12th International Conference on Electrical, Electronics and Computer Engineering (IcETRAN), Čačak, 9-12 June 2025

A Simple Virtual Sensor Approach for Black Carbon Estimation in Sensor Networks

Miloš Davidović
VIDIS centre, Department of Radiation and Environmental Protection, Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade
Belgrade, Serbia
davidovic@vin.bg.ac.rs, 0000-0003-2438-0231

Milena Davidović
Department of mathematics, physics and descriptive geometry
Faculty of Civil Engineering, University of Belgrade,
Belgrade, Serbia
milena@grf.bg.ac.rs, 0000-0003-0267-2751

Maja Jovanović
VIDIS centre, Department of Physical Chemistry, Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade
Belgrade, Serbia
majaj@vin.bg.ac.rs, ORCID

Marija Živković
VIDIS centre, Department of Physical Chemistry, Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade
Belgrade, Serbia
marijaz@vin.bg.ac.rs, 0000-0002-3415-5145

Shahin Tabandeh
VTT Technical Research Centre of Finland Ltd., National Metrology Institute VTT MIKES
Espoo, Finland
shahin.tabandeh@vtt.fi, 0000-0002-2315-0752

Milena Jovašević-Stojanović
VIDIS centre, Department of Radiation and Environmental Protection, Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade
Belgrade, Serbia
mjovst@vin.bg.ac.rs, 0000-0003-0765-6603

***Abstract*—This paper presents a simple virtual sensor predictive model based on multiple linear regression for the estimation of the equivalent black carbon concentration in an air quality automatic monitoring sensor network. The predictive model uses selected pollutants concentration and meteorological parameters as predictor variables, and equivalent black carbon concentration as target variable. Virtual sensor model is assessed for several different training/test periods. Dataset for training and validation of the model is derived using a vast dataset collected in WeBaSOOP project on Ada Marina, Belgrade supersite.**

Keywords—sensor networks metrology, air pollution, equivalent black carbon, monitoring supersite, machine learning

I. Introduction

The prerequisite of sustainable development is the availability of high-quality environmental information data, which can be efficiently utilized to provide basis for safeguarding of urban and natural environment. However, traditionally monitored air quality parameters often do not give a sufficiently detailed and comprehensive picture of air quality, thus limiting the ability of citizens and authorities for information-based decision making.

One such example is a problem of reducing particulate matter atmospheric pollution from relevant anthropogenic sources, which would in consequence result in potential health benefits. The metric that is most commonly available at automatic monitoring sites is averaged mean mass concentration of fine particles, $PM_{2.5}$. Note that the $PM_{2.5}$ is a time- and spatially variable mixture of chemicals (hydrocarbons, salts and other compounds given) which are emitted into ambient air by plethora of diverse sources such as e.g. anthropogenic components in the form of vehicular exhaust and non-exhaust emissions, cooking stoves and industry, and of natural components such as natural dust and microorganisms. Therefore, reducing $PM_{2.5}$ by the same amount of mass concentration in different places will not deliver the same health benefits. Thus, some form of additional PM characterization, i.e. additional metric, in addition to the commonly used mean mass concentration is needed. Carbonaceous aerosols concentration in air is an example of one such useful additional metric.

Carbonaceous aerosols are often the largest component of fine particulate matter. Carbon aerosol particles are composed of light-scattering Organic Carbon (OC), and light-absorbing carbonaceous aerosols, dominated by Black Carbon, (BC) or Elemental Carbon (EC). BC is black material emitted from gas and diesel engines, biomass burning, coal-fired power plants, and other sources that burn fossil fuel. Presence of BC has negative effects for both, human health and our climate [1]. Furthermore, inhalation of BC is associated with health problems including respiratory and cardiovascular disease, cancer, and even birth defects [2,3]. Measuring real time mass concentration of BC requires specialized equipment, most commonly filter absorption photometers, which are typically not a part of routine regulatory monitoring and are thus not readily available throughout the air quality automatic monitoring station sensor networks. They are usually only available at the specialized monitoring sites, where much more comprehensive monitoring is done, so called monitoring supersites.

