Application of non-linear models to black carbon modelling at urban scale

Author: Jordi Rovira Carpi
Supervisor: Mar Viana, mar.viana@idaea.csic.es Yolanda Sola, ysola@meteo.ub.edu
Facultat de Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.*

Abstract: Black carbon (BC) is a health-relevant component of atmospheric particulate matter (PM), present in urban environments. BC is emitted through the incomplete combustion of carbonaceous material and it is typically associated with vehicle exhaust. BC measurements are not always available at urban scale due to the operational cost and complexity of the instrumentation. Therefore, it is advantageous to develop a mathematical model (or BC proxy) to estimate BC concentrations in urban air. This work presents the development and testing of a BC proxy based on a frequentist framework, Support Vector Regression (SVR), using observations of BC, particle mass and number concentrations (N), gaseous pollutants and meteorological variables from a reference air quality monitoring station in Barcelona (Spain) over a 2-year period (2018-2019). Two months of additional data were available from another site in Barcelona, for model validation. The BC concentrations estimated by the adaptive proxy showed a high degree of correlation with the measured BC concentrations (R^2 = 0.838–0.878) with a relatively low error (RMSE = 0.27–0.47 μg/m^3). Model performance was dependent on seasonality and time of the day, due to the influence of new particle formation events on the input variables. When validated at a different station, performance indicators showed a decrease (R^2 = 0.719; RMSE = 1.02 μg/m^3) but still an adequate correlation with BC observations. Due to its flexibility and reliability, it is concluded that the model can act as a virtual sensor to complement on-site measurements, for epidemiological and air quality research.

1. INTRODUCTION

According to the World Health Organization (WHO, 2018) around seven million people die every year from exposure to fine particles in polluted air. Nearly one out of every ten worldwide deaths results from exposure to air pollution (Health Effect Institute, 2019). Poor air quality is associated with increasing a variety of diseases such as stroke, heart disease, lung cancer, chronic obstructive pulmonary disease and respiratory infections, including pneumonia. Currently 91% of the world's population lives in places where air pollution levels exceed WHO guidelines (WHO, 2019). In urban areas, the main air pollution sources are vehicles, followed by domestic and industrial fuel combustion (Cairncross et al., 2007).

Atmospheric particulate matter (PM) has become a key air quality research topic due to its impact on human health, ecosystems and climate (Fuzzi et al., 2015). The Air Quality Directive (2008/50/EC) sets annual and daily limit values for PM_{10} and $PM_{2.5}$ mass concentrations (particles smaller than 10 and 2.5 μm, respectively), as well as the standard procedures to monitor them in reference air quality monitoring networks across Europe. In addition to these regulated parameters, two additional aerosol metrics are especially relevant from a health perspective, even though they are not monitored in reference networks: ultrafine particles (UFPs; particles smaller than 100 nm in diameter), and black carbon (BC). Ultrafine particles penetrate deepest into the respiratory tract (Oberdörster, 2000) as they have very high surface area to mass ratios, and they preferentially deposit in the tracheobronchial and alveolar regions of the human respiratory system. A large fraction of UFPs are derived from emissions associated with traffic, industrial activities, and domestic heating (Evans et al., 2014). UFPs tend to dominate atmospheric particle number size distributions and contribute little to PM mass concentrations (Reche et al., 2011a).

Black carbon (BC) is another health-relevant component of atmospheric PM. It is emitted from the incomplete combustion of carbonaceous material and it is typically associated with vehicle exhaust, coal-fired power plants, and

Figure 1. Location of the Palau Reial air quality monitoring site and the meteorological station, in Barcelona.

* Electronic address: jrovirca12@alumnes.ub.edu

biomass burning for heating and cooking (Singh et al., 2018). BC is a relevant component of PM in European cities, contributing 5%-15% to the total PM mass concentration in urban air (Ding et al., 2016). BC exposure has negative implications for human health (Bond, et al., 2013), as well as for regional and global climate, and extreme weather events (Saide, et al., 2015). Due to its relatively short lifetime in the atmosphere (Zhang, et al., 2015), it has been suggested that mitigation of BC emissions may reduce global warming (Bond, et al., 2013).

In order to understand the nature of urban air pollution, a large number of air quality monitoring stations is available across Europe (Hussein et al., 2012), which monitor the parameters regulated by the Air Quality Directive. Monitoring stations are equipped with reference instruments to monitor a number of gaseous and particulate pollutants, such as carbon monoxide (CO), nitrogen oxides (NOx), sulphur dioxide (SO_2), ozone (O_3) and particles (PM_{10}, $PM_{2.5}$) (Kumar et al., 2015). In addition to these parameters, novel, non-regulated parameters such as BC and UFP (in terms of particle number concentrations, N) are interesting for research and potentially for air quality monitoring as they are tracers of traffic emissions and atmospheric processes (Hamilton & Mansfield, 1991; Pakkanen et al., 2000; Reche et al., 2011a) and therefore may be used to design and test the effectiveness of mitigation strategies. However, the complexity and cost of the instrumentation and the lack of reference monitoring protocols for these parameters result in limited data availability.

To address this issue, different types of modelling approaches have been developed with the aim of increasing the spatial coverage of specific parameters (e.g., BC) and to fill data gaps. Examples of these input-adaptive proxies are land-use regression models (Kerckhoffs et al., 2017, 2021) and black-and white-box models (Fung et al., 2020; Zaidan et al., 2019). Based on a frequentist model, such as the Support Vector Regression (SVR), black box models select air quality variables as input variables based on their correlation coefficients, learn from the dataset, and generate the best result evaluated by the adjusted coefficient of determination with the output variable (Fung et al., 2020). BC is an adequate candidate for the application of black box models, as its concentrations in urban environments correlate with those of traffic-related gaseous pollutants, such as CO, NO, NO_2 and particle mass ($PM_{2.5}$) and number concentrations (N) (Reche et al., 2011a). Therefore, it is hypothesised that it may be possible to estimate BC mass concentrations using other air pollution metrics and indicators such as $PM_{2.5}$, O_3 and NO_2. Thus air quality monitoring datasets may be used to estimate BC concentrations not only in the present, but also historically. This approach was successfully tested for Helsinki and Amman (Zaidan et al., 2019).

This work aimed to apply a black box model to generate a BC proxy for urban environments representative of Mediterranean climates (tested for Barcelona, Spain). Model performance was evaluated using regulated ($PM_{2.5}$, NO_2, O_3) and combinations of regulated and non-regulated air pollutants (e.g., particle number concentrations) as input for the proxy. The aim was to assess the applicability of the proxy for regulatory as well as research purposes.

