King's Research Portal

DOI:

[10.1016/j.atmosenv.2020.117757](https://doi.org/10.1016/j.atmosenv.2020.117757)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Analitis, A., Barratt, B., Green, D., Beddows, A., Samoli, E., Schwartz, J., & Katsouyanni, K. (2020). Prediction of PM_{2.5} concentrations at the locations of monitoring sites measuring PM₁₀ and NO_x using generalized additive models and machine learning methods: A case study in London. *ATMOSPHERIC ENVIRONMENT*, 240, [117757]. <https://doi.org/10.1016/j.atmosenv.2020.117757>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Manuscript Number: ATMENV-D-20-00319R2

Title: Prediction of PM2.5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods: a case study in London

Article Type: Research Paper

Keywords: PM2.5 prediction; environmental exposure; random forest; ensemble methods; London case study

Corresponding Author: Professor Klea Katsouyanni,

Corresponding Author's Institution: University of Athens

First Author: Antonis Analitis

Order of Authors: Antonis Analitis; Ben Barratt; David Green; Andrew Beddows; Evangelia Samoli; Joel Schwartz; Klea Katsouyanni

Abstract: The adverse health effects of air pollutants, especially those of PM2.5, are well documented. However, a lack of adequate monitoring and weaknesses in modelling approaches do not allow a good assessment of health effects in many areas of the World. Advances in computational methods and the availability of new data sets, e.g. satellite remote observations, have enlarged the possibilities of modelling for application in large scale health effects studies. However, PM2.5 monitoring is very recent in most of the World and more limited compared to other pollutants, and understanding how to use PM10 monitors to estimate PM2.5 exposure is therefore important. Since interest in these methods is relatively recent, there is a need for testing their performance against ambient measurements, but long term PM2.5 datasets are less readily available than PM10 in many regions. In the present study we report the methodology and results of using regression modelling and a machine learning method (Random Forest-RF), as well as a combination of the two, to enhance a PM2.5 measurement data base in London using PM10 and NOx measurements as well as other predictors and compare the relative performance of each method. We find that the combination of predictions by the regression model and the RF performs best and we obtain a cross-validation R2 of 99.29% and 98.22% for the 5-year periods 2004-2008 and 2009-2013 respectively and a Mean Square Error near 1. Our enhanced data base for PM2.5 is available for use by other researchers.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
Data will be made available on request

Dear Editor,

Thank you for provisionally accepting our revised manuscript entitled “Prediction of $PM_{2.5}$ concentrations at the locations of monitoring sites measuring PM_{10} and NO_x , using generalized additive models and machine learning methods: a case study in London” by A. Analitis et al., for publication in Atmospheric Environment.

We have now added a map of the monitoring sites network, according to your and the Reviewer’s suggestion. We also identified a mistake in the description of the number of NO_x monitoring sites which we corrected to match the sites used in modelling.

We hope that our paper will now be ready for acceptance.

Sincerely,

K. Katsouyanni on behalf of all authors.

Prediction of PM_{2.5} concentrations at the locations of monitoring sites measuring PM₁₀ and NO_x, using generalized additive models and machine learning methods: a case study in London

Antonis Analitis¹, Ben Barratt^{2,3}, David Green², Andrew Beddows², Evangelia Samoli¹, Joel Schwartz⁴, Klea Katsouyanni^{1,2}

1. National and Kapodistrian University of Athens, Faculty of Medicine, Department of Hygiene, Epidemiology and Medical Statistics, Athens, Greece

2. MRC Centre for Environment and Health, King's College London, UK

3. HPRU in Environmental Exposures and Health, Imperial College London, UK

4. Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

Address correspondence to
Klea Katsouyanni
Professor of Public Health
King's College London
Franklin-Wilkins Building
Room 4.129
150 Stamford street
London SE1 9NH
e-mail: kkatsouy@med.uoa.gr
telephone: +44-207-8483844
mobile tel: +30-6944681057

Abstract

The adverse health effects of air pollutants, especially those of PM_{2.5}, are well documented. However, a lack of adequate monitoring and weaknesses in modelling approaches do not allow a good assessment of health effects in many areas of the World. Advances in computational methods and the availability of new data sets, e.g. satellite remote observations, have enlarged the possibilities of modelling for application in large scale health effects studies. However, PM_{2.5} monitoring is very recent in most of the World and more limited compared to other pollutants, and understanding how to use PM₁₀ monitors to estimate PM_{2.5} exposure is therefore important. Since interest in these methods is relatively recent, there is a need for testing their performance against ambient measurements, but long term PM_{2.5} datasets are less readily available than PM₁₀ in many regions. In the present study we report the methodology and results of using regression modelling and a machine learning method (Random Forest-RF), as well as a combination of the two, to enhance a PM_{2.5} measurement data base in London using PM₁₀ and NO_x measurements as well as other predictors and compare the relative performance of each method.

We find that the combination of predictions by the regression model and the RF performs best and we obtain a cross-validation R² of 99.29% and 98.22% for the 5-year periods 2004-2008 and 2009-2013 respectively and a Mean Square Error near 1. Our enhanced data base for PM_{2.5} is available for use by other researchers.

Key words: PM_{2.5} prediction; environmental exposure; random forest; ensemble methods; London case study

1. Introduction

The adverse health effects of air pollutants, especially those of PM_{2.5}, are well documented (WHO, 2013a, 2013b). Most of the studies reported are from areas of the world where monitoring of pollutants concentrations is routinely performed. In several studies modelling has been used to assess the health effects of exposure to various pollutants in order to enhance the spatial scale (de Hoogh et al 2016; Christidis et al 2019) or to estimate long-term pollutants' levels in the absence of monitoring (Brauer et al 2012). Advances in computational methods and the availability of various data, e.g. satellite remote observations, have led to applications involving spatio-temporal models i.e. predicting daily pollutant concentrations for specific areas at fine spatial scales for a wide range of usage. One objective for developing spatio-temporal prediction models is short term forecasting, as was done by Cortina-Januchs et al (2015) in Salamanca, Mexico and Gocheva-Ilieva and Stoimenova (2018) for the city of Pleven, Bulgaria. Another target is to predict daily concentrations in areas without measurements for use in health studies, as was done in the U.S.A., where Lee et al (2016) predicted PM_{2.5} daily concentrations at 1 km² grid resolution in southeastern US and Di et al (2017) predicted spatiotemporal ozone concentrations in the continental US. Xue et al. (2019) developed spatiotemporal models for PM_{2.5} for China and provided these online. Comparisons of the performance of computationally intensive methods versus more traditional ones have also been reported. Thus, for example, Russo et al. (2015) used stepwise regression procedures and neural networks (NN) based on routine pollutants measurements in Lisbon, Portugal, to develop a prediction model for PM₁₀ concentrations and found no clear superiority of the NN. Mishra and Goyal (2015), based on a relatively short measurement campaign in Agra, India, developed models based on multiple regression, principal components analysis and NN to predict Nitrogen Dioxide (NO₂) levels at the famous Taj Mahal monument, where the combination of principal components and NN performed best. Chen et al (2019) compared the performance of 16 algorithms for predicting the spatial variability of long-term averages of PM_{2.5} and NO₂ across Europe using the AIRBASE data base and found that both the PM_{2.5} and NO₂ models performed similarly across algorithms. It

is evident that models perform better when the area of prediction is relatively homogeneous and when information on relevant predictors is available.

Within the London based, MRC funded, project entitled "Comparative evaluation of Spatio-Temporal Exposure Assessment Methods for estimating the health effects of air pollution" (STEAM), we formed a data base of air pollutants (specifically 24-hour average PM_{10} , $PM_{2.5}$, and NO_x and daily 8-hour maximum O_3) concentrations, including all measurements from sites within the London M25 orbital motorway. The monitoring network in London is exceptionally dense providing a unique opportunity to study the effects of pollutants in a mega-city, observing temporal and spatial differentiations. However, the number of sites measuring $PM_{2.5}$ is significantly smaller than the corresponding for PM_{10} and NO_2 . The objective of the study reported in this paper was to enhance the $PM_{2.5}$ data base for London using other pollutants, meteorological and land-use variables to predict daily $PM_{2.5}$ concentrations at the location of fixed monitors measuring PM_{10} but not $PM_{2.5}$. In our analysis we used regression models and machine learning methods. As $PM_{2.5}$ is the pollutant for which most of the evidence exists for adverse health effects, a larger data base of measurements can be extremely useful to aid the verification of subsequent spatio-temporal exposure models and to be used in health studies conducted in London and elsewhere.

2. Data and Methods

2.1 Data

We used all measurements performed at fixed sites by the London Air Quality Network (LAQN, <http://www.londonair.org.uk/LondonAir/Default.aspx>), during the years 2004-2013 (Figure 1). Not all sites provided measurements for the whole time period and there were missing values during periods of operation.

Specifically, for PM_{10} we compiled daily mean data from 108 sites in 2004-2008 and 115 sites in 2009-2013 (136,874 and 134,230 measurements respectively, range of the number of measurements per site: 50-1,812 and 11-1,808 respectively); for NO_x we compiled data from 127 sites in 2004-2008 and 130 in 2009-2013 (163,778 and

163,088 measurements respectively, range of the number of measurements per site: 38-1,803 and 19-1,817 respectively), whilst for PM_{2.5} data were available only from 24 sites in 2004-2008 and 33 sites in 2009-2013 (16,957 and 40,083 measurements respectively, range of the number of measurements per site 13-1,382 and 95-1,809 respectively). The increased number of PM_{2.5} measurements in later years and the possibility that air pollution determinants may have changed over time, led to the decision to limit the time span of model development to 5 years, thus the model was developed for 2004-2008 and 2009 -2013 separately.

Figure 1. Map of the geographical location of the fixed air pollution monitoring network operated at the Greater London study area

We checked whether all fixed sites that measured PM₁₀ and NOx concurrently reflected the same association between the two pollutants in order to identify sites which may have unique characteristics affecting their association, e.g. be near a local, non-combustion source, such as industry, long-term construction or waste handling. For this purpose, mean NOx vs mean PM₁₀ concentrations from all sites which provided both measurements were plotted and a regression line was estimated. In order to identify and exclude outlier sites, Cook's distance was calculated and a Cook's D value >1 was defined as indicating an influential observation (Cook & Weisberg 1982). Cook's D, by removing the ith data point from the model and re-estimating the regression, summarizes how the values in the regression model change when the ith observation is removed. After applying the method, no fixed site was excluded from the analysis, since no outlier was identified.

Additionally, the different methods of measuring PM_{2.5} and their comparability were taken into account and a correction was adopted. Whilst PM_{2.5} measured by the Filter Dynamics Measurement System (FDMS) instruments (20 sites) were equivalent to the reference method, the PM_{2.5} data obtained with the Tapered Element Oscillating Microbalance (TEOM) instruments (15 sites) had widely acknowledged discrepancies due to the loss of volatile particulate (Allen et al., 1997; Charron et al., 2004). Previously, for PM₁₀ a Volatile Correction Model (VCM) method was developed (Green et al. 2009), which used regional measurements of the volatile particulate made by the FDMS to correct the TEOM measurements; this was optimized for PM_{2.5} by Barratt et al. (2009).

Specifically, the correction applied was:

$$TEOM_{VCM}PM_{2.5} = \left(\frac{T_{d0} - 3}{1.03} \times \frac{P_a}{P_r} \times \frac{T_r}{T_a} \right) - \left((1 + fVCM) \frac{1}{n} \sum_i^n z_i \right)$$

Where:

T_{d0} is the daily mean TEOM PM_{2.5} concentration

Pr is the set reporting pressure of the TEOM

170 P_a is the regional daily mean ambient pressure
 171 T_a is the regional daily mean ambient temperature
 172 T_r is the set reporting temperature of the TEOM
 173 f_{VCM} is the VCM function value (optimised to 0.72 for $PM_{2.5}$ by Barratt et al., 2009)
 174 n is the number of FDMS volatile measurement sites used in that day
 175 z_i is the FDMS daily mean volatile measurement for station i
 176 Data for meteorological variables, specifically for mean daily temperature ($^{\circ}C$),
 177 relative humidity (%), wind speed (ms^{-1}) and wind direction ($^{\circ}N$) were compiled from
 178 all available sites with <25% missing values over the study period ($n=7$).
 179 Meteorological data were obtained from the London Air Quality Network database
 180 (www.londonair.org.uk). Because the number of available measurement sites was
 181 not so dense to allow characterisation at the location of all pollution measurement
 182 sites, but the correlation of variables between sites is very high (r between sites
 183 >0.99 for temperature and >0.85 for the other meteorological variables) for all
 184 meteorological variables, we averaged per day across all meteorological sites and
 185 used one time series for each meteorological variable.

186 ***2.2 Development of predictive models***

187 For the development of the models we included all sites which measured $PM_{2.5}$,
 188 PM_{10} and NO_x concurrently (the number of sites was 18 for 2004-2008 and 25 for
 189 2009-2013) and developed the model using the 2004-2008 and 2009-2013 data
 190 separately, as described above.. We used the $PM_{2.5}$ measurements taken by the
 191 reference method as well as the corrected TEOM measurements. As predictor
 192 variables we included PM_{10} and NO_x as well as other covariates, such as
 193 meteorological, geographical and temporal variables. The model developed was
 194 used to predict $PM_{2.5}$ at fixed sites with no $PM_{2.5}$ measurements (but which
 195 measured PM_{10} & NO_x) and thus provide a greatly enhanced data base of $PM_{2.5}$
 196 estimated concentrations. We used two methods, a Regression model and a Random
 197 Forest, as well as a combination method of the two.

