



# Robust Proxy Sensor Model for Estimating Black Carbon Concentrations Using Low-Cost Sensors

Juan Paredes-Ahumada  
Pau Ferrer-Cid  
Jose M. Barcelo-Ordinas  
Jorge Garcia-Vidal  
juan.antonio.paredes@upc.edu  
pau.ferrer.cid@upc.edu  
jose.maria.barcelo@upc.edu  
jorge.garcia@upc.edu

Universitat Politecnica de Catalunya (UPC)  
Barcelona, Spain

Cristina Reche  
Mar Viana  
cristina.reche@idaea.csic.es  
mar.viana@idaea.csic.es

Environmental Assessment and Water Research, Spanish  
National Research Council (IDAEA-CSIC)  
Barcelona, Spain

## ABSTRACT

Air quality monitoring sensor networks focusing on air pollution measure pollutants that are regulated by the authorities, such as CO, NO<sub>2</sub>, NO, SO<sub>2</sub>, O<sub>3</sub>, and particulate matter (PM<sub>10</sub>, PM<sub>2.5</sub>). However, there are other pollutants, such as black carbon (BC), which are not regulated, have a major impact on health, and are rarely measured. One solution is to use proxies, which consist of creating a mathematical model that infers the measurement of the pollutant from indirect measurements of other pollutants. In this paper, we propose a robust machine learning proxy (RMLP) framework for estimating BC based on nonlinear machine learning methods, calibrating the low-cost sensors (LCSs), and adding robustness against noise and data missing in the LCS. We show the impact of LCS data aggregation, denoising and missing imputation on BC estimation, and how the concentrations estimated by the BC proxy approximate the values obtained by a reference instrument with an accurate BC sensor.

## CCS CONCEPTS

• **Networks** → **Sensor networks**; • **Computing methodologies** → **Machine learning**.

## KEYWORDS

Low-cost sensors, proxy sensors, air quality monitoring networks

### ACM Reference Format:

Juan Paredes-Ahumada, Pau Ferrer-Cid, Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Cristina Reche, and Mar Viana. 2023. Robust Proxy Sensor Model for Estimating Black Carbon Concentrations Using Low-Cost Sensors. In *1st International Workshop on Advances in Environmental Sensing Systems for Smart Cities (EnvSys '23)*, June 18, 2023, Helsinki, Finland. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3597064.3597316>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*EnvSys '23, June 18, 2023, Helsinki, Finland*

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0214-3/23/06...\$15.00

<https://doi.org/10.1145/3597064.3597316>

## 1 INTRODUCTION

Black carbon (BC) is a major component of fine particulate matter, a potent warming agent in the atmosphere which contributes to regional environmental disruption and accelerates glacier melting. BC appears from incomplete combustion and comes mainly from road traffic, it is present in urban aerosols and is linked to cardiovascular and respiratory diseases. The monitoring of BC is not easy and is not regulated by the European Union (EU) Air Quality Directives. In contrast to regulated pollutants, there is not much affordable equipment to monitor BC, which makes the availability of BC measurements operationally expensive and difficult. The new proposal for a Directive of the EU Parliament on ambient air quality and cleaner air for Europe, published in October 2022, states that introducing additional sampling points for unregulated air pollutants of emerging concern, such as ultrafine particles, black carbon, ammonia or particulate matter (PM), will support scientific understanding of their effects on health and the environment. This supports the need to determine BC concentrations, either by direct or indirect methods, in European urban areas. In recent years there has been a great interest in using LCSs to measure regulated pollutants such as CO, NO<sub>2</sub>, NO, SO<sub>2</sub>, O<sub>3</sub>, and PM. Such sensors have a cost ranging from a few tens of Euros to a few hundred Euros, which makes their use and deployment affordable [1, 2]. One way to increase the availability of BC measurements without the need to deploy expensive equipment is the use of virtual sensors. A virtual sensor is defined as a mathematical model that estimates the target phenomenon at a specific location where no physical sensor is available [10]. For example, Zaidan *et al.* [14] propose the use of virtual sensors to calibrate CO<sub>2</sub> sensors and to estimate BC concentrations, demonstrating that black box calibrators are more efficient than white box calibrators. Fung *et al.* [8] also compare the use of white box models with several black box models for building a BC calibrator, including random forest (RF), support vector regression (SVR), artificial neural networks (ANN) and long short-term memory (LSTM) methods, showing the potential use of these methods for building BC virtual sensors. Ferrer-cid *et al.* [3] use signal reconstruction techniques on top of a graph whose relationships are learned using data-driven techniques for building a virtual sensor that imputes missing values in an O<sub>3</sub> monitoring IoT network.