2. METHODS

2.1. Monitoring site

Air quality measurements were carried out at the Palau Reial urban background monitoring site located in Barcelona (41°23′14″ N, 02°06′56″E, 80 m.a.s.l.; Fig. 1). The city is located in NE Spain, on the coast in the western Mediterranean Basin, and it is characterised by a Mediterranean climate. Along its northern and southern margins Barcelona is surrounded by the Llobregat and Besòs river valleys that channel the sea winds. Owing to the topography of the area, the transport and dispersion of atmospheric pollutants within the city are largely controlled by fluctuating coastal winds, which blow in from the sea during the day (diurnal breeze), and, to a lesser extent, from the land at night. Barcelona has one of the highest vehicle densities in Europe (5800 cars km^{-2}) and one of the most relevant harbours in the Mediterranean in terms of ship traffic. This constitutes an additional source of atmospheric pollutants that are transported across the city by the sea breeze.

This monitoring site is influenced by vehicular emissions from one of the city's main traffic avenues (Diagonal Ave.), located at approximately 300 m, with a traffic density of 123990 vehicles day^{-1} (>60% diesel on average in the city's vehicle fleet; Council of Barcelona, *Serveis de Mobilitat*, 2017a, 2017b). Thus, the site is considered representative of urban background air pollutant concentrations while also influenced by the emissions of one of the largest arterial roads of the city.

In addition to the Palau Reial station, air quality data were also collected from a second urban background site in Barcelona (Av. Roma), for subsequent validation of the modelling results. Both stations are part of the Barcelona reference air quality network (XVPCA; http://mediambient.gencat.cat/).

2.2. Air quality measurements

A combination of conventional and novel air quality parameters were monitored at the Palau Reial station for a 2-year period (2018-2019):

Conventional parameters: EU reference analyzers for gaseous pollutants were used to measure tropospheric ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxide (NO), nitrogen dioxide (NO2) and NO_x with a 1-h time resolution, following EU reference protocols (Directive 2008/50/CE).

Mass concentrations for particles with diameter less than 2.5 µm ($PM_{2.5}$) were monitored with an environmental dust monitor Grimm EDM 180, equivalent to reference gravimetric measurements, with a 10-min time resolution.

Novel parameters: Black Carbon (BC; Petzold et al., 2013) mass concentrations were monitored with a multiangle absorption photometer (MAAP, Thermo ESM Andersen Instruments) with a PM_{10} inlet. The instrument operated on a 1-minute time resolution. The MAAP determines absorbance by particles deposited on the filter using measurements of both transmittance and reflectance at different angles. The absorbance was converted to BC mass concentrations by

using a fixed 6.6 m^2/g mass absorption coefficient at 637 nm (Müller et al., 2011), as recommended by the manufacturer.

Total particle number concentrations (N) was 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.

Finally, meteorological variables (atmospheric pressure, wind speed and direction, solar radiation, temperature and relative humidity) were obtained from a meteorological station located on the roof of the Faculty of Physics at Barcelona University, located at approximately 400 m from the Palau Reial station.

In addition to the 2-year dataset from Palau Reial, 2 months of additional data were collected from the Av. Roma station to assess the applicability of the model at a different location. The Av. Roma 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). Ambient temperature (T) and relative humidity (RH) were obtained from the Faculty of Physics, as in the case of Palau Reial.

2.3. Frequentist Model: Black Box

Machine Learning (ML) infers plausible models to explain observed data where the models are capable to make predictions about unseen data and take decisions that are rational given these predictions.

Specifically, black box (BB) is a system which generates predictions based on the inputs it receives and outputs or responses it produces, without taking into account its internal workings. Within the black box models, prediction techniques include linear and nonlinear methods. For the nonlinear methods, hyper-parameters are needed. Hyperparameters are configuration variables whose values are set before the training phase is executed. Hyperparameters are found using a grid search and they have a large impact on the training and testing time execution. In order to find the best set of hyperparameters per each algorithm a tenfold cross-validation strategy is used.

Support Vector Regression (SVR) was used in previous works as a black-box model to calibrate low-cost NO_2 and O_3 air pollution sensors (Ferrer-Cid et al,. 2019, 2020) and is a good candidate for use in black-box BC estimation. Support vector regression (Drucker et al., 1997), a nonlinear model, is a kernel method that is the analogous of support vector machines (SVMs), but using continuous values instead of classifying as SVM. It makes use of the "kernel trick" where the data is implicitly mapped to a higher dimension in order to find a better regression curve but doing all computations in input space via a kernel function *K(x, x')*. The points that are far away from the correct regression plane will be the ones important 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 programming problem. The objective function to solve is obtained with the dual formulation of the problem, minimizing a loss function. We have chosen to work with the radial basis function (RBF) kernel. The RBF kernel is proven to have an implicit map of infinite dimension. Finally, the hyperparameters optimized via cross-validation are the variance of the RBF kernel, the ϵ in the loss function, and a penalization term C. In the model evaluated in this work, the input variables were combinations of $PM_{2.5}$, NO_2, O_3, N, temperature and relative humidity, to model hourly BC concentrations as output.

To run the model, a part of the dataset is used to train it and the remaining part to validate the results. In this case, for training, 80% of the dataset to be used is randomly chosen and the remaining 20% is used as testing.

The model's root mean square error (RMSE) and coefficient of determination (R^2) were used as diagnostic evaluation attributes (Fung et al., 2020). While R^2 measures the amount of variance that the independent variables explain, or in other words how much of the output is explained by the input variables, the RMSE estimates the absolute difference between the modelled and measured concentrations in terms of mass:

$$R^2 = \frac{SSR}{SST}$$
$$SSR = \sum (f_i - \bar{y})^2$$
$$SST = \sum (y_i - \bar{y})^2$$

where *SSR* is Sum of Squared Regression, also known as variation, *SST* is Sum of Squared Total, also known as total variation in the data, $\bar{y}$ is the mean of *y* value, f_i is predicted value of *y* for observation *i* and y_i is the *y* value for observation *i*. The R^2 values range from 0 to 1, where a high value indicates that the model used is able to explain, given the input parameters, the output.

The RMSE is calculated as the square root of the average squared difference between the forecast and the observation pairs:

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

where N is the number of complete data input to the model, and y_i and $\hat{y}_i$ are *i*th measured and *i*th estimated response variable by the model, respectively.

The model was run for 6 different datasets, constructed from the original datasets from the Palau Reial (2 years of data, 2018-2019) and Av. Roma (2 months of data) stations. The purpose was to test model performance under different scenarios where BC concentrations would be influenced by different sources and atmospheric processes. The datasets were:

- Full dataset: 2 years of data from Palau Reial.
- Winter: December, January and February of data from Palau Reial.
- Summer: June, July and August of data from Palau Reial.
- Midday: hourly values between 10:00 and 14:00 of data from Palau Reial.
- Day: hourly values between 14:00 and 10:00 of data from Palau Reial.
- Av. Roma: November and December 2020 of data from Av. Roma.