198

2.2.1 Regression Model

The model was developed with daily values of measured $PM_{2.5}$ as the dependent variable and daily values of the following covariates (all time-varying variables are for lag0, i.e. same day): PM_{10} , measured at the same site as $PM_{2.5}$ (24-hour values, in $\mu g/m^3$); NO_x , measured from the same site as $PM_{2.5}$ (24-hour average, in $\mu g/m^3$); type of monitor (roadside/kerbside, background); Day of the week; Time trend and monthly variations; Mean daily temperature averaged over all available sites; Mean daily relative humidity averaged over all available sites; Mean daily wind speed averaged over all available sites; Mean daily wind direction; Coordinates, longitude and latitude. The above covariates were chosen based on previously reported results (Di et al 2019) and, specifically for the pollutants used as predictors, prior knowledge on emissions and dispersion patterns of each pollutant.

As sensitivity analyses, we tested, smoothed terms for both PM_{10} and NO_x and smoothed terms for all variables with degrees of freedom (df) chosen by the Generalized Cross Validation (GCV) criterion, alternatively. We also tested the interaction between PM_{10} and site classification as well as a bivariate smooth function of wind speed and PM_{10} . Finally, we added a bivariate smooth function for coordinates (latitude and longitude) and its interaction with PM_{10} , to capture any remaining spatial variability of the fitted vs the observed values. In the models with latitude and longitude the variable "monitor type" was excluded.

To obtain the R^2 and Mean Square Error (MSE) of the models we used 10-fold cross-validation (CV) in order to test the model's ability to predict new data that was not used in estimating it. Specifically, we fitted the model based on a training data set (90% of the data) and using 10% of the observations as validation data. We then predicted the omitted observations (the validation data set) and calculated the R^2 and MSE. This was done sequentially 10 times and then the predicted values were used to obtain the R^2 and MSE (Hastie et al 2009; Rodriguez et al 2010).

2.2.2 Random Forest (RF)

We further applied a RF as an alternative method to improve the fit of the prediction for the PM_{2.5} data. RF is a machine learning algorithm, defined as a “combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” (Breiman, 2001). Each tree is a form of nonlinear regression model where samples are partitioned at each node of a binary tree based on the values of randomly selected input variables (features). A RF can be used for both continuous (regression) and categorical outcomes (classification), combining multiple decision trees in determining the final output, that is, the mean prediction of the individual trees in regression or the mode of the classes in classification tasks. Due to its properties, the RF algorithm gives high accuracy prediction and has advantages over other machine learning techniques in terms of ability to handle highly non-linear data, robustness to noise and tuning simplicity (Caruana and Niculescu-Mizil, 2006; Menze et al, 2009). High accuracy in predictions though, comes at the expense of interpretability. The covariate interpretability is not so straightforward, as it is in terms of regression coefficients, and is performed in terms of "variable importance". In the variable importance graphs the predictors are ranked in terms of increase in MSE and in node purity. Additionally, the partial dependence plots show the mean marginal influence of each predictor on the PM_{2.5} estimation.

We used the “randomForest” library in R, version 3.5.1. RF was also applied separately for each time-period of interest, i.e. 2004-2008 and 2009-2013. We used the default parameter values for the number of trees (500) and tree depth (max), and 3 and 4 variables were tested in each step based on the rule “number of variables/3”. We used the same variables as those tested in the regression model (A) described above.

2.2.3 Combination of predictions from the Regression and the Random Forest

We combined the predictions from the regression and the RF method to improve the fit of the model, addressing the situation where the two methods may perform differently at different levels of the PM concentrations. The predictions were used as

independent variables in a Generalized Additive Model (GAM) for each period separately, as described below:

$$PM_{2.5} = s(\text{predictions from the Regression model}) + s(\text{predictions from the RF method})$$

where s is a penalized spline with the number of df chosen by GCV.

Finally we chose the best model based on the CV adjusted R^2 between the Regression, the RF and their GAM combination and used it to 1) fill in missing $PM_{2.5}$ values from stations providing $PM_{2.5}$ measurements both during 2009-2013 and 2004-2008 and 2) provide $PM_{2.5}$ series in stations with only PM_{10} and NO_x measurements.

All analyses were performed using the R statistical package v. 3.5.1 and libraries "mgcv" and "randomForest".

This study is based on publicly available data and therefore no ethics issues are involved.

3. Results

3.1 Predictive model development

Time period 2004-2008

Regression model

Table 1 shows the contribution of each variable to the value of the adjusted R^2 of the regression model, for 2004 - 2008. All variables entered into the model after PM_{10} increased the value of the adjusted R^2 slightly. The best model in terms of adjusted R^2 was Model 14:

$$PM_{2.5} = \text{intercept} + PM_{10} + NO_x + \text{day of the week (6 dummy variables for 7 days, with "Sunday" used as reference category)} + ns(\text{trend, df}=12 \text{ per year}) + ns(\text{Temperature, 4$$

293 $df) + ns(\text{rel. humidity, 2 df}) + ns(\text{wind speed, 2 df}) + ns(\text{wind direction, 3 df}) + s(\text{latitude,}$
 294 $\text{longitude}) + PM_{10} * s(\text{latitude, longitude})$ where:
 295 ns: natural cubic spline
 296 s: bivariate smoothing spline
 297 **PM_{2.5}**: measured PM_{2.5} (24-hour values, $\mu\text{g}/\text{m}^3$)
 298 **PM₁₀**: measured PM₁₀ from the same site as PM_{2.5} (24-hour values, $\mu\text{g}/\text{m}^3$)
 299 **NO_x**: measured NO_x from the same site as PM_{2.5} (24-hour average, $\mu\text{g}/\text{m}^3$),
 300 **Day of the week**: 0=Sunday, 1=Monday,...,6=Saturday (six dummy variables for 7
 301 days; Sunday: reference category)
 302 **Time trend**: a variable denoting the length of the period with values 1,...,N entered
 303 as ns with 12 df/year
 304 **Temperature**: measured in $^{\circ}\text{C}$ and entered as ns with 4 df
 305 **Humidity**: relative humidity % entered as ns with 2df
 306 **Wind speed**: ms^{-1} entered as ns with 2 df
 307 **Wind direction**: $^{\circ}\text{N}$, entered as ns with 4 df
 308 **s(latitude, longitude)**: A bivariate smooth function for the coordinates (longitude,
 309 latitude) of the fixed monitoring sites and
 310 **PM₁₀ * s(latitude, longitude)**: interaction term
 311 The 10-fold CV adjusted R^2 of the model for 2004 – 2008 was 92.10% and the MSE
 312 $10.85 (\mu\text{g}/\text{m}^3)^2$.
 313
 314

Model number	Model covariates	adj R^2
1	PM ₁₀ ($\mu\text{g}/\text{m}^3$)	0.8359
2	Model 1 + NO _x ($\mu\text{g}/\text{m}^3$)	0.8440
3	Model 2 + monitor type (roadside/background)	0.8453
4	Model 3 + week day (six dummy variables for 7 days; Sunday: reference category)	0.8476
5	Model 4 + ns(time trend, 12 df/year)	0.8697
6	Model 5 + ns(temperature $^{\circ}\text{C}$, 4 df)	0.8722
7	Model 6 + ns(relative humidity %, 2 df)	0.8906
8	Model 7 + ns(wind speed ms^{-1} , 2 df)	0.8999
9	Model 8 + ns(wind direction $^{\circ}\text{N}$, 3 df)	0.9000
10	Penalized splines for PM ₁₀ & NO _x (excluding the corresponding linear terms) with df from GCV + all	0.9022

	other covariates from model 9	
11	Penalized splines (instead of natural splines with fixed df) for temperature, humidity, wind speed, wind direction with df from GCV + all other covariates from model 10	0.9037
12	Model 11 + Interaction of PM ₁₀ and monitor type	0.9034
13	Model 11 + Bivar. spline s(PM ₁₀ ,wspeed)	0.9059
14	Model 9 + Main effects s(lat, long) + inter. PM ₁₀ x s(lat,long) excluding monitor type	0.9232

* ns(): natural splines, s(): bivariate smoothing spline, df: degrees of freedom

Table 1. Development of the Regression model: adjusted R² for every set of covariates 2004-2008

Random Forest

The Random Forest R² and MSE for 2004-2008 were 95.03% and 6.88 respectively. Figure 2 shows the relevant partial dependence plot which gives a graphical depiction of the marginal effect of each predictor variable on PM_{2.5}. It can be seen that the associations of the predictors and PM_{2.5} concentrations are described by non-linear functions. In Figure 3 we can see the variable importance of each predictor in terms of both decrease in MSE and remaining error in predictive accuracy after a node split (node impurity).

Figure 2. Partial dependence plot of $PM_{2.5}$ on predictor variables used in the Random Forest, 2004-2008

Figure 3. Variable importance plot in the Random Forest (2004-2008) in terms of % increase in Mean Square Error (%IncMSE) and increase in node purity (IncNodePurity).

Time period 2009-2013

Regression model

Table 2 shows the contribution of each variable to the value of the adjusted R^2 of the regression model, as well as the R^2 of all alternative models applied.

A similar pattern for the contribution to R^2 with the period 2004-2008 was observed. The same model, Model 14, was again the model with the best R^2 equal to 88.40%.

The CV adjusted R^2 of the model was 88.15% and the mean square error (MSE) 12.53 $(\mu\text{g}/\text{m}^3)^2$.

Model #	Model covariates	adj R^2
1	PM ₁₀ ($\mu\text{g}/\text{m}^3$)	0.7921
2	Model 1 +NO _x ($\mu\text{g}/\text{m}^3$)	0.7969
3	Model 2 + monitor type (roadside/background)	0.8003
4	Model 3 +Week day (six dummy variables for 7 days; Sunday: reference category)	0.8010
5	Model 4 +ns(time trend, 12 df/year)	0.8169
6	Model 5 +ns(temperature °C, 4 df)	0.8215
7	Model 6 +ns(relative humidity %,2 df)	0.8400
8	Model 7 +ns(wind speed ms^{-1} , 2 df)	0.8479
9	Model 8 +ns(wind direction °N, 3 df)	0.8486
10	Penalized splines for PM ₁₀ & NO _x (excluding the corresponding linear terms) with df from GCV+ all other covariates from model 9	0.8627
11	Penalized splines (instead of natural splines with fixed df)for temperature, humidity, wind speed, wind directionwith df from GCV+ all other covariates from model 10	0.8663
12	Model 11 + Interaction of PM ₁₀ and monitor type	0.8577
13	Model 11 + Bivar. spline s(PM ₁₀ ,wspeed)	0.8692
14	Model 9 + Main effects s(lat, long) + inter. PM ₁₀ x s(lat,long)+ all other covariates from model 9 excluding monitor type	0.8840

* ns(): natural spline, s(): bivariate smoothing spline, df: degrees of freedom

Table 2. Development of the Regression model: adjusted R^2 for every set of covariates 2009-2013

358

359 **Random Forest**

360 The Random Forest R^2 and MSE for the 2009-2013 period were 92.8% and 7.64.

361 Figure 4 shows the relevant partial dependence plot which gives a graphical
362 depiction of the marginal effect of each predictor variable on $PM_{2.5}$. Although the
363 general pattern shown is consistent between the two 5-year periods, some of the
364 associations of meteorological and spatial variables with $PM_{2.5}$ differ somewhat. This
365 may be attributed to different meteorological conditions across the two periods.

366 Figure 5 presents the variable importance of each predictor in terms of both
367 decrease in MSE and remaining error in predictive accuracy after a node split (node
368 impurity). In both the 5-year time periods the most important predictor is PM_{10} . For
369 the other predictors, in terms of increase in MSE the ranking between variables is
370 slightly different but in terms of increase in node purity it is identical for the two
371 periods.

372

Figure 4. Partial dependence plot of $PM_{2.5}$ on predictor variables used in the Random Forest (2009-2013)

Figure 5. Variable importance plot in the Random Forest (2009-2013) in terms of % increase in Mean Square Error (%IncMSE) and increase in node purity (IncNodePurity).

Time periods 2004-2008 and 2009-2013

Combination of predictions obtained from the linear regression model & the predictions obtained from the Random Forest method

The 10-fold CV adjusted R^2 of the combined GAM model was 99.29% and 98.99% for the time periods 2004-2008 & 2009-2013 while MSE was 0.98 and 1.07, respectively. We also calculated the spatial CV, following the method applied in Hu et al (2017), by excluding 10% of the measurement sites from the training data set and predicting them from the models including the other 90% of sites and repeating this procedure until all predictions were done. The corresponding R^2 and MSE for the two periods

were 99.28%, 1.00 and 98.94%, 1.13. Thus, there is no substantial reduction in R^2 and MSE when the spatial predictions are assessed.