A proxy is a specific type of virtual sensor that estimates a target pollutant from indirect sensor measurements. Zaidan *et al.* [13] propose a mutual information approach to select the most relevant proxy inputs to estimate ozone concentration. Fung *et al.* [7] develop an input-adaptive proxy for BC, which selects input variables of other air quality sensor values based on their correlation coefficients with the output variable. The model uses ordinary least squares as the fitting method. Moreover, a white-box and a black-box model have been compared as a proxy for estimating BC concentrations [15] using a Bayesian neural network, and Rovira *et al.* [12] evaluated the use of precision instrumentation for the creation of BC proxies. In this work, we propose a robust machine learning proxy (RMLP) framework for estimating BC based on nonlinear machine learning methods, calibrating the LCSs, and adding robustness against noise and data missing in the LCS. We focus the goal of the paper on the methodology to build the proxy, highlighting the challenges in the proxy construction process, such as the fact that the data come from different LCS nodes, with different sampling granularity and with gaps at different time stamps. The outline of this paper is as follows: section 2 defines how the proxy is built. Section 3 describes the data set used in the study. Section 4 shows the results obtained with the BC proxy. Finally, section 5 presents the conclusions of the paper.

## 2 ROBUST MACHINE LEARNING PROXY (RMLP) SENSOR MODEL FOR BLACK CARBON ESTIMATION

In the process of creating a data-driven robust proxy for BC estimation, it is necessary to consider the following aspects: i) LCSs calibration, ii) sensors may have sampled each phenomenon involved in the proxy with different time granularity, iii) LCS measurements may be noisy, iv) some of the data captured by the sensors may have gaps, and finally, v) the selection of a model and identification of the features involved in the building of the proxy.

### 2.1 In-situ low-cost sensor calibration

LCSs for air quality monitoring are often not calibrated for the ambient conditions under which they will provide data. This causes that the data provided by these instruments lack sufficient accuracy due to low signal-to-noise ratios or interference from environmental factors. This poses a major problem, as the quality of the sensor data cannot be assessed [5]. To improve the quality of these sensors, LCSs are calibrated in-situ with a supervised machine learning model. To do this, a sensor is co-located with a reference instrument for a period of time. Due to cross-sensitivities and correlations, different sensor calibration models may require different sensor inputs. The response of low-cost particulate matter sensors and gas sensors depends on temperature and relative humidity, and can have other cross-sensitivities with other gas sensors. We refer to papers [1, 2, 4, 5] for in-situ calibration mechanisms in air quality LCSs.

### 2.2 Data aggregation

The set of LCSs used to generate the proxy does not necessarily come from the same node, and therefore each signal may have been sampled at a different frequency and aggregated at a different time interval. In addition, as can be seen in Table 1, to calibrate in-situ each of the LCSs, the period to which the reference values have

been aggregated may also be different from the BC of the reference instrument. Calibrating in-situ with time intervals different from the aggregated time interval of the reference station can have a strong impact on the performance of the calibration method [6]. We will investigate in the results section the impact of aggregating data.

### 2.3 Denoising

The denoising process consists of removing the noise introduced during the data acquisition process, improving the quality of the data. In this sense, during the training phase, the raw LCS measurements are denoised before being corrected by the calibration model. During the estimation phase, the incoming LCS measurements are also corrected by the denoising filter. We have chosen the Savitzky-Golay smoothing filter as the denoising method, since the data streams follow a temporal trend. For this filter, we select a symmetric window around the point to be corrected and fit a polynomial of degree  $p$  in the least-squares sense. The smoothed value will be the estimate of the polynomial at the central point  $t_i$  of the time window [9].

### 2.4 Data imputation

The construction of a BC proxy requires simultaneous measurements of all predictors and no loss of any of them. However, it is possible that for various reasons, such as hardware or communication failures, there may be missing values. In our case the gaps are of the missing completely at random (MCAR) type where the missing values do not depend on any of the variables, either the observed or measured values. Data imputation is the process of filling in missing values with estimated values obtained from the same data set. One approach used in the literature [11] is the multiple imputation process, which uses the entire data set to perform the imputation. This process consists of filling in the missing entries for a variable in the predictor matrix several times based on the other measured variables. Data imputation is an open research field with several approaches proposed in the literature. For imputing the missing values we test multivariate imputation by chained equations (MICE) [11].

