Enhancement of the $PM_{2.5}$ data base for London within the STEAM project

Description of the procedures followed and the enhanced data base

June 2019

Summary

Within the STEAM project we had a large data base of measurements for NOx/NO2, O3 and PM_{10} from fixed site monitors from 2004 to 2013 but a much smaller one for PM $_{2.5}$. We used a method, described below in detail, to enhance the $PM_{2.5}$ data base of daily values in many monitoring sites. Thus, we had $PM_{2.5}$ data available from 24 sites in 2004-08 and 33 sites in 2009-13 (16,957 and 40,083 measurements respectively) and we have now predicted additionally 102,800 and 85,436 daily concentrations for $PM_{2.5}$ in 2004-08 and 2009-13 respectively.

We think that this enhanced data base may be useful for other researchers or policy makers. Below we describe exactly how it was produced and would like to make it available. If you are interested in using it please request a password, briefly explaining your objectives. Address your request to Ben.Barratt@kcl.ac.uk or Klea.Katsouyanni@kcl.ac.uk.

Background and objective

Within the 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_2 and daily 8-hour maximumO_3) concentrations including all measurements from sites within the M25, performed by the London Air Quality Network (LAQN), during the years 2004-13. For reasons related to the number of available measurements, we did all the analyses separately for two periods 2004-08 and 2009-13. Not all sites provided measurements for the whole time period and there were a few missing values during periods of operation.

For PM_{10} we compiled data from 108 sites in 2004-08 and 115 sites in 2009-10 (136,874 and 134,230 measurements respectively, range of the number of measurements per site: 50-1,812 and 11-1,808 respectively); for NOx we compiled data from 197 sites in 2004-08 and 216 in 2009-13 (252,101 and 259,859 measurements respectively, range of the number of measurements per site:21-1,809and 11-1,817 respectively), for O_3 we compiled data from 41 sites in 2004-08 and 42 in 2009-13 (59,893 and 55,121 measurements respectively, range of the number of measurements per site 354-1,819and 59-1,809 respectively),whilst for $PM_{2.5}$ data were available only from 24 sites in 2004-08 and 33 sites in 2009-13 (16,957 and 40,083 measurements respectively, range of the number of measurements per site 13-1,382 and 95-1,809 respectively).

For the purposes of the STEAM project, a larger data base of measurements was necessary and one major reason for implementing the project in London was the large number of operating sites under the LAQN. As there were fewer measurements for $PM_{2.5}$we have decided to run prediction models for $PM_{2.5}$ using available information on correlated variables, such as other pollutants and meteorological, geographical and temporal variables, to enhance the $PM_{2.5}$ data base. We followed an a-priori decided protocol and predicted 102,800 and 85,436 daily concentrations for $PM_{2.5}$ in 2004-08 and 2009-13 respectively. Since this enhanced data base may be useful for other research projects, it will be made available upon request to other researchers. Here we provide a description of the protocol, the performance of the models and descriptive statistics for the resulting $PM_{2.5}$estimates.

Protocol for the estimation of $PM_{2.5}$ concentration data

The general principle was to use all the sites with concurrent measurements of $PM_{2.5}$, PM_{10} and NOx to develop a model associating $PM_{2.5}$ with these pollutants as well as other predictors, such as meteorological, geographical and temporal variables. The model developed was used to predict $PM_{2.5}$at fixed sites with no $PM_{2.5}$ measurements (but which measured PM_{10}&NOx) and thus provide a greatly enhanced data base of $PM_{2.5}$estimated concentrations.

First we checked whether all fixed sites that measure PM_{10}andNOxconcurrently reflect the same association between PM_{10} and NOx. This was done because we hypothesized that some sites may have unique characteristics affecting the association between PM_{10} and NOx, e.g. be near a local, non-exhaust, possibly industrial source, or have outlier meteorological conditions (like high or low wind speed). For this purpose, we plotted mean NOx vs mean PM_{10} concentrations from all sites which provided both measurements and estimated a regression line with the intention to exclude stations identified as outliers in this plot.We calculated the Cook's distance and defined a Cook's D value >1, as indicating an influential observation(Cook, RD & Weisberg, S (1982). Residuals and Influence in Regression. New York, NY: Chapman & Hall). 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. Form this investigation we concluded that no fixed site should be excluded from further 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. Thus, whilst $PM_{2.5}$ measured by FDMS (Filter Dynamics Measurement System) monitors (reference method) were used as such,the $PM_{2.5}$ data from TEOM instruments (non-reference method) were corrected using a method developed by researchers at King's College (Ben Barratt, David Carslaw, Gary Fuller, David Green & AnjaTremper. 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 LeedsMay 2009).

