Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration

Pak L. Fung\textsuperscript{a,\,*}, Martha A. Zaidan\textsuperscript{a,b}, Hilkka Timonen\textsuperscript{c}, Jarkko V. Niemi\textsuperscript{d}, Anu Kousa\textsuperscript{d}, Joel Kuula\textsuperscript{c}, Krista Luoma\textsuperscript{a}, Sasu Tarkoma\textsuperscript{e}, Tuukka Petäjä\textsuperscript{a}, Markku Kulmala\textsuperscript{a}, Tareq Hussein\textsuperscript{a,f,g,**}

\textsuperscript{a}Institute for Atmospheric and Earth System Research / Physics, Faculty of Science, University of Helsinki, Finland
\textsuperscript{b}Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing, 210023, China
\textsuperscript{c}Atmospheric Composition Research, Finnish Meteorological Institute, FI-00560, Helsinki, Finland
\textsuperscript{d}Helsinki Region Environmental Services Authority (HSY), P.O. Box 100, FI-00066, Helsinki, Finland
\textsuperscript{e}Department of Computer Science, University of Helsinki, Finland
\textsuperscript{f}Department Material Analysis and Indoor Chemistry, Fraunhofer WKI, D-38108 Braunschweig, Germany
\textsuperscript{g}Department of Physics, The University of Jordan, Amman, 11942, Jordan

\textbf{ARTICLE INFO}

\textbf{Keywords:} Air quality, Statistical modelling, Random forest, Neural network, Support vector, Input-adaptive proxy

\textbf{ABSTRACT}

Air quality prediction with black-box (BB) modelling is gaining widespread interest in research and industry. This type of data-driven models work generally better in terms of accuracy but are limited to capture physical, chemical and meteorological processes and therefore accountability for interpretation. In this paper, we evaluated different white-box (WB) and BB methods that estimate atmospheric black carbon (BC) concentration by a suite of observations from the same measurement site. This study involves data in the period of 1\textsuperscript{st} January 2017–31\textsuperscript{st} December 2018 from two measurement sites, from a street canyon site in Mäkelänkatu and from an urban background site in Kumpula, in Helsinki, Finland. At the street canyon site, WB models performed ($R^2 = 0.81–0.87$) in a similar way as the BB models did ($R^2 = 0.86–0.87$). The overall performance of the BC concentration estimation methods at the urban background site was much worse probably because of a combination of smaller dynamic variability in the BC values and longer data gaps. However, the difference in WB ($R^2 = 0.44–0.60$) and BB models ($R^2 = 0.41–0.64$) was not significant. Furthermore, the WB models are closer to physics-based models, and it is easier to spot the relative importance of the predictor variable and determine if the model output makes sense. This feature outweighs slightly higher performance of some individual BB models, and inherently the WB models are a better choice due to their transparency in the model architecture. Among all the WB models, IAP and LASSO are recommended due to its flexibility and its efficiency, respectively. Our findings also ascertain the importance of temporal properties in statistical modelling. In the future, the developed BC estimation model could serve as a virtual sensor and complement the current air quality monitoring.

\* Corresponding author.
\** Corresponding author. Institute for Atmospheric and Earth System Research / Physics, Faculty of Science, University of Helsinki, Finland

E-mail addresses: pak.fung@helsinki.fi (P.L. Fung), martha.zaidan@helsinki.fi (M.A. Zaidan), hilkka.timonen@fmi.fi (H. Timonen), jarkko.niemi@hsy.fi (J.V. Niemi), anu.kousa@hsy.fi (A. Kousa), joel.kuula@fmi.fi (J. Kuula), krista.q.luoma@helsinki.fi (K. Luoma), sasu.tarkoma@helsinki.fi (S. Tarkoma), tuukka.petaja@helsinki.fi (T. Petäjä), markku.kulmala@helsinki.fi (M. Kulmala), tareq.hussein@helsinki.fi, t.hussein@ju.edu.jo (T. Hussein).
1. Introduction

Air pollution has been a serious issue with respect to human’s health, especially in densely-populated urban cities, from which over 90% of the population live with poor air quality (World Health Organization, 2019). In Europe, according to European Environment Agency (2019), more than 538,000 premature deaths were estimated to be associated with the exposure to excess air pollutants. These pollutants include a combination of atmospheric aerosol particles with varying concentration (number, volume and mass) and chemical composition as well as a suite of trace gas air pollutants. The negative health effects arise from a combination of exposure and toxicity of the pollutants (World Health Organization, 2017).

To simplify the complex mixture and interaction of different air pollutants, air quality indices (AQI) have been developed to quantify the severity of air pollution around the globe (World Health Organization, 2017). Although different regions have their own AQIs corresponding to different local air quality standards, most organisations consider particulate matter (PM) mass concentration, both thoracic particles (PM$_{10}$) and fine particles (PM$_{2.5}$), nitrogen dioxide (NO$_2$), ozone (O$_3$), sulphur dioxide (SO$_2$) and carbon monoxide (CO). Other variables, such as black carbon (BC), have been suggested to be one of the components alongside with the other air quality parameters in AQI (World Health Organization, 2012) because they can associate better with health risk of aerosol particles than commonly monitored PM, especially for cardiovascular effects (Geng et al., 2013). However, this has not yet been taken into action because not all reference stations in urban area carry out continuous BC measurements, since BC monitoring is not required by many national environmental legislations (Kutzner et al., 2018). In addition, even for the existing measurement stations, due to instrument failure or data corruption (Junger & De Leon, 2015; Zaidan, Wraith, et al., 2019), measurements are not always possible and air quality models are needed for data gap imputation and air quality prediction.

Air quality is extensively investigated by using different modelling methods, namely physics-based and data-driven approaches (Cabaneros et al., 2019; Karroum et al., 2020). The traditional physics-based approaches require knowledge of underlying physical processes related to air quality parameters, including but not limited to emission, transport and cloud interaction. Because of the complexity, some show varying results. For instance, Weather research and forecasting model including chemistry (WRF-Chem) typically and systematically underestimated BC concentration by up to 50% in South Africa (Kuik et al., 2015) and up to 90% in Kathmandu Valley (Mues et al., 2018), respectively. Regional aerosol-climate model (REMO) underestimated the observations significantly by more than 40% in all seasons in Finland (Hiennonla et al., 2013) while a BC mixing state model showed varying correlation ($R^2 = 0.53–0.85$) in China (Zhang et al., 2019). They are typically sensitive to factors like computational resources and the scale of the parameters involved, some of which may not always be available (Zaidan, Wraith, et al., 2019). Data-driven approaches, on the other hand, do not require in-depth understanding of air pollutant dynamics and they have been increasingly utilised as tools for statistical models (Cabaneros et al., 2019). These approaches can be classified into white-box (WB) and black-box models (BB). The classification of the two types of models is a continuum depending on the computational complexity and the ability to interpret the prediction. Generally speaking, WB models, including Input-Adaptive Proxy (IAP) (Fung et al., 2020), Least Absolute Shrinkage and Selection Operator (LASSO) (Van Roode et al., 2019) and Decision Tree method (DT) (Yu et al., 2016), are transparent machine learning (ML) processes for which one can clearly explain how they behave, how they produce predictions and what the influencing variables are (Rudin, 2019). BB models, on the other hand, refer to systems or objects which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings or underlying principles (Rudin, 2019). These include, but are not limited to, Random Forest (RF) (Kang et al., 2018; Masih, 2019), Support Vector Regression (SVR) (Kisi et al., 2017), and Neural Networks (NN) (Cabaneros et al., 2019; Freeman et al., 2018). BB models work generally better in terms of accuracy but provide limited transparency and accountability regarding the results (Zaidan, Wraith, et al., 2019).