### 2.5 Building the proxy sensor model

The BC proxy model consists of a nonlinear data-driven method. We denote by  $y_{BC} \in \mathbb{R}^M$  the BC values provided by the reference instrumentation, where  $M$  is the number of samples. Then, we can group the different LCS calibrated measurements into a sensor matrix  $X_{S_{cal}} \in \mathbb{R}^{M \times P_S}$ , where  $P_S = |S|$  is the dimension of the calibrated sensor set, i.e.,  $S$  and  $P_S$  are respectively the set of sensors and the number of sensors participating in the proxy process. As the number of sensors used as predictors increases, there is a higher probability of having missing values at some point, which makes it costly and difficult to build a successful BC proxy model and also increases the probability of overfitting. For this reason, the set of predictors was iteratively reduced using a backward feature elimination (BFE) algorithm. The BFE algorithm starts with the full set of predictors. At each iteration, the predictor that has the least impact on the model is removed from the model. The process is repeated until only a single predictor remains. We can select the best subset of  $S_{FS}$  sensors to use as predictors, where  $S$  is the set of available sensors:

$$S \xrightarrow[BFE]{} S_{FS} \subset S \quad (1)$$

**Algorithm 1** RMLP for black carbon estimation.

---

**Input:**  $\{S, X_S, Y_{ref}, f_{cal_s}(\cdot), y_{BC}, T_{agg}, f_{proxy}(\cdot)\}$

---

```

    ▶ Obtain LCS calibrated data for the proxy
1:  $X_S \leftarrow \text{Denoising\_Model}(X_S)$ 
2: for  $s \in S$  do
3:   if  $s$  is calibrated then
4:      $x_{s_{cal}} \leftarrow \text{Get\_Sensor}(X_S)$ 
5:   else
6:      $y_s \leftarrow \text{Get\_Ref}(Y_{ref})$ 
7:      $Z_s \leftarrow \text{Select\_Features}(X_S)$ 
8:      $x_{s_{cal}}, \Theta_s \leftarrow \text{Calibrate\_LCS}(Z_s, y_s, f_{cal_s}(\cdot))$ 
9:   end if
10:   $X_{S_{cal}} \leftarrow \text{Add\_To\_Proxy\_Training\_Matrix}(x_{s_{cal}})$ 
11: end for
    ▶ Train proxy model
12:  $t_{max} \leftarrow \text{Max\_TimeInterval\_Proxy}(T_{agg})$ 
13:  $X_{S_{cal}} \leftarrow \text{Aggregation}(X_{S_{cal}}, t_{max})$ 
14:  $y_{BC} \leftarrow \text{Aggregation}(y_{BC}, t_{max})$ 
15:  $\Theta_{imputation} \leftarrow \text{Train\_Imputation\_Model}(S_{FS}, X_{S_{cal}})$ 
16:  $S_{FS}, \Theta_{proxy} \leftarrow \text{BFE}(X_{S_{cal}}, y_{BC}, f_{proxy}(\cdot))$ 
    ▶ Robust proxy model for BC estimation of new measurements
17: while  $x_{new}$  do
18:   for  $s \in S_{FS}$  do
19:     if  $t_s < t_{max}$  then
20:        $x_{new_s} \leftarrow \text{Aggregation}(x_{new_s}, t_{max})$ 
21:     end if
22:   end for
23:    $x_{new} \leftarrow \text{Denoising\_Model}(x_{new})$ 
24:   for  $s \in S_{FS}$  do
25:      $x_{new} \leftarrow f_{cal_s}(x_{new_s}, \Theta_s)$ 
26:   end for
27:    $x_{new} \leftarrow \text{Imputation\_Model}(x_{new}, \Theta_{imputation})$ 
28:    $\tilde{x}_{BC} \leftarrow f_{proxy}(x_{new}, \Theta_{proxy})$ 
29: end while

```

---

Now, the data matrix involved in the design of the proxy model is given by  $X_{S_{FS}} \in \mathbb{R}^{M \times P_{FS}}$ , where  $P_{FS} = |S_{FS}|$  is the dimension of the sensor array selected by the BFE algorithm. We note that  $X_{S_{FS}} \subset X_S$ , as it only includes those features selected by the BFS mechanism. The data-driven proxy model, then, can be defined as:

$$y_{BC_i} \approx f_{proxy}(x_{S_{FS}_i}), \quad i = 1, \dots, M \quad (2)$$

where  $f_{proxy}: \mathbb{R}^{P_{FS}} \rightarrow \mathbb{R}$  is the function that estimates the BC concentrations and depending on the model's assumptions different machine learning models can be applied to learn function  $f_{proxy}(\cdot)$ . For modeling the proxy function  $f_{proxy}(\cdot)$ , we propose to compare three nonlinear models: support vector regression (SVR), random forest (RF), and an artificial neural network (ANN). The SVR is a kernel method that maps the data into a higher feature dimensional space finding the regression curve while the calculations are done in the input space through a kernel function. The RF method combines several decision trees by sampling the data set via bootstrapping. Finally, an ANN consists of layers of interconnected nodes. Each node receives as input a linear combination of the values of the nodes in the previous layer, which is then mapped via a nonlinear activation function. We use a fully connected, feed-forward neural network with a single hidden layer with as many nodes per layer as the number of predictors  $P_{FS}$  and a single output for the regression. ReLu was used as the activation function. We note that the BFE mechanism is linked to the supervised machine learning mechanism used. For each supervised mechanism the BFE result can be different and therefore a different set  $S_{FS}$  can be chosen. When the set  $S_{FS}$  is fixed, the trained model will set the hyperparameters to be used in the estimation process.