Specifically, the correction applied was:

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

Where:

T_{d0} is the TEOM $PM_{2.5}$ concentration ('FINE')

Pr is the set reporting pressure of the TEOM

Pa is the regional mean ambient pressure for the measurement period

Ta is the regional mean ambient temperature for the measurement period

Tr is the set reporting temperature of the TEOM

ƒVCM is the VCM function (optimised in this study to 0.72)

n is the number of FDMS purge measurements used

z_i is the FDMS purge measurement for station i

For the application of the model we included all sites which measured $PM_{2.5}$, PM_{10} and NOx concurrently (number of sites =18 for 2004-08 and 25 for 2009-13) and fit a regression model for 2009-13 and 2004-08 separately, as the capture rate of $PM_{2.5}$ was higher in 2009-13. We used the $PM_{2.5}$measurements done by the reference method as well asthe corrected TEOM measurements.

Generalized Additive Models (GAM)

Specifically we tested Generalized Additive Models (GAM) for each period, with daily values of $PM_{2.5}$ as dependent variable and daily valuesof the variables described below (all variables are for lag0, i.e. same day) as covariates:

PM_{10}: measured PM_{10} from the same site as $PM_{2.5}$ (24- hour values, in µg/m^3)

NOx: measured NOx from the same site as $PM_{2.5}$ (24-hour average, in µg/m^3)

Monitor type: 1=roadside/kerbside, 2=background (categorical)

Day of the week: 0=Sunday, 1=Monday,..., 6=Saturday (categorical, six dummy variables)

Time trend: a continuous variable denoting the length of the period with values 1,…N (N is the total number of days within each period) entered as a natural spline with 12 degrees of freedom (df) per year of data to adjust for (approximately) monthly variations

Temp: mean daily temperature averaged over all available sites (in ^oC), entered as a natural spline with 4 df over the period of analysis.

Humidity: Mean daily relative humidity averaged over all available sites (continuous, %) entered as a naturalspline with 2 df

Wind_speed: mean daily wind speed averaged over all available sites, in 0.1 m s^{-1} ; adjusted with a natural spline with 3 df

Wind_dir: mean daily wind direction in ^oN averaged over all available sites; adjusted with a natural spline with 3 df.

We started with a model including only PM_{10} as independent variable and then we added each term, one at a time. In the full model, we tried alternatively smoothed terms for both PM_{10} and NOx and also smoothed terms for all variables with df chosen by Generalized Cross Validation (GCV) criterion. We also tested 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 & the interaction with PM_{10}. All variables entered into the model after PM_{10} increased the value of the adjusted R^2slightly. We decided to include only the variables that contributed most to the value of the adjusted R^2 of the GAM regression model described above. The final set of covariates used is PM_{10} measurements, NOx measurements, time trend, week day, bivariate smooth function for coordinates & the interaction between the bivariate smooth function of the coordinates and PM_{10} measurements. The removal of variables had a very small impact on the R^2.

The final model was (all variables are for lag0, i.e. same day):

*$PM_{2.5}$= intercept+PM_{10}+NOx+(day of the week/6 dummy variables)+ns(trend, df=12 per year)+s(latitude, longitude) + PM_{10} * s(latitude, longitude)*

where:

$PM_{2.5}$: measured $PM_{2.5}$ (24- hour values,μg/m^3)

PM_{10}: measured PM_{10} from the same site as $PM_{2.5}$ (24- hour values, μg/m^3)

NOx: measured NOx from the same site (24-hour average, μg/m^3),

Day of the week: 0=Sunday, 1=Monday,...,6=Saturday (categorical, six dummy variables)

Time trend: a continuous variable denoting the length of the period with values 1,…N entered as a natural spline with 12 df per year of data to control for (approximately) monthly variations and long-term trend

<u>Plus</u>,

s(latitude, longitude): A bivariate smooth function for the coordinates (longitude, latitude) of the fixed monitoring sites and

PM10 * s(latitude, longitude): an interaction term of this with $PM_{10.}$

The bivariate smooth function for the coordinates & the interaction terms were included in order to capture any remaining spatial variability of the fitted vs the observed values.

