CS-BIGS 3(1): 1-17
http://www.bentley.edu/csbigs/documents/poggi.pdf

Three Non-Linear Statistical Methods for Analyzing PM_{10} Pollution in Rouen Area

François-Xavier Jollois
Laboratoire LIPADE, Université Paris Descartes, France

Jean-Michel Poggi
Laboratoire de Mathématiques, Orsay, & Université Paris Descartes, France

Bruno Portier
Laboratoire LMI, INSA de Rouen, France

The aim of this paper is to illustrate three modern statistical methods through a case study, which arises from a scientific collaboration between Air Normand on the applied side and Université Paris Descartes and the Institut National des Sciences Appliquées (INSA) in Rouen on the academic side. The problem is to analyze PM_{10} pollution during 2004-2006 in Rouen area using six different monitoring sites and to quantify the effects of variables of different types, mainly meteorological versus other pollutant measurements. Three methodologies – random forests, mixtures of linear models, and nonlinear additive models – are used in the analysis. In addition to the statistical interest of the study, we give detailed software oriented results and complete code using three freely available R packages.

Keywords: *PM_{10}; pollution; random forests; regression; classification; variable importance*

1. Introduction

The aim of this paper is to illustrate three modern applied statistical methods through a case study. Let us briefly sketch the context of the work. Suspended particles in the air are of various origins, natural or linked to human activity, and are of variable chemical composition. Air Normand, the observatory of air quality in Haute-Normandie, a coastal Northwest French area, has a network of a dozen stations measuring every quarter of an hour, sometimes going back to 10 years, the concentrations in PM_{10} particles (particles whose diameter is less than 10 μm). Other monitoring stations can have more recent data for smaller particles such $PM_{2.5}$ (with a diameter less than 2.5 μm), and some results on the composition of PM_{10} particles collected during time-limited specific studies.

European regulations stipulate that the PM_{10} daily average cannot exceeds 50 $\mu g/m^3$ more than 35 days per year. The objectives of this paper are organized around two axes: to characterize weather patterns leading to the extent of an exceedance through the joint statistical analysis of PM_{10} concentrations and meteorological parameters, and to distinguish situations in which the origin of the particles is mainly local or on the contrary distant or natural.

The analysis carried out in this paper is based on the so called TEOM (Tapered Element Oscillating Microbalance) PM_{10} concentrations from 2004 to 2006 measured by Air Normand, and the associated weather data provided by Météo France, the French national meteorological service.

Before giving the outline of the paper, let us say a few words of the bibliography about the statistical analysis of PM_{10} particles. It contains hundreds of references, so we prefer here to consider a few typical examples. They differ in their objectives and in the statistical tools used in the analysis.

Salvador *et al.* (2004) identify and characterize PM_{10} sources in Madrid using traditional statistical methods including Principal Component Analysis (PCA), linear regression and PCA regression. The pollution sources are identified via chemical measurements and back trajectories. Related to the same problem, let us mention the paper by Chavent *et al.* (2007) proposing in a case study a methodology for determining the apportionment of air pollution by source in a French urban site. Starting from measurements of chemical composition data, the sources are identified via a factor analysis and then the apportionment by source is completed by a receptor modeling based on a positive matrix factorization.

Karaca *et al.* (2005) deal with the statistical characterization of atmospheric PM_{10} and $PM_{2.5}$ concentrations at a non-impacted suburban site in Istanbul, using robust regression tools. A simple non linear model is fitted and the quality of the model is assessed carefully, identifying seasonal behavior and prevalent meteorological conditions.

Smith *et al.* (2001) focus on a problem close to the one examined in our paper: to identify factors influencing measurements of PM_{10} during 1995-1997 in London, using analysis of variance tools. From this point of view, our paper provides some methodological insights by using three modern non parametric statistical methods (random forests, mixture of linear models and nonlinear additive models) to investigate a similar problem.

The outline of the paper is the following. After this introduction, Section 2 states the context and briefly presents the data. Section 3 introduces and motivates the three main methods used to handle the problem. Section 4 is devoted to the use of random forests focusing on the relative importance of variables and variable selection issues as well as the marginal effects of variables. Section 5 uses two original climatic variables to partition the data and model each cluster using a partially nonlinear additive model. Section 6 explores cluster-wise linear modeling of PM_{10} pollution. Finally, Section 7 focuses on an attempt of a quantification of what we call in a broad sense a local part and a regional part of PM_{10} pollution. In addition the Appendix provides information about online material: full R code as well as the complete data set.

2. Data

2.1 Which pollution stations to select?

Among twelve monitoring stations for PM_{10}, we have selected a small group of six stations reflecting the diversity of situations. For the city of Rouen (see the map in Figure 1 to get an idea of its localization), we consider the urban station JUS, the traffic station GUI, the second most polluted in the region, and GCM which is located in an industrial area in order to have the widest panel. In Le Havre, we have kept the stations REP (the most polluted in the region) and HRI located at the seaside. Last, we focus on the station AIL near Dieppe, because it is rural and coastal, and a priori hardly influenced by social and industrial activity.

Grouping by categories: JUS and HRI are background urban monitoring sites, GUI and REP are urban sites close to traffic, GCM is industrial and AIL is rural. One can find in Figure 1 the map of the Haute-Normandie area and the localization of the different monitoring sites.