## 2.6 Robust machine learning proxy (RMLP) sensor model

The goal of the RMLP sensor model is to estimate BC concentration values from other measurements either taken by reference stations or by LCSs in a robust way, meaning that the algorithm has to calibrate the LCSs if they are not calibrated, it has to eliminate the noise in those sensors, and it has to impute values in those sensors that do not take measurements at that instant before estimating the proxy. Algorithm 1 shows the process of estimating BC concentrations using a robust proxy. The input to the algorithm consists of the set of  $S$  sensors participating in the proxy. These sensors can be accurate instrument sensors, calibrated LCSs or LCSs that need to be in-situ calibrated. The data are organised in a data matrix  $X_S \in \mathbb{R}^{M \times P_S}$  that includes the sensors participating in the process. For those sensors that need to be in-situ calibrated, a matrix of reference values  $Y_{ref} \in \mathbb{R}^M$  and a calibration function  $f_{s_{cal}}: \mathbb{R}^{P_{s_{cal}}} \rightarrow \mathbb{R}$  (e.g., MLR or SVR) for each non-calibrated sensor is included, with  $P_{s_{cal}}$  the number of sensors participating in the calibration of sensor  $s$ . The sensors sample at different frequencies and aggregate samples at different time intervals, so we introduce a vector  $T_{agg} \in \mathbb{R}^{P_S}$  with the time intervals at which each sensor produces values. Finally, we have the BC reference values  $y_{BC} \in \mathbb{R}^M$  to train the proxy and a nonlinear function  $f_{proxy}: \mathbb{R}^{P_{FS}} \rightarrow \mathbb{R}$  or model (e.g., SVR, RF or ANN) to train the proxy.

The RMLP algorithm starts denoising the data and then calibrates the non-calibrated sensors (lines 1-11). In this process, each sensor calibrates in-situ using an array of sensors of size  $P_{s_{cal}}$  including those sensors that have cross-sensitivities or environmental sensors that participate in the calibration. Finally, when the sensors are calibrated, a new matrix  $X_{S_{cal}} \in \mathbb{R}^{M \times P_S}$  of calibrated sensor values is created to participate in the proxy design (line 10). The next phase of the process is to train the proxy model (lines 12-16). Since the sensors participate in the proxy sample at different frequencies, the data must be aggregated (lines 12-14) to the largest time interval. Next, a value imputation model (line 15) is trained for each sensor participating in the process. Then, the BFE algorithm selects (line 16) the final sensors  $S_{FS}$  that participate in the proxy, and that minimize the root mean square error (RMSE) of the BC, obtaining the hyperparameters  $\Theta_{proxy}$  of the proxy model. The last phase represents the estimation of a robust proxy for black carbon estimation (lines 17-29). For each new set of measured values  $x_{new} \in \mathbb{R}^{P_{FS}}$ , first the data is aggregated (lines 18-22) to proxy rate, the denoising (line 23) algorithm is performed to remove noise, and the pollutant concentration is estimated using the hyperparameters obtained in the sensor calibration process (lines 24-26). Then, if any values are missing from the array, an imputation mechanism is performed to fill gaps (line 27). Finally the BC concentration is estimated (line 28).

## 3 SENSOR NODES AND DATA SET

We consider two types of measurements, those obtained by reference instrumentation, and those obtained with nodes deploying LCSs. Reference and LCS values were measured at reference station located in Palau Reial (41°23'14"N, 2°6'56"E, 80 m.a.s.l.), Barcelona, Spain. The reference values for  $O_3$ ,  $NO_2$ ,  $NO$  and  $PM_{10}$  were taken from

the values published by the reference station. The reference station does not provide  $PM_{2.5}$  concentrations. The reference values for BC were taken with a multiangle absorption photometer (MAAP, Thermo ESM Andersen Instrument) and N with a water-based condensation particle counter (WCPC TSI 3785) with 1 min and 5 min intervals respectively. The meteorological variables (temperature and relative humidity) were obtained from a meteorological station located on the roof of the Faculty of Physics of the Univ. of Barcelona, about 400 m from the Palau Reial station. All values were taken in the period from 19/10/2021 to 25/12/2021.