Random Forest

We further applied a random forest, as an alternative method to improve the fit of the GAM regression model used for the estimation of $PM_{2.5}$ data. This was done using the function “randomForest” in the R library “randomForest”. It was also applied separately for each time-period of interest, i.e. 2004-08 and 2009-13. We used the default parameter values concerning 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” described in the manual. We used the same variables as those tested in the GAM described above.

Combination of predictions obtained from the GAM & the predictions obtained from the Random Forest method

We combined predictions from the GAM and the RF method to improve the fit of the model. The predictions were used as independent variables in a new GAM model for each period, as described below:

$$PM_{2.5}= s(\text{predictions from gam model}) + s(\text{predictions from random forest method})$$

Finally we chose the best model based on the Adjusted R^2 among GAM, Random Forest and their combination and used it to 1) fill in missing $PM_{2.5}$ values from stations providing $PM_{2.5}$ measurements both during 2009-13 and 2004-08 and 2) provide $PM_{2.5}$ series in stations with only PM_{10} and NOx measurements.

Results

Time period 2009-13

Generalized Additive Models (GAM)

Table 1 shows the contribution of each variable to the value of the adjusted R^2 of the linear regression model, for the time period 2009 - 2013.

The adjusted R^2 of the final GAM model was 86.5%. Moreover, a model validation was applied using a 10-fold cross validation (CV) method. The CV adjusted R^2 of the model for the time period 2009 – 2013 was 86.2%.

Table 1. Contribution to the model's R^2 per added term after PM_{10}, 2009-13

	R^2	adj R^2	
PM_{10}	0.7921	0.7921	->initial model
+NOx	0.7969	0.7969	->adding terms one by one
+monitor type	0.8003	0.8003	
+Week day	0.8010	0.8010	
+ns*(time trend)	0.8174	0.8169	
+ns(temp)	0.8220	0.8215	
+ns(hum)	0.8404	0.8400	
+ns(wspeed)	0.8483	0.8479	
+ns(wdir)	0.8490	0.8486	->full model
Splines for PM_{10}& NOx (excluding the corresponding linear terms) with df from GCV		0.8627	
Splines for PM_{10} , NOx, time, temp, hum, wspeed, wdir (excluding the corresponding linear terms) with df from GCV		0.8663	
Inter. PM_{10}x monitor type		0.8577	
Bivar. spline s(PM_{10},wspeed)		0.8692	
Main effects plus inter. PM_{10} x s(lat,long)		0.8840	
Final Model**		0.8650	

```
* ns: natural splines
** pm10, nox, s(time), week day, main+int. pm10-s(lat,long)
10-fold cross-val:   MSE: 14.5648    R2adj: 0.8624
```

Random Forest

The Random Forest R^2 and MSE for the 2009-13 period were 92.8% and 7.64, respectively consisting an improvement over the GAM model. Figure 1 shows the relevant partial dependence plot which gives a graphical depiction of the marginal effect of each predictor variable on $PM_{2.5}$. In Figure 2 we can see the variance importance of each predictor in terms of both decrease in MSE and remaining error in predictive accuracy after a node split (node impurity).

