Environmental Research 212 (2022) 113269

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Non-linear models for black carbon exposure modelling using air pollution datasets

J. Rovira [a], J.A. Paredes-Ahumada [b], J.M. Barceló-Ordinas [b], J. García-Vidal [b], C. Reche [c], Y. Sola [a], P.L. Fung [d], T. Petäjä [d], T. Hussein [d,e], M. Viana [c,*]

[a] Barcelona University, Barcelona, Spain
[b] Department of Computer Architecture, Universitat Politècnica de Catalunya, UPC, Barcelona, Spain
[c] Institute of Environmental Assessment and Water Research, Spanish Research Council, IDAEA-CSIC, Barcelona, Spain
[d] University of Helsinki, Institute for Atmospheric and Earth System Research (INAR/Physics), UHEL, Helsinki, Finland
[e] The University of Jordan, School of Science, Department of Physics, Amman, Jordan

ARTICLE INFO

Keywords:
Virtual sensor
Input-adaptive
Data gaps
Novel parameters
Human health
Absorption

ABSTRACT

Black carbon (BC) is a product of incomplete combustion, present in urban aerosols and sourcing mainly from road traffic. Epidemiological evidence reports positive associations between BC and cardiovascular and respiratory disease. Despite this, BC is currently not regulated by the EU Air Quality Directive, and as a result BC data are not available in urban areas from reference air quality monitoring networks in many countries. To fill this gap, a machine learning approach is proposed to develop a BC proxy using air pollution datasets as an input. The proposed BC proxy is based on two machine learning models, support vector regression (SVR) and random forest (RF), using observations of particle mass and number concentrations (N), gaseous pollutants and meteorological variables as the input. Experimental data were collected from a reference station in Barcelona (Spain) over a 2-year period (2018–2019). Two months of additional data were available from a second urban site in Barcelona, for model validation. BC concentrations estimated by SVR showed a high degree of correlation with the measured BC concentrations (R^2 = 0.828) with a relatively low error (RMSE = 0.48 μg/m^3). Model performance was dependent on seasonality and time of the day, due to the influence of new particle formation events. When validated at the second station, performance indicators decreased (R^2 = 0.633; RMSE = 1.19 μg/m^3) due to the lack of N data and $PM_{2.5}$ and the smaller size of the dataset (2 months). New particle formation events critically impacted model performance, suggesting that its application would be optimal in environments where traffic is the main source of ultrafine particles. Due to its flexibility, it is concluded that the model can act as a BC proxy, even based on EU-regulatory air quality parameters only, to complement experimental measurements for exposure assessment in urban areas.

1. Introduction

Exposure to fine particles in polluted air accounts for approximately seven million premature deaths every year (Lelieveld et al., 2015; WHO, 2018), globally. While poor air quality is associated with an increasing variety of cardiovascular and respiratory disease, recent research evidences statistically significant health impacts even at low concentrations (Brunekreef et al., 2021). As a result, air pollution monitoring and mitigation remains a key challenge for urban areas (Viana et al., 2020).

In order to understand the nature of urban air pollution, EU-reference air quality monitoring stations are available across Europe (Hussein et al., 2012), which monitor the parameters regulated by the Air Quality Directive (CO, NOx, SO_2, O_3 and particles PM_{10}, $PM_{2.5}$). Aside from these parameters, two relevant aerosol metrics are so far un-regulated: ultrafine particles (UFPs; particles smaller than 100 nm in diameter), and black carbon (BC). UFPs penetrate deep into the respiratory tract and are especially linked to health impacts due to their high surface area to mass ratios (Oberdörster et al., 2007). BC is emitted from the incomplete combustion of carbonaceous material and it is typically associated with vehicle exhaust, coal-fired power plants, and biomass burning for heating and cooking (Petzold et al., 2013). BC is a relevant component of PM in European cities, contributing 5%–15% to PM mass

* Corresponding author.
E-mail address: mar.viana@idaea.csic.es (M. Viana).

https://doi.org/10.1016/j.envres.2022.113269
Received 14 January 2022; Received in revised form 1 April 2022; Accepted 6 April 2022
Available online 13 April 2022

concentrations in urban air (Cavalli et al., 2016). BC exposure has negative implications for human health as well as for regional and global climate and extreme weather events (Bond et al., 2013; Saide et al., 2015).

Monitoring BC is valuable from a research perspective as well as for air quality and human exposure monitoring, as it is a tracer of traffic emissions and atmospheric processes (Luoma et al., 2021; Pakkanen et al., 2000; Reche et al., 2011a). Monitoring this parameter in urban areas would provide high added value during the design and testing of the effectiveness of mitigation strategies targeting road traffic. However, the complexity and cost of the instrumentation and the lack of reference monitoring protocols for BC results in limited data availability. To address this gap, diverse modelling approaches are applied aiming to improve the spatio-temporal coverage of non-regulated (e.g., BC, UFPs) and regulated metrics (e.g., $PM_{2.5}$, NO_2). Land-use regression (LUR) models are one of the most frequently used tools, especially as input for epidemiological research (Jones et al., 2020; Kerckhoffs et al., 2021, 2017; Tripathy et al., 2019; Vizcaino and Lavalle, 2018). Other approaches are based on GIS (Ma et al., 2019) and spatio-temporal analysis (Simon et al., 2020; van de Beek et al., 2021). As a novel tool, an input-adaptive proxy model was developed (Fung et al., 2021), with which air quality datasets may be used to estimate BC concentrations not only in the present, but also historically or in the future as prediction. Input-adaptive proxies have so far been developed for BC and lung-deposited surface area (LDSA) (Fung et al., 2021, 2019; Martha A. Zaidan et al., 2019), using white and black-box models based on Bayesian and linear mixed-effects models. The models were successfully tested in Helsinki (Finland) and Amman (Jordan).

The present work develops an input-adaptive proxy for BC based on two machine learning models, support vector regression (SVR) and random forest (RF), and tests their performance in a Mediterranean urban environment (Barcelona, Spain). The use of a BC proxy based on data-driven models allows the estimation of BC concentrations where BC is not directly measured, but where other pollutants are measured. BC was considered an adequate candidate for the application of a data-driven model, as its concentrations in urban environments typically correlate with those of traffic-related gaseous pollutants such as CO, NO, NO_2, and fine particulate matter ($PM_{2.5}$) and submicron particle number concentration (or ultrafine particles; N) in this urban environment (Brines et al., 2015; Reche et al., 2011a). It was hypothesised that BC concentrations can be estimated using multipollutant datasets combining regulated ($PM_{2.5}$, O_3, NO_2) and non-regulated (N) concentrations. The application of this proposed input-adaptive proxy model is foreseen relevant and useful for regulatory as well as research (exposure assessment, air quality) purposes, for example to predict BC concentrations and for gap-filling in time series suffering from instrumental failure.