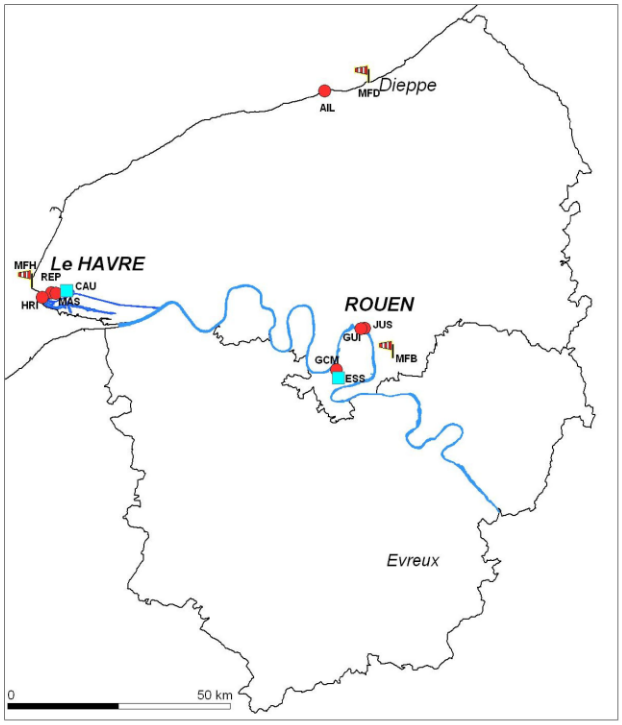

Figure 1. Map of the Haute-Normandie area locating the different monitoring sites of Air Normand and Météo France (blue squares: Air Normand temperature gradient monitoring sites, red dots: Air Normand monitoring sites, windsocks: Météo-France sites).

Table 1. Basic statistics about the daily TEOM PM_{10} concentrations (in $\mu g/m^3$) for 2004-2006

	JUS	GUI	GCM	AIL	HRI	REP
Min.	6	6	5	3	6	7
1st Qu.	16	20	14	13	15	21
Mode	16	23	13	17	14	24
Median	20	25	19	16	19	26
Mean	21.19	26.13	20.69	16.89	21.08	28.21
3rd Qu.	25	31	25	20	24	33
Max.	95	103	90	55	75	76
Std	8.5	9.8	8.9	6.1	9.3	9.9
NaN's	10	7	11	2	42	18

Table 2. Number of daily level exceedances

	JUS	GUI	GCM	AIL	HRI	REP
$\geq 25\mu g/m^3$	253	511	258	95	236	566
$\geq 30\mu g/m^3$	136	320	148	41	144	384
$\geq 50\mu g/m^3$	9	22	7	4	20	42
$\geq 75\mu g/m^3$	4	4	2			1
$\geq 85\mu g/m^3$	2	3	2			

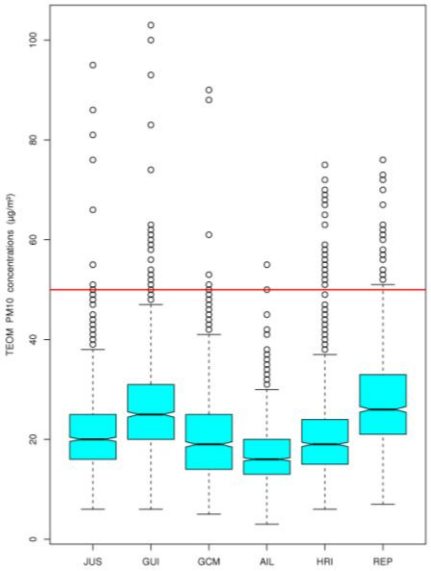

Figure 2. Boxplots of daily TEOM PM_{10} concentrations for the monitoring sites.

The analyzed data are the TEOM PM_{10} daily mean concentrations and concern the period 2004-2006 (1096 days). Table 1 contains some basic statistics about the PM_{10} concentrations coming from the six monitoring sites of Air Normand.

As expected, GUI and REP, which are both urban monitoring sites close to traffic, have a higher average level of pollution and AIL, which is a rural site and a priori not locally polluted, has the smallest one. For each monitoring site, the median is less than the mean which means that the distribution is not symmetric with some high values, as confirmed by the boxplots of Figure 2.

Table 2 of level exceedances shows some high disparities between the monitoring sites and highlights the small number of daily exceedances of 50 $\mu g/m^3$. Finally, let us mention that only the sites of Rouen measure PM_{10} concentrations greater than 85 $\mu g/m^3$.

2.2 Which meteorological predictors to consider?

To analyze PM_{10} concentrations, we consider daily meteorological data coming from three monitoring sites of Météo France, MFB at Rouen, MFD at Dieppe and MFH at Le Havre (see Figure 1). The different meteorological variables, which are calculated from hourly measurements during the period 0h-24h GMT, are the following:

- T.min, T.moy and T.max the daily minimum, mean and maximum temperature (in °Celsius), respectively;
- VV.max and VV.moy the daily maximum and mean wind speed (in m/s);
- PL.som the daily total rain (in mm);
- PA.moy the daily mean atmospheric pressure (in hPa);
- HR.min, HR.moy and HR.max the daily minimum, mean and maximum relative humidity (in %), respectively;
- DV.dom the most frequently observed wind direction (in °);
- DV.maxvv the wind direction associated with VV.max (in °).

We also have the temperature gradients GTrouen and GTlehavre measured by two monitoring sites of Air Normand, respectively denoted by ESS at Rouen and CAU at Le Havre in Figure 1. The temperature gradient is defined as the daily maximum of the hourly differences between the temperature at 3 meters altitude and the temperature at 180 meters altitude for ESS and 110 meters altitude for CAU.

Before entering into details about meteorological variables, let us mention that, in what follows, we associate to each pollution station the nearest meteorological station.