Figure 1. Partial dependence plot (2009-13)

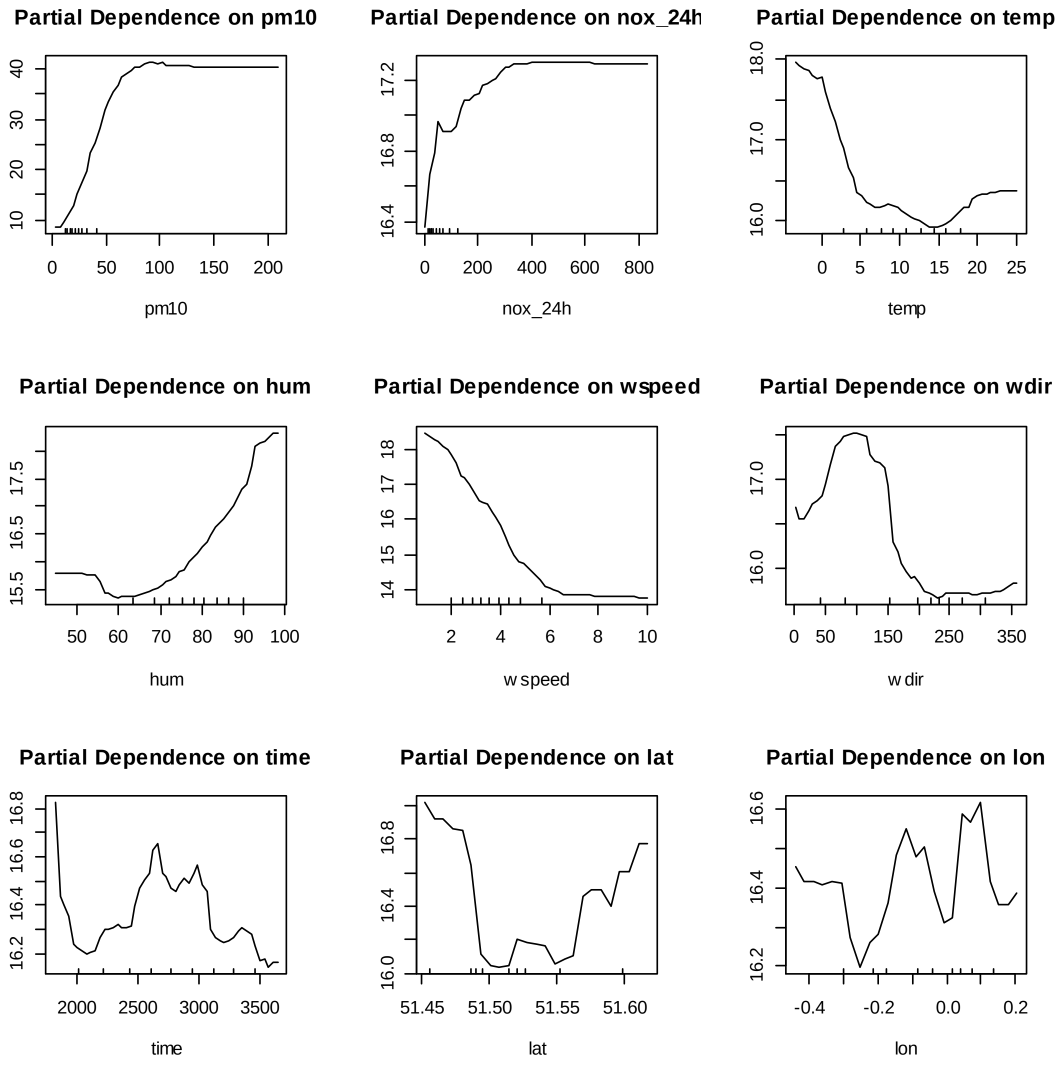
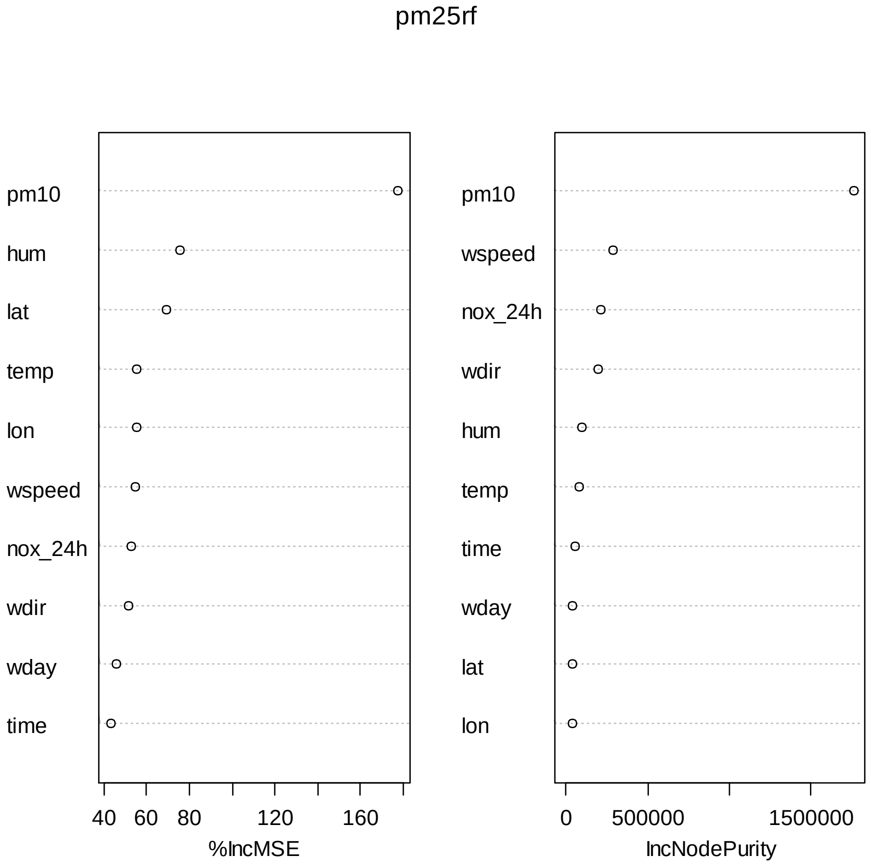