2. Methods

2.1. Monitoring sites

Air quality measurements were carried out at the EU-reference urban background monitoring station Palau Reial, in Barcelona (41°23′14″ N, 02°06′56″E, 80 m a.s.l.;Figure S1). The site is influenced by vehicular emissions, as evidenced by the daily patterns of BC and particle number concentrations (N) (Reche et al., 2011a). Air quality data were also collected from a second EU-reference urban background site (Roma Ave.), for subsequent validation of the modelling results. Both stations are part of the Barcelona reference air quality network (XVPCA; http://mediambi-ent.gencat.cat/).

2.2. Air quality and meteorology datasets

A combination of conventional (regulated) and novel (non-regulated) air quality parameters were monitored at Palau Reial for a 2-year period (2018–2019). Data from EU reference analysers were collected for tropospheric ozone (O_3), nitrogen oxide (NO), and nitrogen dioxide (NO_2), with a 1-h time resolution. Particulate matter concentrations (PM_{10}, $PM_{2.5}$, PM_1) were monitored with an environmental dust monitor Grimm EDM180, corrected against reference gravimetric measurements, with a 10-min time resolution. Black carbon mass concentrations were monitored using a multiangle absorption photometer (MAAP, Thermo ESM Andersen Instruments) fitted with a PM_{10} inlet, operating on a 1-min time resolution. The MAAP determines absorbance by particles deposited on a filter using measurements of transmittance and reflectance at different angles. The absorbance was converted to BC mass concentrations using the default 6.6 m^2/g mass absorption coefficient at 637 nm (Müller et al., 2011; Petzold et al., 2013). Finally, total particle number concentrations (N) were monitored with a water-based condensation particle counter (WCPC TSI 3785), operating on a 5-min time resolution and measuring in the size range 5–1000 nm. The data for all parameters were averaged to a 1-h time resolution, and the entire row of data was removed whenever a missing value was encountered. Data availability was >80% for the period 2018–2019.

Meteorological variables (temperature, relative humidity and boundary layer height, PBL) were obtained from a meteorological station located on the rooftop of the Faculty of Physics at Barcelona University, at approximately 400 m from Palau Reial.

In addition to the Palau Reial dataset, 2 months of data were collected from the Roma Ave. station to assess the applicability of the model at a second urban location. The Roma Ave. dataset included a more limited set of parameters (NO_2 and O_3, monitored with reference instrumentation, and BC, monitored with an AE33 Magee Aerosol doo aethalometer; with 1 h time resolution). Ambient temperature (T) and relative humidity (RH) were obtained from the Faculty of Physics, as in the case of Palau Reial.

2.3. The black carbon proxy model

Machine learning (ML) infers plausible models to explain observed data, which are capable of making predictions about unobserved data and take rational decisions based on these predictions (Hastie et al., 2009). Specifically, machine learning models are systems which generate predictions (output) based on the data inputs they receive. Within the machine learning models, prediction techniques include linear and nonlinear methods, the latter requiring hyperparameters. To find the best set of hyperparameters per each algorithm a N-fold cross-validation strategy is typically used (Hastie et al., 2009).

To test the performance of the BC proxy, we compare two models based on nonlinear supervised machine learning: support vector regression and random forest. The first model is the support vector regression (SVR) (Drucker et al., 1997), used in previous works for calibrating low-cost NO_2 and O_3 sensors (Barcelo-Ordinas et al., 2019; Ferrer-Cid et al., 2019; Ripoll et al., 2019). SVR, a nonlinear model, is a kernel method that is the analogous of support vector machines (SVMs; Cortes and Vapnik, 1997) which uses continuous values instead of classifying as SVM. It maps the data to a higher dimension in order to find a better regression curve while performing computations in input space via a positive-definite kernel function $K(x, x')$. The points which are far away from the correct regression plane will be the ones relevant for the correct model building. This is achieved via the ε-insensitive error loss, where only the points with error greater than ε are considered. The resulting SVR function is as follows:

$$\hat{y}(x) = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) K(x, x_i) + b$$

The values for the parameters $\hat{\alpha}_i^*$ and $\hat{\alpha}_i$ are found by solving a quadratic optimisation problem. The objective function to solve is obtained with the dual formulation of the problem, minimising a loss function. The radial basis function (RBF) kernel was used in this work.

The RBF kernel is proven to have an implicit map of infinite dimension. Finally, the hyperparameters optimised via cross-validation were the variance of the RBF kernel, the ε in the loss function, and a penalisation term C.

The second model was the random forest (RF, Breiman, 2001). It is an ensemble method that uses multiple decision trees to achieve better results. Each decision tree begins with a root node and partitions the data into subsets that contain instances with similar values. The subdivision continues using different model parameters until it reaches a leaf node in which a decision is made on the numerical value of the input. For any new input, the leaf node it falls is determined by starting at the root node and following a path according to the different criteria at each node. The mean squared error function was used to measure the quality of a split at each node. The number of trees in the forest varied from 1 to 1000.

In both models, the input variables were combinations of PM_1, $PM_{2.5}$, PM_{10}, NO_2, O_3, N, PBL, temperature and relative humidity, to model hourly BC concentrations as output.

To run each of the models, a randomly selected fraction of the dataset (80%) was used for training the model and the remaining fraction (20%) for validating the model. Randomising the data is essential to ensure that the model is challenged with the broadest range of concentrations among the input variables, covering different seasons in the year and times of day. Finally, a 10-fold cross-validation strategy was used to obtain the SVR and RF hyperparameters.

The models' root mean square error (RMSE), relative root mean square error (RRMSE) and coefficient of determination (R^2) were used as diagnostic evaluation attributes (Fung et al., 2019). While R^2 measures the amount of variance explained by the independent variables, the RMSE estimates the absolute difference between the modelled and measured mass concentrations. The RMSE is calculated as the square root of the average squared difference between the forecast and the observation pairs, while the RRMSE is calculated as the ratio between the RMSE and the mean BC concentration value for the observations.

The models were run for different datasets, constructed using the original datasets from Palau Reial (2 years of data, 2018–2019) and Roma Ave. (2 months of data). The main comparison between models was carried out for the full dataset, comprising 2 years of data. In addition, the SVR model was run for different scenarios characterised by different sources and atmospheric processes. The 6 datasets included the following data:

- Full dataset: 2 full years of data (2018 and 2019), Palau Reial; full dataset without any filtering (8011 samples).
- Winter: December to February 2018 and 2019, Palau Reial; aiming to avoid new particle formation (NPF) events typical of summer (3049 samples).
- Summer: June to August 2018 and 2019, Palau Reial; aiming to focus on NPF events (674 samples).
- Midday: hourly values between 10:00 and 14:00 UTC, Palau Reial; aiming to focus on midday NPF events (1598 samples).
- Day: hourly values between 14:00 and 10:00 UTC, Palau Reial; aiming to avoid NPF events (6772 samples).
- Roma Ave.: November and December 2020, Roma Ave; for model validation at a location different from where the model was trained. Hourly concentrations for BC, NO_2 and O_3 (105 samples).