One can find in Table 3 some basic statistics about the variables coming from the Météo France monitoring site MFB located at Rouen, and in Table 4 some information about the temperature gradients, coming from the Air Normand monitoring network.

Tables 5 and 6 give the distributions of the wind direction variables DV.dom and DV.maxvv, measured at MFB.

Table 3. Basic Statistics of the Meteorological Variables at MFB in Rouen

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
T.min °C	-8.200	2.300	7.500	6.948	11.600	19.700
T.max °C	-2.00	8.50	14.90	14.44	19.90	34.60
T.moy °C	-5.23	5.47	11.28	10.70	15.63	26.00
VV.max m/s	2.00	5.00	6.00	6.51	8.00	16.00
VV.moy m/s	1.19	2.905	3.792	4.099	5.042	10.17
PL.som, mm	0	0	0	1.722	1.0	43.0
PA.moy, hPa	986	1012	1018	1018	1023	1041
HR.min %	19.00	52.00	64.00	64.17	76.00	99.00
HR.max %	71.00	93.00	95.00	94.05	97.00	99.00
HR.moy %	44.50	75.23	82.88	81.58	88.92	99.00

Table 4. Basic Statistics about the Temperature Gradients

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
GTrouen °C	-1.5	0.1	1.2	1.649	2.8	10.3
GTlehavre °C	-1.1	-0.3	0.3	0.6412	1.3	6.4

Table 5. Distributions of the Wind Direction DV.dom, at MFB

Direction	N 0°	NNE 22.5°	NE 45°	ENE 67.5°	E 90°	ESE 112.5°	SE 135°	SSE 157.5°
Frequency	113	61	74	54	42	36	35	71
Direction	S 180°	SSW 202.5°	SW 225°	WSW 247.5°	W 270°	WNW 292.5°	NW 315°	NNW 337.5°
Frequency	116	77	70	48	149	67	14	44

Table 6. Distribution of the Wind Direction DV.maxvv, at MFB

Direction (°)	10	20	30	40	50	60	70	80	90	100	110	120
Frequency	28	40	30	39	41	26	24	24	16	17	27	7
Direction (°)	130	140	150	160	170	180	190	200	210	220	230	240
Frequency	18	13	40	40	11	18	43	60	50	42	34	32
Direction (°)	250	260	270	280	290	300	310	320	330	340	350	360
Frequency	33	21	41	61	39	26	20	18	23	20	41	30

One can find on the diagonal plots of Figures 3 to 5 the histograms (with in addition the estimated densities except in Figure 4) of the main meteorological variables measured at Rouen. Let us make two comments. The distribution of the total rain PL.som is very asymmetric. The distributions of T.min, T.moy, T.max, HR.min and PA.moy are almost symmetric while the distributions of the other variables are asymmetric.

One can also find scatterplots in the off-diagonal plots of Figures 3 to 5. Let us examine more carefully those involving PM_{10} concentration and the different meteorological variables measured in Rouen. We can note on Figure 3 that the high ***pollution*** episodes (i.e. daily PM_{10} concentrations exceeding $50\mu g/m^3$) only arise for low temperatures. On Figure 4, the higher the rain, the smaller the PM_{10} concentration. This remark also holds for the wind speed. In addition, episodes appear only for low wind speed and when there is no rain. Finally, we can see on Figure 5 that the lower the atmospheric pressure, the smaller the PM_{10} concentration, and episodes only arise for high atmospheric pressure.

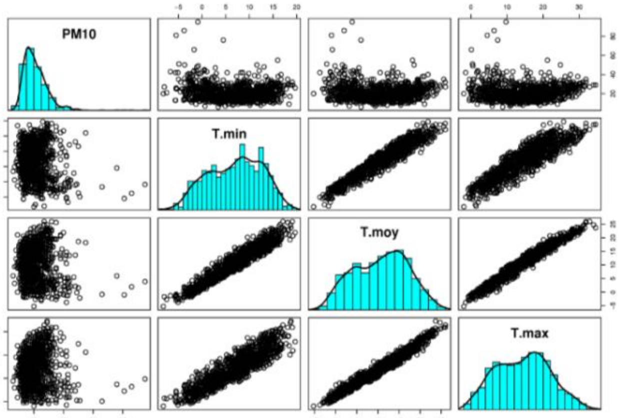

Figure 3. Histograms and scatterplots of PM_{10} at JUS and the temperatures T.min, T.moy and T.max at Rouen.

2.3 Which pollutants in addition to PM_{10}?

In addition to PM_{10}, three other pollutants are measured: NO, NO_2 and SO_2. Nitrogen oxides NO and NO_2 are retained as markers of social activity and are especially related to traffic while sulfur dioxide SO_2 captures the consequences of industrial activity.

In order to get a fair picture of local pollution, it seems interesting to augment the data for three of the six stations by using pollutants measured nearby in order to have a complete picture for pollutant observations. More precisely, at GUI the SO_2 is not measured, so we use the SO_2 data collected at JUS. At Le Havre, we complete the data coming from REP with the SO_2 data measured at MAS, which is a background urban monitoring site located in the center of Le Havre (see Figure 1), and we

complete the data coming from HRI with the NO and NO_2 data also measured at MAS. Finally, let us mention that for the rural station AIL, there is no available measurements of these three pollutants.

