a Research Coordinator with the Department of Computer Science, University of Helsinki. Her research interests include data analysis, distributed learning and inference, Internet of Things, and augmented reality.

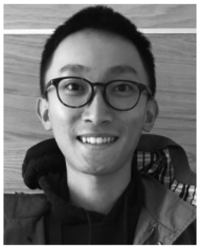
Francesco Concas received the M.Sc. degree in computer science from the University of Helsinki, Helsinki, Finland, in 2018, where he is currently pursuing the Doctoral degree with the Department of Computer Science.

Naser Hossein Motlagh received the D.Sc. degree in networking technology from the School of Electrical Engineering, Aalto University, Espoo, Finland, in 2018.

He is a Senior Researcher with the Department of Computer Science, University of Helsinki, Helsinki, Finland, within the Nokia Center for Advanced Research. His research interests include Internet of Things, wireless sensor networks, environmental sensing, smart buildings, and unmanned aerial and underwater vehicles.

Martha Arbayani Zaidan (Senior Member, IEEE) received the Ph.D. degree in automatic control and systems engineering from Sheffield University, Sheffield, U.K., in 2014.

He currently acts as an Academy Research Fellow and the Data Scientist with the Department of Computer Science and Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland. His research interests include artificial intelligence and machine learning for intelligent control systems, health monitoring technologies, applied physics, atmospheric, and environmental sciences.



Pak Lun Fung received the Ph.D. degree in atmospheric science from the Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland, in 2022.

His dissertation focuses on the proxy derivation of air quality parameters in urban regions. His current research topics include microscopic traffic emission modeling and geospatial analysis on environmental data.

Samu Varjonen received the Ph.D. degree and the Docent title in computer science from the University of Helsinki, Helsinki, Finland, in 2012 and 2022, respectively.

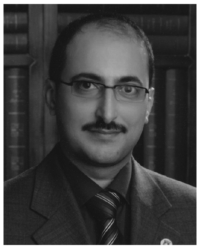
He is a Docent Contract with the Department of Computer Science, University of Helsinki and a Senior Research Scientist with Finnish Meteorological Institute, Helsinki. His research interests include Internet of Things, overlay networks, and sensors.

Jarkko V. Niemi received the Ph.D. degree in environmental sciences from the University of Helsinki, Helsinki, Finland, in 2007.

He is a Senior Specialist of Air Quality with the Helsinki Region Environmental Services Authority, Helsinki. His research interests include urban air quality, physico-chemical characterization and source identification on particles, air quality monitoring methods, and air pollution mitigation measures.

Hilkka Timonen received the Ph.D. degree from the University of Helsinki, Helsinki, Finland, in 2011, and the title of Docent from Tampere University, Tampere, Finland, in 2016.

She is a Senior Research Scientist and the Head of Aerosol Composition Group, Finnish Meteorological Institute, Helsinki. Her research interests include atmospheric aerosol, organic aerosol, black carbon, aerosol mass spectrometry, and emissions measurements.



Tareq Hussein received the Ph.D. degree in atmospheric physics from the Division of Atmospheric Sciences, University of Helsinki, Helsinki, Finland, in 2005, and the Docent title in physics from the University of Helsinki in 2008.

He is a Professor of Atmospheric Sciences with the Institute for Atmospheric and Earth System Research (INAR/Physics), University of Helsinki and also with the Environmental and Atmospheric Research Laboratory, Department of Physics, University of Jordan, Amman, Jordan. His research interests include urban and indoor air quality and exposure.

Tuukka Petäjä received the Ph.D. degree and the Docent title in physics from the University of Helsinki, Helsinki, Finland, in 2006 and 2011, respectively.

He was a Postdoctoral Researcher with the U.S. National Center for Atmospheric Research, Boulder, CO, USA. He is a Professor of Experimental Atmospheric Sciences with the Institute for Atmospheric and Earth System Research, University of Helsinki. His research interest includes atmospheric aerosol particles and their role in climate change and air quality.

Markku Kulmala received the Ph.D. degree in theoretical physics from the University of Helsinki, Helsinki, Finland, in 1988.

He is an Academy Professor and the Head of Institute for Atmospheric and Earth System Research, University of Helsinki. He is the Founder of the International (Station for Measuring Ecosystem-Atmosphere Relations) SMEAR observation networks. He has published more than 800 SCI articles (including more than 40 articles in Science and Nature). His research interests include atmospheric aerosol nucleation and growth mechanisms, kinetics of atmospheric aerosols and clusters, and biosphere–aerosol–cloud–climate interactions.

Dr. Kulmala is currently a member of the Academy of Europe and a Foreign Academician of the Chinese Academy of Sciences.

Petteri Nurmi received the Ph.D. degree in computer science from the Department of Computer Science, University of Helsinki, Helsinki, Finland, in 2009.

He is a Professor of Distributed Systems and Internet of Things with the Department of Computer Science, University of Helsinki. His research interests include distributed systems, pervasive data science, and sensing systems.

Sasu Tarkoma (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Helsinki, Helsinki, Finland, in 2006.

He is a Professor of Computer Science with the University of Helsinki. He is a Visiting Professor with the 6G Flagship, University of Oulu, Oulu, Finland. He has authored four textbooks and has published over 500 scientific articles. He holds ten granted U.S. patents. His research interests include Internet technology, distributed systems, 6G, and mobile and ubiquitous computing.