Recently, a concept of virtual or soft sensors has been introduced [4] as a way of increasing spatial resolution for certain pollutants, and utilizing IoT (low-cost sensors) and AI concepts. Virtual sensors and novel calibration procedures are gaining traction in the metrology of sensor networks [5]. By using similar research vision of virtual sensors we expand the concept to reference air quality automatic monitoring networks, and determine to which extent it would be possible, in principle, to gain information about BC mass concentration for the entire reference grade network, and not just the supersites in the network, where such information is produced by physical sensors. Instead of directly relying on physical on-site sensor hardware, we utilize correlation between (reference) air quality parameters monitored on a supersite network node, and use this

ISBN 978-86-6200-031-6 ©2025 AM & ETRAN

https://doi.org/10.69994/12Ic25024

information to derive predictive models, i.e. virtual sensors for BC mass concentration, which can be potentially used throughout the entire sensor network.

This paper presents and assesses the performance of a simple method for estimation of equivalent BC mass concentration, based on multiple linear regression. The predictive model uses selected particulate matter and gaseous pollutant mass concentration and selected meteorological parameters as predictor variables, and equivalent black carbon concentration as a target variable. Dataset for training and validation of the model is a subset of a vast dataset collected within the framework of ongoing WeBaSOOP project [6], at the newly-established experimental supersite at Ada Marina, Belgrade automatic monitoring station.

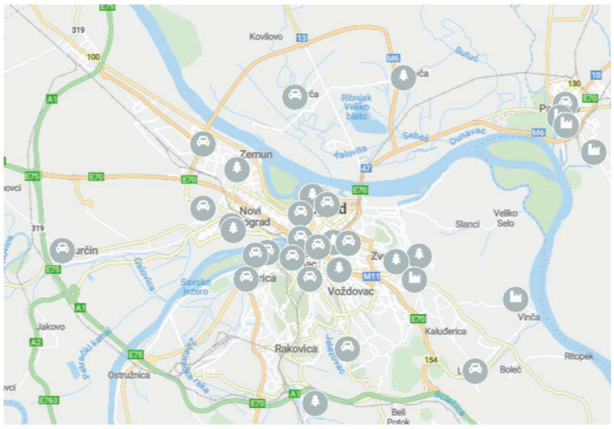
II. Methodology

A. *Serbian National Air Quality monitoring Network and Belgrade Supersite Ada Marina*

Serbian national network of air quality automatic monitoring stations is very well developed, regulatory environmental monitoring is performed by the Serbian Environmental Protection Agency (SEPA), and hourly averages for the monitoring networks are openly available online [7]. Results are reported for 84 monitoring stations, 46 being operated by SEPA, 30 being operated by Institute for Public Health Belgrade, and the remaining 8 being operated by: Serbia Zijin Bor Copper (copper mining and smelting complex in Bor, Serbia, 2 stations), Autonomous Province of Vojvodina in city of Pančevo (4 stations) and 2 being operated by local municipalities in the cities of Pirot and Negotin. With few exceptions all of the stations report $PM_{2.5}$ mass concentration. Figure 1a shows SEPA stations in Belgrade, including the Ada Marina station, and Figure 1b shows the Ada Marina station from the outside.

Within the framework of WeBaSOOP project, a monitoring supersite was established at the Ada Marina, Belgrade automatic monitoring station, and it is operational since June 2023. In addition to the standard set of air quality parameters present at the automatic monitoring station (SO_2, NO, NO_2, NO_x, PM_{10}, $PM_{2.5}$, O_3, CO, meteorological data), supersite features instruments for more comprehensive characterization of PM, namely: scanning mobility particle sizer (TSI 3082), condensation particle sizer (TSI 3775), optical particle size spectrometer (TSI 3330) and most relevant for this study filter absorption photometer (Magee AE33 aethalometer), capable of estimating real-time BC mass concentration. In addition to these online measurement capable instruments, supersites features extensive gravimetric sampling for offsite laboratory analysis.