The predictions from the combined model were included in the final enhanced $PM_{2.5}$ database.

3.2 Predicted values for $PM_{2.5}$

Based on the described methodology we predicted 102,800 and 85,436 daily concentrations in specific sites for $PM_{2.5}$ in 2004-2008 and 2009-2013, respectively.

Table 3 presents the descriptive statistics for measured and predicted values of $PM_{2.5}$ overall, as well as predicted values separately for sites with measured $PM_{2.5}$ (directly comparable with measured values) and for sites without $PM_{2.5}$ measurements (which provided only PM_{10} and NO_x) for 2004-2008 and 2009-2013. During 2004-2008 there were 16,957 $PM_{2.5}$ measurements and 102,800 additional estimates are made available whilst for 2009-2013 the corresponding numbers are 40,083 and 85,436. Measured and predicted means and medians are identical to the first decimal for 2009-2013. It is also noted that concerning the prediction of concentrations in the relatively higher range of values, those >15 , >25 and $>50 \mu g/m^3$, the percentages of measured and predicted values above those thresholds are remarkably similar (Table 3).

Figure 6 shows the agreement between the measured $PM_{2.5}$ and the predicted from each method and their combination. The predictions from the combined method have a better agreement with measured values, and adequate at both the low and high end of the exposure range. The intercepts are very near 0 and the slopes very close to 1 in both time periods.

421

	25 th percentile	Median	Mean	75 th percentile	% values >15 µg/m ³	% values >25 µg/m ³	% values >50 µg/m ³	N
2004 – 2008								
A. PM_{2.5} measurements	9.29	13.99	17.45	22.17	45.9	19.6	2.1	16957
B. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were available	9.49	14.45	17.81	22.57	48.0	20.7	2.1	11695
C. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were NOT available but there were PM₁₀ and NOx	8.60	12.77	15.99	19.87	40.1	16.4	1.7	102800
D. all predicted PM_{2.5} concentrations (sum of B & C)	8.68	12.93	16.18	20.2	40.9	26.8	1.8	114495
2009 – 2013								
A. PM_{2.5} measurements	9.33	13.35	16.39	20.10	42.1	15.4	1.6	40083
B. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were available	9.46	13.37	16.35	19.95	41.9	15.0	1.6	26645
C. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were NOT available but there were PM₁₀ and NOx	9.98	13.46	15.97	19.11	41.4	12.5	1.0	85436
D. all predicted PM_{2.5} concentrations (sum of B & C)	9.87	13.44	16.06	19.30	41.5	13.1	1.1	112081

422

423 **Table 3.** Descriptive statistics for the observed & predicted PM_{2.5} (µg/m³) concentrations applying the combination of both methods model for the
424 time periods 2004 - 2008 and 2009 - 2013.

425

Figure 6. Graph of observed vs predicted $PM_{2.5}$ for each period, by method (Regression, Random Forest and GAM combination).

4 Discussion

We predicted a large number of daily PM_{2.5} concentrations ($\mu\text{g}/\text{m}^3$) in London at the location of monitoring sites from measured PM₁₀ and NO_x and several other covariates. Our predictions were done to fill in missing PM_{2.5} values in monitors that normally measure PM_{2.5} concentrations and to predict PM_{2.5} in monitors which do not measure PM_{2.5} but do measure PM₁₀ and NO_x. We developed a Regression and a RF model and also used a GAM to ensemble average the Regression and RF predictions. All models performed well in terms of adjusted R², but the combined model improved the fit and had an adjusted R² of 99.29% for 2004-2008 and 98.22% for 2009-2013. Additionally, the combined model had a much smaller MSE, 0.98 for 2004-2008 and 1.07 for 2009-2013. Our model predicted better than using PM₁₀ alone as predictor: about 10% of the variability of PM_{2.5} is explained by adding the other variables in the regression model and most of the other variables are important in the RF as well. The enhanced data base is intended for use in health studies, either as such or as input for use in exposure assessment models.

London is unique globally for having a very large density of fixed site monitors measuring pollutants routinely in a relatively limited area. For example in London from 2009 to 2013 there are data available for NO_x from 216 sites within the M25 orbital motorway, an area of approximately 1100 square miles, whilst the AIRBASE data set provides NO₂ measurements from 2399 sites for the whole of Europe, an area larger by 3.5 thousand times (Chen et al 2019). Thus, it is possible to use the London data to study the spatiotemporal variation of air pollutants concentrations and the consequences of measurement error in assessing human exposure, since the large data set allows for testing various situations. However the data in London are less abundant for PM_{2.5}, which is measured in a much smaller number of sites, and thus enhancing the PM_{2.5} data base can be very useful for methodological and health studies, especially since the reported predictions for PM_{2.5} daily concentrations in this study are excellent.

Several studies have made comparisons between regression models and machine learning techniques for predicting pollutant concentrations either for detecting

spatial or spatio-temporal variability. Chen et al (2019) compared the performance of 16 algorithms for predicting the spatial variability of long-term averages of PM_{2.5} and NO₂ across Europe using the AIRBASE data base. They report that the PM_{2.5} and NO₂ models performed similarly across algorithms with an average CV R² of 0.53. However, the RF among the machine learning techniques performed somewhat better (CV R² 0.63), whilst the artificial neural network performed less well. Several publications report spatio-temporal predictions, more comparable to the present study. Generally, they are using measurements from smaller density networks and often develop complex models, possibly using machine learning methods and/or incorporating AOD satellite data. To focus on PM_{2.5}, Lee et al (2016) used a regression based approach to estimate PM_{2.5} at a 1km² resolution in 7 Southeastern States in the US and obtained a 10-fold CV R² between 0.70 and 0.81 and RMSE between 2 and 3 µg/m³. Xue et al (2019) developed a multi-dimensional model to predict PM_{2.5} across China, with an objective similar in some way to ours, namely to provide a data base for use in health studies. Typically, they report R²<0.80. Brokamp et al (2018) use a RF to predict PM_{2.5} in the Cincinnati, Ohio metropolitan area from 2000 to 2015, using several available variables including AOD and report a CV R² of 0.91 with a RMSE equal to 2.2 µg/m³. Hu et al (2017) develop a RF approach to predict PM_{2.5} in the 48 conterminous US, an area of 3.12 million square miles, and report a CV R² of 0.80.

Beloconi et al (2016) developed spatiotemporal models for PM₁₀ and PM_{2.5} for London using satellite AOD data, measurements, meteorological and land use variables. Their purpose was to estimate concentrations at 1km² grid using sophisticated modelling to obtain the best possible predictability. However, the fact that remote sensing data have missing values and measurements of PM_{2.5} are less dense compared to PM₁₀, resulted in a maximum R² of 90% for PM_{2.5} models. The objective of our study is not directly comparable to that of Beloconi et al: we aimed at predicting the best possible spatiotemporal estimate for PM_{2.5} concentrations at the geographical locations of the monitoring sites in London. Since we have excellent predictors for the site locations (such as PM₁₀ measurements), our models perform almost perfectly and further predictors would not add to the model performance.

Our combination model performs better than those reported in the literature mainly because of the availability of much more dense measurements on the predictors and the fact that we had a closely related proxy measurement (PM_{10}) at the prediction points. In some published studies, various machine learning methods predicting $PM_{2.5}$ have been combined (Di et al 2019; Xiao et al 2018) using a GAM function, and the combination (ensemble) models also performed best.

The enhanced $PM_{2.5}$ data base for London can be used in other health studies and provide improved input for estimating concentrations in points of interest and for modelling. Our method demonstrates that the combination of more traditional regression and machine learning methods may improve the predictions of pollutants concentrations for various applications.

Our enhanced $PM_{2.5}$ data base for the year 2004-2013 is available on request from the corresponding author.

Conclusions: We present a methodology to enhance the $PM_{2.5}$ database of measurements for use in health studies and apply it in the case of London. The GAM combination (ensemble) of the two separate model predictions performs best, demonstrating that the combination of more traditional regression and machine learning methods leads to improved predictions throughout the range of concentrations. A larger data base of measurements can be extremely useful to aid the verification of subsequent spatio-temporal exposure models and to be used in health studies conducted in London and elsewhere.

Acknowledgements: We are grateful to the UK Government and Local Authorities providing air pollution measurements used in this study, managed by the Environmental Research Group at King's College London and Ricardo Energy and Environment.

Funding: Research in this paper as part of the STEAM project was funded under the MRC UK Grant ref.: MR/N014464/1. The Funding Source (MRC UK) had no role in the

study design, the collection, analysis or interpretation of data, in writing the paper
and in the decision to submit the article for publication.

Author contributions: AA: Formal Analysis, Data Curation, Methodology, Writing -
Original Draft, Review & Editing; BB: Data Curation, Methodology, Writing - Review &
Editing; DG: Methodology, Writing - Review & Editing; AB: Methodology, Writing -
Review & Editing; ES: Conceptualization, Methodology, Writing - Review & Editing;
JS: Conceptualization, Methodology, Writing - Review & Editing; KK:
Conceptualization, Funding acquisition, Methodology, Writing - Original Draft,
Review & Editing

Conflicts of Interest: The authors declare that they have no conflict of interest

541 REFERENCES

- 542 Allen G, Sioutas C, Koutrakis P, Reiss R, LurmannFWand, Roberts PT. Evaluation of
 543 the TEOM® Method for Measurement of Ambient Particulate Mass in Urban Areas,
 544 Journal of the Air & Waste Management Association1997; 47(6); 682-689. DOI:
 545 10.1080/10473289.1997.10463923
- 546 Beloconi A, Kamarianakis Y, Chrysoulakis N. Estimating urban PM₁₀ and PM_{2.5}
 547 concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover
 548 and morphology data. Remote Sensing of Environment 2016; 172: 148-164
- 549 Barratt B, Carslaw D, Fuller G, Green D, Tremper A. Analysis of Air Quality Data – Low
 550 Emission Zone Year 1 Results. Prepared for Transport for London by King’s College
 551 London Environmental Research Group & Institute for Transport Studies, University
 552 of Leeds, May 2009
- 553 Brauer M, Amann M, Burnett RT, Cohen A, Dentener F, Ezzati M, Henderson SB,
 554 Krzyzanowski M, Martin RV, Van Dingenen R, van Donkelaar A, Thurston
 555 GD.Exposureassessment for estimation of the global burden of disease attributable
 556 to outdoor air pollution. Environ Sci Technol. 2012 Jan; 17;46(2):652-60. doi:
 557 10.1021/es2025752. Epub 2012 Jan 6.
 558
- 559 Breiman, L.Random Forests. Machine Learning 2001; 45: 5-32.
 560 doi.org/10.1023/A:1010933404324
- 561 Brokamp C, Jandarov R, Hossain M, Ryan P. Predicting Daily Urban Fine Particulate
 562 Matter Concentrations Using a Random Forest Model. Environ Sci Technol. 2018 Apr;
 563 3;52(7):4173-4179. doi: 10.1021/acs.est.7b05381.
- 564 Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, Bauwelinck M, van
 565 Donkelaar A, Hvidtfeldt UA, Katsouyanni K, Janssen NAH, Martin RV, Samoli E,
 566 Schwartz PE, Stafoggia M, Bellander T, Strak M, Wolf K, Vienneau D, Vermeulen R,
 567 Brunekreef B, Hoek G. A comparison of linear regression, regularization, and
 568 machine learning algorithms to develop Europe-wide spatial models of fine particles
 569 and nitrogen dioxide.Environ Int. 2019 Sep;130:104934. doi:
 570 10.1016/j.envint.2019.104934. Epub 2019 Jun 20
- 571 Christidis T, Erickson AC, Pappin AJ, Crouse DL, Pinault LL, Weichenthal SA, Brook JR,
 572 van Donkelaar A, Hystad P, Martin RV, Tjepkema M, Burnett RT, Brauer M. Low
 573 concentrations of fine particle air pollution and mortality in the Canadian Community
 574 Health Survey cohort. Environ Health. 2019; 18: 84. Published online 2019 Oct 10.
 575 doi: 10.1186/s12940-019-0518-y
- 576 Caruana R, Niculescu-Mizil A, 2006; An empirical comparison of supervised learning
 577 algorithms. ICML '06: Proceedings of the 23rd international conference on Machine
 578 learning. June 2006;pp 161-68https://doi.org/10.1145/1143844.1143865

579 CharronA, Harrison RM, Moorcroft S, Booker J. Quantitative interpretation of
 580 divergence between PM₁₀ and PM_{2.5} mass measurement by TEOM and gravimetric
 581 (Partisol) instruments. *Atmospheric Environment* 2004; 38(3) 415-423.
 582 <https://doi.org/10.1016/j.atmosenv.2003.09.072>

583 Cook RD, Weisberg S. *Residuals and Influence in Regression*. New York, NY: Chapman
 584 & Hall 1982

585 Cortina-Januchs MG, Quintanilla-Dominguez J, Vega-Corona A, Andina D.
 586 Development of a model for forecasting of PM₁₀ concentrations in Salamanca,
 587 Mexico. *Atmospheric Pollution Research* 2015; Volume 6 (4).
 588

589 de Hoogh K, Gulliver J, Donkelaar AV, Martin RV, Marshall JD, Bechle MJ, Cesaroni G,
 590 Pradas MC, Dedele A, Eeftens M, Forsberg B, Galassi C, Heinrich J, Hoffmann B,
 591 Jacquemin B, Katsouyanni K, Korek M, Künzli N, Lindley SJ, Lepeule J, Meleux F, de
 592 Nazelle A, Nieuwenhuijsen M, Nystad W, Raaschou-Nielsen O, Peters A, Peuch VH,
 593 Rouil L, Udvardy O, Slama R, Stempfelet M, Stephanou EG, Tsai MY, Yli-Tuomi T,
 594 Weinmayr G, Brunekreef B, Vienneau D, Hoek G. Development of West-European
 595 PM_{2.5} and NO₂ land use regression models incorporating satellite-derived and
 596 chemical transport modelling data. *Environ Res.* 2016 Nov;151:1-10. doi:
 597 10.1016/j.envres.2016.07.005.