Figure 2. Variance importance plot (2009-13).



Time period 2004-08

Generalized Additive Models (GAM)

Table 2 shows the contribution of each variable to the value of the adjusted R^2 of the GAM regression model, for 2004 - 2008. A similar pattern for the contribution to R^2 with the period 2009-13 was observed. The adjusted R^2 of the final model (described previously) was 90.1%. The 10-fold CV adjusted R^2 of the model for the time period 2004 – 2008 was 89.9%.

Table 2. Contribution to the model's R^2 per added term after PM_{10}, 2004-08

	R^2	**adj R^2**	
PM_{10}	0.8359	0.8359	->initialmodel
+NOx	0.8440	0.8440	->addingterms one by one
+monitor type	0.8453	0.8453	
+Week day	0.8477	0.8476	
+ns*(time trend)	0.8705	0.8697	
+ns(temp)	0.8730	0.8722	
+ns(hum)	0.8913	0.8906	
+ns(wspeed)	0.9006	0.8999	
+ns(wdir)	0.9007	0.9000	->fullmodel
Splines for PM_{10}&NOx (excluding the corresponding linear terms) with df from GCV		0.9022	
Splines for PM_{10} , NOx, time, temp, hum, wspeed, wdir (excluding the corresponding linear terms) with df from GCV		0.9037	
Inter. PM_{10}x monitor type		0.9034	
Bivar. spline s(PM_{10},wspeed)		0.9059	
Main effects plus inter. PM_{10} x s(lat,long)		0.9232	
Final Model**		0.9015	

* ns: natural splines

** pm10, nox, s(time), week day, main+int pm10-s(lat,long)
10-fold cross-val: MSE: 13.9067 R^2adj: 0.8988

Random Forest

The Random Forest R^2 and MSE for the 2004-08 period were 95.03% and 6.88, respectively, consisting again of an improvement over the GAM model. Figure 3 shows the relevant partial dependence plot which gives a graphical depiction of the marginal effect of each predictor variable on $PM_{2.5}$. In Figure 4 we can see the variance importance of each predictor in terms of both decrease in MSE and remaining error in predictive accuracy after a node split (node impurity).