3. Results and discussion

3.1. Daily variability of atmospheric pollutants

Prior to BC modelling, the mean daily patterns of gaseous and particulate pollutants at the Palau Reial site were evaluated, for the period 2018–2019 (Fig. 1). The purpose was to understand the daily pollutant trends and to assess their representativity. As observed in previous works (Brines et al., 2015; Carnerero et al., 2021; Reche et al., 2011a, 2011b, 2015), NO_2 and BC concentrations followed a diurnal pattern influenced by traffic activity, with maxima during the morning and evening rush hours (06:00–08:00 and 18:00–21:00 UTC), decreasing during midday mainly due to lower emissions and atmospheric dilution (Fig. 1). Ozone levels showed the characteristic inverse trend, with an increase at midday coinciding with the maximum photochemistry and solar radiation during the central hours of the day. Submicron particle number concentrations (N) were also closely related to traffic emissions, with maxima during the morning and evening rush hours. However, this parameter is known to be highly influenced in Barcelona by new particle formation (NPF) during the central hours of the day, especially in summer (photochemically induced nucleation; Carnerero et al., 2019; Cheung et al., 2011). Therefore, N and BC maxima were typically detected during morning and evening rush-hours linked to traffic (Maricq, 2007; Wehner et al., 2009), while N showed an additional maximum at midday, coinciding with a decrease in BC concentration, resulting from photochemical nucleation processes. This midday nucleation takes place in Barcelona as a consequence of the high solar radiation, the growth of the mixing layer, the increase in wind speed and the consequent decrease in pollutant concentrations.

As a result, the dataset selected was considered representative. BC and N patterns observed in this study for Barcelona were in agreement with results shown by previous authors (Brines et al., 2015; Reche et al., 2011a), who observed similarities between N and BC daily variability in North-European cities in contrast with the differences observed in high-insolation climates.

3.2. Data-driven modelling – predicting BC concentrations

An initial assessment of the relationship between each of the input parameters and BC concentrations was carried out based on Spearman's coefficient (Table S1). $PM_{2.5}$ and NO_2 showed the highest correlation

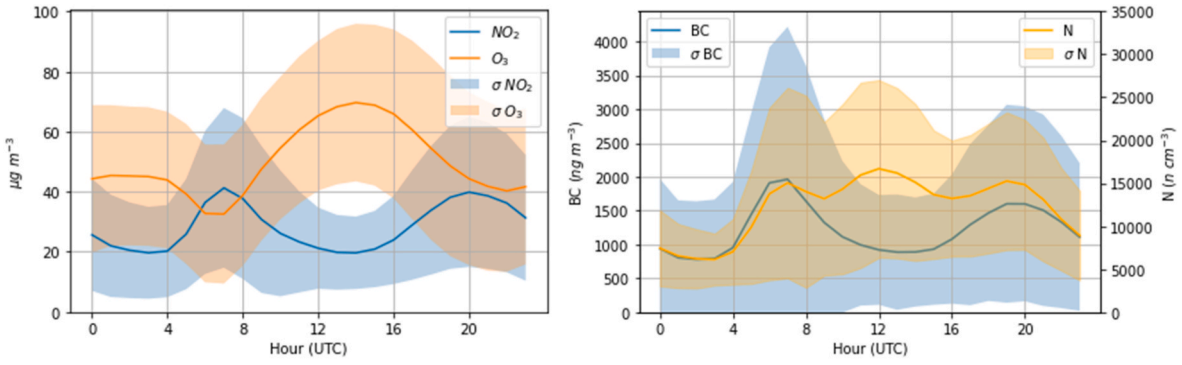

Fig. 1. Mean hourly variation of gaseous pollutant concentrations (NO_2 and O_3; left) and black carbon (BC) and particle number (N) concentrations (right) at the Barcelona Palau Reial station for the period 2018–2019.

with hourly BC concentrations (>0.60), while ozone was anti-correlated with BC (−0.51). N showed a moderate positive correlation (0.54). The relationship with meteorological parameters (T, RH, PBL) was not especially relevant. PBL, included in the analysis as an indicator of atmospheric dilution processes, showed an inverse relationship with BC although minor in terms of Spearman's coefficient (−0.077).

Fig. 2 shows the results (in terms of R^2) obtained after applying SVR models to the full dataset from Palau Reial (PR), for a two-year period (2018–2019). The full dataset was divided into a training and a test set and each variable was used individually as input to the model. The process of splitting the dataset was repeated 10 times for each variable. The predictions were compared with the BC concentration in terms of R^2. In univariable models NO_2 and O_3 showed better results than the other variables, but NO_2 was the only parameter which achieved a fit with $R^2 > 0.5$ (median $R^2 = 0.537$).

Table 1 summarises the results of multivariable SVR on the full dataset from PR. The Table shows the independent variables used as input, and the results obtained when compared to the dependent variable (BC) in terms of R^2 and RMSE for the testing dataset. The order of the input variables was obtained by forward selection. In this method, the model starts with a single input variable. In each iteration, the variable which improved the model the most was added to the model until all variables were considered. The first variable selected was NO_2, as since it showed the best performance in univariable models Table S1 and Fig. 2).

As shown in Table 1, the model performance varied as a function of the input variables selected, with R^2 for the testing dataset spanning between 0.537 and 0.828. After including 6 variables, the R^2 for the testing dataset was >0.800, suggesting that the model was able to reproduce observed BC concentrations with a reasonably high degree of correlation. In terms of RMSE, results were also promising given that errors were <0.69 μg/m^3 when using at least 2 variables. The highest model fit for the full dataset (R^2 testing = 0.828; RMSE = 0.478 μg/m^3; RRMSE = 36%) was obtained using all the variables as input (NO_2, N, PM_{10}, $PM_{2.5}$, PM_1, O_3, T, RH, PBL). In addition, if we select an arbitrary threshold of $R^2 = 0.800$, this was already achieved with the combination of 6 variables (NO_2, N, $PM_{2.5}$, O_3, T, PM_1). As the number of input variables was reduced, the fit between the modelled and measured BC concentrations decreased (Table 1). The lowest R^2 obtained (R^2 testing = 0.537) used only NO_2 as input. It is relevant to take into account that adding a large number of parameters in a machine learning model may lead to overfitting, which in this case was not observed. Adding certain input parameters may result in significant quantitative leaps in R^2 or RMSE, while others may result in only a marginal leap: for example, including PM_{10}, and PBL only improved R^2 and RMSE by a few in a few

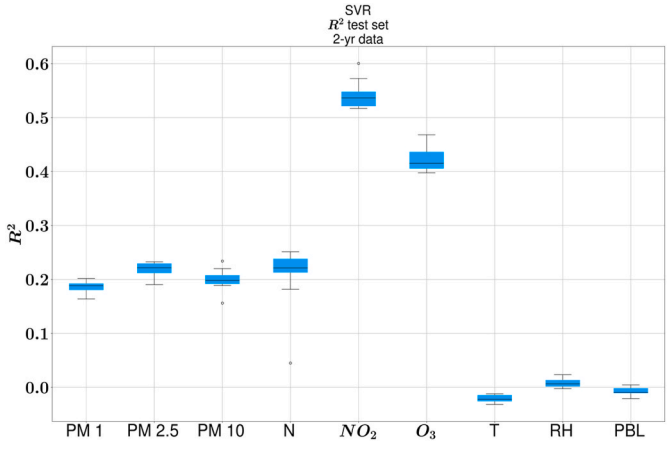

Fig. 2. Box plot of R^2 on the testing dataset between measured and modelled hourly BC concentrations, for each variable individually. Results obtained after dividing the dataset 10 different times into training and test datasets.