598 Di Q, Rowland S, Koutrakis P, Schwartz J. A hybrid model for spatially and temporally
 599 resolved ozone exposures in the continental United States. *J Air Waste Manag Assoc.*
 600 2017 Jan;67(1):39-52. doi: 10.1080/10962247.2016.1200159.

601 Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, Sabath MB, Choirat C, Koutrakis P,
 602 Lyapustin A, Wang Y, Mickley LJ, Schwartz J. An ensemble-based model of PM_{2.5}
 603 concentration across the contiguous United States with high spatiotemporal
 604 resolution. *Environ Int* 2019; 130: 104909. doi: 10.106/j.envint.2019.104909

605 Gocheva-Ilieva SG, Stoimenova MP. PM₁₀ Prediction and Forecasting Using CART: A
 606 Case Study for Pleven, Bulgaria. *International Journal of Environmental and*
 607 *Ecological Engineering* 2018; 12(9).
 608

609 Green DC, Fuller GW, Baker T. Development and validation of the volatile correction
 610 model for PM₁₀ – An empirical method for adjusting TEOM measurements for their
 611 loss of volatile particulate matter. *Atmospheric Environment* 2009; 43(12);2132-
 612 2141. <https://doi.org/10.1016/j.atmosenv.2009.01.024>
 613

614 Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning:*
 615 *data mining, inference, and prediction*. 2nd ed. New York: Springer.

616 Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, Liu Y. Estimating PM_{2.5}
 617 Concentrations in the Conterminous United States Using the Random Forest
 618 Approach. *Environ. Sci. Technol.* 2017; 51(12);6936-6944. Publication Date: May 23,
 619 2017. <https://doi.org/10.1021/acs.est.7b01210>

620 Lee M, Kloog I, Chudnovsky A, Lyapustin A, Wang Y, Melly S, Coull B, Koutrakis P,
 621 Schwartz J. Spatiotemporal prediction of fine particulate matter using high-
 622 resolution satellite images in the Southeastern US 2003-2011. *J Expo Sci Environ*
 623 *Epidemiol.* 2016 Jun;26(4):377-84. doi: 10.1038/jes.2015.41.

624 LONDON AIR QUALITY NETWORK (LAQN)
 625 <http://www.londonair.org.uk/LondonAir/Default.aspx>

626 Menze, B.H., Kelm, B.M., Masuch, R. et al. A comparison of random forest and its
 627 Gini importance with standard chemometric methods for the feature selection and
 628 classification of spectral data. *BMC Bioinformatics* 2009; 10: 213.
 629 <https://doi.org/10.1186/1471-2105-10-213>

630 Mishra D, Goyal P. Development of artificial intelligence based NO₂ forecasting
 631 models at Taj Mahal, Agra. *Atmospheric Pollution Research* 2015; 6(1): 99-106.

632 Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in
 633 prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2010;32, 569–575.

634 Russo A, Lind PG, Raische IF, Trigo R, Mendes M. Neural network forecast of daily
 635 pollution concentration using optimal meteorological data at synoptic and local
 636 scales. *Atmospheric Pollution Research* 2015;6(3)
 637

638 WHO, 2013. Review of evidence on health aspects of air pollution – REVIHAAP
 639 Project, Technical Report, World Health Organization, Regional Office for Europe.
 640

641 WHO, 2013b. Health risks of air pollution in Europe – HRAPIE project.
 642 Recommendations for concentration–response functions for cost–benefit analysis of
 643 particulate matter, ozone and nitrogen dioxide. World Health Organization, Regional
 644 Office for Europe.
 645

646 Xiao Q, Chang HH, Geng G, Liu Y. An Ensemble machine learning model to predict
 647 historical PM_{2.5} concentrations in China from satellite data. *Environ Sci Technol* 2018;
 648 52: 13260-13269.

649 Xue W, Zhang J, Zhong C, Ji D, Huang W. Satellite-derived spatiotemporal PM_{2.5}
 650 concentrations and variations from 2006 to 2017 in China. *Sci Total Environ.* 2019
 651 Nov; 20:134577. doi: 10.1016/j.scitotenv.2019.134577.

652

Prediction of PM_{2.5} concentrations at the locations of monitoring sites measuring PM₁₀ and NO_x using generalized additive models and machine learning methods: a case study in London

Antonis Analitis¹, Ben Barratt^{2,3}, David Green², Andrew Beddows², Evangelia Samoli¹, Joel Schwartz⁴, Klea Katsouyanni^{1,2}

1. National and Kapodistrian University of Athens, Faculty of Medicine, Department of Hygiene, Epidemiology and Medical Statistics, Athens, Greece

2. MRC Centre for Environment and Health, King's College London, UK

3. HPRU in Environmental Exposures and Health, Imperial College London, UK

4. Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

Address correspondence to
Klea Katsouyanni
Professor of Public Health
King's College London
Franklin-Wilkins Building
Room 4.129
150 Stamford street
London SE1 9NH
e-mail: kkatsouy@med.uoa.gr
telephone: +44-207-8483844
mobile tel: +30-6944681057

Abstract

The adverse health effects of air pollutants, especially those of PM_{2.5}, are well documented. However, a lack of adequate monitoring and weaknesses in modelling approaches do not allow a good assessment of health effects in many areas of the World. Advances in computational methods and the availability of new data sets, e.g. satellite remote observations, have enlarged the possibilities of modelling for application in large scale health effects studies. However, PM_{2.5} monitoring is very recent in most of the World and more limited compared to other pollutants, and understanding how to use PM₁₀ monitors to estimate PM_{2.5} exposure is therefore important. Since interest in these methods is relatively recent, there is a need for testing their performance against ambient measurements, but long term PM_{2.5} datasets are less readily available than PM₁₀ in many regions. In the present study we report the methodology and results of using regression modelling and a machine learning method (Random Forest-RF), as well as a combination of the two, to enhance a PM_{2.5} measurement data base in London using PM₁₀ and NO_x measurements as well as other predictors and compare the relative performance of each method.

We find that the combination of predictions by the regression model and the RF performs best and we obtain a cross-validation R² of 99.29% and 98.22% for the 5-year periods 2004-2008 and 2009-2013 respectively and a Mean Square Error near 1. Our enhanced data base for PM_{2.5} is available for use by other researchers.

Key words: PM_{2.5} prediction; environmental exposure; random forest; ensemble methods; London case study

1. Introduction

The adverse health effects of air pollutants, especially those of PM_{2.5}, are well documented (WHO, 2013a, 2013b). Most of the studies reported are from areas of the world where monitoring of pollutants concentrations is routinely performed. In several studies modelling has been used to assess the health effects of exposure to various pollutants in order to enhance the spatial scale (de Hoogh et al 2016; Christidis et al 2019) or to estimate long-term pollutants' levels in the absence of monitoring (Brauer et al 2012). Advances in computational methods and the availability of various data, e.g. satellite remote observations, have led to applications involving spatio-temporal models i.e. predicting daily pollutant concentrations for specific areas at fine spatial scales for a wide range of usage. One objective for developing spatio-temporal prediction models is short term forecasting, as was done by Cortina-Januchs et al (2015) in Salamanca, Mexico and Gocheva-Ilieva and Stoimenova (2018) for the city of Pleven, Bulgaria. Another target is to predict daily concentrations in areas without measurements for use in health studies, as was done in the U.S.A., where Lee et al (2016) predicted PM_{2.5} daily concentrations at 1 km² grid resolution in southeastern US and Di et al (2017) predicted spatiotemporal ozone concentrations in the continental US. Xue et al. (2019) developed spatiotemporal models for PM_{2.5} for China and provided these online. Comparisons of the performance of computationally intensive methods versus more traditional ones have also been reported. Thus, for example, Russo et al. (2015) used stepwise regression procedures and neural networks (NN) based on routine pollutants measurements in Lisbon, Portugal, to develop a prediction model for PM₁₀ concentrations and found no clear superiority of the NN. Mishra and Goyal (2015), based on a relatively short measurement campaign in Agra, India, developed models based on multiple regression, principal components analysis and NN to predict Nitrogen Dioxide (NO₂) levels at the famous Taj Mahal monument, where the combination of principal components and NN performed best. Chen et al (2019) compared the performance of 16 algorithms for predicting the spatial variability of long-term averages of PM_{2.5} and NO₂ across Europe using the AIRBASE data base and found that both the PM_{2.5} and NO₂ models performed similarly across algorithms. It

is evident that models perform better when the area of prediction is relatively homogeneous and when information on relevant predictors is available.

Within the London based, MRC funded, project entitled "Comparative evaluation of Spatio-Temporal Exposure Assessment Methods for estimating the health effects of air pollution" (STEAM), we formed a data base of air pollutants (specifically 24-hour average PM_{10} , $PM_{2.5}$, and NO_x and daily 8-hour maximum O_3) concentrations, including all measurements from sites within the London M25 orbital motorway. The monitoring network in London is exceptionally dense providing a unique opportunity to study the effects of pollutants in a mega-city, observing temporal and spatial differentiations. However, the number of sites measuring $PM_{2.5}$ is significantly smaller than the corresponding for PM_{10} and NO_2 . The objective of the study reported in this paper was to enhance the $PM_{2.5}$ data base for London using other pollutants, meteorological and land-use variables to predict daily $PM_{2.5}$ concentrations at the location of fixed monitors measuring PM_{10} but not $PM_{2.5}$. In our analysis we used regression models and machine learning methods. As $PM_{2.5}$ is the pollutant for which most of the evidence exists for adverse health effects, a larger data base of measurements can be extremely useful to aid the verification of subsequent spatio-temporal exposure models and to be used in health studies conducted in London and elsewhere.

2. Data and Methods

2.1 Data

We used all measurements performed at fixed sites by the London Air Quality Network (LAQN, <http://www.londonair.org.uk/LondonAir/Default.aspx>), during the years 2004-2013 ([Figure 1](#)). Not all sites provided measurements for the whole time period and there were missing values during periods of operation.

Specifically, for PM_{10} we compiled daily mean data from 108 sites in 2004-2008 and 115 sites in 2009-2013 (136,874 and 134,230 measurements respectively, range of the number of measurements per site: 50-1,812 and 11-1,808 respectively); for NO_x we compiled data from 1927 sites in 2004-2008 and ~~216~~ 130 in 2009-2013 (~~163,778~~

252,101 and 163,088 259,859 measurements respectively, range of the number of measurements per site: 2438-1,8093 and 119-1,817 respectively), whilst for PM_{2.5} data were available only from 24 sites in 2004-2008 and 33 sites in 2009-2013 (16,957 and 40,083 measurements respectively, range of the number of measurements per site 13-1,382 and 95-1,809 respectively). The increased number of PM_{2.5} measurements in later years and the possibility that air pollution determinants may have changed over time, led to the decision to limit the time span of model development to 5 years, thus the model was developed for 2004-2008 and 2009 -2013 separately.

Figure 1. Map of the geographical location of the fixed air pollution monitoring network operated at the Greater London study area

Formatted: Font: +Body, Bold

We checked whether all fixed sites that measured PM₁₀ and NO_x concurrently reflected the same association between the two pollutants in order to identify sites which may have unique characteristics affecting their association, e.g. be near a local, non-combustion source, such as industry, long-term construction or waste handling. For this purpose, mean NO_x vs mean PM₁₀ concentrations from all sites which provided both measurements were plotted and a regression line was estimated. In order to identify and exclude outlier sites, Cook's distance was calculated and a Cook's D value >1 was defined as indicating an influential observation (Cook & Weisberg 1982). Cook's D, by removing the *i*th data point from the model and re-estimating the regression, summarizes how the values in the regression model change when the *i*th observation is removed. After applying the method, no fixed site was excluded from the analysis, since no outlier was identified.

Additionally, the different methods of measuring PM_{2.5} and their comparability were taken into account and a correction was adopted. Whilst PM_{2.5} measured by the Filter Dynamics Measurement System (FDMS) instruments (20 sites) were equivalent to the reference method, the PM_{2.5} data obtained with the Tapered Element Oscillating Microbalance (TEOM) instruments (15 sites) had widely acknowledged discrepancies due to the loss of volatile particulate (Allen et al., 1997; Charron et al., 2004). Previously, for PM₁₀ a Volatile Correction Model (VCM) method was developed (Green et al. 2009), which used regional measurements of the volatile particulate made by the FDMS to correct the TEOM measurements; this was optimized for PM_{2.5} by Barratt et al. (2009).