In this paper, we study black carbon (BC) as a model compound because it could be an important addition to AQI. A reliable BC model can act as a virtual sensor in support with other variables in air quality monitoring. The implementation of virtual sensors will lower the costs by proving added value of the already measured parameters without a physical observation site upgrade, operation costs or maintenance costs. Furthermore, a generalised BC model will enable us to estimate BC mass concentrations in retrospect providing tools for health effect re-evaluation and assessment. The objective of this paper is to evaluate WB and BB models of BC mass concentration in terms of their accuracy and sensitivity against observations at two different urban observation sites in Helsinki, Finland. We also illustrate how seasonal variation, the weekend effect and the diurnal pattern influence the machine learning process for both types of box models. Finally, we assess the most robust BC estimation model.

2. Resources

2.1. Measurement sites

We retrieved aerosol, gaseous and meteorological data for two years (1st January 2017–31st December 2018) for the modelling.
purpose from two measurement sites in Helsinki, which is situated on a relatively flat land at the coast of Gulf of Finland. According to National Land Survey of Finland (2019) and Official Statistics of Finland (2020), Helsinki, together with Vantaa and Espoo, which forms the Helsinki Metropolitan Area (HMA), has a total area of 1484 km² and population of about 1.18 million inhabitants as of 2019. The two measurement stations, Mäkelänkatu and Kumpula, are located approximately 1 km from each other (Fig. 1). They represent environments of a street canyon and an urban background, respectively. The two environments possess the common features in urban areas and they will give feedback on the model performance.

Street canyon site: Mäkelänkatu street canyon site (street address Mäkelänkatu 50; 60°11’N, 24°57’E) is operated by the Helsinki Region Environmental Services Authority (HSY). This station (Hietikko et al., 2018; Kuuluvainen et al., 2018) is located 3 km from the city centre in a street canyon in the immediate vicinity to one of the main roads leading to downtown Helsinki. The street consists of six lanes and two tramlines and the annual mean traffic volume is approximately 28,000 vehicles per workday. The proportion of heavy duty vehicles is 11% and the speed limit is set to 50 km h⁻¹. The traffic loads are especially high during rush hours at 8 a.m. and 5 p.m.. Both sides of the street have rows of buildings which weaken the dispersion process of the direct vehicular emissions (Dos Santos-Juusela et al., 2013). All the inlets for the measuring devices are positioned approximately at a height of 4 m from the ground level (Kuuluvainen et al., 2018).

Urban background site: The Station for Measuring Ecosystem-Atmosphere Relations III (SMEAR III) in Kumpula (60°12’N, 24°57’E, 26 m a.s.l.) is built on a rocky hill about 4 km northeast from the Helsinki centre (Järvi et al., 2009). The surroundings of this urban background station are heterogeneous at different directions, constituting of patchy forest and low vegetation (180°–325°), buildings and parking lots (325°–35°) and a main road (35°–180°) at 150 m from the site (Järvi et al., 2009). The main road Hämeentie leads to the centre of Helsinki. The average daily traffic intensity is 50,000 vehicles and the amount of heavy duty vehicles on that road is considerable. Secondary aerosol formation, traffic, wood combustion and long-range transport account for a significant portion of atmospheric pollutant sources (Carbone et al., 2014; Timonen et al., 2013). Trace gases are measured at a 31-m-high triangular lattice tower, while aerosol measurements are taken at a measurement container next to the tower.

In addition, traffic rates are monitored by the City of Helsinki. The nearest continuous counter station is about 500 m northwest and 1 km southwest from Mäkelänkatu and Kumpula measurement sites, respectively (Fig. 1). The traffic monitoring does not split between light- and heavy-duty vehicles. It logs the number of vehicles along Mäkelänkatu but it is estimated that the actual traffic is about 30% less at the Mäkelänkatu site because traffic may have been diverted to Kumpulanlankatu before reaching the Mäkelänkatu site.

2.2. Data collection

In this study, BC mass concentration (µg m⁻³) is measured by a multi-angle absorption photometer (MAAP; Thermo Scientific model 5012) with a PM₁ inlet at a 1-min time resolution (Petzold & Schönlinner, 2004) based on its optical properties. Using measurements of both transmittance and reflectance at different angles, the MAAP determines the light absorbance from particles deposited on the filter. The absorbance is converted to BC mass concentration by using a fixed 6.6 m² g⁻¹ mass absorption coefficient at wavelength of 637 nm (Müller et al., 2011).

Particle mass concentration of diameter less than 2.5 µm (PM₂.₅) and less than 10 µm (PM₁₀) are measured continuously with ambient particulate monitor TEOM 1405 at Mäkelänkatu street canyon site and TEOM 1405-D at Kumpula urban background site. Both are recorded in µg m⁻³.

A differential mobility particle sizer (DMPS), which is a combination of a differential mobility analyser (DMA) and a condensation

---

**Fig. 1.** Location of our measurement sites in Helsinki metropolitan region.
Fig. 2. The simplified workflow and the architecture of the different methods used in this paper. White-box models include IAP, LASSO and DT, which stand for input-adaptive proxy, least absolute shrinkage and selection operator and decision tree, respectively. Black-box models are RF, SVR, SNN and LSTM, which denotes random forest, support vector regression, shallow neural network and long short-term memory, respectively.
particle counter (CPC), measures aerosol size distribution. A Vienna DMA and an Airmodus A20 CPC are used at the street canyon site
while at the urban background site a twin DMPS (Hauke-type DMA and TSI Model 3025 CPC + Hauke-type DMA and TSI Model 3010
CPC) are used. The two DMPS systems measure particles with diameter 6–800 nm and 3–950 nm, respectively. The technique makes
use of the bipolar charging of aerosol particles, followed by classification of particles into size classes according to their electrical
mobility (Aalto et al., 2001; Järvi et al., 2009).

From the DMPS size distributions, we calculated the lung deposited surface areas (LDSA) in μm² cm⁻³ by a particle size-dependent
deposition fraction conversion equation first introduced in the human respiratory tract model of the ICRP report by Bair (1994). This
conversion involves the calculation of alveolar deposition efficiency and the summation of the LDSA concentration at the corre-
sponding size.

The total particle number concentration (N₁₀₀) is extracted from the DMPS size distribution data, which represents the total number
concentration of particles with diameters between 6–800 nm and 3–950 nm at the street canyon and at the urban background site,
respectively. We also extracted three particle modes from DMPS data. They include nucleation mode, Aitken mode and accumulation
mode. The nucleation mode sums up the particle number concentration of size bins below 25 nm (PN₀.025). Aitken mode is defined to
include particles with diameter between 25 and 90 nm (PN₀.09–0.025). Accumulation mode is also computed as size equal to or larger
than 90 nm (PN₁–0.09) at both sites. All the modal aerosol particle number concentrations are reported in number concentration cm⁻³.

Gaseous variables (in mass concentration μg m⁻³ at the street canyon site and in part per billion, ppb, at the urban background site),
such as nitrogen oxide (NO), NO₂, their sum NOₓ, O₃ and CO measurements are measured with standard instruments at both sites. In
addition, meteorological variables, including air temperature (Temp), relative humidity (RH), air pressure (P), wind speed (WS) and
wind direction (WD), are measured in both locations. Since wind direction is a circular variable, it is resolved into North-South and
East-West vector components by trigonometric function (Fernández-Guisuraga et al., 2016; Järvi et al., 2009), labelled as WD–N and
WD–E respectively. Solar radiation, including photosynthetically active radiation (PAR), is measured only at the rooftop in the uni-
versity building 100 m away from the urban background site. In summary, a total of 21 variables, including 1 output variable (BC) and
20 predictor variables, are utilised in the analysis and modelling.