Figure 3. Partial dependence plot (2004-08)

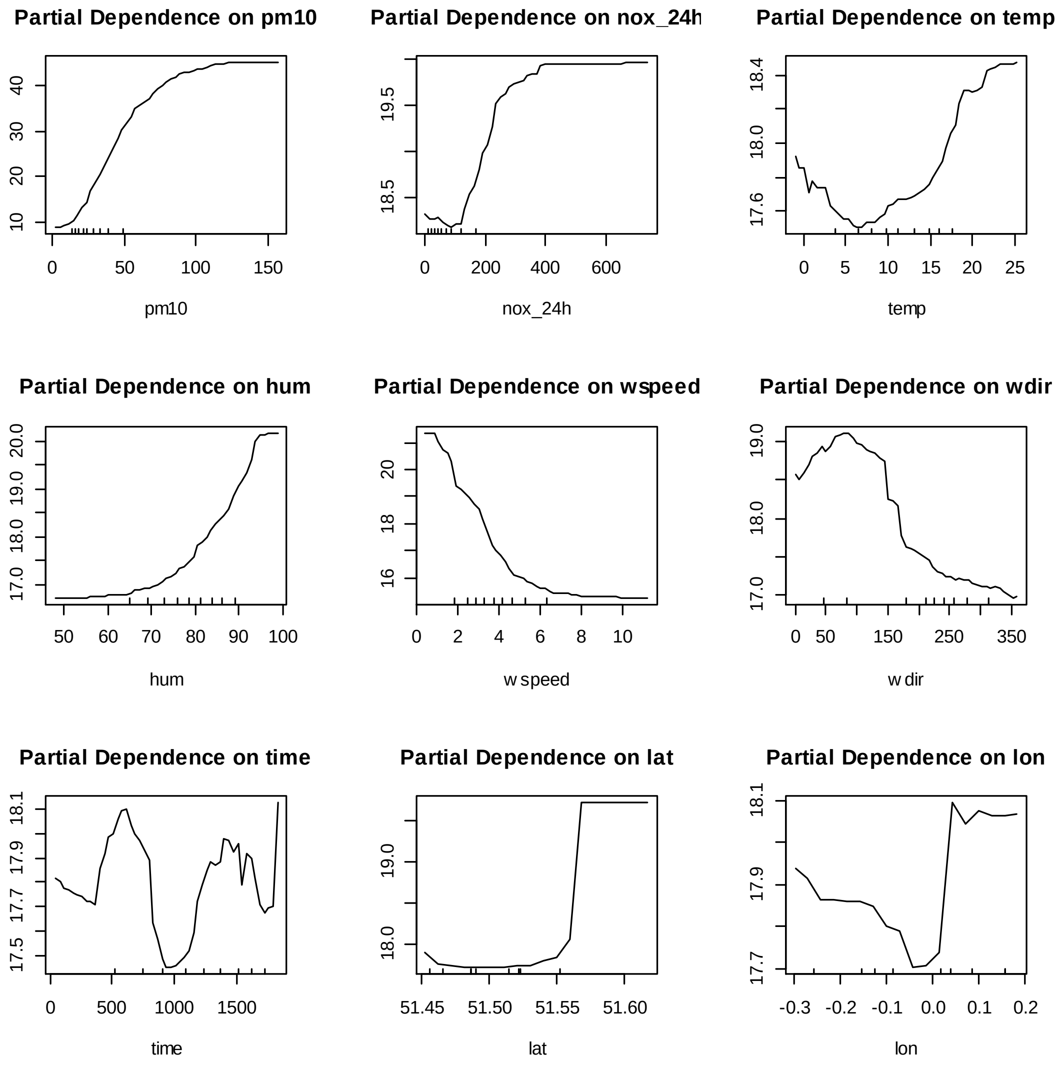
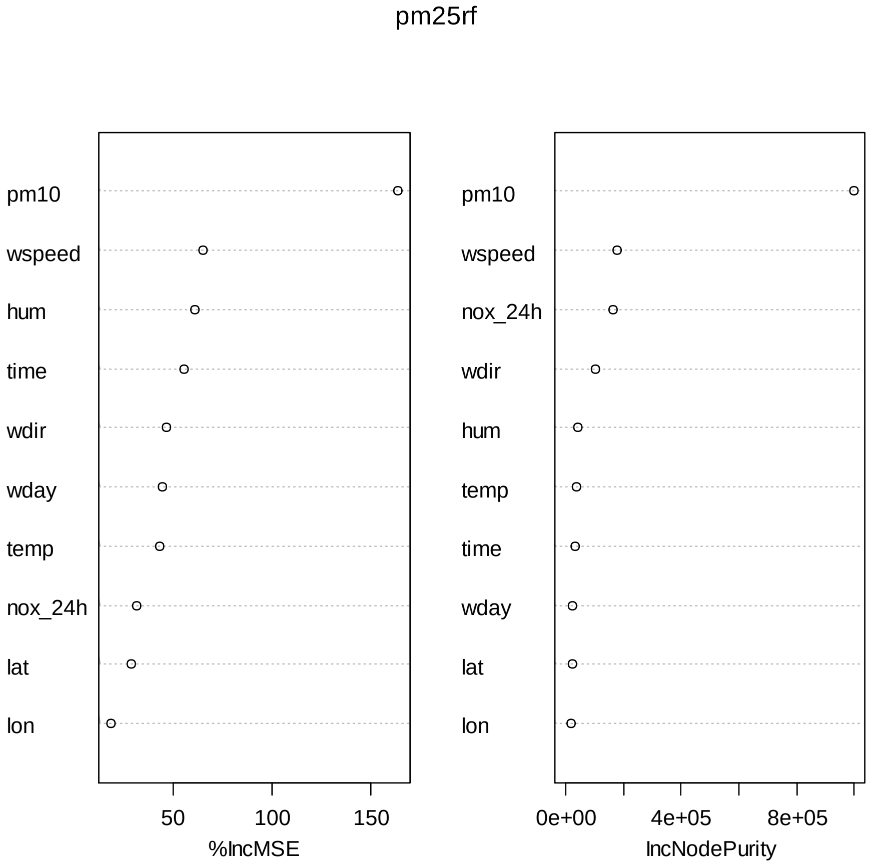