Table 1
SVR model performance (RMSE in μg/m^3, RRMSE in % and R^2 for the testing dataset) for the full dataset, with data collected over the full 2-year period (2018–2019) at the Palau Reial reference station.

Input parameters	R^2	RMSE	RRMSE
NO_2	0.537	0.805	60.0
NO_2, N	0.660	0.671	50.0
NO_2, N, $PM_{2.5}$	0.725	0.604	45.0
NO_2, N, $PM_{2.5}$, O_3	0.761	0.562	41.9
NO_2, N, $PM_{2.5}$, O_3, T	0.780	0.541	40.3
NO_2, N, $PM_{2.5}$, O_3, T, PM_1	0.810	0.501	37.3
NO_2, N, $PM_{2.5}$, O_3, T, PM_1, RH	0.822	0.485	36.1
NO_2, N, $PM_{2.5}$, O_3, T, PM_1, RH, PM_{10}	0.826	0.480	35.8
NO_2, N, $PM_{2.5}$, O_3, T, PM_1, RH, PM_{10}, PBL	0.828	0.478	35.6

thousandths (Table 1). In accordance with the results from Table 1, the parameters which contributed the most (and therefore were considered critical) to estimate BC in this work, for the full dataset, were NO_2, N and $PM_{2.5}$. Parameters considered as correctors, with only marginal improvements to model performance, were PM_{10}, PBL and RH (Table 1).

The scatter plot and the time series for 1 month of the measured *vs.* modelled BC concentrations for the most optimal combination of input variables (resulting in $R^2 = 0.828$, Table 1) for the full dataset are shown in Fig. 3. The Figure suggests better model performance for low when compared to high BC concentrations, given that data dispersion increased significantly for measured BC concentrations >7 μg/m^3. In addition, the model tends to underestimate the highest measured BC concentrations as well as the seasonal variability of the model's performance.

These results are aligned with those reported by Zaidan et al. (2019) and Fung et al. (2019), who also modelled BC using black- and white-box approaches and obtained R^2 ranging between 0.74 and 0.94 for measured *vs.* modelled BC concentrations. Conversely, higher RMSE were obtained (0.19–2.3 μg/m^3) than in the present study (0.48–0.81 μg/m^3).

In addition to the SVR model, a RF model was also applied to the full dataset aiming to compare the performance of two machine-learning models. Results from the RF model did not significantly differ from those obtained with SVR (R^2 = 0.544–0.808; Table S2 and Figure S2). The order in which the variables were included in the forward selection algorithm was different for each of the models, and this is why the parameter combinations are not shown in Table S2. The relative weight of $PM_{2.5}$ and N was different for both models, and PM_1 was not as critical in RF as it was with SVR. Figure S2 shows the performance of SVR and RF in terms of R^2 and RMSE, as the number of parameters considered according to forward selection increased. The SVR model was slightly better at predicting hourly BC concentrations, but the difference was not significant, and we conclude that both models can be used to build the BC proxy.

3.3. Model performance as a function of air pollutant emissions and meteorology

After assessing model performance for the full dataset, the model was challenged with different subsets of data where BC concentrations were influenced by different emission sources and atmospheric processes. The full dataset was divided into the summer and winter periods with the aim of assessing the influence of new particle formation events on model performance, given that N was one of the parameters identified as critical. Similarly, subsets of data were also created for the midday hours of the day *vs.* the rest of the day. For each subset of data, a forward selection method was used. The results for the different datasets are shown in Fig. 4. The box plots describe the interquartile range (IQR) of the R^2 and RMSE solutions obtained for the testing datasets, respectively. The results used to generate the box plots (9 variables considered for the forward selection) can be found in Table 1 and in Supporting

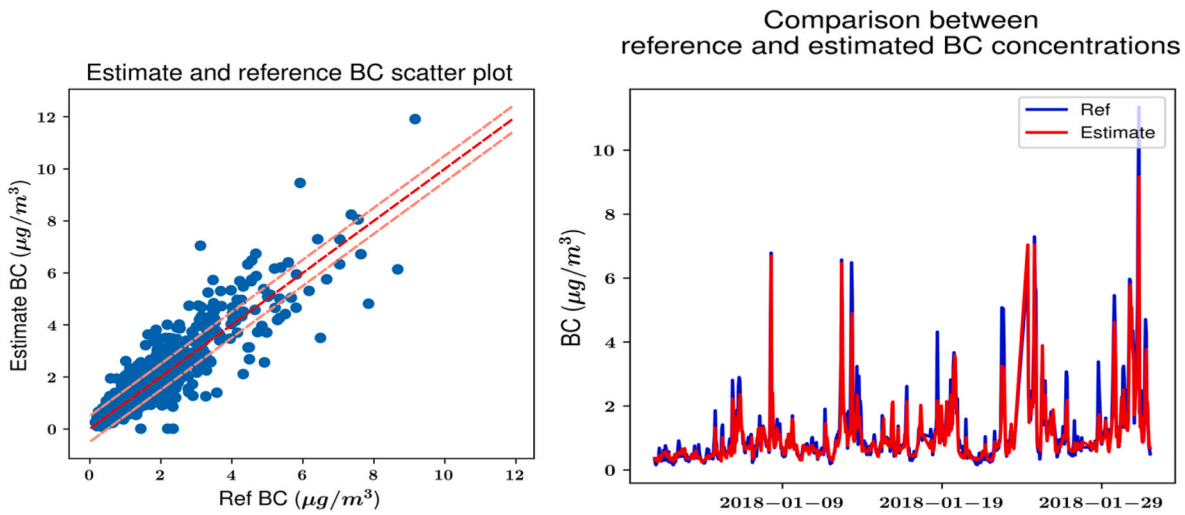

Fig. 3. Left: scatter plot of estimated BC *vs.* measured (Ref) BC for the optimal parametrisation for the full dataset (with NO_2, N, PM_{10}, $PM_{2.5}$, PM_1, O_3, T, RH and PBL as input variables), for the period 2018–2019 at the Palau Reial site (full dataset). The red line indicated 1:1, while the thinner dotted lines indicate a 0.5 μg/m^3 uncertainty range. Right: time series of estimated BC and measured (ref) BC for this parametrisation, for the month of January 2019, as an example.