**Table 1: Data sets used in the proxy (19/10/2021 to 25/12/2021).**

Variable	# Samples	Time interval	Measurement Source
BC	89361	1 min	Ref. Stat.
O <sub>3</sub>	9527	10 min	Ref. Stat.
NO <sub>2</sub>	9527	10 min	Ref. Stat.
NO	9527	10 min	Ref. Stat.
PM <sub>10</sub>	9527	10 min	Ref. Stat.
N	19040	5 min	Ref. Stat.
T	9521	10 min	Met. Stat.
RH	9521	10 min	Met. Stat.
O <sub>3</sub>	3292242	2 s	LCS
NO <sub>2</sub>	3292242	2 s	LCS
NO	3308335	2 s	LCS
PM <sub>10</sub>	85895	2 min	LCS
PM <sub>2.5</sub>	85895	2 min	LCS
N <sub>i</sub>	85895	2 min	LCS

We used two nodes with LCSs for measuring O<sub>3</sub>, NO<sub>x</sub>, PM<sub>x</sub>. The Captor node is a node built at Universitat Politècnica de Catalunya (UPC), and includes three electrochemical Alphasense sensors; one OX-B431 O<sub>3</sub> sensor, one NO2-B43F NO<sub>2</sub> sensor and one NO-B4 NO sensor, and one DHT1 Grove air temperature (T) and air relative humidity (RH) sensor to measure the internal box environmental temperature and relative humidity. The sampling rate can be reconfigured, and for our experiments it was set to 2 s. The second LCS node is a PurpleAir PA-II node that measures PM<sub>x</sub> concentrations. The PA-II node uses PMS5003 dual laser particle counters that count suspended particles in sizes of N<sub>i</sub> with  $i=\{0.3, 0.5, 1.0, 2.5, 5.0, 10\}$   $\mu m$ . These particle counts are converted to PM mass concentrations in  $\mu g/m^3$  or to the total ultrafine particle number concentration (N). The data can be downloaded with a 2 min time interval. Table 1 summarizes the data sets indicating the time interval of each variable considered for the BC proxy model.

## 4 RESULTS

The methodology for training the proxy is as follows: i) a randomly selected fraction of the data set (75%) was used for training to ensure that the model copes with a wide range of concentrations among the input variables, eliminating variations due to different seasons and times of day, and the remaining fraction (25%) was used to validate the model, ii) a 10-fold cross-validation strategy was used to obtain the model hyperparameters by averaging the root-mean square error (RMSE).

### 4.1 Low-cost sensor calibration

Table 2 shows the RMSE and R<sup>2</sup> values obtained in the calibration of the O<sub>3</sub>, NO<sub>2</sub>, NO, and PM<sub>10</sub> sensors with R<sup>2</sup> ranging from 0.8 to

0.9. Since the SVR calibration results are the best, we will use these estimated values as input values for the BC proxy.

**Table 2: LCS calibration results using MLR and SVR.**

	MLR		SVR	
	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
O <sub>3</sub>	8.14	0.86	7.08	0.90
NO <sub>2</sub>	7.98	0.81	6.70	0.86
NO	7.07	0.83	5.99	0.88
PM <sub>10</sub>	5.22	0.70	4.34	0.79

### 4.2 Machine learning BC proxy performance

As shown in table 1, the pollutants used as predictors for BC concentration have different time interval. Hence, the first step is to add the data to the most restrictive feature, which samples every 10 min. For the period under consideration, 7.6% of BC measurements were missed completely at random, whereas the BC concentration has a mean value of  $0.82 \pm 0.79 \mu g/m^3$ . In this section, we will consider a dataset where we do not denoise and do not consider measurements when any of the variables are missing. We use the BFE algorithm on the reference station data set to determine the optimal subset of predictors  $S_{FS}$  and build a baseline BC proxy (SVR, RF and ANN). Using this subset of predictors we then build a BC proxy using the LCSs included in Captor and commercial nodes, Table 3. The first thing we can observe is that the BFE algorithm chooses different features when using different models. For example SVR chooses as optimal features {O<sub>3</sub>, PM<sub>10</sub>, N, RH, T} and RF selects {O<sub>3</sub>, PM<sub>10</sub>, N, T}, removing RH. On the other hand, ANN adds NO and NO<sub>2</sub> as features to its optimal set {O<sub>3</sub>, NO<sub>2</sub>, NO, PM<sub>10</sub>, N, T}. The second observation is that the BC proxy using LCSs performs close to the BC proxy using features measured by the reference station. Among the models used, we see that SVR offers the best performance with a RMSE=0.37  $\mu g/m^3$  and R<sup>2</sup>=0.76 if we use the data from the reference stations versus a RMSE=0.41  $\mu g/m^3$  and R<sup>2</sup>=0.71 if we use LCSs. These results are in agreement with the results obtained in the same area during 2 years (2018 and 2019 [12]), with reference stations, and where seasonality was also studied.