3. RESULTS AND DISCUSSION

3.1. Temporal variability of atmospheric pollutants

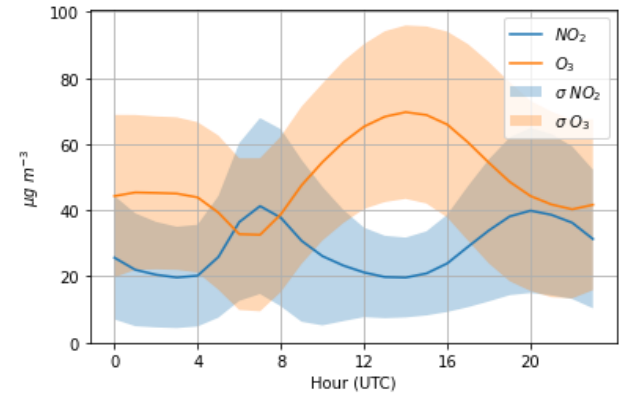

Figure 2. Mean hourly cycle of gaseous pollutant concentrations (NO_2 and O_3) at the Barcelona Palau Reial station for the period 2018-2019.

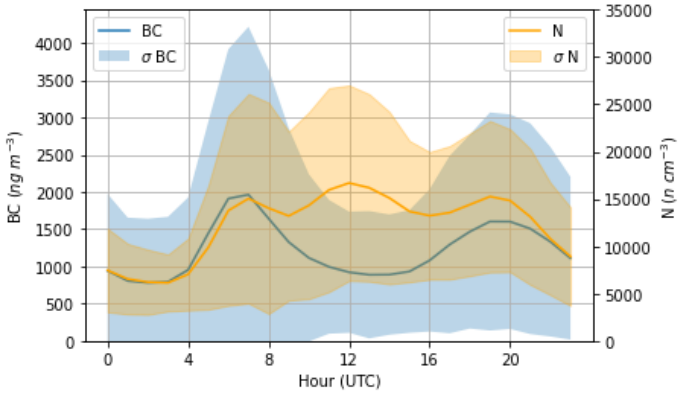

Figure 3. Mean hourly cycle of Black Carbon (BC) and particle number (N) concentrations at the Barcelona Palau Reial station for the period 2018-2019.

Prior to BC modelling, the mean daily cycles of gaseous and particulate pollutants at the Palau Reial site were evaluated, for the period 2018-2019 (Figures 2 and 3). The purpose was to understand the daily pollutant trends and to assess their representativity. As observed in previous works (Reche et al., 2011a, 2011b, 2015), NO_2 concentrations followed a diurnal pattern influenced by traffic intensity, with maxima during the morning rush hour (06:00–08:00 UTC), decreasing during the day due to atmospheric dilution, and increasing again in the evening (18:00–21:00 UTC). Inversely O_3 levels showed a typical diurnal pattern, with an increase at midday coinciding with the maximum photochemistry and solar radiation during the central hours of the day.

The same daily evolution described for NO_2 was observed for BC concentrations (Figure 3), given that vehicular traffic is the main source of the BC in the Barcelona urban area.

Particle number concentrations (N) were also largely controlled by traffic emissions at Palau Reial, with maxima, during the morning and evening rush hour periods. However, this parameter is known to be highly influenced in Barcelona by new-particle formation mechanisms during the central hours of the day, especially in the summer months (photochemically induced nucleation: Reche et al,. 2011a; Pey et al., 2008; Perez et al., 2010; Fernández-Camacho et al, 2010; Cheung et al., 2011). Therefore, N and BC maxima are typically detected during morning and evening rush-hours (06:00–08:00 and 18:00–21:00 UTC), with N being mainly influenced by primary aerosol emissions and by new particle formation through the dilution and cooling of the vehicle exhaust (Maricq, 2007, Wehner et al., 2009).In addition, N shows a secondary maximum at midday, coinciding with a decrease in BC concentration, resulting from photochemical nucleation processes. This midday nucleation takes place as a

Y	X1	X2	X3	X4	X5	X6	RMSE testing ($\mu g/m^3$)	R^2 training	R^2 testing
BC norm	$PM_{2.5}$ norm	NO_2 norm	O_3 norm	N norm	RH norm	T norm	0.44	0.893	0.865
log BC	log $PM_{2.5}$	log NO_2	log O_3	log N	log RH		0.46	0.851	0.842
log BC	log $PM_{2.5}$	log NO_2	log O_3	log N	log RH	log T	0.45	0.863	0.841
BC norm	$PM_{2.5}$ norm	NO_2 norm	O_3 norm	N norm	RH norm		0.46	0.853	0.823
log BC	log $PM_{2.5}$	log NO_2	log O_3	log N			0.50	0.822	0.813
BC norm	$PM_{2.5}$ norm	NO_2 norm	O_3 norm	N norm			0.49	0.860	0.800
BC norm	$PM_{2.5}$ norm	NO_2 norm	O_3 norm	RH norm	T norm.		0.50	0.812	0.799
log BC	log $PM_{2.5}$	log PM_{10}	log NO_2	log O_3			0.57	0.771	0.748
log BC	log $PM_{2.5}$	log NO_2	log O_3	log RH	log T		0.63	0.797	0.745
log BC	log PM_{10}	log NO_2	log O_3				0.58	0.765	0.744
log BC	log $PM_{2.5}$	log NO_2	log O_3				0.59	0.747	0.744
log BC	log $PM_{2.5}$	log NO_2	log O_3	log T			0.64	0.789	0.734
log BC	log $PM_{2.5}$	log NO_2	log O_3	log RH			0.64	0.778	0.732
log BC	log $PM_{2.5}$	log PM_{10}	log NO_2				0.61	0.743	0.716
log BC	log $PM_{2.5}$	log PM_{10}	log O_3				0.69	0.672	0.634

Table 1. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for data collected over the full 2-year period (2018-2019) at the Palau Reial reference station. Using logarithmic and normalized values.

consequence of the high solar radiation, the growth of the mixing layer, the increase in wind speed and the consequent decrease in pollutant concentrations. The different patterns between BC and N observed in this work for Barcelona are in agreement with the results presented by Reche et al. (2011a), who reported on the similarities between N and BC daily trends in North-European cities in contrast with the differences observed in Southern-European climates.

3.2. Black box modelling of BC concentrations

Table 1 summarises the results obtained after applying the box model to the full dataset from the Palau Reial (PR) station, for a two-year period (2018-2019). The Table shows the different combinations of independent variables used as input, and the results obtained when compared to the dependent variable (BC) in terms of R^2 for the training and testing datasets, and RMSE for the testing dataset. The input variables were used normalised and as logarithmic.