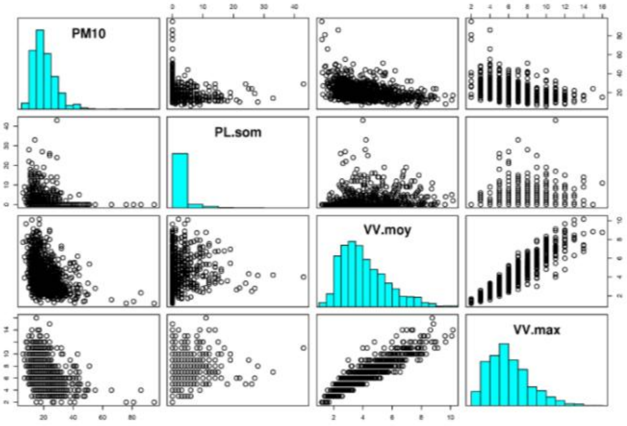

Figure 4. Histograms and scatterplots of PM_{10} at JUS and the meteorogical variables PL.som, VV.moy and VV.max at Rouen.

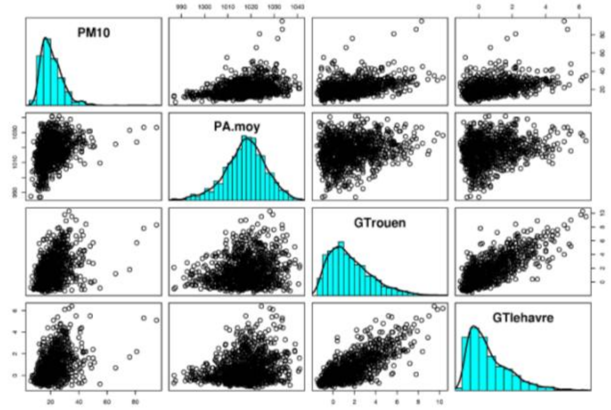

Figure 5. Histograms and scatterplots of PM_{10} at JUS and the meteorological variables PA.moy at Rouen, GTrouen and GTlehavre.

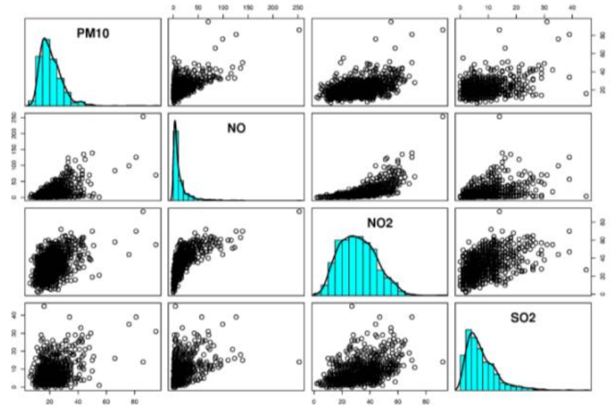

Figure 6. Histograms and scatterplots of the daily concentrations of the pollutants at JUS.

2.3 Which pollutants in addition to PM_{10}?

In addition to PM_{10}, three other pollutants are measured: NO, NO_2 and SO_2. Nitrogen oxides NO and NO_2 are retained as markers of social activity and are especially related to traffic while sulfur dioxide SO_2 captures the consequences of industrial activity.

In order to get a fair picture of local pollution, it seems interesting to augment the data for three of the six stations by using pollutants measured nearby in order to have a complete picture for pollutant observations. More precisely, at GUI the SO_2 is not measured, so we use the SO_2 data collected at JUS. At Le Havre, we complete the data coming from REP with the SO_2 data measured at MAS, which is a background urban monitoring site located in the center of Le Havre (see Figure 1), and we complete the data coming from HRI with the NO and NO_2 data also measured at MAS. Finally, let us mention that for the rural station AIL, there is no available measurements of these three pollutants.

Let us give some elements about the distributions of the different pollutants at the station JUS of Rouen, as well as the correlations with the PM_{10}. As it can be seen in Table 7 and Figure 6, the pollutant distributions are very asymmetric with few extreme values. Figure 6 shows the different scatterplots. In particular, we observe that if the pollution levels of NO, NO_2 and SO_2 are high, then the pollution level of PM_{10} is also high.

Table 7. Basic Statistics of the Pollutant Concentrations (in µg/m3) Collected at the Urban Monitoring Site JUS of Rouen

	PM_{10}	NO	NO_2	SO_2
Min.	6.00	0.00	2.00	0.00
1st Qu.	16.00	3.00	21.00	4.00
Median	20.00	7.00	31.00	6.00
Mean	21.19	13.43	31.45	7.83
3rd Qu.	25.00	16.00	40.00	10.25
Max.	95.00	253.00	92.00	45.00
NA's	10	6	7	12

3. Models

3.1 Which models to use?

Random forests are a very powerful method for prediction and variable importance quantification. By computing the marginal effects of each variable on the PM_{10} pollution, we get a rough idea of the shape of its influence, distinguishing pollutants and climatic variables. In addition, variable importance scores allow one to identify the most influential variables. However a random forest does not define an explicit model since it builds a prediction model which is an aggregation of regression trees. So two additional models are then considered. Both are regression models by classes built according to different principles.

The first one is based on generalized additive models and proposes weather type dependent nonlinear additive models (in fact partially linear if some components are linearizable). The classes are explicit and related to weather types (three in general) but they are rigid since they are based on only two variables selected *a priori:* rain and wind direction. Indeed they lead to explicit and easy to understand classes with predictors that have a highly nonlinear effect on PM_{10}.

The second one is based on mixtures of linear models and builds also clusterwise linear models but the building strategy combines more closely clustering and regression fitting: the classes are unknown as well the model in each class and the whole model is optimized using an iterative algorithm. This model makes for a more flexible classification as well as simpler models within a class but of course the classes are less directly interpretable.