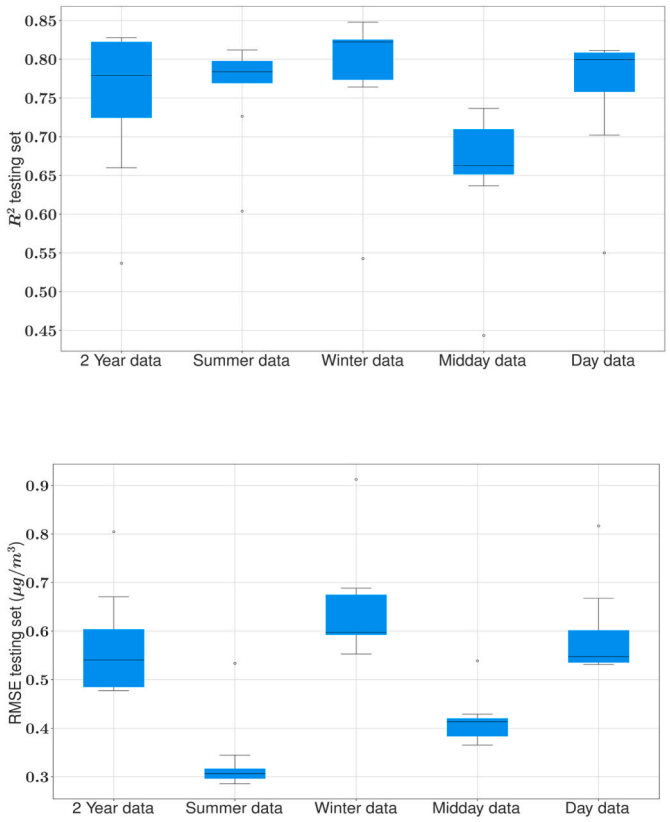

Fig. 4. Box plot of R^2 (top) and RMSE (bottom) between measured and modelled hourly BC concentrations for the different subsets of data. Results shown only for the testing datasets. Year_PR: full dataset (hourly values for 2018 and 2019); Winter_PR: hourly data for December, January and February 2018 and 2019; Summer_PR: hourly data for June, July and August 2018–2019; Midday_PR: all hourly values between 10:00 and 14:00 UTC; Day_PR: all hourly values from 14:00 to 10:00 UTC; Roma Ave.: dataset for model validation, including data from this site for November and December 2020.

Information (Tables S3, S4, S5, S6 and S7).

As in the case of the full dataset, the correlation between modelled and measured BC evidenced the model's capability to predict BC concentrations in the urban environment selected, with the optimal R^2 coefficients for the testing datasets >0.800 for the seasonal datasets. Model performance was the highest for winter (R^2 = 0.848, Fig. 4 and Table S4 in Supporting Information) and decreased in summer (R^2 = 0.812; Table S5), as expected due to the influence of nucleation peaks in summer which are uncorrelated with those of BC. Model performance was poorest for the daily datasets, with the lowest R^2 obtained for the midday dataset (R^2 = 0.737, Fig. 4 and Table S6), and a higher R^2 for the dataset excluding the midday hours (R^2 = 0.811; Table S7). Once again, this was explained by the influence of new particle formation during midday hours. The fit was better for the dataset excluding the midday hours as during those hours BC was mainly driven by road traffic emissions.

Overall, the optimal parametrisation for all of the subsets of data included N as a critical variable. These results confirm that the absence or presence of NPF events is key for the model's success. This has implications regarding the potential for application of the model especially in environments where traffic is the main driver of N concentrations, and with low influence of NPF events (e.g., highly polluted environments, colder climates, etc.). Despite this, it should be noted that, for Barcelona, the model was able to adequately predict BC concentrations during traffic rush hour periods, when population exposure is typically the highest (Reche et al., 2015; Rivas et al., 2014).

In terms of RMSE (Fig. 4, bottom) results ranged between 0.286 and 0.553 μg/m^3, with the lower RMSE being obtained for the summer dataset (0.286 μg/m^3). Contrarily to the R^2 coefficients, the best performance (lowest RMSE) was obtained for the summer dataset, probably linked to the smaller size of the dataset: fewer values imply fewer large errors, resulting in lower RMSEs. The largest RMSE (for the winter and full day datasets, RMSE = 0.553 μg/m^3 and 0.478 μg/m^3, respectively) were at the low end of the values reported in the literature (0.19–2.3 μg/m^3, Fung et al., 2019; Zaidan et al., 2019). When calculating RMSE, larger errors have a disproportionately large effect, which means that having a limited number of outliers has a strong impact on the calculated RMSE.

Finally, the model was validated with data from a second reference station (Roma Ave.) after it had been trained with the winter dataset from Palau Reial. This period was selected to match the environmental conditions for the period of data available from Roma Ave. (Nov–Dec 2020), even though due to the impact of the COVID-19 pandemic the year 2020 should not be considered as fully representative of traffic emission patterns. Prior to this validation at a different station, the model was tested to predict BC concentrations at Palau Reial from the

winter of 2020 when trained with data from 2018 to 2019 (2 full years), to confirm the capability of the model to predict towards future years at a given station. Because the results for Palau Reial were positive (R^2 = 0.837, RMSE = 0.478 μg/m^3), the same analysis was applied for Roma Ave. For this validation it was not possible to apply the optimal combination of input parameters selected for Palau Reial (NO_2, N, PM_{10}, $PM_{2.5}$, PM_1, O_3, T, RH, PBL), given that particle number concentrations (N) were not monitored at Roma Ave. and $PM_{2.5}$ concentrations were only available as daily means. Instead, only regulatory (NO_2, O_3) and meteorological (T, RH, PBL) parameters were used. Five combinations of input variables were tested, based on the experience obtained from the Palau Reial dataset and on the parameters available (Table S8). The optimal model parametrisation (with NO_2, O_3, T, RH and PBL) achieved an R^2 coefficient for the testing dataset of 0.633, with RMSE = 1.19 μg/m^3, significantly lower than the results for the Palau Reial site. However, the Roma Ave. dataset was strongly limited by its size (only two months of data) and the lack of critical variables such as N and $PM_{2.5}$. Also, as expected, the fact that the model had been trained at a different location clearly impacted the performance indicators, as well as the different BC emission patterns during the COVID pandemic period (winter 2020, while the model had been trained with data from a pre-pandemic period).

3.4. Parametrisation with regulatory vs. non-regulatory air quality variables

In the last stage of this assessment we evaluated the applicability of the model in urban scenarios where different combinations of input parameters may be available, depending on the infrastructure available at the air quality networks. For example, while networks in certain cities (e.g., Paris or Copenhagen) include several stations with particle number monitors, others currently cover strictly the regulatory parameters (e.g., Madrid or Barcelona in the majority of stations). Thus, it was considered useful to test the model using only regulatory parameters as input, and compare the results with parametrisations using combinations of regulatory and non-regulatory parameters (mainly, N). The Roma Ave. dataset was not included in this assessment due to the lack on N data.

As shown in Table 1, the most optimal model parametrisation for the full dataset included N (non-regulatory) as input. The range of R^2 coefficients obtained when including N was 0.660–0.828 (Table 1), while it decreased to 0.537–0.727 using only regulatory parameters (Table 2). Despite this decrease, the optimal model solution using regulatory parameters still obtained $R^2 > 0.700$.