The model fits the normalised and logarithmic values similarly, although with certain differences. The purpose of data pre-processing prior to model application is to facilitate model convergence in finding the estimation parameters. While normalising the input parameters is more time, and resource, consuming, it eliminates the impacts derived from processing data with different orders of magnitude. When normalising, the data are scaled to the variance and the orders of magnitude are similar. Using logarithms, on the other hand, results in a certain normalisation (as the order of magnitude of the large values is reduced) but also in linearization of the data, that is, the output is a linear combination of the inputs, which makes the model run faster. It has certain implications on the results, but they are minimal. As shown in Table 1, the best model fit (R^2 testing = 0.865) was obtained for the full dataset in Palau Reial using normalised input variables, with NO_2, $PM_{2.5}$, O_3, N, T and RH as input variables.

The second best fit (R^2 testing = 0.842) was obtained with logarithmic values, for the same combination of variables (with and without T). As the number of input variables was decreased, and different combinations were tested, the fit between the modelled and measured BC concentrations decreased. The lowest R^2 obtained (for the testing period) was 0.634, using only PM_x and O_3 as inputs. 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 temperature only makes R^2 and RMSE improve a few tenths (Table 1). The parameters which contributed the most (and therefore were more critical) to estimate BC in this work, for the full dataset, were $PM_{2.5}$, NO_2 and N. Parameters considered as correctors, with only marginal improvements to model performance, were O_3 , RH and T (Table 1).

In general, almost all of the combinations tested for the full dataset at Palau Reial (Table 1) provided acceptable results, with R^2_{test}> 0.7. Using only regulatory air quality parameters ($PM_{2.5}$, NO_2, O_3, temperature and humidity) as input, the best result reached 0.799, with normalised values. Adding particle number concentrations, a non-regulatory parameter, provided a significant improvement. The model's performance improved with particle number concentrations because, as shown in section 3.1, N and BC show a very similar hourly evolution (therefore, very high correlation), with the exception of the midday peak for N. When evaluating the RMSE, results were also quite promising as errors were always lower than 0.69 μg/m^3 and even reached 0.44 μg/m^3 for the best solution (Table 1)

In relation to other studies in the literature, our results comparable to those presented by Zaidan et al., (2019) and Fung et al., (2020) who also modelled BC using black- and white-box approaches. The results obtained by these authors were R^2 ranging between 0.74 and 0.94 for measured vs. modelled BC concentrations, which is in the same range as the results obtained in this work (Table 1). Conversely, the RMSE reported by these authors (from 0.19 to 2.36 μg/m3) were larger than in the present work. Land Use Regression (LUR) models, as discussed above, are also used for BC and ultrafine particle modelling at urban scale. According to the literature (Montagne et al., 2015, Dons et al., 2013, Martenies et al., 2019, Masiol et al., 2018, Kerckhoffs et al., 2017) the performance of LUR models has improved in recent years, with R^2 coefficients between modelled and measured concentrations ranging from 0.4 (for BC; Montagne et al., 2015) to 0.64-0.80 (for ultrafine particles, Kerckhoffs et al., 2017) and even 0.94-0.99 (for ultrafine particles; Kerckhoffs et al., 2021). As observed for black- and white-box models, the results in this work are comparable to those obtained by LUR modelling. Thus, these results support the validation of this methodology in Southern European urban environments, and would suggest the applicability of this kind of models for exposure assessment in epidemiological studies (Kerckhoffs et al., 2017, 2021).

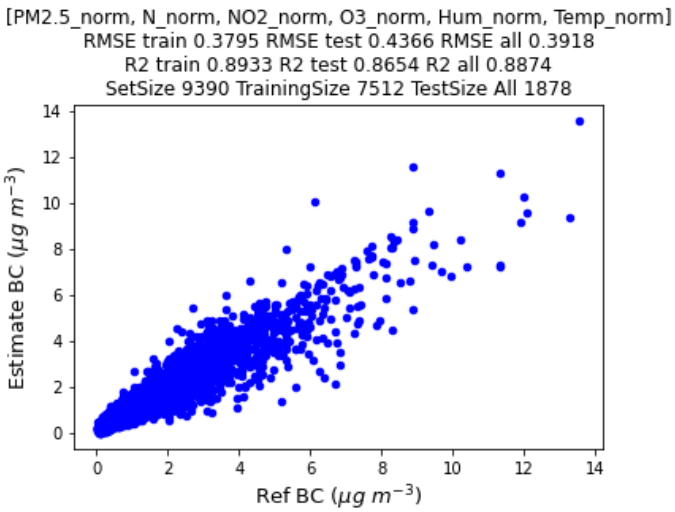

Figure 4. Scatter plot of estimated BC vs. reference (Ref) BC for the model using normalised $PM_{2.5}$, N, NO_2, O_3, temperature and relative humidity as input, for the period 2018-2019 at the Palau Reial site.

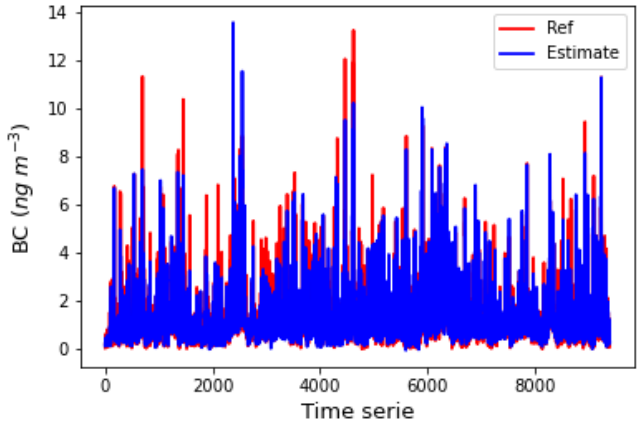

Figure 5. Time series of estimated BC and measured BC for the normalised $PM_{2.5}$, N, NO_2, O_3, temperature and relative humidity, for the period 2018-2019 at the Palau Reial site.

The time series and scatter plot of the measured vs. modelled BC concentrations for this combination of input variables are shown in Figures 4 and 5. Figure 4 shows the better lower dispersion of the data for measured BC concentrations < 7 µg/m^3, which is the most populated data range. For higher concentrations the dispersion is higher. Figure 5 evidences that the model tends to underestimate the highest reference BC concentrations.