Here a quick note about Magee AE-33 aethalometer, an industry standard for filter absorption photometers, is in place. In a nutshell, this instrument collects air sample on a filter tape, and measures changes in light absorption at 7 different wavelengths through sample laden filter tape to obtain an estimate of BC mass concentration. The spectral dependence of the sample absorption at 7 wavelengths can be used to derive absorption Ångström exponent for different aerosol samples. During a longer sampling campaign, absorption Ångström exponent (AAE) will vary, going from lower values (AAE around 1) corresponding to sample dominated by incomplete liquid fuel combustion, to higher values (AAE around 2) corresponding to sample dominated by incomplete solid fuel (biomass) combustion.



a)

b)

c)

Fig. 1. a) SEPA automatic monitoring stations in Belgrade b) outside of the Ada Marina, Belgrade supersite c) sampling inlet detail for black carbon and total carbon instruments

Magee AE-33 reports this biomass burning percentage (BB), using a simple aethalometer model [8], and this value will be also used in initial correlation analysis. Also, an interesting sidenote from the metrological standpoint is that the inlet structure for the $PM_{2.5}$ and BC measurements is different (Fig. 1b and Fig. 1c), and they use completely separate sample tubing system.

The dataset obtained via this comprehensive sensor suite available at the supersite can be used to prepare training and validation data for the virtual sensor model. Here we will utilize a limited subset of the complete dataset since the goal is to derive a simple linear regression model for BC concentration estimation.

B. Virtual BC sensor multiple linear regression model

Multiple linear regression model is a generalization of univariate linear regression, and a widely used technique for modelling dependencies between two or more predictor variables and a single target variable. After the choice of predictors in the linear model, modelling equation is given by (1), for each data point in the training set $i = 1, .., n$:

$$y_i = a_p \cdot x_{ip} + \cdots + a_1 \cdot x_{i1} + a_0 + \epsilon_i \quad (1)$$

MLR model can be represented in a matrix form:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} = X \cdot A + E, \quad (2)$$

where $X \cdot A$ is a system component, X containing predictor values for individual data points, A containing model parameters (to be determined in the process of training) and E is the residual. Model parameters are determined to minimize the residual sum of squares between the observed targets in the dataset [9]:

$$S = \sum_{i=1}^{n} (\epsilon_i)^2 = E^T \cdot E \rightarrow min. \quad (3)$$

In ordinary least squares (unlike the orthogonal least squares) the assumption is that predictor variables are observed without errors, and that error term refers to target variable, with variance of error being constant (homoscedasticity). Since we are applying the virtual sensor model to the reference grade instruments such assumption can be justified, since their error is low. Although the methods that account for uncertainty in both predictor and target, such as total least squares are available, we opted for use ordinary least squares, for simplicity and practicality reasons.

Simple multiple linear regression models in the context of air quality monitoring and data driven modeling have shown certain advantages compared to the more complex ones, such as machine learning (ML) based models. Linear regression models are typically less prone to overfitting, more stable across multiple seasons or longer time spans, and have model parameters, and model equation that offers much more simple and straightforward interpretation compared to the parameters in a more complex ML models.

Since we will be using Scikit-learn library [9] for the modeling, for validation of the models and estimation of their predictive power we will use training and validation split. This will not be done as is implemented in the Scikit-learn library by default, i.e. by random training-test split, but by sequential training-test split, which is more appropriate for time-series based modeling.

As for the choice of predictor variables, we are guided by the method developed for low-cost sensors calibration approach used in [10] where authors have used a combination meteorological data from the reference (temperature and relative humidity) and signal from the low-cost sensor. Note the very important distinction that in this study we exclusively use only the signals from the reference instruments and target BC concentration determined by Magee AE-33 reference instrument.