Let us briefly introduce the principle of each model in the next three paragraphs before using them to model PM_{10}.

3.2 Random forests

Random forests (RF henceforth) is a popular and very efficient algorithm, based on model aggregation ideas, for both classification and regression problems. Introduced by Breiman (2001), the principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the set of observations and choosing randomly at each node a subset of explanatory variables. More precisely, with respect to the well-known CART (see Breiman *et al.* 1984) model building strategy performing a growing step followed by a pruning one, the two main differences are: first, at each node, a given number of input variables are randomly chosen and the best split is calculated only within this subset and second, no pruning step is performed so all the trees are maximal trees.

In the random forest framework, the most widely used score of importance of a given variable is the increase in the mean of the error of a tree in the forest when the observed values of this variable are randomly permuted in the so-called out-of-bag sample. The higher the importance, the stronger the variable influence.

The associated R package is `randomForest` which is based on the initial contribution of Breiman and Cutler (2005) and is described in Liaw and Wiener (2002). It appears to be very powerful in a lot of applications (see Genuer *et al.* 2008).

This method is applied to the pollution data set in Section 4.

3.3 Non-linear additive models

Nonlinear additive regression models of the form

$$Z_n = \sum_{i=1}^{d} f^i\left(X_n^i\right) + \mu + \varepsilon_n \qquad (1)$$

where Z is a real-valued dependent variable, $X^1, \dots, X^d$ are the explanatory variables, μ is a constant and ε is an unobservable noise, have been widely used and studied since the pioneer work by Breiman and Friedman (1985), Buja *et al.* (1989) and Hastie and Tibshirani (1990). Such models are particularly attractive since they represent an interesting compromise between the classical linear model $Z_n = \sum_{i=1}^{d} \theta^i X_n^i + \mu + \varepsilon_n$, and the fully non-parametric one $Z_n = \psi(X_n^1, \dots, X_n^d) + \varepsilon_n$. Indeed with respect to the linear model, considerable additional flexibility is given by the allowed nonlinear effect of each explanatory variable without losing ease of interpretation. In addition with respect to the fully nonparametric model, the separable model (1) is more explicit and can be estimated without suffering from the so-called curse of dimensionality (Stone 1986) which is the main drawback of the unstructured nonparametric regression model.

The associated R package is `mgcv` developed by Wood (2006). The nonlinear functions are estimated using penalized regression splines. This method is applied to the pollution data set in Section 5.

3.4 Mixture of linear models

Finite mixture models are classical in statistics (see McLachlan and Peel 2000) and have been recently extended by mixing standard linear regression models. The main hypothesis is that observations come in some unknown proportions from a mixture of s components, which are modeled by linear models. Then, the purpose is to estimate the parameters of each linear model. In the clustering context, each object is supposed to be generated by one of the components of the mixture model being fitted. The partition is derived from these parameters using the maximum *a posteriori* (MAP) principle from the posterior probabilities for an object to belong to a component.

Finite mixture models with a fixed number of components are usually estimated with the expectation-maximization (EM) algorithm within a maximum likelihood framework (Dempster *et al.* 1977). This algorithm iteratively repeats two steps until convergence. The first step *E* computes the conditional expectation of the complete log-likelihood, and the second one M computes the parameters maximizing the complete log-likelihood.

In real applications the number of components is unknown and has to be estimated. A classical approach is to then fit models with an increasing number of components and to compare them using the BIC criterion (Schwarz 1978).

To compute mixture of linear regressions, we use the `flexmix` R package, described in Leisch (2004), and Gruen and Leisch (2007). This method is applied to the pollution data set in Section 6.

4. Random Forests

For each model, one can find in sections 4 to 6 a quite similar outline. First, we briefly introduce the general strategy and the main commands to fit the model using the corresponding R package, with data from the station JUS. Then, the estimated models across stations are considered for description and comparison purposes.

4.1 Random forest for JUS

The first command allows to build a random forest from the data of the JUS station using $mtry=\sqrt{p}$ input variables randomly chosen among the p original variables at each split and using $ntree = 500$ trees in the forest:

```
res  <-  randomForest  (formula  (jus_comp),
       data = jus_comp , importance = TRUE )
```

The training performance is given by the explained variance which is about 58%.

The second command allows one to compute the estimated marginal effect of the variable x_i. It is obtained by mimicking partial integration of the forest, *i.e.* the predicted regression function, using a sum along k grid-points as:

$$\tilde{f}_i(x) = \frac{1}{k}\sum_{i=1}^{k} f\left(x, \underline{x_i}\right)$$

where x_i stands for the variable for which partial dependence is evaluated, and $\underline{x_i}$ stands for the other variables. The following command computes the marginal effect for NO variable, presented in Figure 7:

```
partialPlot  (res,  jus_comp  ,  'NO',  main  =
       'Marginal effect - NO')
```

The third command (`importance(res)`) allows to extract from the forest the score of permutation importance of a given variable as the increase in the mean of the prediction error of a tree in the forest (MSE) or a second measure (not used here) based on the total decrease in node impurities (residual sum of squares in regression). Figure 8 is obtained by:

```
varImpPlot  (res,  sort  =T,  main  =  'RF  -
       Variable importance ')
```

4.2 Marginal effects across the stations

The typical useful estimated marginal effects of the main explanatory variables are collected in Figure 9. The marginal effects of pollutants are still increasing as expected, looking roughly linear, or slightly convex or concave depending on the pollutant and the station. The marginal effect of rain is decreasing and reflects a washing effect. The effects of temperature are nonlinear: positive when it is cold or hot, zero or negative for medium temperature.