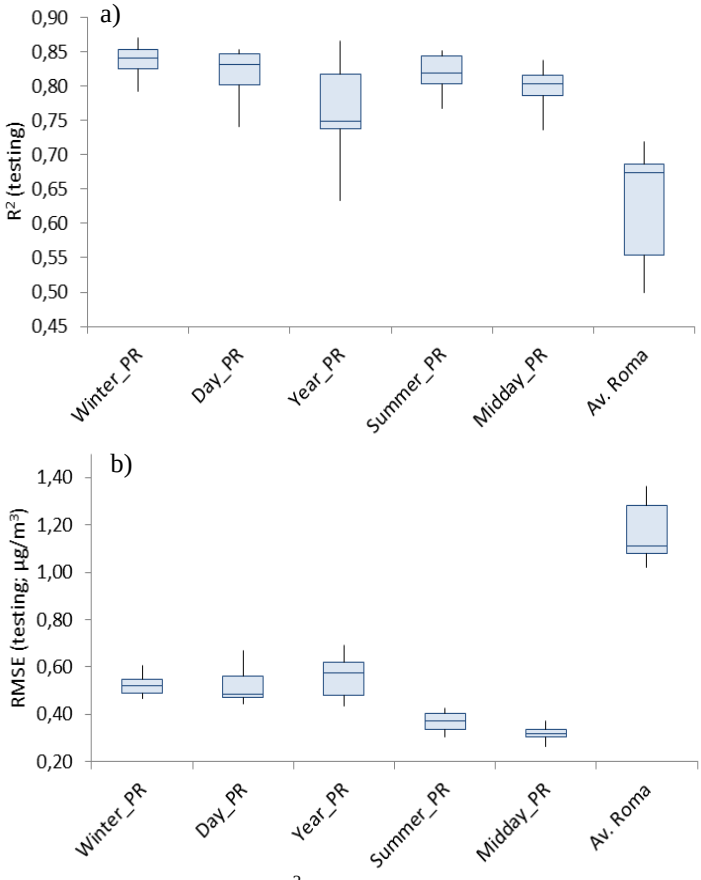

Figure 6 a) Box plot R^2, **b)** RMSE between measured and modelled hourly BC concentrations for the different datasets. Results shown only for the testing datasets.

After assessing model performance for the full dataset (Table 1, Figures 4 and 5), the model was challenged with different subsets of data where BC concentrations were influenced by diverse emission sources and atmospheric processes. The combination of input parameters used was the one with the best fit for the full dataset (normalised $PM_{2.5}$, N, NO_2, O_3, temperature and relative humidity; Table 1). For example, the full dataset was divided into the summer and winter periods to assess the influence of the midday N peak (in the summer subset, but not in the winter subset) on model performance indicators (R^2 and RMSE). The same approach was applied for the full day subset vs. the midday subset (again, assessing the influence of the midday nucleation peak). The results are shown in Figures 6a and 6b. The box plot describes the interquartile range (IQR), which is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values in Table 1 and Tables A1, A2, A3, A4 and A5 (see Appendix) have been used to generate the box plot for each subset of data.

The Year_PR is the dataset with all the hourly values for 2018 and 2019; the Winter_PR is with the values for the months of December, January and February; the Summer_PR for the months of June, July and August; the Midday_PR are the hourly values between 10:00 and 14:00; the Day_PR are all the values without the midday, that is, from 14:00 to 10:00; and Av. Roma is the dataset for model validation and uses the data from this site for November and December 2020.

The best results were obtained for the Winter_PR (R^2 = 0.871, Table A1) dataset and the worst with Midday_PR (R^2 = 0.838, Table A3). Despite this ranking of the solutions obtained, it should be noted that the correlation between modelled and measured BC concentrations was always high (> 0.8), for all of the subsets of data. The lower R^2 obtained for the Midday-PR subset was due to the influence of the nucleation peak during the central hours of the day, when the air pollutant concentrations are diluted in a thicker mixing layer and insolation and O_3 concentrations are highest. Therefore, it may be expected that the Midday dataset should show worse results, since at these hours the correlation between the input parameters is the poorest and therefore model fit is the lowest. This midday peak is especially relevant during the summer period as high temperatures favour nucleation, and consequently the Summer dataset (R^2 = 0.853, Table A2) provided results similar to that of the Midday subset. Conversely, the influence of the nucleation peak is minimal in winter, which implies that the model is not impacted by this trend and a better fit is obtained (R^2 = 0.871, Table A1). Similarly, in the Day dataset (R^2 = 0.854, Table A4), as it does not include the midday period and thus BC is impacted mainly by the morning and evening traffic rush hour peaks, the fit improves again. Finally, for the the Year_PR dataset (R^2 = 0.865, Table 1 the performance of the model is slightly lower due to the greater variability in the concentrations of the input variables (including the meteorological variables), as well as of the measured BC concentrations. When modelling annual datasets the performance in terms of R^2 is lower, but the model has a better predictive capacity for more variable conditions given that the training dataset includes a greater variability of conditions (Barcelo-Ordinas et al., 2019; Ripoll et al., 2019).

The results were relatively similar for the different subsets in terms of RMSE (Figure 6b). The lowest RMSE

was obtained for the Midday_PR (RMSE = 0.27 μg/m^3) dataset, and the largest for Winter_PR (RMSE = 0.47 μg/m^3) and Day_PR (RMSE = 0.46 μg/m^3). The differences between the Summer_PR dataset (RMSE = 0.30 μg/m^3) and Year_PR (RMSE = 0.44 μg/m^3) were also not especially large. In 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. The Midday_PR dataset, with the lowest RMSE, was also the one with the smallest dataset size; this indicates that it has fewer large errors, because it has fewer values.

Finally, the model was validated with data from a different reference station (Av. Roma), with measured BC data for a shorter period of time (2 months; November-December 2020) (Table A5). For this validation it was not possible to apply the combination of input parameters selected for PR (normalised $PM_{2.5}$, N, NO_2, O_3, temperature and relative humidity), given that particle number concentrations were not monitored at this site. Instead, only the regulatory parameters (NO_2, O_3) were used. The model was trained with the Winter_PR dataset, with the most similar environmental conditions to the validation dataset, and tested it with the entire Av. Roma dataset. Different combinations of the normalised and logarithmic input variables were tested. Overall, R^2 coefficients (Table A5) were lower than for the PR dataset (Table 1), ranging between 0.499-0.719, with the best model fit obtained for the parameterization with NO_2, O_3, normalised temperature and humidity (R^2 = 0.719 and RMSE = 1.02 μg/m^3). In relation to the literature, as discussed above, these results are still comparable and in the same order of magnitude as those obtained with other models (black/white box, LUR) and for different air pollutants (BC, ultrafine particles).