III. Results and Discussion

A. Correlation of air quality parameters and choice of predictor variables

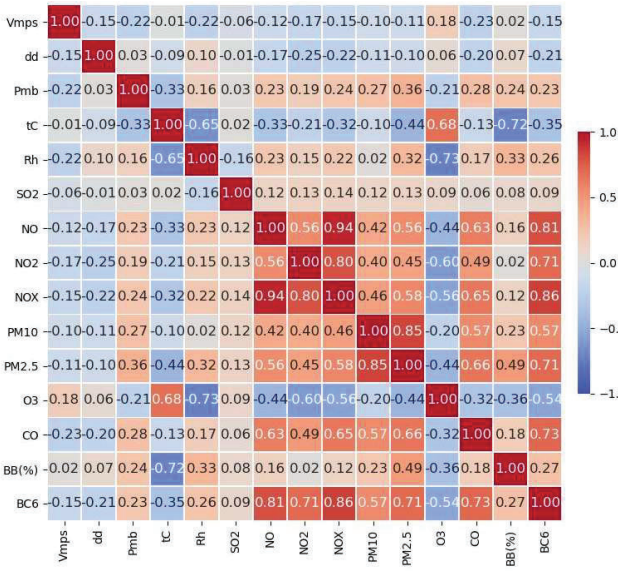


Fig. 2. Correlation matrix (Pearson correlation coefficient) between relevant air quality parameters at the Ada Marina, Belgrade supersite during the period between October 1st 2023 to August 19th 2024, calculated using 1-hour averages.

Air quality parameters used for calculation of correlation matrix depicted in Fig. 2 are meteorological (wind speed, direction, pressure, temperature, relative humidity), gaseous pollutants (SO_2, NO, NO_2, NO_x, O_3, CO), particulate matter fractions (PM_{10}, $PM_{2.5}$) and aethalometer data (biomass burning percentage and black carbon concentration). This correlation matrix can be used to guide the selection of predictor variables for the virtual sensor model. Based on the last column of the matrix, we observe that strongest correlation amount meteorological parameters is observed for temperature (-0.35), for gaseous pollutants strongest correlation is for NO_x (0.86) and for particulate matter fraction $PM_{2.5}$ (0.71). We will use those as predictors to arrive at our virtual sensor model, but note that other choices could also be possible if other criteria are applied. Thus we arrive at the virtual sensor model equation:

$$BC_{virtual} = a_t \cdot t + a_{NOx} \cdot NO_x + a_{PM2.5} \cdot PM_{2.5} + a_0 \quad (4)$$

B. *Performance of the virtual sensor models*

Now that we have established a set of predictors, we can evaluate the performance of the virtual sensor. The metrics that we can use are the same as the metrics we use for real physical sensors, i.e. suitable metrics are R^2 and RMSE as calculated on the test portion of the dataset.

As mentioned above, training and test split for the models based on time series should be performed in a sequential manner. This also closely mimics real world scenarios of calibration, and similarly virtual sensor models, in which model data is collected for some time, and then tested over prolonged periods of time. It is worth noting that the overall span of air quality parameters for which the models are constructed will need to sufficiently capture the conditions in which models will be used, otherwise performance drops of the models are to be expected. Table 1 summarizes the performance of the models, where the training percentage (compared to the complete dataset) goes from 10% to 50% of a complete dataset in increments of 10%.

TABLE I. PERFORMANCE OF THE VIRTUAL SENSOR MODEL EQ. (4) DEPENDING ON THE TRAINING AND TEST SPLIT

Training/test percentage	**R^2**	**RMSE** [µg/m3]
10/90	0.0212	2.5300
20/80	0.8044	1.1342
30/70	0.7788	0.9977
40/60	0.7388	0.9479
50/50	0.5565	0.9788

Based on results in Table 1 it is evident that the calibration/deployment mismatch, which is modeled by the training/test split can easily occur. For example, if we pair the results in Table 1 with Fig. 3 it is evident that data campaign period starts with small values of BC. 10/90 split therefore has the worst performance, since it trains the virtual sensor on very small values of BC, but then tests it on a very different remainder of the dataset.