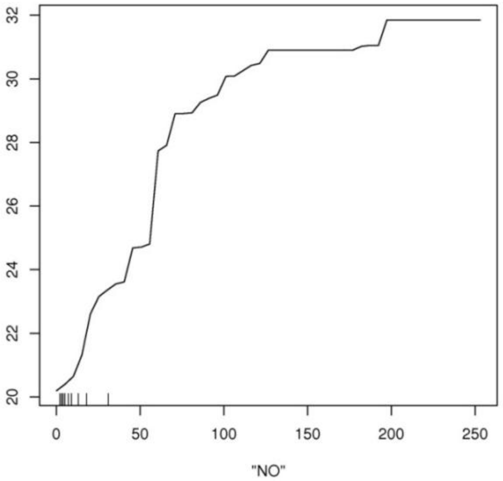

Figure 7. Marginal effect of NO on PM_{10}, measured at JUS

For wind direction effects, there are two situations (see Figure 10). For the stations GCM, JUS, AIL, it shows a major east/west axis, while for the stations GUI, REP, HRI, it highlights a north/south inhibitor axis.

For the effects of wind speed, we must distinguish three situations (see Figure 11): inhibitor for stations downtown Rouen JUS, GUI, decreasing (then slightly increasing outside the city of Rouen) for GCM, REP, HRI and finally increasing for the rural station AIL. The behavior for AIL is the only one which is, at first sight, surprising. It is due to the lack of local pollution sources combined with a low global level of pollution which lead to consider imported pollution as the major source of pollution at AIL, explaining the amazing increasing effect of wind speed.

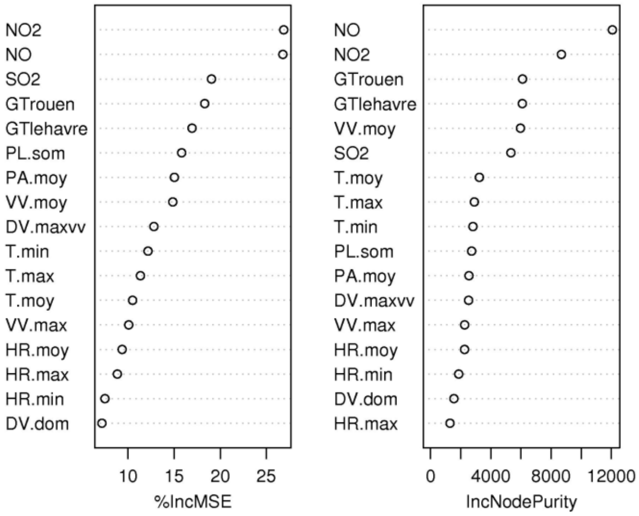

Figure 8. Variable importance for JUS

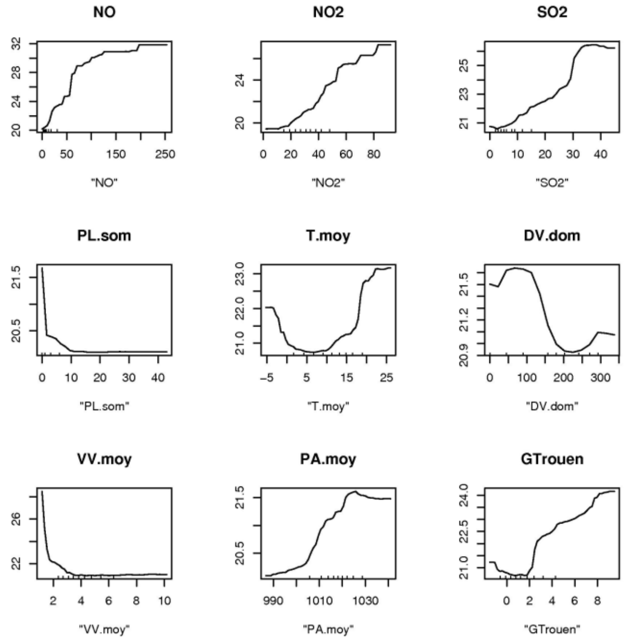

Figure 9. Typical marginal effects on PM_{10} of main explanatory variables

The effects of relative humidity are of very low importance and are very rarely useful in the remainder of this paper. Note that the minimum and maximum effects are different: increasing and decreasing respectively.

The marginal effect of atmospheric pressure is increasing as well as the ones of temperature gradient as expected. In addition, the temperature gradients of Le Havre and Rouen give very similar information and are equally important.

So the main conclusion at this stage is twofold. The marginal effects of pollutants are increasing, and can be considered as good markers of local pollution effects useful for PM_{10} modeling. For the meteorological variables we distinguish (eliminating HR which is unimportant) those who have a rather decreasing effect on pollution: PL and VV (for all stations except AIL); those who have a favorable effect (increasing): GT, PA and VV (for AIL); and those having a non monotonic effect: T and DV.

4.3 Random forest variable importance across stations

For each of the six stations we calculate the individual variable importance considering a random forest model involving all the variables. One can find here a synthesis organized following the results of the previous section.

To define the importance of a group of variables, we use the following rule (proposed by Genuer *et al.* 2008): if the variables are sufficiently redundant, or more weakly, have similar effects on PM_{10} concentrations then the importance of a group is defined as the maximum importance. Thus, we can directly compare groups of variables. So, considering the four groups given by the pollutants (separating NO_x from SO_2) and the meteorological variables, distinguishing those having a negative impact or positive or mixed respectively, we obtain the results of Table 8:

The importance of the three groups of weather variables are nearly equal to 20, both for a given station and between the stations, except for that of PA at GUI which reached 30. However, the importance of pollutants fluctuates much more: from 29 to 49 except 0 for AIL since no measurement of pollutants is done.