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

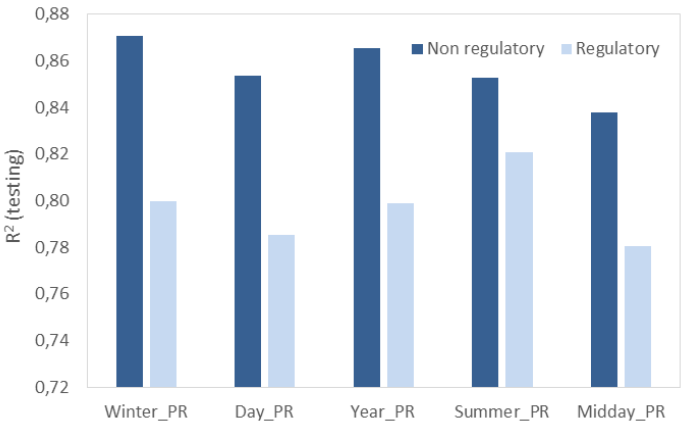

Figure 7. Comparison between model results (R^2 for the testing datasets) when non-regulatory vs. regulatory air quality parameters were used as input

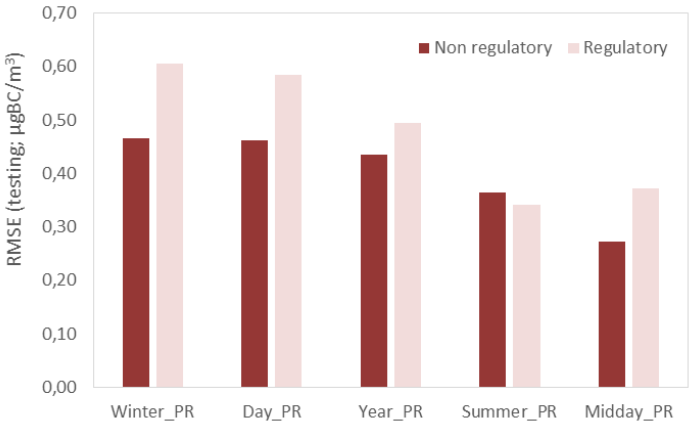

Figure 8. Comparison between model results (RMSE for the testing datasets) when non-regulatory vs. regulatory air quality parameters were used as input.

The last stage of this assessment aimed to evaluate the applicability of the model in urban scenarios where different combinations of input parameters may be available, depending on the air quality networks. For example, while networks in certain cities include particle number monitors at several sites (e.g., in Paris or Copenhagen), others cover strictly the regulatory parameters (e.g., Madrid or Barcelona in the majority of stations). Thus, it was considered useful to compare the results of the model using only regulatory parameters as input, and using combinations of regulatory and non-regulatory parameters (mainly, N). This was implemented for the different subsets of data described in the previous section.

As shown in Table 1, the best model fit was always obtained when including N (non-regulatory) in the analysis (Figure 7 and 8). The range of R^2 coefficients obtained when including N was 0.838-0.871, while it decreased to 0.781-0.821 with the regulatory parameters only (combinations of NO_2, O_3, PM_{10}, $PM_{2.5}$, T, RH). Despite this decrease, the correlation between modelled and measured BC concentrations was considered high, in both cases. In addition, the difference between both parametrisations was not high, evidencing once again adequate model performance.

Comparing the different subsets of data, it is observed that the Summer_PR was the one with the smallest difference between regulatory and non-regulatory parameters, and Winter_PR was the one with the largest difference. This is related to the correlation between BC and N, which was not the same (by definition) for all the subsets of data and resulted in better performance for the datasets with higher correlation between both pollutants. For example, in summer, when the midday N peak is most relevant, the model does not improve as much as in the Winter_PR where N and BC have the same hourly evolution.

With regard to the RMSE (Figure 8), 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 0.27-0.47 $\mu g/m^3$ slightly lower than when considering only regulated pollutants (0.34-0.61 $\mu g/m^3$). The Summer_PR dataset was the only one in which the RMSE of the non-regulatory values was greater than that of the regulatory parametrisation, due to the nucleation peak at midday which causes BC estimates with very large errors.

The Av. Roma dataset could not be included in this assessment due to the lack on N data. In sum, based on these results it may be concluded that modelling results improved with N as input variable, but that model performance maybe considered adequate when only regulated pollutants are used, thus validating the application of this methodology in locations where only EU air quality reference data are available.

4. CONCLUSIONS

- This paper presents the development of a BC proxy based on a frequentist framework using black-box (BB) models. A non-linear method based on SVR was used as BB. The method was evaluated on BC data obtained from a reference air quality monitoring station in Barcelona (Spain), representative of Mediterranean air pollutant sources and dynamics.
- The model provided adequate results, obtaining a coefficient of determination (R^2) between measured and modelled BC concentrations equal to 0.865, for a 2-year dataset (2018-2019). The calculated RMSE was 0.44 $\mu g/m^3$. The R^2 obtained was comparable to those reported in the literature for BC and ultrafine particle monitoring with white/black box and land-use regression (LUR) models. Conversely, the RMSE obtained was lower than those reported for black/white box models, which is considered an improvement. These results validate the use of the SVR method for BC modelling in Mediterranean urban environments..
- Model performance was dependent on seasonality (for winter R^2 = 0.871 and for summer R^2 = 0.853) as a result of the seasonality of the air pollutant sources and processes (e.g, new particle formation in summer in Mediterranean climates).
- The potential input parameters were evaluated to assess which ones played a more relevant role in the estimation of BC. Results showed that $PM_{2.5}$, NO_2, and N were the most relevant predictors, whereas O_3, RH and T played a more secondary role (as correctors).
- Model performance was highest when N was included as input variable. However, performance was also considered adequate when only regulatory parameters were considered (R^2 = 0.781).
- The results obtained demonstrate the applicability of this methodology to predict BC concentrations in Mediterranean urban environments where this parameter is not measured, which is of special interest for urban air quality and epidemiological research.

Acknowledgments

I would like to give my sincere gratitude to José María Barceló, Yolanda Sola and Mar Viana for his guidance and motivation while working in this project.

Acknowledgments are due to IDAEA-CSIC for hosting this project and to the Generalitat de Catalunya (Departament de Medi Ambient) for providing the AQ data.

5. REFERENCES

Barcelo-Ordinas, J.M., Ferrer-Cid, P., Garcia-Vidal, J., Ripoll, A., Viana, M.: Distributed multi-scale calibration of low-cost ozone sensors in wireless sensor networks. Sensors (Switzerland) 19, 2019.

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, P., Shindell, D., Storelvmo, T., Warren, S.G., Zender, C.S.: Bounding the role of black carbon in the climate system: A scientific assessment. J. Geophys. Res. Atmos., 118, 5380–5552, 2013.