3. Methods

A simplified workflow for the work is outlined in Fig. 2. The raw data stored in our database first went through pre-processing
procedures, including outlier removal, normalisation and data partition, which are described in detail in Section 3.1. The pre-
processed data were trained for establishing models through WB and BB models. We select and showcase seven methods (three for
WB and four for BB) in this study due to their proven performance addressed by several researchers (Cabaneros et al., 2019; Fung et al.,
2020; Kisi et al., 2017; Van Roode et al., 2019; Yu et al., 2016). The architecture of the methods and their optimisation criteria are
elaborated for WB and BB models in Section 3.2 and 3.3, respectively. In addition, the characteristics and benefits of the individual
models are stressed in the same sections. Finally, the optimised models are tested and compared by evaluation attributes further
explained in Section 3.4.

![Histogram showing the distribution of BC concentrations at the street canyon (in blue) and at the urban background site (in red). The histogram contains 90 size bins from BC concentration of 0.01–10 μg m⁻³ in logarithm scale. 95% confidence intervals (CI), mean and median of the BC concentrations (in μg m⁻³) were calculated. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)](image)
3.1. Data pre-processing

We first synchronised the data into hourly median time resolution, which is less influenced by extreme values, when compared with mean, and we subsequently converted the data into same unit for both Mäkelänkatu and Kumpula sites for better comparison. We identified and eliminated the outliers by calculating their z-scores with a threshold of $|z| = 3$ such that 99.7% of data points were retained. We then substituted values below detection limit (DL), including negative values, with $0.5 \times \text{DL}$. In order to undergo ML regression, the data has to follow a normal distribution; therefore, we examined this by plotting histograms for each variable. We applied appropriate transformations, and in this case, the frequently used logarithm transformation was applied to all aerosol and gaseous variables (Maciejewska et al., 2015; Teinilä et al., 2019). No transformation was done to the meteorological data.

The mean of BC are $(1.03 \pm 0.88 \ \mu g \ m^{-3})$ and $(0.47 \pm 0.46 \ \mu g \ m^{-3})$ at the street canyon and at the urban background site, respectively, and the distribution of the BC concentration is graphically illustrated in Fig. 3. The overall data coverage of all parameters used is 94% and 87% in the two sites. For the operation of WB and BB models, missing data was gap-filled with two methods, linear interpolation and nearest neighbour method, based on the length and the extent of the missing gap in datasets (Junninen et al., 2004). Both datasets were then normalised into a new variable (mean = 0, standard deviation (SD) = 1) (Cabaneros et al., 2019; Järvi et al., 2009) as the ML processes are sensitive to variable scales. Some selected models take autocorrelation into account themselves; therefore, to keep the comparison simple, no autoregressive terms were included.

To increase the reliability of our comparison, we partitioned the first 80% of the time series as training set and the last 20% as testing set. Flag vectors were created for workdays and weekends (also includes holidays) for comparison. Seasons were also classified thermally (Fung et al., 2020; Järvi et al., 2009). Since the testing set consists of the last 20% of the measurement period, i.e., 7th August–31st December 2018, only winter, summer and autumn were classified in the testing set.

3.2. White-box (WB) models

WB models refer to transparent ML processes that produce accountable predictions. They put emphasis on introspection and processes. Users are able to interpret the models and identify the relative importance of the corresponding predictor variables. It allows users to spot possible errors when the model coefficients contradict our physics-based knowledge.

3.2.1. Input-adaptive proxy (IAP)

IAP was first described by Fung et al. (2020). This method managed to estimate continuous BC concentration with coefficient of determination ($R^2$) higher than 0.8. IAP first examines the correlation of output variable with the input features by Pearson’s correlation coefficient. It pre-selects the most correlated input features and creates sub-models with maximum three input features with ordinary least-squares (OLS) linear regression (Fernández-Guisuraga et al., 2016):

![Fig. 4. Time-series of daily-averaged air quality variables, including (a) black carbon (BC), (b) nitrogen oxides (NOx), (c) ozone (O3), (d) carbon monoxide (CO) and (e) fine particles (PM2.5), during measurement period 1st January 2017–31st December 2018. Vertical lines tell the thermal separation of seasons.](image-url)
where $\hat{y}$ is the output variable, $\beta_0$ is the intercept of the model, $\beta_i$ and $x_i$ correspond to the $i^{th}$ regression coefficient and $i^{th}$ input variable of the model, and $\varepsilon$ is the random error. The model applies an extra regularisation by using ‘bisquare’ weight function, which depends on the residuals, leverages from OLS fits and the estimates of the standard deviation of the error terms, with a tuning factor 4.685 (Gross, 1977). This robust fitting serves as an alternative to the traditional OLS regression (Holland & Welsch, 1977) when data are contaminated with outliers that often takes place in field measurements. We run the regression and evaluate every sub-model. According to the evaluation metrics used, we rank the sub-models based on their performance in descending order. In practice, datasets are typically incomplete; in case of missing data, some data points in the input variables in the best sub-model can possibly be missing, hence the imputation cannot be fully achieved. IAP further imputes the missing data with the second best performing sub-model, and so on, until all the voids are filled up. Since IAP has been developed to fill up the missing data itself, no missing data imputation is done before modelling, unlike the other models we use.

### 3.3.1. Random forest – bagging ensemble method (RF)

RF is constructed on each subset by aggregating the results generated from all individual decision trees (Kamińska, 2019; Kang et al., 2018; Masih, 2019), as individual decision tree tends to over-fit (Yu et al., 2016). There are different ways for the aggregation: bagging (Breiman, 1996), boosting (Drucker et al., 1994) or stacking (Ting & Witten, 1997). In the bagging approach, different random subsets from the original dataset with replacement are generated. The same learning method on each sample is trained later and finally outputs of each model are simply weighted. This bootstrap-aggregated ensemble method reduces bias and error variance and improves generalisation (Van Roode et al., 2019). Same curvature test is used for determining split decision. Breiman’s random forest algorithm is also applied to determine the number of variables to be selected at random for each decision split (Breiman, 1996). Out-of-bag error is also calculated as a metric of inaccuracy. As the outcome is obtained by aggregation, after the aggregation, the comprehensive
3.3.2. Support vector regression (SVR)

SVR is a statistical learning theory that was introduced by Vapnik et al. (1997) and extensively implemented in air quality prediction (Kisi et al., 2017; Leong et al., 2019; Moazami et al., 2016; Siwek & Osowski, 2016). SVR performs regression by finding the kernel function that maximises the margin of tolerance $\varepsilon$ of the regression fit. The vectors that define the kernel function are the support vectors. The goal of the SVR model is to find a function $f(x)$ that deviates from output variables in the training data by a value no greater than $\varepsilon$ for each training point $x$. Another goal is to minimise the flatness of $f(x)$, determined by box constraint $C$, a positive numeric value that controls the penalty imposed on observations that lie outside the predefined $\varepsilon$, by using two Lagrange multipliers of support vectors:

$$f(x) = \sum_{i=1}^{N} (a_i - a_i^*)K(x_i, x) + b, \forall i: 0 \leq a_i, a_i^* \leq C$$

where $a_i$ and $a_i^*$ are the two Lagrange multipliers of support vectors in upper and lower bounds, respectively. $N$ and $b$ are the total number of data points and the bias of the model, respectively. $K(x_i, x)$ represents radial basis kernel function, which is recently used by Leong et al. (2019):

$$K(x_i, x_j) = \exp\left(-||x_i - x_j||^2\right)$$

where $x_i$ and $x_j$ represent two inputs to the kernel function, respectively.

In this paper, we use sequential minimal optimisation to perform a series of two-point optimisations. In each iteration, a set of two points are chosen based on a selection rule that uses second-order information. The final output is resulted from the addition of the summation of the difference of the support vectors terms and a bias.

3.3.3. Shallow neural network (SNN)

Artificial neural network models have been utilised in predicting air quality (Cabaneros et al., 2019; Freeman et al., 2018; Maleki et al., 2019). The architecture of neural networks consists of nodes which generate a signal or remain silent as activation function. Activation function in each layer determines the output value of each neuron that becomes the input values for neurons in the next hidden layer connected to it. In this paper, feed-forward multi-perceptron network is used. This type of network usually consists of a series of layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the network’s output. According to the review paper by Cabaneros et al. (2019), a shallow neural network (SNN) with one hidden layer and enough neurons in the hidden layers can fit any finite input-output mapping problem for non-linear relationship. In order to keep the simplicity, one hidden layer of five neurons are implemented.