In both traffic stations (GUI and REP) NO_2 dominates and its importance takes similar values (41 and 45). For the industrial station (GCM), the importance of pollutant SO_2 is about 39, which is high. For both urban background stations (JUS and HRI), we observe a different importance for SO_2: 28 for JUS (where NO_2 slightly dominates) and 49 for HRI (where NO_2 measured at MAS is less important).

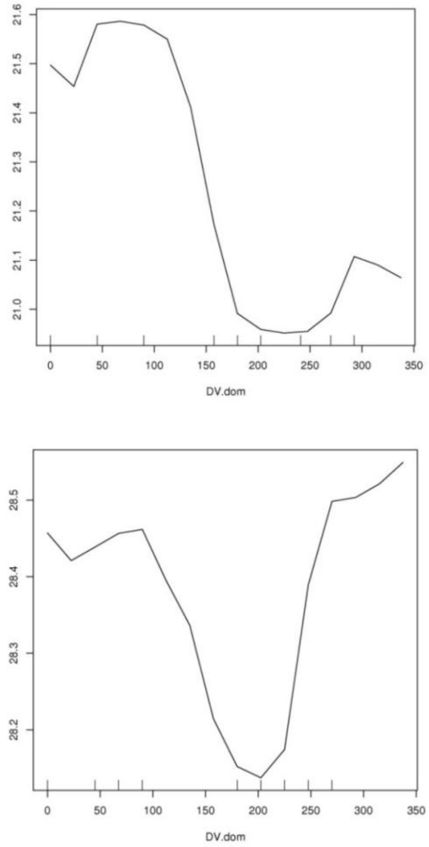

Figure 10. Marginal effects on PM_{10} of wind direction. On the top JUS, on the bottom REP.

Table 8. Random forest variable: importance across stations

	Pollutant	Neg.	Meteo. Pos.	Mixed
GCM Rouen, industrial	39 (SO_2)	23 (PL)	20 (GT)	21 (T)
JUS Rouen, urban	28 (NO_2) 22 (SO_2)	19 (PL)	21 (GT)	19 (DV)
GUI Rouen, traffic	41 (NO_2) 18 (SO_2)	18 (PL)	32 (PA)	19 (T)
AIL Dieppe, rural		15 (PL)	23 (PA)	21 (DV)
REP Le Havre, traffic	45 (NO_2) 31 (SO_2)	20 (VV)	24 (GT)	19 (T)
HRI Le Havre, urban	49 (SO_2) 22 (NO_2)	16 (VV)	18 (GT)	22 (T)

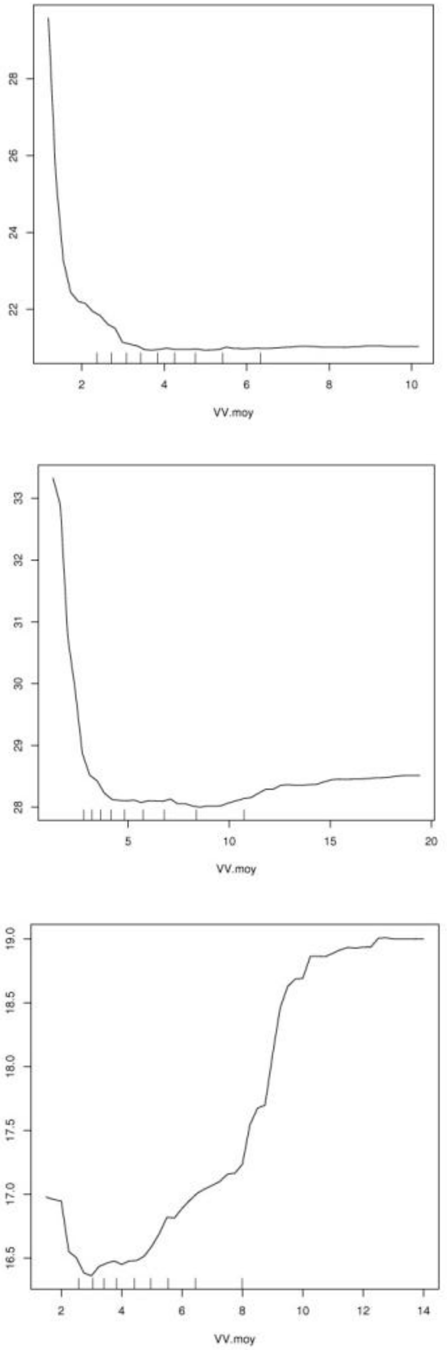

Figure 11. Marginal effects on PM_{10} of wind speed (from top to bottom). Graphs show, in turn, JUS, REP and AIL.

5 Conditional nonlinear additive models

5.1 Conditional nonlinear additive models for JUS

For each station, the approach is at first to partition the days of years 2004 to 2006, depending on weather patterns based on the quantity of rain (variable PL.som) and the wind direction (variable DV.maxvv). Recall that these two meteorological variables are preferred on the one hand for their segmentation power on the average level of PM_{10} pollution and on the other hand for their highly nonlinear effect on PM_{10}. This segmentation step is performed using the CART method and the associated R package rpart to build a binary regression tree explaining PM_{10} concentration by the two selected weather variables. The optimal splits obtained by maximizing the decrease in explained variance define weather types as well as a partition of the set of days. Three classes are defined but they may differ between stations.