Cairncross, E.K.; John, J.; Zunckel, M.: A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants, Atmos. Environ., 41, 38, 8442-8454, 2007.

Cheung, H.C.; Morawska, L.; Ristovski, Z.D.: Observation of new particle formation in subtropical urban environment, Atmos. Chem. Phys., 11, 3823–3833, 2011.

Council of Barcelona. Caracterització dels vehicles i les seves emissions a l'àrea metropolitana de Barcelona. Avalaible at: https://ajuntament.barcelona.cat/premsa/wp-content/uploads/2017/09/Dossier_estudi_RSD_lowres2.pdf, 2017a. Last access: 14/06/2021.

Council of Barcelona. Dades bàsiques de mobilitat. Available at:. https://www.barcelona.cat/mobilitat/sites/default/files/documentacio/dadesbasiquesmobilitat-2017.pdf, 2017b. Last access: 14/06/2021.

Ding, A.J., Huang, X., Nie, W., Sun, J., Kerminen, V.M., Petäjä, T., Su, H., Cheng, Y.F., Yang, X.Q., Wang, M.H., Chi, X.G., Wang, P., Virkkula, A., Guo, D., Yuan,. J., Wang, S.Y., Zhang, R.J., Wu, Y.F., Song, Y., Zhu, T., Zilitinkecich, S., Kulmala, M., Fu, C.B.: Enhanced haze pollution by black carbon in megacities in China, Geophys. Res. Lett., 43, 2873–2879, 2016.

Dons, E., Van Poppel, M., Kochan, B., Wets, G., Int Panis, L.: Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon, Atmos. Environ., 74, 237-246, 1352-2310, 2013.

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

Evans, K.A., Halterman, J.S., Hopke, P.K., Fagnano, M., Rich, D.Q.: Increased ultrafine particles and carbon monoxide concentrations are associated with asthma exacerbation among urban children, Environ. Res., 129, 11–19, 2014.

Fernández-Camacho, R., Rodríguez, S., de la Rosa, J., Sánchez de la Campa, A. M., Viana, M., Alastuey, A., and Querol, X.: Ultrafine particle formation in the inland sea breeze airflow in Southwest Europe, Atmos. Chem. Phys., 10, 9615-9630, 2010.

Ferrer-Cid, P., Barcelo-Ordinas, J.M., Garcia-Vidal, J., Ripoll, A., Viana, M.: A comparative study of calibration methods for low-cost ozone sensors in IoT platforms, IEEE Internet of Things Journal, 6(6), 9563-9571, 2019.

Ferrer-Cid, P., Barcelo-Ordinas, J.M., Garcia-Vidal, J., Ripoll, A., Viana, M.: Multisensor data fusion calibration in IoT air pollution platforms, IEEE Internet of Things Journal, 7(4), 3124-3132, 2020.

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. Input-Adaptive Proxy for Black Carbon as a Virtual Sensor, Sensors 20, 182, 2020.

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., Nemitz, E., Pandis, S., Riipinen, I., Rudich, Y., Schaap, M., Slowik, J. G., Spracklen, D. V., Vignati, E., Wild, M., Williams, M., and Gilardoni, S.: Particulate matter, air quality and climate: lessons learned and future needs, Atmos. Chem. Phys., 15, 8217–8299, 2015

Hamilton, R.S., and Mansfield, T.A.: Airborne particulate elemental carbon: its sources, transport and contribution to dark smoke and soiling, Atmos. Environ., 25, 715–723, 1991.

Health Effect Institute. State of Global Air 2019 Special Report; Health Effect Institute: Boston, MA, USA, 2019.

Hussein, T., Johansson, C., Morawska, L.: Forecasting Urban Air Quality, Adv. Meteorol., 2012, 5–7, 2012.

Kerckhoffs, J., Hoek, G., Vlaanderen, J., Van Nunen, E., Messier, K., Brunekreef, B., Gulliver, J., Vermeulen, R.: Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring, Environ. Res., 159:500-508, 2017.

Kerckhoffs, J., Hoek, G., Gehring, U., Vermeulen, R.: Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ. Int. 154, 106569, 2021.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., Britter, R.: The rise of low-cost sensing for managing air pollution in cities, Environ. Int., 75, 199–205, 2015.

Maricq, M.M.: Chemical characterization of particulate emissions from diesel engines: A review, Aerosol Sci., 38, 1079–1118, 2007.

Martenies, S., WeMott, S., Kuiper, G., Lorber, K., Dawson, C., Andresen, K., Allshouse, W., Starling, A., Adgate, J., Dabelea, D., Magzamen, S.: Developing a black carbon land use regression model for the Denver, CO metropolitan area, Environ. Epidemi., 3, 261-262, 2019.

Masiol, M., Zíková, N., Chalupa, D.C., Rich, D.Q., Ferro, A.R., Hopke, P.K.: Hourly land-use regression models based on low-cost PM monitor data, Environ. Res., 167, 7-14, 0013-9351, 2018.

Montagne, D.R., Hoek, G., Klompmaker, J.O., Wang, M., Meliefste, K., Brunekreef, B.: Land Use Regression Models for Ultrafine Particles and Black Carbon Based on Short-Term Monitoring Predict Past Spatial Variation, Environ. Sci. Technol., 49, 14, 8712-8720, 2015.

Müller, T., Henzing, J.S., de Leeuw, G., Wiedensohler, A., Alastuey, A., et al.: Characterization and intercomparison of aerosol absorption photometers: result of two intercomparison workshops, Atmos. Meas. Tech. 4, 245–268, 2011.

Oberdörster, G.: Pulmonary effects of inhaled ultrafine particles, Int. Arch. Occup. Environ. Health, 74, 1–8, 2000

Pakkanen, T. A., Kerminen, V. M., Ojanena, C. H., Hillamo, R. E., Aarnio, P., and Koskentalo, T.: Atmospheric Black Carbon in Helsinki, Atmos. Environ., 34, 1497–1506, 2000.

Perez, N., Pey, J., Cusack, M., Reche, C., Querol, X., Alastuey, A., and Viana, M.: Variability of Particle Number, Black Carbon, and PM10, PM2.5, and PM1 Levels and Speciation: Influence of Road Traffic Emissions on Urban Air Quality, Aerosol Sci. Technol., 44, 487–499, 2010.

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

Pey, J., Rodríguez, S., Querol, X, Alastuey, A., Moreno, T., Putaud, J. P., and Van Dingenen, R.: Events and cycles of urban aerosols in the western Mediterranean, Atmos. Environ., 42, 9052–9062, 2008.