A neuron can be thought as a combination of two parts:

$$z_n = w_n(x_i) + b_n$$

where $z_n$ and $b_n$ are the intermediate output and the bias term for the $n^{th}$ neuron, respectively. $w_n$ is the $n^{th}$ weight for each data points $x_i$. The second part performs the sigmoid activation function on $z_n$ to give out the final output $\hat{y}_n$ of the neuron:

$$\hat{y}_n = \sigma(z_n) = \frac{1}{1 + \exp^{-z_n}}$$

We initialise the weights randomly and the weights have to be updated through gradient descent optimisation, which might cause vanishing gradient problem. We perform five training cycles and several iterations within a cycle to minimise the training loss with Bayesian regularisation.

3.3.4. Long short-term memory (LSTM)

LSTM was first described by Hochreiter and Schmidhuber (1997). Since then, LSTM has been investigated intensively in estimating air quality (Cabaneros et al., 2019; Freeman et al., 2018; Kim et al., 2019; Wang & Song, 2018). LSTM units partially solve the vanishing gradient and long-term dependency problem because LSTM units allow gradients to also flow unchanged (Wang & Song, 2018). A common architecture is composed of a cell (the memory part of the LSTM unit) and three regulators of the flow of information inside the LSTM unit: an input gate, an output gate and a forget gate. The cell is responsible for keeping track of the dependencies between the elements in the input sequence. The input gate controls the extent to which a new value flows into the cell. The forget gate gives the information to throw away from the cell state. The output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit block at timestamp $t$. The activation function of the three LSTM gates is the logistic sigmoid function as in equation (6).

There are connections into and out of the LSTM gates, a few of which are recurrent. The weights of these connections, which need to be learned during training, determine how the gates operate as follows (Hochreiter & Schmidhuber, 1997):
\[ \tilde{c}_t = \tanh \left( \mathbf{w}_{c,h} \times \mathbf{x}_t + \mathbf{w}_{c,h-1} \times h_{t-1} + \mathbf{b}_c \right) \]  

(7)

\[ c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \]  

(8)

\[ h_t = \sigma_t \times \tanh(c_t) \]  

(9)

where \( \tilde{c}_t \) is the candidate calculated for cell state (memory vector) at the current timestamp \( t \), which comprises weights \( \mathbf{w}_{c,h} \) and \( \mathbf{w}_{c,h-1} \) of the layer for its connections to the input vector \( \mathbf{x}_t \) and to the previous short-term state \( h_{t-1} \), respectively. A bias term \( b_c \) is also added to equation (7). \( c_t \) is the cell state (memory vector) at the current timestamp, which is the addition of the outcome from forget gate \( (f_t) \) times the previous timestamp and the outcome from the input gate \( (i_t) \) times the candidate at the current timestamp. Lastly, we filter the cell state and then the outcome from the output gate \( (h_t) \) is passed through another activation function ‘tanh’ which predicts what portion should appear as the output of current LSTM unit at timestamp \( t \). We can pass the final outcome \( (h_t) \) in the current LSTM block through the other two layers, fully connected layer and regression layer, to get the predicted output from the current block. In this paper, we use adaptive momentum (‘adam’) (Freeman et al., 2018) as an optimiser to minimise loss function and reduce generalisation errors.

3.4. Evaluation attributes

In order to evaluate and compare the models, coefficient of determination \( (R^2) \), together with mean absolute error \( (MAE) \) and root mean square error \( (RMSE) \), are used as diagnostic evaluation attributes, as follows (Lee Rodgers & Nicewander, 1988):

\[ R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \]  

(10)

\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \]  

(11)

\[ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \]  

(12)

where \( y_i \) and \( \hat{y}_i \) are \( i^{th} \) measured and \( i^{th} \) estimated output variable by the model, respectively, and \( \bar{y} \) is the expected value of the measured valuable. \( N \) is the number of complete data input to the model. \( R^2 \) is a measure of how close the data lie to the fitted regression line. It, however, does not consider the biases in the estimation. Therefore, we further validate the models with \( MAE \) and \( RMSE \).

### Table 1

Table shows the coefficients \( (\beta) \) of corresponding input variables \( (x) \) and intercept \( (\beta_0) \) of IAP (best 5) and LASSO linear models at the street canyon site and urban background site. The asterisk signs indicate that the input variables are in their logarithm form.

| Input Variables \( (x) \) | Street canyon | | Urban background |
|---------------------------|---------------|-----------------|-----------------|
|                           | IAP (best 5 ranks, \( \beta \)) | LASSO (best 5 ranks, \( \beta \)) | LASSO (best 5 ranks, \( \beta \)) |
|                           | 1   | 2   | 3   | 4   | 5   | 1   | 2   | 3   | 4   | 5   | 1   | 2   | 3   | 4   | 5   |
| Traffic                   | 0.11| 0.12| 0.10|     |     | 0.09|     |     |     |     |     |     |     |     |     |
| *NO₂                      | 0.42| 0.47| 0.49| 0.49| 0.37|     | 0.04|     |     |     |     |     |     |     |     |
| NO                        |     |     |     |     |     | 0.06| 0.24| 0.19|     |     |     |     |     |     |
| *NO₃                      |     |     |     |     |     | –0.07|     |     |     |     |     |     |     |     |
| *O₃                       |     |     |     |     |     | 0.34| 1.01| 1.35| 1.43| 1.58| 1.02|     |     |     |
| *PM₁₀                     |     |     |     |     |     | 0.12| 0.13|     |     |     |     |     |     |     |
| *PM₂₅                     |     |     |     |     |     |     |     |     |     |     |     | 0.19|     |     |
| *PM₀.₀₂₅                  |     |     |     |     |     |     |     |     |     |     |     | 0.10|     |     |
| *PM₀.₀₀⁹-₀.₀₂₅            |     |     |     |     |     |     |     |     |     |     |     | 0.08|     |     |
| *PM₀.₀₀⁹                   |     |     |     |     |     | 0.68| 0.60| 0.66| 0.67| 0.33| 0.52| 0.44|     |     |
| *Nₐ₉₅                     |     |     |     |     |     | 0.75| 0.03| 0.54| 0.54| 0.61| 0.03|     |     |
| *LDSA                     |     |     |     |     |     |     |     |     |     |     |     | 0.42|     |     |
| Temp                      |     |     |     |     |     |     |     |     |     |     |     |     | 0.01|     |
| RH                        |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.01|
| P                         |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.06|
| WS                        |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| WD-E                      |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| WD-N                      |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| PAR                       |     |     |     |     |     |     |     |     |     |     |     |     |     | 0.01|
| \( \beta_0 \)            | –6.97| –7.29| –2.70| –6.48| –4.74| –5.85| –9.66| –8.95| –9.52| –12.09| –7.54| –4.89|     |     |


4. Results and discussion

This section first illustrates the general characteristics of BC in both locations and the relative importance of predictor variables in using WB models in Section 4.1. Section 4.2 analyses the model performance with respect to accuracy and sensitivity. In the latter part of the section, we entail the possible influence of seasonal variation, weekend effects and diurnal pattern on the model performance.

4.1. Variable characterisation

Before moving on to the model evaluation, it is important to familiarise ourselves with the pattern of the output variable, BC. We also briefly describe the correlation of BC with other input variables and sort out the relative importance in the ML process.