Specifically, the correction applied was:

$$TEOM_{VCM}PM_{2.5} = \left(\frac{T_{d0} - 3}{1.03} \times \frac{P_a}{P_r} \times \frac{T_r}{T_a} \right) - \left((1 + fVCM) \frac{1}{n} \sum_i^n z_i \right)$$

Where:

T_{d0} is the daily mean TEOM PM_{2.5} concentration

P_r is the set reporting pressure of the TEOM

171 Pa is the regional daily mean ambient pressure
 172 Ta is the regional daily mean ambient temperature
 173 Tr is the set reporting temperature of the TEOM
 174 fVCM is the VCM function value (optimised to 0.72 for PM_{2.5} by Barratt et al., 2009)
 175 n is the number of FDMS volatile measurement sites used in that day
 176 z_i is the FDMS daily mean volatile measurement for station i
 177 Data for meteorological variables, specifically for mean daily temperature (°C),
 178 relative humidity (%), wind speed (ms⁻¹) and wind direction (°N) were compiled from
 179 all available sites with <25% missing values over the study period (n=7).
 180 Meteorological data were obtained from the London Air Quality Network database
 181 (www.londonair.org.uk). Because the number of available measurement sites was
 182 not so dense to allow characterisation at the location of all pollution measurement
 183 sites, but the correlation of variables between sites is very high (r between sites
 184 >0.99 for temperature and >0.85 for the other meteorological variables) for all
 185 meteorological variables, we averaged per day across all meteorological sites and
 186 used one time series for each meteorological variable.

187 **2.2 Development of predictive models**

188 For the development of the models we included all sites which measured PM_{2.5},
 189 PM₁₀ and NOx concurrently (the number of sites was 18 for 2004-2008 and 25 for
 190 2009-2013) and developed the model using the 2004-2008 and 2009-2013 data
 191 separately, as described above.. We used the PM_{2.5} measurements taken by the
 192 reference method as well as the corrected TEOM measurements. As predictor
 193 variables we included PM₁₀ and NOx as well as other covariates, such as
 194 meteorological, geographical and temporal variables. The model developed was
 195 used to predict PM_{2.5} at fixed sites with no PM_{2.5} measurements (but which
 196 measured PM₁₀ & NOx) and thus provide a greatly enhanced data base of PM_{2.5}
 197 estimated concentrations. We used two methods, a Regression model and a Random
 198 Forest, as well as a combination method of the two.
 199

2.2.1 Regression Model

The model was developed with daily values of measured $PM_{2.5}$ as the dependent variable and daily values of the following covariates (all time-varying variables are for lag0, i.e. same day): PM_{10} , measured at the same site as $PM_{2.5}$ (24-hour values, in $\mu g/m^3$); NO_x , measured from the same site as $PM_{2.5}$ (24-hour average, in $\mu g/m^3$); type of monitor (roadside/kerbside, background); Day of the week; Time trend and monthly variations; Mean daily temperature averaged over all available sites; Mean daily relative humidity averaged over all available sites; Mean daily wind speed averaged over all available sites; Mean daily wind direction; Coordinates, longitude and latitude. The above covariates were chosen based on previously reported results (Di et al 2019) and, specifically for the pollutants used as predictors, prior knowledge on emissions and dispersion patterns of each pollutant.

As sensitivity analyses, we tested, smoothed terms for both PM_{10} and NO_x and smoothed terms for all variables with degrees of freedom (df) chosen by the Generalized Cross Validation (GCV) criterion, alternatively. We also tested the interaction between PM_{10} and site classification as well as a bivariate smooth function of wind speed and PM_{10} . Finally, we added a bivariate smooth function for coordinates (latitude and longitude) and its interaction with PM_{10} , to capture any remaining spatial variability of the fitted vs the observed values. In the models with latitude and longitude the variable "monitor type" was excluded.

To obtain the R^2 and Mean Square Error (MSE) of the models we used 10-fold cross-validation (CV) in order to test the model's ability to predict new data that was not used in estimating it. Specifically, we fitted the model based on a training data set (90% of the data) and using 10% of the observations as validation data. We then predicted the omitted observations (the validation data set) and calculated the R^2 and MSE. This was done sequentially 10 times and then the predicted values were used to obtain the R^2 and MSE (Hastie et al 2009; Rodriguez et al 2010).

2.2.2 Random Forest (RF)

We further applied a RF as an alternative method to improve the fit of the prediction for the $PM_{2.5}$ data. RF is a machine learning algorithm, defined as a “combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” (Breiman, 2001). Each tree is a form of nonlinear regression model where samples are partitioned at each node of a binary tree based on the values of randomly selected input variables (features). A RF can be used for both continuous (regression) and categorical outcomes (classification), combining multiple decision trees in determining the final output, that is, the mean prediction of the individual trees in regression or the mode of the classes in classification tasks. Due to its properties, the RF algorithm gives high accuracy prediction and has advantages over other machine learning techniques in terms of ability to handle highly non-linear data, robustness to noise and tuning simplicity (Caruana and Niculescu-Mizil, 2006; Menze et al, 2009). High accuracy in predictions though, comes at the expense of interpretability. The covariate interpretability is not so straightforward, as it is in terms of regression coefficients, and is performed in terms of "variable importance". In the variable importance graphs the predictors are ranked in terms of increase in MSE and in node purity. Additionally, the partial dependence plots show the mean marginal influence of each predictor on the $PM_{2.5}$ estimation.

We used the “randomForest” library in R, version 3.5.1. RF was also applied separately for each time-period of interest, i.e. 2004-2008 and 2009-2013. We used the default parameter values for the number of trees (500) and tree depth (max), and 3 and 4 variables were tested in each step based on the rule “number of variables/3”. We used the same variables as those tested in the regression model (A) described above.

2.2.3 Combination of predictions from the Regression and the Random Forest

We combined the predictions from the regression and the RF method to improve the fit of the model, addressing the situation where the two methods may perform differently at different levels of the PM concentrations. The predictions were used as

Field Code Changed

independent variables in a Generalized Additive Model (GAM) for each period separately, as described below:

$$PM_{2.5} = s(\text{predictions from the Regression model}) + s(\text{predictions from the RF method})$$

where s is a penalized spline with the number of df chosen by GCV.

Finally we chose the best model based on the CV adjusted R^2 between the Regression, the RF and their GAM combination and used it to 1) fill in missing $PM_{2.5}$ values from stations providing $PM_{2.5}$ measurements both during 2009-2013 and 2004-2008 and 2) provide $PM_{2.5}$ series in stations with only PM_{10} and NO_x measurements.

All analyses were performed using the R statistical package v. 3.5.1 and libraries "mgcv" and "randomForest".

This study is based on publicly available data and therefore no ethics issues are involved.

3. Results

3.1 Predictive model development

Time period 2004-2008

Regression model

Table 1 shows the contribution of each variable to the value of the adjusted R^2 of the regression model, for 2004 - 2008. All variables entered into the model after PM_{10} increased the value of the adjusted R^2 slightly. The best model in terms of adjusted R^2 was Model 14:

$$PM_{2.5} = \text{intercept} + PM_{10} + NO_x + \text{day of the week (6 dummy variables for 7 days, with "Sunday" used as reference category)} + ns(\text{trend, df}=12 \text{ per year}) + ns(\text{Temperature, 4$$

294 $df) + ns(rel. humidity, 2 df) + ns(wind speed, 2 df) + ns(wind direction, 3 df) + s(latitude,$
 295 $longitude) + PM_{10} * s(latitude, longitude)$ where:
 296 ns: natural cubic spline
 297 s: bivariate smoothing spline
 298 **PM_{2.5}**: measured PM_{2.5} (24-hour values, $\mu g/m^3$)
 299 **PM₁₀**: measured PM₁₀ from the same site as PM_{2.5} (24-hour values, $\mu g/m^3$)
 300 **NO_x**: measured NO_x from the same site as PM_{2.5} (24-hour average, $\mu g/m^3$),
 301 **Day of the week**: 0=Sunday, 1=Monday,...,6=Saturday (six dummy variables for 7
 302 days; Sunday: reference category)
 303 **Time trend**: a variable denoting the length of the period with values 1,...,N entered
 304 as ns with 12 df/year
 305 **Temperature**: measured in °C and entered as ns with 4 df
 306 **Humidity**: relative humidity % entered as ns with 2df
 307 **Wind speed**: ms^{-1} entered as ns with 2 df
 308 **Wind direction**: °N, entered as ns with 4 df
 309 **s(latitude, longitude)**: A bivariate smooth function for the coordinates (longitude,
 310 latitude) of the fixed monitoring sites and
 311 **PM₁₀ * s(latitude, longitude)**: interaction term
 312 The 10-fold CV adjusted R^2 of the model for 2004 – 2008 was 92.10% and the MSE
 313 $10.85 (\mu g/m^3)^2$.

Model number	Model covariates	adj R^2
1	PM ₁₀ ($\mu g/m^3$)	0.8359
2	Model 1 +NO _x ($\mu g/m^3$)	0.8440
3	Model 2 + monitor type (roadside/background)	0.8453
4	Model 3 + week day (six dummy variables for 7 days; Sunday: reference category)	0.8476
5	Model 4 +ns(time trend, 12 df/year)	0.8697
6	Model 5 +ns(temperature °C, 4 df)	0.8722
7	Model 6 +ns(relative humidity %, 2 df)	0.8906
8	Model 7 +ns(wind speed ms^{-1} , 2 df)	0.8999
9	Model 8 +ns(wind direction °N, 3 df)	0.9000
10	Penalized splines for PM ₁₀ & NO _x (excluding the corresponding linear terms) with df from GCV + all	0.9022

	other covariates from model 9	
11	Penalized splines (instead of natural splines with fixed df) for temperature, humidity, wind speed, wind direction with df from GCV + all other covariates from model 10	0.9037
12	Model 11 + Interaction of PM ₁₀ and monitor type	0.9034
13	Model 11 + Bivar. spline s(PM ₁₀ ,wspeed)	0.9059
14	Model 9 + Main effects s(lat, long) + inter. PM ₁₀ x s(lat,long) excluding monitor type	0.9232

* ns(): natural splines, s(): bivariate smoothing spline, df: degrees of freedom

Table 1. Development of the Regression model: adjusted R² for every set of covariates 2004-2008

Random Forest

The Random Forest R² and MSE for 2004-2008 were 95.03% and 6.88 respectively. Figure 4-2 shows the relevant partial dependence plot which gives a graphical depiction of the marginal effect of each predictor variable on PM_{2.5}. It can be seen that the associations of the predictors and PM_{2.5} concentrations are described by non-linear functions. In Figure 2-3 we can see the variable importance of each predictor in terms of both decrease in MSE and remaining error in predictive accuracy after a node split (node impurity).

329

330

331 | **Figure 42.** Partial dependence plot of PM_{2.5} on predictor variables used in the
 332 Random Forest, 2004-2008

333

Figure 23. Variable importance plot in the Random Forest (2004-2008) in terms of % increase in Mean Square Error (%IncMSE) and increase in node purity (IncNodePurity).

Time period 2009-2013

Regression model

Table 2 shows the contribution of each variable to the value of the adjusted R^2 of the regression model, as well as the R^2 of all alternative models applied.

A similar pattern for the contribution to R^2 with the period 2004-2008 was observed. The same model, Model 14, was again the model with the best R^2 equal to 88.40%.

The CV adjusted R^2 of the model was 88.15% and the mean square error (MSE) 12.53 $(\mu\text{g}/\text{m}^3)^2$.

Model #	Model covariates	adj R^2
1	PM ₁₀ ($\mu\text{g}/\text{m}^3$)	0.7921
2	Model 1 +NO _x ($\mu\text{g}/\text{m}^3$)	0.7969
3	Model 2 + monitor type (roadside/background)	0.8003
4	Model 3 +Week day (six dummy variables for 7 days; Sunday: reference category)	0.8010
5	Model 4 +ns(time trend, 12 df/year)	0.8169
6	Model 5 +ns(temperature °C, 4 df)	0.8215
7	Model 6 +ns(relative humidity %,2 df)	0.8400
8	Model 7 +ns(wind speed ms^{-1} , 2 df)	0.8479
9	Model 8 +ns(wind direction °N, 3 df)	0.8486
10	Penalized splines for PM ₁₀ & NO _x (excluding the corresponding linear terms) with df from GCV+ all other covariates from model 9	0.8627
11	Penalized splines (instead of natural splines with fixed df)for temperature, humidity, wind speed, wind directionwith df from GCV+ all other covariates from model 10	0.8663
12	Model 11 + Interaction of PM ₁₀ and monitor type	0.8577
13	Model 11 + Bivar. spline s(PM ₁₀ ,wspeed)	0.8692
14	Model 9 + Main effects s(lat, long) + inter. PM ₁₀ x s(lat,long)+ all other covariates from model 9 excluding monitor type	0.8840

* ns(): natural spline, s(): bivariate smoothing spline, df: degrees of freedom

Table 2. Development of the Regression model: adjusted R^2 for every set of covariates 2009-2013

359

360 **Random Forest**

361 The Random Forest R^2 and MSE for the 2009-2013 period were 92.8% and 7.64.

362 | Figure [3-4](#) shows the relevant partial dependence plot which gives a graphical
363 depiction of the marginal effect of each predictor variable on $PM_{2.5}$. Although the
364 general pattern shown is consistent between the two 5-year periods, some of the
365 associations of meteorological and spatial variables with $PM_{2.5}$ differ somewhat. This
366 may be attributed to different meteorological conditions across the two periods.