**Table 3: BC proxy comparison after backwards feature selection.**

	Predictors subset	Ref. Station		LCS	
		RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
SVR	O <sub>3</sub> , PM <sub>10</sub> , N, T, RH	0.37	0.76	0.41	0.71
RF	O <sub>3</sub> , PM <sub>10</sub> , N, T	0.41	0.71	0.47	0.68
ANN	O <sub>3</sub> , NO <sub>2</sub> , NO, PM <sub>10</sub> , N, T	0.41	0.71	0.41	0.71

A third observation is whether some overfitting is present in the BC proxy calculations with LCSs because the optimal model using BFE is calculated on the reference station data which does perform a cross-validation process to avoid overfitting. We use reference data given that it acts as a baseline case since we know that they are accurate data, whereas if we perform a BFE on the LCS data, the selection of the BFE will be very dependent on the quality of each sensor at every moment. We have run a BFE with SVR on the LCSs to see how different the choice of features is, doing cross-validation as is done with the BFE on reference data. The results

of this experiment showed that the set of predictors is comprised of  $\{O_3, NO_2, PM_{2.5}, N, T, \text{ and } RH\}$ , where in the set appears  $NO_2$ , and  $PM_{10}$  is replaced by  $PM_{2.5}$ , obtaining a slightly better performance ( $RMSE=0.39 \mu g/m^3$ ,  $R^2=0.76$ ) than when using the set fixed by the reference station<sup>1</sup>. Finally, in the case of ANN, a higher number of features have been chosen, namely  $NO_2$  and  $NO$ , which means a higher number of sensors in the array implying a higher energy consumption and node cost. On the other hand, between SVR and RF there is a difference between the RH sensor that is usually integrated with the T sensor. We ran the ANN with the features chosen by SVR, and observed an  $RMSE=0.41 \mu g/m^3$  and  $R^2=0.72$  with data from the reference station and  $RMSE=0.44 \mu g/m^3$  and  $R^2=0.68$  with LCS data. The performance is a little worse, since it does not use its optimal parameters, but the three models have a similar behavior in terms of RMSE and  $R^2$ . To analyse the robustness of the BC proxy in the following sections we will use the SVR model with  $\{O_3, PM_{10}, N, RH, T\}$  features as baseline case.

### 4.3 Impact of data aggregation

Typically, monitoring reference stations report air quality values by aggregating samples at different time intervals, such as 30 and 60 min. LCSs aggregate samples in a similar way and are in-situ calibrated over these time intervals. We study the impact of aggregating the data of the variables involved in the proxy. We construct a BC proxy model using a data set averaged at 30 and 60 min for both the baseline and LCS data. As shown in table 4, for the 30 min averaged data set the performance of the reference station proxy remains almost unchanged, while the performance of the LCS proxy model decreases slightly. For the 60 min averaged data set, the performance of both proxy models worsens, mainly due to the decrease in samples.

**Table 4: Impact of sensor data aggregation on the SVR BC proxy.**

Aggr. time interval	Ref. Station Proxy		LCS Proxy	
	RMSE	$R^2$	RMSE	$R^2$
10 min	0.37	0.76	0.41	0.71
30 min	0.38	0.75	0.43	0.68
60 min	0.43	0.69	0.50	0.58

### 4.4 Robust machine learning proxy performance

LCS and reference stations suffer from missing data and LCS from noise. In order to achieve a robust proxy, we investigate the impact of these two problems in the following sections.

#### 4.4.1 Noise reduction filter for building a BC proxy model.

We apply a Savitzky-Golay noise reduction filter on the raw measurements of the optimal predictor set of LCSs. We use time windows of 1, 2 and 6 hours, in conjunction with low-order polynomials (order  $p=1$  and  $p=2$ ) to smooth the data. The filtered data streams are then calibrated and a new BC proxy model is built via the SVR method. During the estimation phase the new measurements are filtered using the same parameters obtained during the training phase. Table 5 shows the testing phase performance of the denoising procedure. The model mitigates the effects of noise on the measured signals and

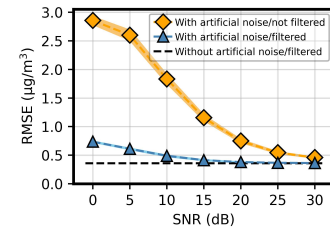
<sup>1</sup>We cannot compare with the reference station using  $PM_{2.5}$  since the reference station did not measure it.

thus has improved performance, with little variation in the specific filter parameters. We can observe that low order filters, e.g.,  $p=1$  or  $p=2$  with small windows of 1 h or 2 h, are the best performers. We choose for the rest of the study a filter of order  $p=2$  and window 1 h.