4.1.1. Typical features of BC concentration at the two sites

Table 2 shows the statistical description in measured BC concentration. Mean and median are presented with SD and 25\textsuperscript{th} and 75\textsuperscript{th} percentiles as the dispersion of the dataset, respectively. Figs. 3, 4a and 5 graphically portray the BC concentration distribution, trend and the dispersion in terms of month of the year and hour of the day at both sites, respectively. The typical features of BC in this study are in alignment with Luoma et al. (in review, 2020), who performed long-term spatio-temporal trend analysis for BC concentration in HMA. Overall, the BC concentration at the street canyon site in Mäkelänkatu (1.03±0.88 μg m\textsuperscript{-3}) is almost twice as high as that at the urban background site in Kumpula (0.47±0.46 μg m\textsuperscript{-3}) because the street canyon site is in proximity of a heavily-trafficked road while the urban background records the source from residential area situated 500 m to the north and roads with moderate traffic separated by a forest. The long distance from the source could increase dilution rate of BC and hence lower BC levels. When we consider seasonal variation, BC concentration in the summer at the street canyon site appears to be only 10% higher than the other seasons. The small variation agrees with Helin et al. (2018) who observed the lack of BC seasonal variability in traffic environment. Also, Teinilä et al. (2019) suggested that the seasonal changing mixing height does not show correlation with the dilution of local pollution in the street canyon. In urban background environment, BC concentration in winter appears to be the highest (0.65±0.6 μg m\textsuperscript{-3}) due to the lower mixing height (Teinilä et al., 2019) and elevated wood combustion by domestic heating (Hellén et al., 2017). Seasonal meteorological conditions can be seen in Fig. 6.

In street canyon, the mean of BC on weekdays (1.17±0.95 μg m\textsuperscript{-3}) are 60% higher than on weekends (0.72±0.57 μg m\textsuperscript{-3}). The discrepancy between weekdays and weekends in the urban background is smaller than in the street canyon. The discrepancy is less than 30% because the traffic emissions at the urban background site are less pronounced than at the street canyon site.

At both sites, the diurnal cycle is bimodal on weekdays due to the elevated traffic counts (Fig. 7) during working peak hours at 7–9 a.m. and at 4–6 p.m. in all months as typically observed in Helsinki (Timonen et al., 2014). The evening peak during workday is smaller than the morning peak because of the higher mixing height in the evening (Teinilä et al., 2019), which dilutes the BC pollutants from the surface. In addition, there is a clear seasonal variation in diurnal pattern in the street canyon site. Due to low boundary layer height in summer, BC concentrations are higher in summer mornings than in winter mornings (Luoma et al., in review, 2020). During weekends, only one peak is observed in the late evening. The evening peak during weekends appears at 5–8 p.m. in the wintertime and at 8–10 p.m. in the summertime. The boosted nocturnal BC concentration might be attributed to the increasing traffic rates along the daytime, reaching a peak approaching sunset when residents in the city return home (Fung et al., 2020).

| Location          | Period   | Mean   | SD    | P25   | P50   | P75   | Coverage (%) |
|-------------------|----------|--------|-------|-------|-------|-------|--------------|
| Street canyon     | Winter   | 0.99   | 0.85  | 0.42  | 0.75  | 1.29  | 100          |
|                   | Spring   | 0.93   | 0.83  | 0.36  | 0.70  | 1.28  | 100          |
|                   | Summer   | 1.13   | 0.93  | 0.49  | 0.88  | 1.49  | 99           |
|                   | Autumn   | 0.98   | 0.84  | 0.39  | 0.74  | 1.35  | 100          |
|                   | Workdays | 1.17   | 0.95  | 0.47  | 0.95  | 1.60  | 99           |
|                   | Weekends | 0.72   | 0.57  | 0.36  | 0.60  | 0.92  | 100          |
| Urban background  | Winter   | 1.03   | 0.88  | 0.43  | 0.79  | 1.37  | 99           |
|                   | Spring   | 0.65   | 0.6   | 0.27  | 0.51  | 0.82  | 61           |
|                   | Summer   | 0.50   | 0.49  | 0.19  | 0.34  | 0.64  | 34           |
|                   | Autumn   | 0.39   | 0.37  | 0.15  | 0.29  | 0.50  | 76           |
|                   | Workdays | 0.44   | 0.43  | 0.16  | 0.32  | 0.55  | 89           |
|                   | Weekends | 0.41   | 0.44  | 0.13  | 0.28  | 0.53  | 71           |
|                   | All      | 0.47   | 0.46  | 0.18  | 0.34  | 0.60  | 70           |
The completeness of the dataset in Mäkelänkatu is close to 99% while only 70% of data in Kumpula is valid during the whole measurement period. The reason for the missing data is owing to instrument failure due to connectivity issues, data corruption or human error in data acquisition (Junninen et al., 2004; Zaidan, Dada, et al., 2019).

The trends of the other air pollutants are illustrated in Fig. 4 b–e. Without taking into diurnal patterns into account, the seasonal variations are not obvious for NO\textsubscript{x}, PM\textsubscript{2.5}, CO appears to have higher concentrations during winter and lower amounts in the summer due to the same reasons for BC, i.e. variation in the height of boundary layer and the amount of local combustion. O\textsubscript{3} varies in a rather narrow range; however, it fluctuates more in the winter. These can serve as additional information to better review the site condition.

4.1.2. Relative importance of the different predictors to the modelled BC concentration

In this study, the database is comprised of 20 predictor variables for BC. The importance for each predictor variable is determined...
by various metrics for different WB models (Fig. 8). We limited the number of predictor variables to up to three in all sub-models for IAP, hence there is no simple way to compare all the 20 variables. Here we used simple Pearson’s correlation for IAP. For LASSO, since the dataset has been normalised, we took the coefficients of each standardised predictor variable to be the relative importance. DT uses curvature tests to split at nodes and decide which predictor variable is more preferred for the next branch. Since different metrics were used, we normalised them into the range of 0–1 for comparison. The resulting values indicate the relative importance with respect to the most important variable. A similar analysis was carried out by Singh et al. (2013).

The most important predictor variables illustrated by IAP are NO\textsubscript{x} and LDSA at street canyon and urban background, respectively. Many of the other aerosol and gaseous predictors have also relatively strong importance at both sites because they both are influenced by combustion which emits gaseous and aerosol pollutants at the same time (e.g. Helin et al., 2018; Rönkkö & Timonen, 2019). However, IAP puts low importance to most meteorological variables. At the street canyon site, wind speed influences the model most, with relative importance of about 0.4. In the urban background, on the other hand, meteorological variables generally influence

![Fig. 7. Diurnal pattern of traffic counts at the nearest traffic counter from the two measurement stations (a) during workdays and (b) during weekends.](image1)

![Fig. 8. Relative importance of input variables to the output variable, the logarithm of BC, (a) at Mäkelänkatu street canyon site and (b) at Kumpula urban background site. Only WB models are shown (black circle: IAP, red square: LASSO and blue cross: DT). The input variables on the x-axis are arranged in descending order of the average relative importance of the three methods. The asterisk signs denote that the variable is in its logarithm form. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)](image2)
stronger, compared to the street canyon site (Fung et al., 2020). This is because there are various BC sources from a distance at the urban background site than at the street canyon site. The longer distance from the source allows meteorological factors to impose a relatively strong influence on the estimation of BC, for instance dispersion and dilution. Wind speed, pressure and temperature are the three strongest predictor variables, with relative importance of about 0.6, 0.4 and 0.3, respectively. The LASSO regularisation has eliminated most of the meteorological variables, except temperature. It has the highest relative importance for NOx and accumulation mode (PN1.09) at street canyon and urban background, respectively. The second strongest variables, which are accumulation mode at the street canyon site and temperature at the urban background site, have the relative importance of about 0.5. The optimised λ in LASSO regularisation is 0.0082 and 0.265 at the street canyon site and at the urban background site, respectively. The higher the λ, the smaller the number of input predictors. The coefficients and intercepts of both IAP and LASSO linear models can be seen in Table 1 for street canyon and urban background.