Comparing the different subsets of data (Fig. 5), results showed that the midday dataset produced the most similar result for parametrisations with regulatory and non-regulatory parameters, while the winter dataset showed the largest difference. This was due to the different sizes of the data subsets, and to the seasonality of the correlation between BC and N: in summer, when the midday N peak was most relevant, model performance did not improve as much as in winter, when N and BC show a highly correlated hourly evolution. The errors were larger for the parametrisations using only regulatory parameters. This is in agreement with the fact that the R^2 was higher when N (non-regulatory) was included. The range of RMSE estimated when including N was 0.286–0.553 μg/m^3 slightly lower than when considering only regulated pollutants (0.339–0.756 μg/m^3).

Table 2
Model performance (RMSE in μg/m^3, RRMSE in % and R^2 for the testing dataset) for the full dataset, with data collected over the full 2-year period (2018–2019) at the Palau Reial reference station. Only regulatory parameters used as input.

Input parameters (only regulatory)	R^2	RMSE	RRMSE
NO_2	0.537	0.805	60.0
NO_2, $PM_{2.5}$	0.652	0.678	50.5
NO_2, $PM_{2.5}$, O_3	0.699	0.630	46.9
NO_2, $PM_{2.5}$, O_3, RH	0.700	0.630	46.9
NO_2, $PM_{2.5}$, O_3, RH, PM_{10}	0.719	0.610	45.4
NO_2, $PM_{2.5}$, O_3, RH, PM_{10}, T	0.727	0.601	44.8

In summary, the modelling results improved with N as input variable. The model's performance using only regulated pollutants supports the application of this methodology in locations where only EU air quality reference data are available.

4. Conclusions

This work presents the development of a BC proxy based on supervised machine learning frameworks using data-driven models. The validity of the proxy approach was evaluated on a 2-year BC dataset obtained from two reference air quality monitoring stations in Barcelona (Spain), representative of Mediterranean climate, air pollutant mix and main emission sources.

After testing diverse combinations of input variables, the optimal model parametrisation using SVR was found to be using NO_2, N, PM_{10}, $PM_{2.5}$, PM_1, O_3, T, RH and PBL as input variables. With this parametrisation, the model was able to estimate BC concentrations with a correlation coefficient between modelled and measured concentrations of R^2 = 0.828 and low errors (RMSE = 0.48 μg/m^3). The correlation coefficient was comparable to those reported in the literature with white/black box models. Conversely, the RMSE obtained was lower than those reported for black/white box models, which was considered an improvement. Within the optimal parametrisation, critical variables were N, $PM_{2.5}$ and NO_2, while O_3, RH, PBL and T were considered correctors with only marginal improvements to model performance. The relevance of N as input variable was linked to the influence of new particle formation (NPF) events, which mainly occur at midday and during the summer months in Mediterranean climates (e.g., Carnerero et al., 2018; Casquero-Vera et al., 2020; Petäjä et al., 2007). As a result, model performance was dependent on seasonality (maximum R^2 for winter = 0.848, *vs.* 0.812 for summer). This has implications regarding the model's applicability, which may have the highest potential in environments where traffic is the main driver of N concentrations and where NPF events are scarce (e.g., highly polluted environments, colder climates, etc.). Finally, aiming to support the use of the model by air quality monitoring networks, the model was challenged with a dataset consisting of only EU-regulatory parameters ($PM_{2.5}$, NO_2, O_3). While performance was lower in comparison to other parametrisations, results evidenced a relatively high degree of correlation between modelled and measured BC concentrations (R^2 = 0.727; RMSE = 0.601 μg/m^3). Future research will involve including solar radiation as indicator of NPF events for locations where N data are not available.

To conclude, it is evident that experimental monitoring of BC concentrations in urban areas is highly advisable. However, high infrastructure costs and the lack of specific legislation limit the deployment of BC monitors across urban areas. In this framework, the data-driven model presented, together with other proxies being developed (Fung et al., 2021, 2019; Zaidan et al., 2019), may constitute a useful addition to urban air quality monitoring, which can contribute to the relatively scarce datasets currently available (mostly, from research teams). The proxy model proposed may provide value for exposure assessment in terms of gap filling in air quality time series (e.g., instrumental failures), or of predicting BC concentrations at locations where this parameter was once monitored (so that sufficient data are available to train the model). Under these considerations, we conclude that this methodology is applicable in Mediterranean urban environments for exposure assessment of BC concentrations, provided that the model can be adequately trained. It may be of interest for urban air quality research and management.

Declaration of competing interest

The authors declare that they have no known competing financial

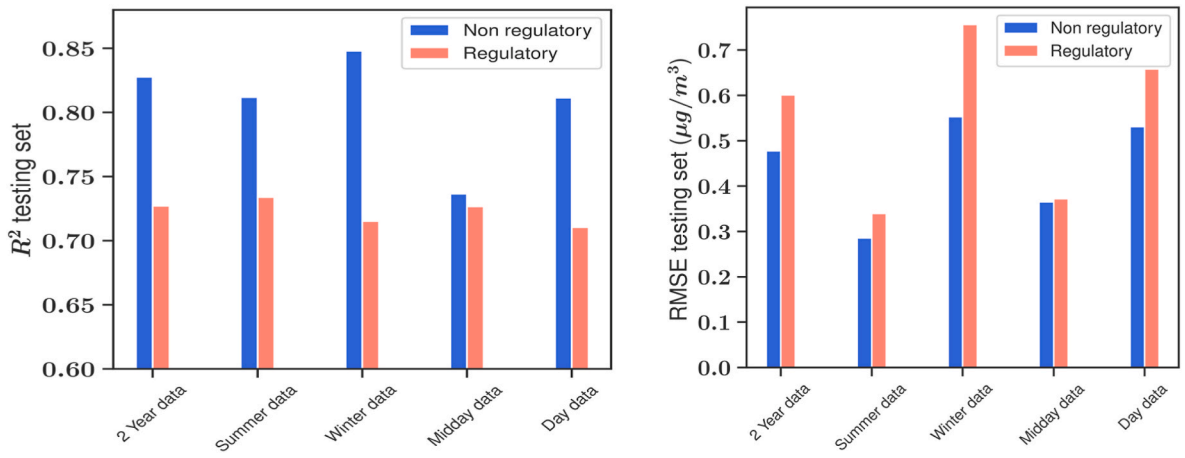

Fig. 5. Comparison between model results (R^2 and RMSE for the testing datasets) when non-regulatory *vs.* regulatory air quality parameters were used as input.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the support from the Generalitat de Catalunya (Dept. Medi Ambient) by providing the air quality data. This work was partly supported by H2020 project RI-URBANS (H2020-LC-GD-2020-6, reference 101036245), the Spanish Ministry of Science and Innovation (projects CEX2018-000794-S and PID2019-107910RB-I00), Academy of Finland via flagship on Atmosphere and Climate Competence Center (ACCC, project number 337549) and by AGAUR (project 2017 SGR41 and 2017 SGR 990). It was carried out in the framework of a joint collaboration between IDAEA-CSIC and University of Barcelona (Physics Faculty).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2022.113269.