367 | Figure [4-5](#) presents the variable importance of each predictor in terms of both
368 decrease in MSE and remaining error in predictive accuracy after a node split (node
369 impurity). In both the 5-year time periods the most important predictor is PM_{10} . For
370 the other predictors, in terms of increase in MSE the ranking between variables is
371 slightly different but in terms of increase in node purity it is identical for the two
372 periods.

373

Figure 34. Partial dependence plot of $PM_{2.5}$ on predictor variables used in the Random Forest (2009-2013)

Figure 45. Variable importance plot in the Random Forest (2009-2013) in terms of % increase in Mean Square Error (%IncMSE) and increase in node purity (IncNodePurity).

Time periods 2004-2008 and 2009-2013

Combination of predictions obtained from the linear regression model & the predictions obtained from the Random Forest method

The 10-fold CV adjusted R^2 of the combined GAM model was 99.29% and 98.99% for the time periods 2004-2008 & 2009-2013 while MSE was 0.98 and 1.07, respectively. We also calculated the spatial CV, following the method applied in Hu et al (2017), by excluding 10% of the measurement sites from the training data set and predicting them from the models including the other 90% of sites and repeating this procedure until all predictions were done. The corresponding R^2 and MSE for the two periods

were 99.28%, 1.00 and 98.94%, 1.13. Thus, there is no substantial reduction in R^2 and MSE when the spatial predictions are assessed.

The predictions from the combined model were included in the final enhanced $PM_{2.5}$ database.

3.2 Predicted values for $PM_{2.5}$

Based on the described methodology we predicted 102,800 and 85,436 daily concentrations in specific sites for $PM_{2.5}$ in 2004-2008 and 2009-2013, respectively.

Table 3 presents the descriptive statistics for measured and predicted values of $PM_{2.5}$ overall, as well as predicted values separately for sites with measured $PM_{2.5}$ (directly comparable with measured values) and for sites without $PM_{2.5}$ measurements (which provided only PM_{10} and NO_x) for 2004-2008 and 2009-2013. During 2004-2008 there were 16,957 $PM_{2.5}$ measurements and 102,800 additional estimates are made available whilst for 2009-2013 the corresponding numbers are 40,083 and 85,436. Measured and predicted means and medians are identical to the first decimal for 2009-2013. It is also noted that concerning the prediction of concentrations in the relatively higher range of values, those >15 , >25 and $>50 \mu g/m^3$, the percentages of measured and predicted values above those thresholds are remarkably similar (Table 3).

Figure 5-6 shows the agreement between the measured $PM_{2.5}$ and the predicted from each method and their combination. The predictions from the combined method have a better agreement with measured values, and adequate at both the low and high end of the exposure range. The intercepts are very near 0 and the slopes very close to 1 in both time periods.

422

	25 th percentile	Median	Mean	75 th percentile	% values >15 µg/m ³	% values >25 µg/m ³	% values >50 µg/m ³	N
2004 – 2008								
A. PM_{2.5} measurements	9.29	13.99	17.45	22.17	45.9	19.6	2.1	16957
B. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were available	9.49	14.45	17.81	22.57	48.0	20.7	2.1	11695
C. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were NOT available but there were PM₁₀ and NOx	8.60	12.77	15.99	19.87	40.1	16.4	1.7	102800
D. all predicted PM_{2.5} concentrations (sum of B & C)	8.68	12.93	16.18	20.2	40.9	26.8	1.8	114495
2009 – 2013								
A. PM_{2.5} measurements	9.33	13.35	16.39	20.10	42.1	15.4	1.6	40083
B. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were available	9.46	13.37	16.35	19.95	41.9	15.0	1.6	26645
C. predicted PM_{2.5} concentrations for days and sites where PM_{2.5} measurements were NOT available but there were PM₁₀ and NOx	9.98	13.46	15.97	19.11	41.4	12.5	1.0	85436
D. all predicted PM_{2.5} concentrations (sum of B & C)	9.87	13.44	16.06	19.30	41.5	13.1	1.1	112081

423

424 **Table 3.** Descriptive statistics for the observed & predicted PM_{2.5} (µg/m³) concentrations applying the combination of both methods model for the
425 time periods 2004 - 2008 and 2009 - 2013.

426

Figure 56. Graph of observed vs predicted $PM_{2.5}$ for each period, by method (Regression, Random Forest and GAM combination).

4 Discussion

We predicted a large number of daily PM_{2.5} concentrations ($\mu\text{g}/\text{m}^3$) in London at the location of monitoring sites from measured PM₁₀ and NO_x and several other covariates. Our predictions were done to fill in missing PM_{2.5} values in monitors that normally measure PM_{2.5} concentrations and to predict PM_{2.5} in monitors which do not measure PM_{2.5} but do measure PM₁₀ and NO_x. We developed a Regression and a RF model and also used a GAM to ensemble average the Regression and RF predictions. All models performed well in terms of adjusted R², but the combined model improved the fit and had an adjusted R² of 99.29% for 2004-2008 and 98.22% for 2009-2013. Additionally, the combined model had a much smaller MSE, 0.98 for 2004-2008 and 1.07 for 2009-2013. Our model predicted better than using PM₁₀ alone as predictor: about 10% of the variability of PM_{2.5} is explained by adding the other variables in the regression model and most of the other variables are important in the RF as well. The enhanced data base is intended for use in health studies, either as such or as input for use in exposure assessment models.

London is unique globally for having a very large density of fixed site monitors measuring pollutants routinely in a relatively limited area. For example in London from 2009 to 2013 there are data available for NO_x from 216 sites within the M25 orbital motorway, an area of approximately 1100 square miles, whilst the AIRBASE data set provides NO₂ measurements from 2399 sites for the whole of Europe, an area larger by 3.5 thousand times (Chen et al 2019). Thus, it is possible to use the London data to study the spatiotemporal variation of air pollutants concentrations and the consequences of measurement error in assessing human exposure, since the large data set allows for testing various situations. However the data in London are less abundant for PM_{2.5}, which is measured in a much smaller number of sites, and thus enhancing the PM_{2.5} data base can be very useful for methodological and health studies, especially since the reported predictions for PM_{2.5} daily concentrations in this study are excellent.

Several studies have made comparisons between regression models and machine learning techniques for predicting pollutant concentrations either for detecting

spatial or spatio-temporal variability. Chen et al (2019) compared the performance of 16 algorithms for predicting the spatial variability of long-term averages of PM_{2.5} and NO₂ across Europe using the AIRBASE data base. They report that the PM_{2.5} and NO₂ models performed similarly across algorithms with an average CV R² of 0.53. However, the RF among the machine learning techniques performed somewhat better (CV R² 0.63), whilst the artificial neural network performed less well. Several publications report spatio-temporal predictions, more comparable to the present study. Generally, they are using measurements from smaller density networks and often develop complex models, possibly using machine learning methods and/or incorporating AOD satellite data. To focus on PM_{2.5}, Lee et al (2016) used a regression based approach to estimate PM_{2.5} at a 1km² resolution in 7 Southeastern States in the US and obtained a 10-fold CV R² between 0.70 and 0.81 and RMSE between 2 and 3 µg/m³. Xue et al (2019) developed a multi-dimensional model to predict PM_{2.5} across China, with an objective similar in some way to ours, namely to provide a data base for use in health studies. Typically, they report R²<0.80. Brokamp et al (2018) use a RF to predict PM_{2.5} in the Cincinnati, Ohio metropolitan area from 2000 to 2015, using several available variables including AOD and report a CV R² of 0.91 with a RMSE equal to 2.2 µg/m³. Hu et al (2017) develop a RF approach to predict PM_{2.5} in the 48 conterminous US, an area of 3.12 million square miles, and report a CV R² of 0.80.

Beloconi et al (2016) developed spatiotemporal models for PM₁₀ and PM_{2.5} for London using satellite AOD data, measurements, meteorological and land use variables. Their purpose was to estimate concentrations at 1km² grid using sophisticated modelling to obtain the best possible predictability. However, the fact that remote sensing data have missing values and measurements of PM_{2.5} are less dense compared to PM₁₀, resulted in a maximum R² of 90% for PM_{2.5} models. The objective of our study is not directly comparable to that of Beloconi et al: we aimed at predicting the best possible spatiotemporal estimate for PM_{2.5} concentrations at the geographical locations of the monitoring sites in London. Since we have excellent predictors for the site locations (such as PM₁₀ measurements), our models perform almost perfectly and further predictors would not add to the model performance.

Our combination model performs better than those reported in the literature mainly because of the availability of much more dense measurements on the predictors and the fact that we had a closely related proxy measurement (PM_{10}) at the prediction points. In some published studies, various machine learning methods predicting $PM_{2.5}$ have been combined (Di et al 2019; Xiao et al 2018) using a GAM function, and the combination (ensemble) models also performed best.

The enhanced $PM_{2.5}$ data base for London can be used in other health studies and provide improved input for estimating concentrations in points of interest and for modelling. Our method demonstrates that the combination of more traditional regression and machine learning methods may improve the predictions of pollutants concentrations for various applications.

Our enhanced $PM_{2.5}$ data base for the year 2004-2013 is available on request from the corresponding author.

Conclusions: We present a methodology to enhance the $PM_{2.5}$ database of measurements for use in health studies and apply it in the case of London. The GAM combination (ensemble) of the two separate model predictions performs best, demonstrating that the combination of more traditional regression and machine learning methods leads to improved predictions throughout the range of concentrations. A larger data base of measurements can be extremely useful to aid the verification of subsequent spatio-temporal exposure models and to be used in health studies conducted in London and elsewhere.

Acknowledgements: We are grateful to the UK Government and Local Authorities providing air pollution measurements used in this study, managed by the Environmental Research Group at King's College London and Ricardo Energy and Environment.

Funding: Research in this paper as part of the STEAM project was funded under the MRC UK Grant ref.: MR/N014464/1. The Funding Source (MRC UK) had no role in the

527 study design, the collection, analysis or interpretation of data, in writing the paper
528 and in the decision to submit the article for publication.

529

530 **Author contributions:** AA: Formal Analysis, Data Curation, Methodology, Writing -
531 Original Draft, Review & Editing; BB: Data Curation, Methodology, Writing - Review &
532 Editing; DG: Methodology, Writing - Review & Editing; AB: Methodology, Writing -
533 Review & Editing; ES: Conceptualization, Methodology, Writing - Review & Editing;
534 JS: Conceptualization, Methodology, Writing - Review & Editing; KK:
535 Conceptualization, Funding acquisition, Methodology, Writing - Original Draft,
536 Review & Editing

537

538 **Conflicts of Interest:** The authors declare that they have no conflict of interest

539

540

541

542 REFERENCES

- Allen G, Sioutas C, Koutrakis P, Reiss R, LurmannFWand, Roberts PT. Evaluation of the TEOM® Method for Measurement of Ambient Particulate Mass in Urban Areas, Journal of the Air & Waste Management Association 1997; 47(6); 682-689. DOI: 10.1080/10473289.1997.10463923
- Beloconi A, Kamarianakis Y, Chrysoulakis N. Estimating urban PM₁₀ and PM_{2.5} concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data. Remote Sensing of Environment 2016; 172: 148-164
- Barratt B, Carslaw D, Fuller G, Green D, Tremper A. Analysis of Air Quality Data – Low Emission Zone Year 1 Results. Prepared for Transport for London by King's College London Environmental Research Group & Institute for Transport Studies, University of Leeds, May 2009
- Brauer M, Amann M, Burnett RT, Cohen A, Dentener F, Ezzati M, Henderson SB, Krzyzanowski M, Martin RV, Van Dingenen R, van Donkelaar A, Thurston GD. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. Environ Sci Technol. 2012 Jan; 17;46(2):652-60. doi: 10.1021/es2025752. Epub 2012 Jan 6.
- Breiman, L. Random Forests. Machine Learning 2001; 45: 5-32. doi.org/10.1023/A:1010933404324
- Brokamp C, Jandarov R, Hossain M, Ryan P. Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model. Environ Sci Technol. 2018 Apr; 3;52(7):4173-4179. doi: 10.1021/acs.est.7b05381.
- Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, Bauwelinck M, van Donkelaar A, Hvidtfeldt UA, Katsouyanni K, Janssen NAH, Martin RV, Samoli E, Schwartz PE, Stafoggia M, Bellander T, Strak M, Wolf K, Vienneau D, Vermeulen R, Brunekreef B, Hoek G. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ Int. 2019 Sep;130:104934. doi: 10.1016/j.envint.2019.104934. Epub 2019 Jun 20
- Christidis T, Erickson AC, Pappin AJ, Crouse DL, Pinault LL, Weichenthal SA, Brook JR, van Donkelaar A, Hystad P, Martin RV, Tjepkema M, Burnett RT, Brauer M. Low concentrations of fine particle air pollution and mortality in the Canadian Community Health Survey cohort. Environ Health. 2019; 18: 84. Published online 2019 Oct 10. doi: 10.1186/s12940-019-0518-y
- Caruana R, Niculescu-Mizil A, 2006; An empirical comparison of supervised learning algorithms. ICML '06: Proceedings of the 23rd international conference on Machine learning. June 2006; pp 161-68 <https://doi.org/10.1145/1143844.1143865>