Figure 4. Variance importance plot (2004-08).



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 model was 99.2% and 98.9% for the time periods 2004-2008 & 2009-2013, respectively.

These predictions were included in the final enhanced $PM_{2.5}$ database that is made available.

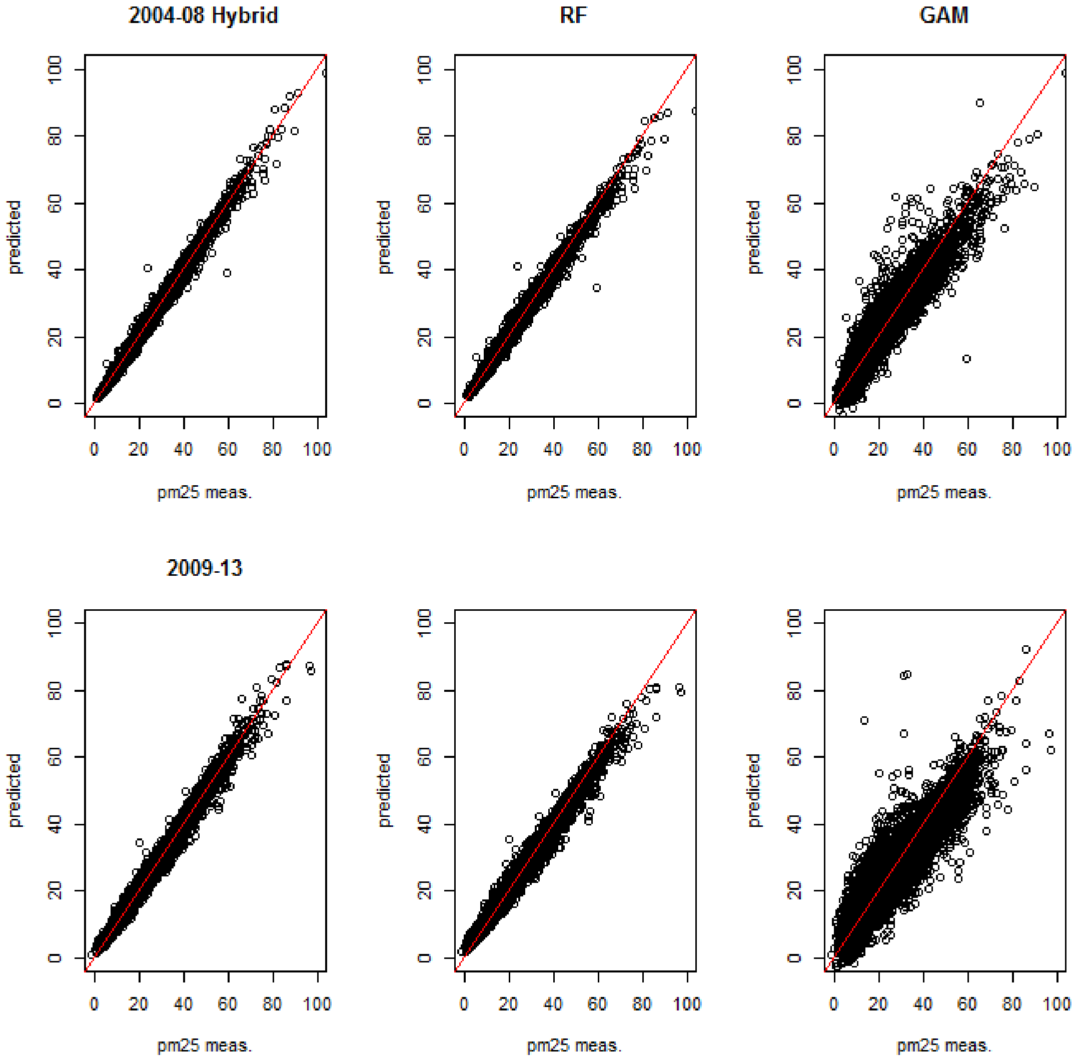
In Table 3 we can see the descriptive statistics for all predicted values of $PM_{2.5}$, all measured values and separately predicted values for sites with measured $PM_{2.5}$ (directly comparable with measured values) and predicted values for sites without $PM_{2.5}$ measurements (which provided only PM_{10} and NOx) for 2004-08 and 2009-13. We can see that during 2004-08 there were 16,957 $PM_{2.5}$ measurements and 102,800 additional estimates are made available whilst for 2009-13 the corresponding numbers are 40,083 and 85,436. It can be seen that measured and predicted means and medians are identical to the first decimal for 2009-13.

Figure 5 shows the agreement between the measured $PM_{2.5}$ and the predicted from each method and their combination. It can be seen that the combined methods predictions have a better agreement with measured values.

Table 3. Descriptive statistics for the observed & predicted $PM_{2.5}$ ($\mu g/m^3$) concentrations, when applying the random forest method, the regression model (GAM) & when combining the 2 methods into one model, for the time periods 2004 - 2008 and 2009 - 2013.