**Table 5: SVR BC proxy results using a Savitzky-Golay filter.**

	Window (1 h)		Window (2 h)		Window (6 h)	
	$p=1$	$p=2$	$p=1$	$p=2$	$p=1$	$p=2$
RMSE	0.39	0.36	0.37	0.40	0.42	0.38
$R^2$	0.74	0.78	0.77	0.72	0.70	0.75

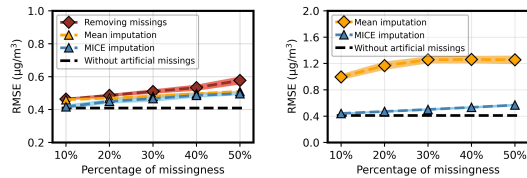
We simulate artificial noise on all predictors in the test set with a signal-to-noise ratio (SNR) ranging from 0 to 30 dB. The RMLP trained on a filtered data set (window=1 h,  $p=2$ ), is used to estimate the BC concentration. Figure 1 shows the RMLP performance with noise. For all noise values studied, the performance improves after applying the noise reduction filter. A degradation of the BC estimation is observed when the SNR is below 25 dB. In contrast, applying the RMLP, it is observed that this degradation occurs with noise that produces SNRs below 10 dB, indicating the filter's ability to improve the robustness of the RMLP model in estimating BC in the presence of noise in the sensor streams.



**Figure 1: Impact of noise in the BC proxy estimation.**

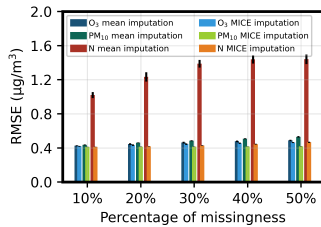
**4.4.2 Impact of missing data imputation.** We simulate missing data completely at random (MCAR) where the missingness is independent on either the observed values or the missing values. The missingness occurs simultaneously to all the predictors with percentages ranging from 10% to 50% of the data size for each predictor. The values of each sensor in the training data set are imputed using the MICE algorithm, and then a BC proxy model is trained using the imputed data set. First, to test the suitability of such a proxy, the performance was measured on a test data set with no missing inputs, Figure 2(a). It is observed that by imputing values with MICE losses below 20% the performance is similar to the baseline case (without artificial missings). The performance slightly worsens between 20-50% of missing values. MICE is better than performing a simple mean imputation for each predictor, specially for small missing percentages. It is also observed that removing missing data worsens the performance of the proxy. To test the stability of the RMLP model in the long run, we train an RMLP on a set of MICE imputed data with 10% of missing values and simulate missing entries in measurements during the estimation phase, with varying missing percentages. Figure 2(b) shows the performance of a BC proxy model with no artificial inputs missing during the training

phase. In the latter case, as no relationship between predictors was learned during training, a simple mean value imputation was applied in the estimation phase to fill in the missing inputs. As can be seen in figure 2(b), the performance of the BC proxy model suffers in the presence of missing values with RMSE ranging from  $0.4 \mu\text{g}/\text{m}^3$  in the case of the reference BC proxy to RMSE values ranging from  $1.0$  to  $1.2 \mu\text{g}/\text{m}^3$ . In contrast, the RMLP with MICE imputation has a good performance in a wide range of missing value percentages.



(a) Impact of missing imputation over the train data set. (b) Impact of missing imputation over the test data set.

**Figure 2: Impact of missing data imputation on RMLP.**



**Figure 3: Impact of missing on different predictors.**

Finally, Figure 3 shows the performance of the RMLP during the estimation phase as different LCSs fail with varying missing percentages. It is observed that the impact of missing sensor values impacts the performance of the proxy depending on which sensor is missing those values. Missing  $\text{O}_3$  and  $\text{PM}_{10}$  has less impact than missing  $\text{N}$ . In the case of imputation using only the mean of the  $\text{N}$ , the performance worsens, while using RMLP with MICE improves with values close to the baseline with an RMSE of  $0.41 \mu\text{g}/\text{m}^3$ .

## 5 CONCLUSIONS

In this paper, we have proposed a mechanism called robust machine learning proxy (RMLP) for black carbon (BC) estimation from indirect measurements of low-cost sensors (LCSs) taken by IoT nodes in an air quality monitoring network. RMLP uses the concept of proxy using machine learning techniques (e.g. SVR, RF or ANN). The RMLP considers several aspects in the design of the proxy such as the calibration of the LCSs, the aggregation of data, the elimination of noise in the different sensor streams, the imputation of missing values and the selection of the sensors that can participate in the proxy. Among the results obtained we can observe that the three models used to create the proxy have similar performances, although SVR performs slightly better than RF and ANN. We have also

observed that filtering the sensor streams with a Savitzky-Golay low-pass filter improves the proxy results and that high aggregation of the data worsens the proxy performance. Finally, we have studied the impact of missing data by adding a missing value imputation method based on MICE, observing that MICE improves proxy estimation with losses up to 50% with respect to not using imputations. The results obtained show a different impact depending on the type of feature that has gaps in the data, and it is observed that the imputation method mitigates these losses when estimating BC.