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

- Charron A, Harrison RM, Moorcroft S, Booker J. Quantitative interpretation of divergence between PM₁₀ and PM_{2.5} mass measurement by TEOM and gravimetric (Partisol) instruments. *Atmospheric Environment* 2004; 38(3) 415-423. <https://doi.org/10.1016/j.atmosenv.2003.09.072>
- Cook RD, Weisberg S. *Residuals and Influence in Regression*. New York, NY: Chapman & Hall 1982
- Cortina-Januchs MG, Quintanilla-Dominguez J, Vega-Corona A, Andina D. Development of a model for forecasting of PM₁₀ concentrations in Salamanca, Mexico. *Atmospheric Pollution Research* 2015; Volume 6 (4).
- de Hoogh K, Gulliver J, Donkelaar AV, Martin RV, Marshall JD, Bechle MJ, Cesaroni G, Pradas MC, Dedele A, Eeftens M, Forsberg B, Galassi C, Heinrich J, Hoffmann B, Jacquemin B, Katsouyanni K, Korek M, Künzli N, Lindley SJ, Lepeule J, Meleux F, de Nazelle A, Nieuwenhuijsen M, Nystad W, Raaschou-Nielsen O, Peters A, Peuch VH, Rouil L, Udvady O, Slama R, Stempfelet M, Stephanou EG, Tsai MY, Yli-Tuomi T, Weinmayr G, Brunekreef B, Vienneau D, Hoek G. Development of West-European PM_{2.5} and NO₂ land use regression models incorporating satellite-derived and chemical transport modelling data. *Environ Res*. 2016 Nov;151:1-10. doi: 10.1016/j.envres.2016.07.005.
- Di Q, Rowland S, Koutrakis P, Schwartz J. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. *J Air Waste Manag Assoc*. 2017 Jan;67(1):39-52. doi: 10.1080/10962247.2016.1200159.
- Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, Sabath MB, Choirat C, Koutrakis P, Lyapustin A, Wang Y, Mickley LJ, Schwartz J. An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution. *Environ Int* 2019; 130: 104909. doi: 10.106/j.envint.2019.104909
- Gocheva-Ilieva SG, Stoimenova MP. PM₁₀ Prediction and Forecasting Using CART: A Case Study for Pleven, Bulgaria. *International Journal of Environmental and Ecological Engineering* 2018; 12(9).
- Green DC, Fuller GW, Baker T. Development and validation of the volatile correction model for PM₁₀ – An empirical method for adjusting TEOM measurements for their loss of volatile particulate matter. *Atmospheric Environment* 2009; 43(12);2132-2141. <https://doi.org/10.1016/j.atmosenv.2009.01.024>
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer.
- Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, Liu Y. Estimating PM_{2.5} Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol*. 2017; 51(12);6936-6944. Publication Date: May 23, 2017. <https://doi.org/10.1021/acs.est.7b01210>

[illegible]

621 Lee M, Kloog I, Chudnovsky A, Lyapustin A, Wang Y, Melly S, Coull B, Koutrakis P,
622 Schwartz J. Spatiotemporal prediction of fine particulate matter using high-
623 resolution satellite images in the Southeastern US 2003-2011. *J Expo Sci Environ*
624 *Epidemiol.* 2016 Jun; 26(4):377-84. doi: 10.1038/jes.2015.41.

625 LONDON AIR QUALITY NETWORK (LAQN)
626 <http://www.londonair.org.uk/LondonAir/Default.aspx>

627 Menze, B.H., Kelm, B.M., Masuch, R. et al. A comparison of random forest and its
628 Gini importance with standard chemometric methods for the feature selection and
629 classification of spectral data. *BMC Bioinformatics* 2009; 10: 213.
630 <https://doi.org/10.1186/1471-2105-10-213>

631 Mishra D, Goyal P. Development of artificial intelligence based NO₂ forecasting
632 models at Taj Mahal, Agra. *Atmospheric Pollution Research* 2015; 6(1): 99-106.

633 Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in
634 prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2010; 32, 569–575.

635 Russo A, Lind PG, Raische IF, Trigo R, Mendes M. Neural network forecast of daily
636 pollution concentration using optimal meteorological data at synoptic and local
637 scales. *Atmospheric Pollution Research* 2015; 6(3)
638

639 WHO, 2013. Review of evidence on health aspects of air pollution – REVIHAAP
640 Project, Technical Report, World Health Organization, Regional Office for Europe.
641

642 WHO, 2013b. Health risks of air pollution in Europe – HRAPIE project.
643 Recommendations for concentration–response functions for cost–benefit analysis of
644 particulate matter, ozone and nitrogen dioxide. World Health Organization, Regional
645 Office for Europe.
646

647 Xiao Q, Chang HH, Geng G, Liu Y. An Ensemble machine learning model to predict
648 historical PM_{2.5} concentrations in China from satellite data. *Environ Sci Technol* 2018;
649 52: 13260-13269.

650 Xue W, Zhang J, Zhong C, Ji D, Huang W. Satellite-derived spatiotemporal PM_{2.5}
651 concentrations and variations from 2006 to 2017 in China. *Sci Total Environ.* 2019
652 Nov; 20:134577. doi: 10.1016/j.scitotenv.2019.134577.

653

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Highlights

- A methodology based on traditional regression, machine learning and ensemble models to enhance the availability of a PM_{2.5} measurement data base was developed.
- The exceptionally dense London monitoring network was used to predict PM_{2.5} concentrations, using other pollutants, meteorological and land-use variables, at the location of fixed monitors measuring PM₁₀ and NO_x but not PM_{2.5}.
- The Generalized Additive Model combination (ensemble) of the two separate model predictions performs best, demonstrating that the combination of more traditional regression and machine learning methods leads to improved predictions throughout the range of concentrations.
- Given the relative scarcity of long-term PM_{2.5} measurements in many cities across the world, this methodology could be extremely useful to aid the verification of spatio-temporal exposure models and to be used in health studies.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Title: Prediction of PM_{2.5} concentrations at the locations of monitoring sites measuring PM₁₀ and NOx, using generalized additive models and machine learning methods: a case study in London

Response to Reviewers' comments

(Following the Editor's advice, we first responded to the comments of Reviewers 2 and 3 and the responses to Reviewer 1 follow last).

Page number and lines refer to the highlighted manuscript.

Reviewer #2: Reviewer comments were copied from an attachment.

Comment 2.1. Enclosed please find my review of the paper ATMENV-D-20-00319 "Prediction of PM_{2.5} concentrations at the locations of monitoring sites measuring PM₁₀ and NOx, using generalized additive models and machine learning methods, for enhancement of the PM_{2.5} database: a case study in London." by A. Analitis, B. Barratt, D. Green, A. Beddows, E. Samoli, J. Schwartz, and K. Katsouyanni.

The paper is devoted to the analysis of air quality data including PM_{2.5}, PM₁₀, NOx and Ozone in London. Since PM₁₀ and NOx are monitored much more intensively compared to PM_{2.5}, the objective is to enhance the PM_{2.5} data at the locations where only PM₁₀ and NOx measurements are available. For this purpose, several methods are used including regression model and Neural Network (NN) based method.

Since PM_{2.5} is directly associated with a threat to health, the topic raised in this article is very relevant. Indeed, PM_{2.5} is measured much less compared to PM₁₀ and NOx over M25 London Orbital Motorway area, therefore to get the reliable estimations of PM_{2.5} at many additional locations is a great purpose. Still, there are a couple of questions of the analysis performance.

Response: We thank the Reviewer for summarizing the purpose and for agreeing that the work is worthwhile.

Comment 2.2. First, for regression, several covariates are used to compare to dependent variable of PM_{2.5}, then the calculated values are compared to the measured values to get R² and MSE. This approach is doubtful. You have to get the regression values at the subset of PM_{2.5} volume, and check the calculated PM_{2.5} with the rest of PM_{2.5} measurements. Or, you may randomly take off a certain percentage of PM_{2.5} sites, for ex. 10%, use regression model on 90% PM_{2.5} sites, calculate PM_{2.5} for all 100% sites and compare what you get at that 10% independent sites.

Could you get the same 99.29% R²?

Response: We agree with the Reviewer. What he suggests is precisely what we have done, and what we mean when we say that "The 10-fold CV adjusted R²....". However, apparently we have not explained the procedure clearly in our "Data and Methods" section. So, we have added a paragraph in the revised manuscript to clarify the procedure.

The text added (page 8, lines 222 to 228): "To obtain the R² and MSE of the models we used 10-fold cross-validation (CV) in order to test the model's ability to predict new data that was not used in estimating it. Specifically we fitted the model based on a training data set (90% of the data) and used 10% of the observations as validation data. We then predicted the

omitted observations (the validation data set) and calculated the R^2 and MSE. This was done sequentially 10 times and then the predicted values were used to obtain the R^2 and MSE ".

References: 1.Hastie, T., Tibshirani, R., & Friedman, J. H. (2009).The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer."

2. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation.IEEE Trans. Pattern Anal.Mach. Intell.2010,32, 569–575."

Comment 2.3. Second, to get the estimated $PM_{2.5}$ at many sites is the very good task. The only problem is to evaluate the quality of calculated values. You use just two estimators, R^2 and MSE. They are very important estimators. Still, for air quality problems it is more important to have rigorous estimations of high concentration $PM_{2.5}$. So, it is very important to include categorical estimators of $PM_{2.5}$ calculations like percent of $PM_{2.5} > 15$, or 25, or $50 \mu g/m^3$.

Response: Following the advice of the Reviewer we have now provided the corresponding calculations for our combined model. Indeed the observed and predicted percentage of $PM_{2.5}$ concentrations >15 , 25, and 50 were 42.1%, 15.4%, 1.6% and 41.9%, 15.0%, 1.6% in 2009-2013 and 45.9%, 19.6%, 2.1% and 48%, 20.7%, 2.1% in 2004-2008 correspondingly. We included this information in the revised manuscript, in Table 3 and added text in the Results (page 18, lines 417 to 420).

The text added is:

"It is also noted that concerning the prediction of concentrations in the relatively higher range of values, those >15 , 25 and $50 \mu g/m^3$, the percentages of measured and predicted values above those thresholds are remarkably similar (Table 3)."

Comment 2.4. This article is suitable for the Journal of Atmospheric Environment and could be published after major revision.

Response: Thank you

Comment 2.5. Several comments may be helpful to improve the paper presentation:

1. Lines 122-132: what do you mean by "range of the number of measurements per site"? Why do you include O_3 ? Do you use ozone as a covariate?

Response: The Reviewer is right, we did not use O_3 for predicting $PM_{2.5}$ and therefore we deleted this reference to O_3 . By "range of the number of measurements per site" we refer to the number of measurements provided by each site, which varied.

2. Lines 140-144: you try to find the outlier sites to calculate Cook's distance, then on lines 250-252 you declare that no one site was excluded. You have to put this statement back to lines 140-144 if it's relevant.

Response: We moved the text from lines 250-2 to page 5 lines 149-150, section "Data and Methods", as suggested by the Reviewer.

3. Lines 146-155: $PM_{2.5}$ TEOM measurements have been corrected using regional and local values of pressure, temperature and averaged value of all FDMS measurements. It is interesting to see how much these corrections are. Does it depend on year, location, season?

Response: As the method primarily acts to correct for secondary particles lost by the TEOM method's heated inlet technique, the correction factor does vary temporally, but not spatially (secondary typically PM being a regional pollutant). The calculated correction factor varies day by day, as described in detail in Green et al., 2009 -

<https://doi.org/10.1016/j.atmosenv.2009.01.024>. We have amended the explanation of variables in the TEOM_{VC}M equation to make this temporal variation clear.

4. Lines 173: You work with two segments separately, years 2004-2008 and 2009-2013 because “the capture rate of PM_{2.5} was different for those two periods”. How different? One percent or 50 percent?

Response: The increase in the PM₁₀ and NO₂ sites for measurements between the 2 five-year periods was moderate, 5% and 10% respectively. However, the available PM_{2.5} sites increased substantially (by 38%). This latter increase, together with the fact that conditions may have changed, made us feel that it is appropriate to use a maximum of 5 years for fitting a model. We have now explained this better in the text (see page 5, lines 133 to 137):

“The increased number of PM_{2.5} measurements in later years and the possibility that air pollution determinants may have changed over time, led to the decision to limit the time span of model development to 5 years, thus the model was developed for 2004-2008 and 2009 -2013 separately.”