References

Barcelo-Ordinas, J.M., Ferrer-Cid, P., Garcia-Vidal, J., Ripoll, A., Viana, M., 2019. Distributed multi-scale calibration of low-cost ozone sensors in wireless sensor networks. Sensors 19. https://doi.org/10.3390/s19112503.

Bond, T.C., Doherty, S.J., Fahey, D.W., Forster, P.M., Berntsen, T., Deangelo, B.J., Flanner, M.G., Ghan, S., Kärcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P.K., Sarofim, M.C., Schultz, M.G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S.K., Hopke, P.K., Jacobson, M.Z., Kaiser, J.W., Klimont, Z., Lohmann, U., Schwarz, J.P., Shindell, D., Storelvmo, T., Warren, S.G., Zender, C.S., 2013. Bounding the role of black carbon in the climate system: a scientific assessment. J. Geophys. Res. Atmos. 118, 5380–5552. https://doi.org/10.1002/jgrd.50171.

Brines, M., Dall'Osto, M., Beddows, D.C.S., Harrison, R.M., Gómez-Moreno, F., Núñez, L., Artíñano, B., Costabile, F., Gobbi, G.P., Salimi, F., Morawska, L., Sioutas, C., Querol, X., 2015. Traffic and nucleation events as main sources of ultrafine particles in high-insolation developed world cities. Atmos. Chem. Phys. 15, 5929–5945. https://doi.org/10.5194/acp-15-5929-2015.

Brunekreef, B., Strak, M., Chen, J., Andersen, Z.J., Atkinson, R., Bauwelinck, M., 2021. Mortality and Morbidity Effects of Long-Term Exposure to Low-Level PM2.5, BC, NO2, and O3: an Analysis of European Cohorts in the ELAPSE Project.

Carnerero, C., Pérez, N., Petäjä, T., Laurila, T.M., Ahonen, L.R., Kontkanen, J., Ahn, K.-H., Alastuey, A., Querol, X., 2019. Relating high ozone, ultrafine particles, and new particle formation episodes using cluster analysis. Atmos. Environ. X 4, 100051. https://doi.org/10.1016/j.aeaoa.2019.100051.

Carnerero, C., Pérez, N., Reche, C., Ealo, M., Titos, G., Lee, H.-K., Eun, H.-R., Park, Y.-H., Dada, L., Paasonen, P., Kerminen, V.-M., Mantilla, E., Escudero, M., Gómez-Moreno, F.J., Alonso-Blanco, E., Coz, E., Saiz-Lopez, A., Temime-Roussel, B., Marchand, N., Beddows, D.C.S., Harrison, R.M., Petäjä, T., Kulmala, M., Ahn, K.-H., Alastuey, A., Querol, X., 2018. Vertical and horizontal distribution of regional new particle formation events in Madrid. Atmos. Chem. Phys. 18, 16601–16618. https://doi.org/10.5194/acp-18-16601-2018.

Carnerero, C., Rivas, I., Reche, C., Pérez, N., Alastuey, A., Querol, X., 2021. Trends in primary and secondary particle number concentrations in urban and regional environments in NE Spain. Atmos. Environ. 244, 117982. https://doi.org/10.1016/j.atmosenv.2020.117982.

Casquero-Vera, J.A., Lyamani, H., Dada, L., Hakala, S., Paasonen, P., Román, R., Fraile, R., Petäjä, T., Olmo-Reyes, F.J., Alados-Arboledas, L., 2020. New particle formation at urban and high-altitude remote sites in the south-eastern Iberian Peninsula. Atmos. Chem. Phys. 20, 14253–14271. https://doi.org/10.5194/acp-20-14253-2020.

Cavalli, F., Alastuey, A., Areskoug, H., Ceburnis, D., Čech, J., Genberg, J., Harrison, R.M., Jaffrezo, J.L., Kiss, G., Laj, P., Mihalopoulos, N., Perez, N., Quincey, P., Schwarz, J., Sellegri, K., Spindler, G., Swietlicki, E., Theodosi, C., Yttri, K.E., Aas, W., Putaud, J. P., 2016. A European aerosol phenomenology -4: harmonized concentrations of carbonaceous aerosol at 10 regional background sites across Europe. Atmos. Environ. 144, 133–145. https://doi.org/10.1016/j.atmosenv.2016.07.050.

Cheung, H.C., Morawska, L., Ristovski, Z.D., 2011. Observation of new particle formation in subtropical urban environment. Atmos. Chem. Phys. 11, 3823–3833. https://doi.org/10.5194/acp-11-3823-2011.

Cortes, C., Vapnik, V., 1997. Support-vector networks. Mach. Learn. 20, 273–297, 3.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V., 1997. Support vector regression machines. Adv. Neural Inf. Process. Syst. 155–161.

Ferrer-Cid, P., Barcelo-Ordinas, J.M., Garcia-Vidal, J., Ripoll, A., Viana, M., 2019. A comparative study of calibration methods for low-cost ozone sensors in IoT platforms. IEEE Internet Things J. 6 https://doi.org/10.1109/JIOT.2019.2929594.

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

Fung, P.L., Zaidan, M.A., Sillanpää, S., Kousa, A., Niemi, J.V., Timonen, H., Kuula, J., Saukko, E., Luoma, K., Petäjä, T., Tarkoma, S., Kulmala, M., Hussein, T., 2019. Input-adaptive proxy for black carbon as a virtual sensor. Sensors 20.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer E.

Hussein, T., Johansson, C., Morawska, L., 2012. Forecasting urban air quality. Adv. Meteorol. 243603. https://doi.org/10.1155/2012/243603, 2012.

Jones, R.R., Hoek, G., Fisher, J.A., Hasheminassab, S., Wang, D., Ward, M.H., Sioutas, C., Vermeulen, R., Silverman, D.T., 2020. Land use regression models for ultrafine particles, fine particles, and black carbon in Southern California. Sci. Total Environ. 699, 134234. https://doi.org/10.1016/j.scitotenv.2019.134234.

Kerckhoffs, J., Hoek, G., Gehring, U., Vermeulen, R., 2021. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ. Int. 154, 106569. https://doi.org/10.1016/j.envint.2021.106569.

Kerckhoffs, J., Hoek, G., Vlaanderen, J., van Nunen, E., Messier, K., Brunekreef, B., Gulliver, J., Vermeulen, R., 2017. Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ. Res. 159, 500–508. https://doi.org/10.1016/j.envres.2017.08.040.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525, 367.

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

Ma, X., Longley, I., Gao, J., Kachhara, A., Salmond, J., 2019. A site-optimised multi-scale GIS based land use regression model for simulating local scale patterns in air pollution. Sci. Total Environ. 685, 134–149. https://doi.org/10.1016/j.scitotenv.2019.05.408.

Maricq, M., 2007. Chemical characterization of particulate emissions from diesel engines: a review. J. Aerosol Sci. 38, 1079–1118. https://doi.org/10.1016/j.jaerosci.2007.08.001.