As for DT, the strongest predictor variables are LDSA and NOx at street canyon and urban background, respectively. Generally, the relative importance by DT resembles the other WB models. However, PM2.5 and PM10 have almost zero influence at urban background unexpectedly, which might be owing to the high data incompleteness. As for meteorological data, they have least impacts on the fitting. PAR and wind speed are relatively stronger variables at street canyon and urban background, respectively.

Combining the relative importance by the three WB models, BC at street canyon has high correlation with NOx and LDSA, which mainly come from vehicular combustion. It appears that the BC concentration measured at the street canyon site has similar sources as of NOx. At the urban background site, similar parameters, NOx and LDSA, correlate well with BC, for the same reason. PM10, which is believed to come from longer range transport, and temperature also impose an influential factor to the concentration of BC. The more predictor variables with high relative importance further prove that BC at the urban background site comes from different sources.

4.2. Model evaluation

WB and BB models are evaluated by the attributes described in Section 3.4. Scatter plots of the fitting, histograms of residuals and diurnal cycle plots are presented to delineate the model accuracy and sensitivity. The latter part of this section further analyses the variation in model performance in different seasons, days of the week and hours of the day at both sites.

4.2.1. General analysis of model evaluation

Table 3 displays the overall performance of training set and testing set. R² for all the WB and BB models are higher than 0.8 (WB: R² = 0.81–0.87; BB: R² = 0.86–0.87) at the street canyon site in Mäkelänkatu. Among all methods, DT has a relatively low R² and high MAE. By observing the scatter plots, the fitting in DT gives many discrete levels due to its model architecture. A single tree with a decision split is not able to give spectrum of estimates. Increasing the tree depth might be able to solve the problem, but over-fitting issue might arise (Kang et al., 2018). On the other hand, RF grows decision trees to generalise the model and its R² is apparently higher than DT. In spite of a higher R², the fitting for RF, as well as SVR, at the lower tail diverges from the 1:1 line (Fig. 9a).

The model performance at the urban background site in Kumpula is generally worse (WB: R² = 0.44–0.60; BB: R² = 0.41–0.64) and shows more marked variation among all WB and BB models than in Mäkelänkatu (Table 3). LSTM performs best (R² = 0.64, RMSE = 0.2646 µg m⁻³) and LASSO’s MAE is the lowest. SVR has the lowest R² (R² = 0.41) while IAP generates the highest MAE. DT fails to give spectrum of estimates, likewise in street canyon. The extent of the discretion is more severe than that in street canyon. LASSO, RF, SNN and SVR tend to underestimate the extreme values and the fitting slope is less than 1 (Fig. 9b). In addition, IAP appears to overestimate all data points generally.

RF performs well for training data, but is not able to maintain a high R² for testing data. With the hyperparameters used, RF over-
fits datasets at both sites. DT is the single decision tree version of RF and apparently it shows poorer performance than RF. LASSO, on the other hand, is the only model that has a lower $R^2$ for training data than testing data. The shrinkage of the number of predictor variables seems to work well to avoid over-fitting. Overall, LSTM performs quite well in all cases, because it treats also data from the previous timestep to cope with the time dependency. SNN and LSTM models have resembling architecture where SNN has only one hidden layer and does not consider short-term memory. SNN turns out to perform generally worse than LSTM. IAP and SVR have fair fitting accuracy, but IAP is still recommended because it manages to fill in missing data by other predictor variables without extra interpolation.

The overall performance difference at Mäkelänkatu street canyon site and Kumpula urban background site indicates that the models are location-specific because of their different pollutant sources. The location-specific property of the models imposes

![Fig. 9. Scatter plots comparing the fitting performance of WB (upper row from left to right: IAP, LASSO and DT) and BB models (lower row from left to right: RF, SVR, SNN and LSTM) with testing set (a) at Mäkelänkatu street canyon site and (b) at Kumpula urban background site. Blue and red circles denote workdays and weekends, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)](image)
limitation to the use of the models as virtual sensors, which will be elaborated in Section 4.3. The street canyon has been suggested to have a relatively constant BC source (Helin et al., 2018; Luoma et al., in review, 2020), so the regression performance at the street canyon site is overall better than at the urban background where the large temporal variation is large on the impact of different BC sources, such as local traffic and residential wood combustion as well as long-range transported pollutants. The model performance also depends on the missing data patterns, such as completeness of the dataset and the length of the data gap, suggested by Junninen et al. (2004).

In short, the difference between WB and BB models are insignificant at both sites, which can further be justified by the time-series of measured BC and BC modelled by WB and BB (Fig. 10). The average absolute difference of WB (represented by LASSO) and BB (represented by LSTM) is 0.08 μg m⁻³ for both sites. The difference for more than 80% of the data points is less than 0.2 μg m⁻³. Due to the transparency property in WB models, one could relate the models with physics-based knowledge and be able to spot the possible errors. Based on the relative importance of predictors described in Section 4.1.2, one could recognise the variables that alter the results. The estimation of BB models, on the contrary, might contain misconceptions (Rudin, 2019), which are hard to discover.

4.2.2. Verification of model performance

We then investigate how the seasonal variation, the weekend effects and the diurnal cycle affect the model performance. Fig. 11a portrays the distribution of the residual (difference in measured BC and modelled BC) at Mäkelänkatu street canyon site. IAP, RF and DT have positive median residuals for all seasons, and especially high in autumn (0.035–0.038 μg m⁻³). For SNN, LSTM and SVR, summer has a relatively high positive median residuals (0.049–0.074 μg m⁻³), compared to other seasons. LASSO, unlike other models, underestimates the measurements in winter. Most models generate slightly more scattered distribution in summer, as indicated by the higher value of SD (0.32–0.42 μg m⁻³). This might be contributed to the lowered traffic during summer. The relative contribution of other sources is stronger and this adds up the uncertainty of the modelling.

In Fig. 11b, IAP and LSTM exhibit a remarkably high positive median residual (skewed normal distribution), which indicates overestimation in winter (0.072–0.155 μg m⁻³) at Kumpula urban background site. Not only in winter, IAP also overestimates BC concentration in summer and autumn, which adds up to an overall overestimation. DT shows strong overestimation in the summer (0.041 μg m⁻³) while SNN and SVR underestimate BC concentration in the same period (−0.028 μg m⁻³). Although the strong derivation in summer is found in many models aforementioned, the SD in summer is the lowest compared to the other seasons. The more varied performance in all models might be resulted from the lower completeness of data in urban background. Although we have interpolated the missing data, the data gap length is not considered, which make the interpolation unsatisfactory and lead to the poor model performance.

Despite the clear difference of BC concentration between workdays and weekends, weekend effect on model performance is not

![Fig. 10. Time-series of measured BC (black), BC modelled by LASSO (blue) and BC modelled by LSTM (red) in a selected timeframe (1st August – 15th August 2018) during the measurement period (a) at Mäkelänkatu street canyon and (b) at Kumpula urban background site. Shaded areas represent weekends. The dotted lines denote the partition of training and testing sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.]

observable in street canyon (Fig. 9a). In urban background, the fitting of LASSO, RF, SNN and LSTM on weekdays performs differently as on weekends. Their underestimation at extreme values are more severe on weekdays than that on weekends (Fig. 9b).

Fig. 12a shows diurnal pattern of measured BC and featured WB models (IAP and LASSO), together with BB models (RF and LSTM) in street canyon. Apart from seasonal variation and weekend effect, the models follow the diurnal pattern quite well. Most of them catch the two peaks on workdays at 6–8 a.m. and at 3–4 p.m.. However, all the featured models fail to predict the sharp increasing trend in the morning at 3–6 a.m. perfectly. LASSO slightly underestimates the period between the two peaks. On weekends, the featured models overestimate the declining trend after the only peak at 6–8 p.m..