## ACKNOWLEDGMENTS

This work is supported by projects H2020 FIRE-RES, PID 2019-107910RB-I00, CEX 2018-000794-S, 2017 SGR41, 2021 SGR-01059, and the support of Sec. d'Universitats i Recerca de la Generalitat de Catalunya i del Fons Social Europeu. We would also like to acknowledge the support of Gen. de Catalunya for providing the air quality data.

## REFERENCES

- [1] Jose M Barcelo-Ordinas, Messaud Doudou, Jorge Garcia-Vidal, and Nadjib Badache. 2019. Self-Calibration Methods for Uncontrolled Environments in Sensor Networks: a Reference Survey. *Ad Hoc Networks* 88 (2019), 142–159.
- [2] Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Xiaoli Liu, Kai Puolamäki, Petteri Nurmi, and Sasu Tarkoma. 2021. Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration: A Survey and Critical Analysis. *ACM Transactions on Sensor Networks* 17, 2, Article 20 (may 2021), 44 pages.
- [3] Pau Ferrer-Cid, Jose M Barcelo-Ordinas, and Jorge Garcia-Vidal. 2022. Data reconstruction applications for IoT air pollution sensor networks using graph signal processing. *Journal of Network and Computer Applications* 205 (2022), 103434.
- [4] Pau Ferrer-Cid, Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Anna Ripoll, and Mar Viana. 2019. A Comparative Study of Calibration Methods for Low-Cost Ozone Sensors in IoT Platforms. *IEEE Internet of Things Journal* 6, 6 (2019), 9563–9571.
- [5] Pau Ferrer-Cid, Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Anna Ripoll, and Mar Viana. 2020. Multisensor Data Fusion Calibration in IoT Air Pollution Platforms. *IEEE Internet of Things Journal* 7, 4 (2020), 3124–3132.
- [6] Pau Ferrer-Cid, Julio Garcia-Calvete, Aina Main-Nadal, Zhe Ye, Jose M Barcelo-Ordinas, and Jorge Garcia-Vidal. 2022. Sampling Trade-Offs in Duty-Cycled Systems for Air Quality Low-Cost Sensors. *Sensors* 22, 10 (2022), 3964.
- [7] Pak Lun Fung, Martha A Zaidan, Salla Sillanpää, Anu Kousa, Jarkko V Niemi, Timonen, et al. 2019. Input-adaptive proxy for black carbon as a virtual sensor. *Sensors* 20, 1 (2019), 182.
- [8] Pak L Fung, Martha A Zaidan, Hilkka Timonen, Jarkko V Niemi, Anu Kousa, Joel Kuula, Krista Luoma, Sasu Tarkoma, Tuukka Petäjä, et al. 2021. Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. *Journal of Aerosol Science* 152 (2021), 105694.
- [9] Sunder Ram Krishnan and Chandra Sekhar Seelamantula. 2012. On the selection of optimum Savitzky-Golay filters. *IEEE transactions on signal processing* 61, 2 (2012), 380–391.
- [10] Lichuan Liu, Sen M Kuo, and MengChu Zhou. 2009. Virtual sensing techniques and their applications. In *2009 International Conference on Networking, Sensing and Control*. IEEE, 31–36.
- [11] Nwamaka U Okafor and Declan T Delaney. 2021. Missing data imputation on IoT sensor networks: Implications for on-site sensor calibration. *IEEE Sensors Journal* 21, 20 (2021), 22833–22845.
- [12] J Rovira, Juan Antonio Paredes-Ahumada, Jose Maria Barcelo-Ordinas, Jorge Garcia-Vidal, Cristina Reche, Y Sola, Pak Lun Fung, Tuukka Petäjä, Tareq Hussein, and Mar Viana. 2022. Non-linear models for black carbon exposure modelling using air pollution datasets. *Environmental research* 212 (2022), 113269.
- [13] Martha A Zaidan, Lubna Dada, Mansour A Alghamdi, Hisham Al-Jeelani, Heikki Lihavainen, Antti Hyvärinen, and Tareq Hussein. 2019. Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies. *Applied Sciences* 9, 20 (2019), 4475.
- [14] Martha Arbayani Zaidan, Naser Hossein Motlagh, Pak L Fung, David Lu, Hilkka Timonen, Joel Kuula, Jarkko V Niemi, Sasu Tarkoma, Tuukka Petäjä, et al. 2020. Intelligent calibration and virtual sensing for integrated low-cost air quality sensors. *IEEE Sensors Journal* 20, 22 (2020), 13638–13652.
- [15] Martha A Zaidan, Darren Wraith, Brandon E Boor, and Tareq Hussein. 2019. Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models. *Applied sciences* 9, 22 (2019), 4976.