Müller, T., Henzing, J.S., De Leeuw, G., Wiedensohler, A., Alastuey, A., Angelov, H., Bizjak, M., Collaud Coen, M., Engström, J.E., Gruening, C., Hillamo, R., Hoffer, A., Imre, K., Ivanow, P., Jennings, G., Sun, J.Y., Kalivitis, N., Karlsson, H., Komppula, M., Laj, P., Li, S.-M., Lunder, C., Marinoni, A., Martins Dos Santos, S., Moerman, M., Nowak, A., Ogren, J.A., Petzold, A., Pichon, J.M., Rodriquez, S., Sharma, S., Sheridan, P.J., Teinilä, K., Tuch, T., Viana, M., Virkkula, A., Weingartner, E., Wilhelm, R., Wang, Y.Q., 2011. Characterization and intercomparison of aerosol absorption photometers: result of two intercomparison workshops. Atmos. Meas. Tech. 4 https://doi.org/10.5194/amt-4-245-2011.

Oberdörster, G., Stone, V., Donaldson, K., 2007. Toxicology of nanoparticles: a historical perspective. Nanotoxicology 1, 2–25. https://doi.org/10.1080/17435390701314761.

Pakkanen, T.A., Kerminen, V.-M., Ojanen, C.H., Hillamo, R.E., Aarnio, P., Koskentalo, T., 2000. Atmospheric black carbon in Helsinki. Atmos. Environ. 34, 1497–1506. https://doi.org/10.1016/S1352-2310(99)00344-1.

Petäjä, T., Kerminen, V.-M., Dal Maso, M., Junninen, H., Koponen, I.K., Hussein, T., Aalto, P.P., Andronopoulos, S., Robin, D., Hämeri, K., Bartzis, J.G., Kulmala, M., 2007. Sub-micron atmospheric aerosols in the surroundings of Marseille and Athens: physical characterization and new particle formation. Atmos. Chem. Phys. 7, 2705–2720. https://doi.org/10.5194/acp-7-2705-2007.

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

Reche, C., Querol, X., Alastuey, A., Viana, M., Pey, J., Moreno, T., Rodríguez, S., González, Y., Fernández-Camacho, R., De La Campa, A.M.S., De La Rosa, J., Dall'Osto, M., Prevt, A.S.H., Hueglin, C., Harrison, R.M., Quincey, P., 2011a. New considerations for PM, Black Carbon and particle number concentration for air quality monitoring across different European cities. Atmos. Chem. Phys. 11 https://doi.org/10.5194/acp-11-6207-2011.

Reche, C., Rivas, I., Pandolfi, M., Viana, M., Bouso, L., Àlvarez-Pedrerol, M., Alastuey, A., Sunyer, J., Querol, X., 2015. Real-time indoor and outdoor measurements of black carbon at primary schools. Atmos. Environ. 120 https://doi.org/10.1016/j.atmosenv.2015.08.044.

Reche, C., Viana, M., Moreno, T., Querol, X., Alastuey, A., Pey, J., Pandolfi, M., Prévôt, A., Mohr, C., Richard, A., Artiñano, B., Gomez-Moreno, F.J., Cots, N., 2011b. Peculiarities in atmospheric particle number and size-resolved speciation in an urban area in the western Mediterranean: results from the DAURE campaign. Atmos. Environ. 45 https://doi.org/10.1016/j.atmosenv.2011.06.059.

Ripoll, A., Viana, M., Padrosa, M., Querol, X., Minutolo, A., Hou, K.M., Barcelo-Ordinas, J.M., Garcia-Vidal, J., 2019. Testing the performance of sensors for ozone pollution monitoring in a citizen science approach. Sci. Total Environ. 651, 1166–1179.

Rivas, I., Viana, M., Moreno, T., Pandolfi, M., Amato, F., Reche, C., Bouso, L., Àlvarez-Pedrerol, M., Alastuey, A., Sunyer, J., Querol, X., 2014. Child exposure to indoor and outdoor air pollutants in schools in Barcelona, Spain. Environ. Int. 69 https://doi.org/10.1016/j.envint.2014.04.009.

Saide, P.E., Spak, S.N., Pierce, R.B., Otkin, J.A., Schaack, T.K., Heidinger, A.K., da Silva, A.M., Kacenelenbogen, M., Redemann, J., Carmichael, G.R., 2015. Central American biomass burning smoke can increase tornado severity in the. U.S. Geophys. Res. Lett. 42, 956–965. https://doi.org/10.1002/2014GL062826.

Simon, M.C., Naumova, E.N., Levy, J.I., Brugge, D., Durant, J.L., 2020. Ultrafine particle number concentration model for estimating retrospective and prospective long-term ambient exposures in urban neighborhoods. Environ. Sci. Technol. 54, 1677–1686. https://doi.org/10.1021/acs.est.9b03369.

Tripathy, S., Tunno, B.J., Michanowicz, D.R., Kinnee, E., Shmool, J.L.C., Gillooly, S., Clougherty, J.E., 2019. Hybrid land use regression modeling for estimating spatio-temporal exposures to PM2.5, BC, and metal components across a metropolitan area of complex terrain and industrial sources. Sci. Total Environ. 673, 54–63. https://doi.org/10.1016/j.scitotenv.2019.03.453.

van de Beek, E., Kerckhoffs, J., Hoek, G., Sterk, G., Meliefste, K., Gehring, U., Vermeulen, R., 2021. Spatial and spatiotemporal variability of regional background ultrafine particle concentrations in The Netherlands. Environ. Sci. Technol. 55, 1067–1075. https://doi.org/10.1021/acs.est.0c06806.

Viana, M., Leeuw de, F., Bartonova, A., Castell, N., Ozturk, E., González Ortiz, A., 2020. Air quality mitigation in European cities: status and challenges ahead. Environ. Int. 143, 105907. https://doi.org/10.1016/j.envint.2020.105907.

Vizcaino, P., Lavalle, C., 2018. Development of European NO2 Land Use Regression Model for present and future exposure assessment: implications for policy analysis. Environ. Pollut. 240, 140–154. https://doi.org/10.1016/j.envpol.2018.03.075.

Wehner, B., Uhrner, U., von Löwis, S., Zallinger, M., Wiedensohler, A., 2009. Aerosol number size distributions within the exhaust plume of a diesel and a gasoline passenger car under on-road conditions and determination of emission factors. Atmos. Environ. 43, 1235–1245. https://doi.org/10.1016/j.atmosenv.2008.11.023.

WHO, 2018. World Health Statistics 2018: Monitoring Health for the SDGs.

Zaidan, Martha A., Dada, L., Alghamdi, M.A., Al-Jeelani, H., Lihavainen, H., Hyvärinen, A., Hussein, T., 2019. Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies. Appl. Sci. 9 https://doi.org/10.3390/app9204475.

Zaidan, Martha A., Wraith, D., Boor, B.E., Hussein, T., 2019. Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models. Appl. Sci. 9, 1–18. https://doi.org/10.3390/APP9224976.