At the urban background site, IAP tends to overestimate in all diurnal time-stamps on workdays (Fig. 12b). The other models also fail to catch the tips of the trough and peak at 4–9 a.m.. Among the featured models, RF performs best and manages to fit tightly the diurnal cycle. LSTM is the second best for workdays case. During weekends, the diurnal curve is comparatively flat. The medians of the
calculated values by WB and BB models further flatten the curve. Again, compared to the abovementioned methods, RF manages to best estimate the characteristics of the diurnal pattern.

4.2.3. Model evaluation in other aspects

Apart from the model accuracy (Table 3), other evaluation aspects, such as flexibility, complexity and efficiency, are also taken into consideration. The architecture of the WB models is transparent, which allows users to understand and modify, if necessary, the influencing variables. Among all WB models, IAP takes maximum three input predictors in an adaptive way, while the number of input predictors for LASSO depends on $\lambda$. DT takes all input predictors. For BB models, the inner components or logic are inaccessible. In case of anomalies, it is difficult to inspect and locate the problems from the model structure. Some BB models, for example SNN and LSTM, have many hyperparameters, which requires extensive efforts for model optimisation. In terms of efficiency, LASSO, DT and RF require the least computational resources among all the models we tested in this study. SVR and SNN take 30–50% longer time for the modelling process. Since IAP has to search for the best combination of input predictors and fill in missing data, its computational time can be five times higher than the most efficient ones. LSTM takes the longest computation time, still under 10 min, because it also deal with time delay in the modelling process.

Combining all the evaluation criteria, we deduce that IAP could be the best model despite the slightly higher computational time. It has relatively good accuracy at both sites. The model structure is transparent and the inner components can be inspected easily. The flexibility to select input predictors is also one of the advantages when using this model. The second best model could be LASSO since it performs evenly in both locations and could be considered to be the most generalised model. Its fast modelling process and transparent model structure could be also able to validate the high rank of this model.

4.3. BC models as virtual sensors

This section discusses the possibility of turning BC models into virtual sensors to provide additional information in support with physical sensing data. We further discuss the benefits and limitations of using BC models as virtual sensors, and how the virtual sensors are used to supplement the current AQI.
Virtual sensors use physical sensing data and calculate the outputs by suitable model to grasp the information of interested areas without placing real sensors there. Wilson (1997) suggested that virtual sensors should be able to provide 1) virtual measurements, 2) predictive capability, 3) continuous output based on periodic real measurements, and 4) robustness in case of physical sensor failure. These four criteria are used to assess the feasibility of using BC models as virtual sensors (Liu et al., 2009; Woo et al., 2016). On the ground that we deduced that IAP and LASSO could be the best two models, the assessment in this section only consider these two models. Based on the model evaluation of the BC models in Section 4.2, provided with adequate trained data in advance, virtual measurements can be given continuously with fairly high accuracy and predictive capability at the street canyon site. In case of physical sensor failure, models which take all the predictors might not work. IAP and LASSO still manage to predict BC by fewer input predictors. In extreme situations, where only meteorological parameters are available, IAP is still able to work thanks to its input adaptability.

However, there are limitations in using BC models as virtual sensors (Zaidan et al., 2020). In this paper, the two measurement sites can be characterised as street canyon and urban background. Each characterisation represents places with similar air pollutant sources. The corresponding BC models are also location-specific, but it might be possible to extend the models from one street canyon site to another with a calibration factor. Similar ideas could also be applied to urban background sites. Once we gather more training data from different locations, a network of database can be established. Within the network, selection of predictors can be interchanged. These calibration factors and the estimations from the virtual sensors might drift due to environmental changes or physical sensors degradation. These models might not be accurate anymore in case of occasional events, such as wildfire, extreme weather, etc. Therefore, the application of virtual sensors require a long-term drift monitoring and online-adaptive models could be used to adjust the models accordingly (Zaidan et al., 2020).

Once the virtual sensors are validated and approved, the continuous measurements can be utilised to update the current AQI. While there is no universally accepted method, most organisations consider PM\(_{10}\), PM\(_{2.5}\), NO\(_2\), O\(_3\), SO\(_2\) and CO as parameters for AQI calculation. From the scientific point of view, these parameters are insufficient to show association with health risk of aerosol particles, especially for cardiovascular effects (Geng et al., 2013). Although World Health Organization (2012) has recommended the inclusion of BC as one of the components in AQI alongside with the other air quality parameters, this has not yet been taken into action due to the unavailability of continuous BC measurements. This is attributed to the lack of national environmental legislations (Kutzner et al., 2018), instrument failure or data corruption (Junger & De Leon, 2015; Zaidan, Wraith, et al., 2019). For this, Monteiro et al. (2017) suggested to use modelling data to complement non-existent physical sensing data in AQI computation. Our BC models can then provide continuous modelling data with fairly good accuracy as virtual sensors. Similarly, other aerosols parameters with data unavailability, such as LDSA, can also be estimated by virtual sensors.

5. Conclusion

This paper presents the evaluation of white-box (WB) and black-box (BB) model performance for air quality monitoring, focusing in particular on black carbon (BC) measurement. The evaluation involves data in the period of 1\(^{st}\) January 2017–31\(^{st}\) December 2018 from the two measurement sites, the street canyon site in Mäkelänkatu and the urban background site in Kumpula, in Helsinki, Finland. The two selected environments possess the common features in urban areas and they will give feedback on the model performance. A total of 20 variables including aerosol particles, gaseous and meteorological parameters are used as predictor variables.

In street canyon, WB models perform (R\(_2\) = 0.81–0.87) in a similar way as BB models do (R\(_2\) = 0.86–0.87). The overall performance in urban background is much worse probably because of the greater variation in BC sources (Helin et al., 2018; Luoma et al., in review, 2020) and the greater extent of missing gap (Junninen et al., 2004). Similarly, the difference in WB (R\(_2\) = 0.44–0.60) and BB models (R\(_2\) = 0.41–0.64) is not noticeable in urban background. As for seasonal factor, fitted data in street canyon tend to overestimate in summer and autumn while this occurs in winter for data in urban background. The standard deviation of the residual distribution is the highest in summer in street canyon, but the lowest in urban background. In terms of the weekend effect, particular models, like Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Shallow Neural Network (SNN) and Long Short-Term Memory (LSTM), underestimate extreme values at both ends at a larger extent on weekdays than on weekends. RF appears to closely follow the diurnal pattern, followed by LSTM in both locations. Input-Adaptive Proxy (IAP), however, shows the worst fitting diurnally.

Among all models in this paper, LSTM surpasses others in term of accuracy at both the street canyon (R\(_2\) = 0.87) and the urban background sites (R\(_2\) = 0.64) probably due to its ability to consider the time dependency. RF has a high R\(_2\) for training sets and it manages to fit the diurnal pattern well. IAP performs fairly compared with other methods; however, it is still recommended to serve as a virtual sensor owing to its flexibility and its ability to fill in missing data by other variables in the same dataset. In spite of its longer computational time, it is easy to use with only three input predictors at most. LASSO could be considered to be the most generalised model because it performs evenly in both locations. It also has a high modelling speed and, therefore, it could be the second best model. Overall, despite the slightly higher performance for some individual BB models, the prediction might contain and perpetuate critical misconceptions that have gone unnoticed. The widespread use of machine learning models which are not comprehensive could impose a lasting negative impact in the field we are applying to (Rudin, 2019). WB models are still preferred because of its transparency of the architecture. Moreover, WB models are closer to physics-based models so that we can easily recognise the relative importance of the predictor variables and determine whether the model makes sense.

For our future step, we could combine the benefits of both models and improve the overall accuracy of the performance. Since we know that BC concentrations perform a mild seasonal cycle and clear diurnal cycle and that LSTM, which includes short-term memory,
outperforms the other models, it is of great interest to see how time dependency improves the precision of WB models. We could consider extracting the time dependency ability from LSTM and apply it to our selected WB models by, for instance, adding an autoregressive term. With the transparency properties of WB models, we could determine the lag of the autoregressive term, and verify the use of it.