5. Line 264 and 270 and Table 1: why do you use SIX days of the week? Should be seven.

Response: To adjust for a categorical variable with 7 categories, typically 6 dummy or indicator variables are constructed, as one is the "reference" category for which all six variables take the value 0. We understand that this was not properly explained, so we have now added the explanation in all relevant places, as indicated by the Reviewer (see page 10, lines 296 to 297; lines 305 to 306; and Tables 1 and 2).

6. Figure 1 vs Figure 3: why is partial dependence on temperature opposite for two time periods? Is it partially explained by (4)?

Response: In both models the partial dependence on temperature follows a U-shape but the curves are not identical. A similar difference is partly evident for other meteorological and geographical variables. This can be attributed to different meteorological conditions across years. We have added a comment in the text to indicate that (page 15, lines 368 to 371).

7. Table 3: The maximum value of PM_{2.5}, especially for the second time period of 2009-2013, is noticeably different for PM_{2.5} measurements and PM_{2.5} calculations, therefore it is important to show the categorical comparison.

Response: Following the suggestion of the Reviewer, we now included this information in Table 3 and in the text, please see response to Comment 2.3.

8. Line 421: CV R².

Response: Thank you for pointing out this typo. We have corrected it.

Reviewer #3: This is a case study in London conducted to predict PM_{2.5} levels at the monitoring sites using PM₁₀ and NO_x. Model performances were enhanced by combining generalized additive and random forest models. I recommend accepting this paper after addressing the issues as outlined below in the specific comments.

Major points:

Comment 3.1. Given the relatively dense coverage of PM₁₀ and NO_x monitors in the study region, the authors' prediction models were able to provide daily PM_{2.5} concentrations at more than a hundred monitoring locations and may further serve for health studies. However, the problem in some highly polluted or less developed regions is lacking of ground monitors or the uneven distributed monitors. Since this method can only predict PM_{2.5} concentrations at locations of fixed monitors where PM₁₀ and NO_x have to be measured, availability of input data limits the application to other regions. Therefore, the sentence regarding the importance of your work for "the World with high levels of air pollution" in the abstract section was not proper.

Response: We deleted this sentence following the Reviewer's comment.

Comment 3.2. Satellite-based techniques have been applied to predict long-term PM_{2.5} levels globally (van Donkelaar et al. Environ Sci Technol 2016; 50: 3762-72) and in London (Beloconi et al. Remote Sensing of Environment, 2016, 172:148-164.). Using machine learning or ensemble models for PM_{2.5} prediction is also not innovative. What is the advantage of your PM_{2.5} predictions compared with the high-resolution ones?

Response: We thank the Reviewer for asking this question and prompting us to respond. The objective of our study is different from the objective of studies like the one mentioned above. The studies using satellite information or combined information (like the Beloconi paper) often have the objective to provide a spatio-temporal map from which information can be obtained for PM concentrations at locations of interest, e.g. the residences of subjects participating in a health study. Beloconi et al applied their models specifically to London and noted the problem of the relative sparsity of PM_{2.5} measurements. Our purpose was different: it was to predict the PM_{2.5} concentrations at the location of sites measuring PM₁₀ and other pollutants, thus enhancing the PM_{2.5} data base that can be used for modelling or other purposes. Using satellite data would not add predictive value: our model R² is around 99% whilst the models reported by Beloconi et al for PM_{2.5} have a maximum of 90% R².

Prompted by the Reviewer, we have now added a paragraph in the Discussion to discuss and compare the Beloconi et al paper (page 22, lines 489 to 499).

The text added is: "Beloconi et al (2016) developed spatiotemporal models for PM₁₀ and PM_{2.5} for London using satellite AOD data, measurements, meteorological and land use variables. Their purpose was to estimate concentrations at 1km² grid using sophisticated modelling to obtain the best possible predictability. However, the fact that remote sensing data have missing values and measurements of PM_{2.5} are less dense compared to PM₁₀, resulted in a maximum R² of 90% for PM_{2.5} models. The objective of our study is not directly comparable to that of Beloconi et al: we aimed at predicting the best possible spatiotemporal estimate for PM_{2.5} concentrations at the geographical locations of the monitoring sites in London. Since we have excellent predictors for the site locations (such as PM₁₀ measurements), our models perform almost perfectly and further predictors would not add to the model performance."

Comment 3.3. How did the authors choose the predictors? Did you consider other pollutants, such as SO₂ or O₃ in your model?

Response: We chose predictors based on previous literature and knowledge of the pollutants emissions and dispersion. Among the predictors chosen a priori we evaluated their predictive value in the Regression model. Ozone has not been often used as a predictor and SO₂ levels are low in London we therefore chose not to use them. We have added an explanation in the text concerning selection of predictors (page 7, lines 209 to 211).

Comment 3.4. Please further conduct the spatial and temporal cross validation to fully evaluate your models, as shown in Hu, et al. Environ Sci Technol. 2017; 51:6936-44.

Response: We applied the Hu et al method for spatial cross-validation of our final model. We think that the temporal CV is not relevant for our study (in any case the temporal CV is better compared to the spatial) because there are no situations where we predict a PM_{2.5} concentration in the absence of any measurement on a particular day. The spatial CV R² and MSE were 99.28% and 1.00 for 2004-2008 and 98.94% and 1.13 for 2009-2013 respectively. This information has now been added in page 17, lines 396 to 401.

The text added is: "We also calculated the spatial CV, following the method applied in Hu et al (2017), by excluding 10% of the measurement sites from the training data set and predicting them from the models including the other 90% of sites and repeating this procedure until all predictions were done. The corresponding R² and MSE for the two periods were 99.28%, 1.00 and 98.94%, 1.13. Thus, there is no substantial reduction in R² and MSE when the spatial predictions are assessed.

Comment 3.5. Why site-wide averages of meteorological factors instead of site-specific values were used (lines 188-191). These averages would not contribute to the daily spatial distribution of PM_{2.5}.

Response: Measurements of meteorological variables are not so dense in London and therefore the geographical resolution needed to account for spatial differences was not available. Since daily values at different locations for temperature and other meteorological variables are very highly correlated, we aimed at explaining temporal variability by using the area specific values and explaining spatial variability by variables characterizing the location (such as traffic or background). We have added a description of the meteorological data used in section 2.1 Data (page 6, lines 176 to 185).

The text added is: "Data for meteorological variables, specifically for mean daily temperature (°C), relative humidity (%), wind speed (ms⁻¹) and wind direction (°N) were compiled from all available sites with <25% missing values over the study period (n=7). Because the number of available measurement sites was not so dense to allow characterization at the location of all pollution measurement sites, but the correlation of variables between sites is very high (r between sites >0.99 for temperature and >0.85 for the other meteorological variables) for all meteorological variables), we averaged per day across all meteorological sites and used one time series for each meteorological variable."

Comment 3.6. PM_{2.5} and PM₁₀ measurements were highly correlated in most cases and their ratios were generally adopted to predict PM_{2.5} in some previous studies. PM₁₀ is also the dominant predictor in your model. A comparison of PM_{2.5} concentrations predicted using PM₁₀ (Model 1, Table2) and your full model may help to show the improvement of your method.

Response: We thank the Reviewer for mentioning this point. We have added a comment in the Discussion to bring the point to the attention of the reader (page 21, lines 446 to 449).

The text added is: "Our model predicts better than using PM₁₀ alone as predictor: about 10% of the variability of PM_{2.5} is explained by adding the other variables in the regression model and most of the other variables are important in the RF as well"

Comment 3.7. Ensemble models for GAM and random forest models were reported previously (Di, et al. Environ Sci Technol. 2020;54(3):1372-84; Xiao, et al. Environ Sci Technol. 2018;52(22):13260-9). Please modify the related discussions (Lines: 423-427).

Response: In our study we combined predictions of a regression model and a RF with a GAM function. The papers mentioned by the Reviewer combine various machine learning methods with a GAM function (the specific Di et al paper for NO₂). We modified the Discussion accordingly and we mentioned the papers and their results (page 23, lines 503 to 505).

The modified text now is: In some published studies, various machine learning methods predicting PM_{2.5} or NO₂ have been combined (Di et al 2019; Xiao et al 2018) using a GAM function, and the combination (ensemble) models also performed best.

Minor points:

1. Please provide the data sources of meteorological variables

Response: We provided the sources, please see page 6, lines 179 to 180.

2. Line 316, RMSE or MSE? Please check it and add its unit?

Response: It is MSE (mean square error) and the units were inserted.

3. Please explain the x axis in the legend of Figure 2.

Response: We have explained it both in the methods section (page 8, lines 246 - 249) and in the legends of Figures 2 and 4.

4. Please provide the fitted slopes and intercepts in Figure 5.

Response: We have added the slopes and intercepts in the graphs of Figure 5.

Reviewer #1: In this work, the authors used two models - a regression model and a machine learning method - to predict PM_{2.5} concentrations based on co-located measurements of PM₁₀ and NO₂. They also used a GAM model to combine the results from the two above models to further improve the prediction accuracy. Although they got quite good results for their models (i.e., ten-fold cross-validation R² above 0.98), I do not feel the manuscript could be published at its current form. The written needs improvement and the authors should clearly clarify why they choose such method over the satellite-based method.

Major comments:

Comment 1.1. As mentioned by the authors in the Introduction, estimation of PM_{2.5} concentrations is a hot topic recently and a lot of works have been done in such area. Previous works have used satellite AOD as a key predictor to estimate PM_{2.5}, which have greatly enlarged the spatial coverage of PM_{2.5}, though satellite AOD has missing values. There are also some studies trying to fill in the gaps in missing AOD by incorporating chemical transport model simulations or other parameters, resulting in full spatial-coverage of PM_{2.5} predictions. However, in this work, the PM_{2.5} predictions are only made in certain monitoring locations, which are much, much sparser than satellite-based PM_{2.5} predictions. So I am wondering what the point is of this method as the satellite AOD are available globally.

Response: The objective of our study is different from the usual objective of studies using satellite information. Please also see our response to comment 3.2 (of Reviewer 3) above. The studies using satellite information or combined information (like van Donkelaar et al. Environ Sci Technol 2016; 50: 3762-72 and Beloconi et al. Remote Sensing of Environment, 2016, 172:148-164 referred to by Reviewer 3) often have the objective to provide a spatio-temporal map from which information can be obtained for PM concentrations at locations of interest, e.g. the residences of subjects participating in a health study. In several cases these are developed in the absence of dense measurements. Beloconi et al applied their models specifically to London and noted the problem of the relative sparsity of PM_{2.5} measurements. Our purpose was to predict the PM_{2.5} concentrations at the location of sites measuring PM₁₀ and other pollutants. Using satellite data would not add predictive value: our model R² is around 99% whilst the models reported by Beloconi et al for PM_{2.5} have a maximum of 90% R².

Comment 1.2. A spatial cross-validation should be included to test the model's ability in spatial interpolation since predictions are made outside the locations used for model training. Since nearby locations might have auto-correlations in their measurements, a more strict cross-validation, the spatial-clustered cross-validation, which is proposed by Murray et al. (2019), should also be considered to avoid influence from nearby located monitors.

Response: We applied the method by Hu et al. Please see our response to comment 3.4.

Comment 1.3. The Results section are simple and the authors should interpret more about their results.

Response: We understand the Reviewer's comment. We have enriched our Results section. Please see also our responses to comments 2.3, 2.5 and 3.4.

Minor comments:

1. The title is a bit too long.

Response: we shortened the title.

2. All the periods shown in this work such as 2004-08 should be 2004-2008.

Response: Done.

3. Line 172: Two models were built for 2004-2008 and 2009-2013 separately. Will the discrepancies between the two models affect the predicted trend?

Response: The predictions are very good and the trends are preserved.

4. Line 183: What is the regression model used in this study?

Response: We have chosen to describe the predictors in the Methods section and mention the performance of various models in the Results. The final model that was chosen is described in detail in the Results.

5. Table 1&2: I am not sure what could these figures tell us. I do see R² increase with more variables included, but can we tell which one is more important? Please clarify.

Response: we have added a related comment, please see our response to comment 3.6

6. Figure 1&3: The authors should explain more about these two figures. What's their implications?

Response: We have added explanation on the Figures, see also response to Comment 2.5.6.

7. Figure 2&4: The importance ranking of parameters differed in the two periods. For example, wind speed ranked 2nd in 2004-2008 but ranked 6th in 2009-2013. How do the authors explain these differences?

Response: Please see our response to Comment 2.5.5

References:

Murray, N. L.; Holmes, H. A.; Liu, Y.; Chang, H. H., A Bayesian ensemble approach to combine PM_{2.5} estimates from statistical models using satellite imagery and numerical model simulation. *Environmental Research* 2019, 178, 108601.

Figure6

Figure1

[Click here to download high resolution image](#)

Author contributions: AA: Formal Analysis, Data Curation, Methodology, Writing - Original Draft, Review & Editing; BB: Data Curation, Methodology, Writing - Review & Editing; DG: Methodology, Writing - Review & Editing; AB: Methodology, Writing - Review & Editing; ES: Conceptualization, Methodology, Writing - Review & Editing; JS: Conceptualization, Methodology, Writing - Review & Editing; KK: Conceptualization, Funding acquisition, Methodology, Writing - Original Draft, Review & Editing