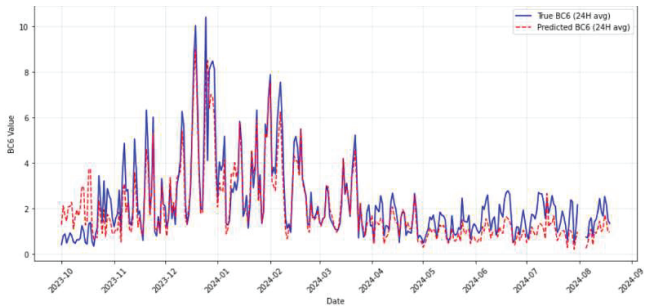


Fig. 3. True vs predicted equivalent BC values [µg/m^3] (24 hour average), using the model defined by Eq. 4, 20/80 training/test split. Model coefficients NO*x*: 0.0312, $PM_{2.5}$: 0.0482, t: -0.0263 µg/(m^3 °C) and intercept: 0.2491 µg/m^3

Similarly, as we increase the training set, now the test set becomes a problem, since it starts to lack the larger values of BC, which are characteristic during the heating season, but not present in the rest of the dataset.

IV. CONCLUSION

In this paper we have derived a simple virtual sensor model for black carbon mass concentration in ambient air, using as a target the data from a supersite, and input data that is widely available at air quality automatic monitoring stations. It was shown that for a suitable chosen training period RMSE as low as ~1µg/m^3 can be achieved, thus, in principle, enabling black carbon mass concentration estimates on sensor network sites with no dedicated black carbon measuring instruments. Future work will be focused on assessing virtual sensor model performance for scenarios with different biomass burning and fossil fuel contribution to black carbon concentrations.

ACKNOWLEDGMENT

This work was funded by European Union's Horizon Europe Research and Innovation Program under GA 101060170 (WeBaSOOP project https://webasoop.vinca.rs/), the project 22DIT02 FunSNM (https://www.funsnm.eu/) has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States; and the Ministry of Education, Science and Technological Development of the Republic of Serbia under GA 451-03-136/2025-03/200017.

REFERENCES

[1] Novakov, Tica, and Hal Rosen. "The black carbon story: early history and new perspectives." *Ambio* 42, no. 7 (2013): 840-851.

[2] Janssen, Nicole AH, Gerard Hoek, Milena Simic-Lawson, Paul Fischer, Leendert Van Bree, Harry Ten Brink, Menno Keuken et al. "Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM_{10} and $PM_{2.5}$" *Environmental health perspectives* 119, no. 12 (2011): 1691-1699.

[3] Janssen, N. A. H., M. E. Gerlofs-Nijland, T. Lanki, R. O. Salonen, F. Cassee, G. Hoek, P. Fischer, B. Brunekreef, and M. Krzyzanowski. "Health effects of black carbon, The WHO European Centre for Environment and Health, Bonn, Germany." *World Health Organisation Regional Office for Europe, Copenhagen, Denmark* (2012).

[4] Zaidan, Martha Arbayani, Naser Hossein Motlagh, Brandon E. Boor, David Lu, Petteri Nurmi, Tuukka Petäjä, Aijun Ding, Markku Kulmala, Sasu Tarkoma, and Tareq Hussein. "Virtual Sensors: toward High-resolution air pollution monitoring using AI and IoT." *IEEE Internet of Things Magazine* 6, no. 1 (2023): 76-81.

[5] Tabandeh, Shahin, Anupam Prasad Vedurmudi, Henrik Söderblom, Sara Pourjamal, Peter Harris, Yuhui Luo, Maximilian Gruber et al. "Sensor network metrology: Current state and future directions." *Measurement: Sensors* (2025): 101798.

[6] HE WeBaSOOP project web page, https://webasoop.vinca.rs/

[7] Serbian Environmental Protection Agency, https://sepa.gov.rs/

[8] Sandradewi, J., Prevot, A. S. H., Weingartner, E., Schmidhauser, R., Gysel, M., and Baltensperger, U.: A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength Aethalometer, Atmos. Environ., 42, 101–112, 2008b.

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Topalović, Dušan B., Miloš D. Davidović, Maja Jovanović, Alena Bartonova, Zoran Ristovski, and Milena Jovašević-Stojanović. "In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches." Atmospheric Environment 213 (2019): 640-658.