Another possible improvement could be to utilise the proposed models to emulate existing physics-based models, such as ENFUSER (Johansson et al., 2015), which is expensive in computation. The developed emulator can speed up some modules in those expensive physics-based models. Reliable spatio-temporal air quality map can then be created.

Since the extent of missing data might be a factor that degrades the fitting performance, the existing network of several measurement stations within the Helsinki metropolitan region (Luoma et al., in review, 2020) is beneficial. In case of missing data, we could fill up the voids with the information from the other stations after conscientious calibration. For example, in this paper, the two measurement sites are characterised as street canyon and urban background. In a different setup, we may assume the similarity of the same type of environment and utilise the measurements as replacement.

Due to the comparatively scarce amount of continuous urban BC measurements (Kutzner et al., 2018), this BC estimation can be used as a virtual sensor in support with the reference measurement. The BC virtual sensor empowers a reliable continuous BC measurements. This continuous BC measurement could be useful in updating the current air quality indices so as to be better associated with health effect by air pollutants (World Health Organization, 2012).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work is supported by MegaSense program, the City of Helsinki Innovation Fund, Business Finland, the European Union through the Urban Innovative Action Healthy Outdoor Premises for Everyone (HOPE, project No. UIA03-240). Grants are also received from the European Research Council (ERC, grant No. 742206), the European Union’s Horizon 2020 Research and Innovation Framework Program (grant agreement No. 654109), and ERA-PLANET (www.era-planet.eu) and its trans-national project SMURBS (www.smurbs.eu) funded under the same program (grant agreement No. 689443). The authors show gratitude to Academy of Finland for the funding via the Centre of Excellence in Atmospheric sciences (project No. 272041) and NanoBioMass (project No. 1307537). The authors acknowledge the City of Helsinki for providing traffic data.

References

Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mikaela, J. M., Hoell, C., O’dowd, C. D., Hansson, H. C., Väkevä, M., et al. (2001). Physical characterization of aerosol particles during nucleation events. Tellus B: Chemical and Physical Meteorology, 53, 344–358. https://doi.org/10.1034/j.1600-0870.2001.00011.x
Bair, W. (1994). The revised international commission on radiological protection (ICRP) dosimetric model for the human respiratory tract—an overview. Annals of Occupational Hygiene, 38, 251–256.
Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123–140. https://doi.org/10.1007/BF00058655
Cabaneros, S. M. S., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. Environmental Modelling & Software. https://doi.org/10.1016/j.envsoft.2019.06.014
Carbone, S., Aurela, M., Saario, K., Saarikoski, S., Timonen, H., Frey, A., Supeer, D., Ulbrich, I. M., Jimenez, J. L., Kulmala, M., et al. (2014). Wintertime aerosol chemistry in sub-Arctic urban air. Aerosol Science and Technology, 48, 313–323. https://doi.org/10.1080/02786826.2013.875115
Davalos, A. D., Luben, T. J., Herring, A. H., & Sacks, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. Annals of Epidemiology, 27, 145–153. https://doi.org/10.1016/j.amepi.2016.11.016
Dos Santos-Juusela, V., Petäjä, T., Kousa, A., & Hämeri, K. (2013). Spatial–temporal variations of particle number concentrations between a busy street and the urban background. Atmospheric Environment, 79, 324–333. https://doi.org/10.1016/j.atmosenv.2013.05.077
Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and other ensemble methods. Neural Computation, 6, 1289–1301. https://doi.org/10.1162/neco.1994.6.6.1289
European Environment Agency. (2019). Air quality in Europe — 2019 report. European Environment Agency. (2019). Air quality in Europe — 2019 report.
Fernández-Guisuraga, J. M., Castro, A., Alves, C., Calvo, A., Alonso-Blanco, E., Blanco-Alegre, C., Rocha, A., & Fraile, R. (2016). Nitrogen oxides and ozone in Portugal: Trends and ozone estimation in an urban and a rural site. Environmental Science and Pollution Research, 23, 17171–17182. https://doi.org/10.1007/s11356-016-6880-6
Freeman, B. S., Taylor, G., Bharaghbi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 68, 866–886. https://doi.org/10.1080/10962247.2018.1549956
Fung, P. L., Zaidan, M. A., Sillanpää, S., Kousa, A., Niemi, J. V., Timonen, H., Kuula, J., Saulkko, E., Luoma, K., Petäjä, T., et al. (2020). Input-adaptive proxy for black carbon as a virtual sensor. Sensors, 20, 188. https://doi.org/10.3390/s20010182
Geng, F., Hua, J., Mu, Z., Peng, L., Xu, X., Chen, R., & Kan, H. (2013). Differentiating the associations of black carbon and fine particle with daily mortality in a Chinese city. Environmental Research, 120, 27–32. https://doi.org/10.1016/j.envres.2012.08.007
Gross, A. M. (1977). Confidence intervals for bisquare regression estimates. Journal of the American Statistical Association, 72, 341–354. https://doi.org/10.1080/01621459.1977.10481001
Helin, A., Niemi, J. V., Virkkula, A., Pirjola, L., Teinilä, K., Backman, J., Aurela, M., Saarikoski, S., Rönkkö, T., Aarni, E., et al. (2018). Characteristics and source apportionment of black carbon in the Helsinki metropolitan area, Finland. Atmospheric Environment, 190, 87–98. https://doi.org/10.1016/j.atmosenv.2018.07.022
Hellén, H., Kangas, L., Kousa, A., Vestenius, M., Teinilä, K., Karppinen, A., Kukkonen, J., & Niemi, J. V. (2017). Evaluation of the impact of wood combustion on benz[a]pyrene (BaP) concentrations; ambient measurements and dispersion modeling in Helsinki, Finland. Atmospheric Chemistry and Physics, 17, 3475–3487. https://doi.org/10.5194/acp-17-3475-2017
Timonen, H., Aurela, M., Carbone, S., Saarnio, K., Frey, A., Saarikoski, S., Teiniä, K., Kulmala, M., & Hillamo, R. (2014). Seasonal and diurnal changes in inorganic ions, carbonaceous matter and mass in ambient aerosol particles in an urban, background area. Boreal Environment Research, 19, 71–86.

Timonen, H., Carbone, S., Aurela, M., Saarnio, K., Saarikoski, S., Ng, N. L., Canagaratna, M. R., Kulmala, M., Kerminen, V. M., Worsnop, D. R., et al. (2013). Characteristics, sources and water-solubility of ambient submicron organic aerosol in springtime in Helsinki, Finland. Journal of Aerosol Science, 56, 61–77. https://doi.org/10.1016/j.jaerosci.2012.06.005

Ting, K. M., & Witten, I. H. (1997). Stacked generalization: When does it work?.

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

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. In Advances in Neural Information Processing Systems, 281–287.

Wang, J., & Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing, 314, 198–206. https://doi.org/10.1016/j.neucom.2018.06.049

Wilson, E. (1997). Virtual sensor technology for process optimization. In Symposium on computers and controls in the metals industry in iron and steel society.

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

Zhang, Y., Li, M., Cheng, Y., Geng, G., Hong, C., Li, H., Li, X., Tong, D., Wu, N., Zhang, X., et al. (2019). Modeling the aging process of black carbon during atmospheric transport using a new approach: A case study in Beijing. Atmospheric Chemistry and Physics, 19, 9663–9680. https://doi.org/10.5194/acp-19-9663-2019

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

Zaidan, M. A., Motlagh, N. H., Fung, P. L., Lu, D., Timonen, H., Kuula, J., Niemelä, J., Tarkoma, S., Pere, J., Kajala, T., Kulmala, M., et al. (2020). Intelligent calibration and virtual sensing for integrated low-cost air quality sensors. IEEE Sensors Journal. https://doi.org/10.1109/JSEN.2020.3010316

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