	Min	**25th percentile**	**Median**	**Mean**	**75th percentile**	**Max**	**N**
2004 – 2008							
A. $PM_{2.5}$ measurements	-0.77	9.29	13.99	17.45	22.17	113.61	16957
B. For days and sites where$PM_{2.5}$ measurements were available							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	1.49	9.50	14.47	17.81	22.65	108.78	11695
C. For days and sites where$PM_{2.5}$ measurements were NOT available but there were PM_{10} and NOx							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	-0.46	8.90	13.01	16.22	19.97	118.29	102800
D. All available predictions (sum of B & C)							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	-0.46	8.95	13.14	16.38	20.29	118.29	114495
2009 – 2013							
A. $PM_{2.5}$ measurements	-1.26	9.33	13.35	16.39	20.10	115.11	40083
B. For days and sites where $PM_{2.5}$ measurements were available							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	0.89	9.47	13.38	16.35	19.96	87.91	26645
C. For days and sites where $PM_{2.5}$ measurements were NOT available but there were PM_{10} and NOx							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	2.09	9.97	13.44	15.93	19.02	78.98	85436
D. All available predictions (sum of B & C)							
predicted $PM_{2.5}$ concentrations, applying the combination of both methods	0.89	9.86	13.42	16.03	19.26	87.91	112081

Figure 5. Observed vs predicted $PM_{2.5}$ for each period, by method (GAM, Random Forest and combined (hybrid)).


2004-08 Hybrid
RF
GAM
predicted
pm25 meas.