Reche, C., Querol, X., Alastuey, A., Viana, M., Pey, J., Moreno, T., Rodríguez, S., González, Y., Ferneández-Camacho, R., Sánchez de la Campa, A.M., de la Rosa, J., Dall'Osto, M., Prévôt, A.S.H., Hueglin C., Harrison R.M., Quincey, P..: New considerations for PM, Black Carbon and particle number concentration for air quality monitoring across different European cities, Atmos. Chem. Phys., 11, 6207–6227, https://doi.org/10.5194/acp-11-6207-2011, 2011a.

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.: Peculiarities in atmospheric particle number and size-resolved speciation in an urban area in the western Mediterranean: Results from the DAURE campaign, Atmos. Environ., 43, 5282-5293, 2011b.

Reche, C., Viana, M., Brines, M., Pérez, N., Beddows, D., Querol, X., Alastuey,: Determinants of aerosol lung-deposited surface area variation in an urban environment, Science of The Total Environment, 517, 38-47, 2015.

Ripoll, A., Viana, M., Padrosa, M., Querol, X., Minutolo, A., Hou, K.M., Barcelo-Ordinas, J.M., Garcia-Vidal, J.: Testing the performance of sensors for ozone pollution

monitoring in a citizen science approach. Sci. Total Environ. 651, 1166–1179, 2019.

Saide, P., Spak, S., Pierce, R., Otkin, J., Schaack, T., Heidinger, A., da Silva, A., Kacenelenbogen, M., Redemann, J., Carmichael, G.: Central American biomass burning smoke can increase tornado severity in the US, Geophys. Res. Lett. 42, 956–965, 2015.

Singh, V., Ravindra, K., Sahu, L., Sokhi, R.: Trends of atmospheric black carbon concentration over the United Kingdom. Atmos. Environ. 178, 148–157, 2018.

Wehner, B., Uhrner, U., von Löwis, S., Zallinger, M., and Wiedensohler, A.: 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, 2009.

World Health Organization Global Ambient Air Quality Database. Available at: https://www.who.int/airpollution/data/en/ Last access: 17/05/2021.

World Health Organization. World health statistics 2019: monitoring health for the SDGs, sustainable development goals. World Health Organization, 2019.

Zaidan, M.A., Wraith, D., Boor, B.E., Hussein, T.: Bayesian Proxy Modelling for Estimating Black Carbon Concentrations using White-Box and Black-Box Models, Appl. Sci. 9, 4976, 2019.

Zhang, J., Liu, J., Tao, S., Ban-Weiss, G.: Long-range transport of black carbon to the Pacific Ocean and its dependence on aging timescale, Atmos. Chem. Phys. 15, 11521–11535, 2015.

APPENDIX

Y	X1	X2	X3	X4	X5	X6	RMSE testing (µg/m3)	R2 training	R2 testing
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	RH norm	T norm	0.47	0.882	0.871
log BC	log PM2,5	log NO2	log O3	log N	log RH	log T	0.48	0.886	0.857
log BC	log PM2,5	log NO2	log O3	log N			0.52	0.850	0.853
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	RH norm		0.54	0.864	0.842
log BC	log PM2,5	log NO2	log O3	log N	log RH		0.52	0.874	0.842
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm			0.49	0.888	0.835
BC norm	PM2,5 norm	NO2 norm	O3 norm	RH norm	T norm.		0.61	0.824	0.800
log BC	log PM2,5	log NO2	log O3	log RH	log T		0.57	0.819	0.793

Table A1. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for Winter_PR.

Y	X1	X2	X3	X4	X5	X6	RMSE testing (µg/m3)	R2 training	R2 testing
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm	T norm	0.37	0.873	0.853
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm			0.30	0.860	0.848
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm		0.31	0.862	0.842
BC norm	PM2,5 norm	NO2 norm	O3 norm	HR norm	T norm		0.34	0.827	0.821
log BC	log PM2,5	log NO2	log O3	log N			0.38	0.859	0.817
log BC	log PM2,5	log NO2	log O3	log HR	log T		0.40	0.826	0.808
log BC	log PM2,5	log NO2	log O3	log N	log HR		0.40	0.863	0.792
log BC	log PM2,5	log NO2	log O3	log N	log HR	log T	0.42	0.862	0.768

Table A2. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for Summer_PR.

Y	X1	X2	X3	X4	X5	X6	RMSE testing (µg/m3)	R2 training	R2 testing
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm	T norm	0.27	0.883	0.838
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm		0.31	0.842	0.829
log BC	log PM2,5	log NO2	log O3	log N	log HR		0.31	0.816	0.812
log BC	log PM2,5	log NO2	log O3	log N	log HR	log T	0.26	0.850	0.810
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm			0.34	0.808	0.800
log BC	log PM2,5	log NO2	log O3	log N			0.32	0.811	0.789
BC norm	PM2,5 norm	NO2 norm	O3 norm	HR norm	T norm		0.37	0.836	0.781
log BC	log PM2,5	log NO2	log O3	log HR	log T		0.33	0.834	0.736

Table A3. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for Midday_PR.

Y	X1	X2	X3	X4	X5	X6	RMSE testing (µg/m3)	R2 training	R2 testing
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm	T norm	0.46	0.874	0.854
log BC	log PM2,5	log NO2	log O3	log N	log HR	log T	0.44	0.864	0.853
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm	HR norm		0.50	0.846	0.846
log BC	log PM2,5	log NO2	log O3	log N	log HR		0.48	0.839	0.832
log BC	log PM2,5	log NO2	log O3	log N			0.48	0.821	0.832
BC norm	PM2,5 norm	NO2 norm	O3 norm	N norm			0.56	0.861	0.807
BC norm	PM2,5 norm	NO2 norm	O3 norm	HR norm	T norm		0.58	0.797	0.787
log BC	log PM2,5	log NO2	log O3	log HR	log T		0.67	0.799	0.741

Table A4. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for Day_PR.

Y	X1	X2	X3	X4	X5	RMSE testing (µg/m3)	R2 training	R2 testing
BC norm	NO2 norm	O3 norm	HR norm	T norm		1.02	0.748	0.719
BC norm	NO2 norm	HR norm	T norm			1.07	0.725	0.692
log BC	log NO2	log O3	log HR	log T		1.08	0.750	0.686
BC norm	PM2,5 norm	NO2 norm	O3 norm	HR norm	T norm	1.09	0.808	0.685
log BC	log PM2,5	log NO2	log O3	log HR	log T	1.13	0.786	0.664
log BC	log NO2	log HR	log T			1.26	0.697	0.568
BC norm	O3 norm	HR norm	T norm			1.35	0.531	0.512
log BC	log O3	log HR	log T			1.36	0.545	0.499

Table A5. Model performance (RMSE and R^2 for the testing dataset, and R^2 for the training dataset) for Av. Roma.