The classes obtained for JUS come from the regression tree given by Figure 12.

Since the class of days with rain and wind direction DV.maxvv less than 185° is too small to correctly fit a nonlinear additive model, we prefer not to split the class of rainy days and therefore only consider three classes. The resulting clusters are characterized by the elements given in Table 9. The distributions of PM_{10} across the classes, given in Figure 13, appear to be well separated.

In a second step, a nonlinear additive model is fitted for each weather type and the global model is given by the three conditional models together with the weather type definition.

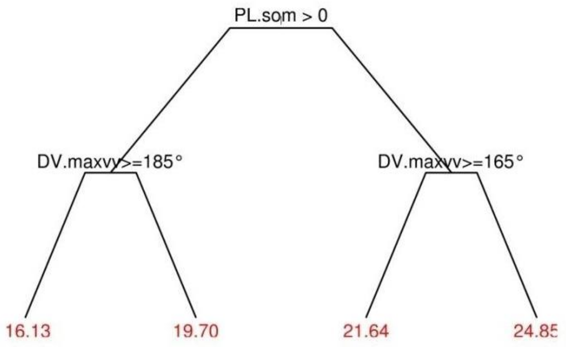

Figure 12. Station JUS: regression tree fitting the PM_{10} concentration using PL.som and DV.maxvv

Table 9. Station JUS – means and frequencies of PM_{10} in CART clusters

	PL.som>0	PL.som=0	
		165° ≤ DV.maxvv	DV.maxvv < 165°
Mean	17.05	21.64	24.85
Frequency	349	353	330

For a given weather type, a sub-model is built according to the following strategy. We start by fitting a nonlinear additive model involving all the explanatory variables and then apply the following descending strategy:

1. elimination of variables whose effects are considered as non significant;
2. elimination of redundant variables (even if they have a significant effect);
3. iterate steps 1 and 2 if necessary;
4. linear modeling of weakly nonlinear effects.

Of course, such strategy cannot be applied automatically without caution, as usual in such descending variable selection approach.

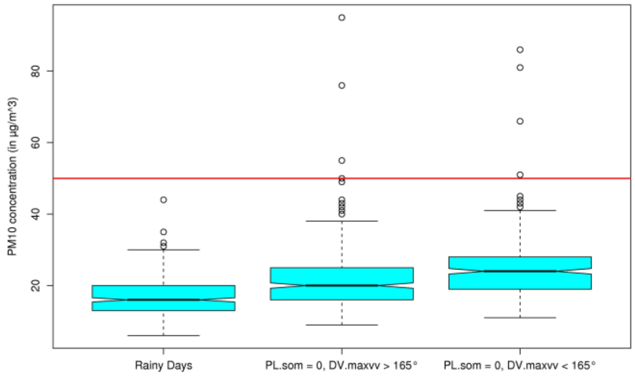

Figure 13. Boxplots of PM10 concentrations across clusters, JUS station, Rouen.

For example, let us consider the class of days without rain (PL.som=0) and a wind direction DV.maxvv less than 165°. The construction strategy is as follows. We begin with the R instruction:

```
res <- gam (PM10 ~ s(NO) + s(NO2) + s(SO2) +
s(T.max) + s(T.min) + s(VV. moy) + s(VV. max) +
s(PA. moy) + s(GTlehavre) + s(GTrouen) + s(HR.max)
+ s(HR. moy ) + s(HR.min) + s(DV. maxvv) + s(DV.
dom), data = jus_comp, subset = (DV.maxvv <165) &
(PL. som ==0))
```

and we proceed as follows:

1. successively eliminate the explanatory variables GTlehavre, GTrouen, VV.moy, HR.max, T.min, DV.dom, HR.min;
2. linearly model the effects of VV.max and NO2.

The final model is obtained using the R commands:

```
res2  <-  gam  (PM10  ~  s(NO)  +  NO2  +  s(SO2)  +
s(T.max)  +  VV.max  +  s(PA.  moy)  +  s(HR.  moy)  +
s(DV.  maxvv),data  =  jus_comp,  subset  =  (DV.maxvv
<165) & (PL. som ==0))
```

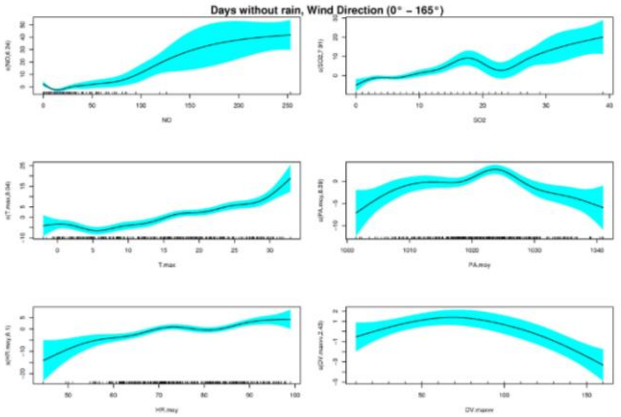

Figure 14. Station JUS: days without rain and wind direction DV.maxvv<165°. Nonlinear effects of NO, SO2, T.max, PA.moy, HR.moy and DV.maxvv.

The estimated nonlinear effects of the variables NO, SO_2, T.max, PA.moy, HR.moy and DV.maxvv are presented in Figure 14. Except for HR.moy, they are comparable to the marginal effects obtained with random forests.

The performance of the overall model is evaluated using two criteria: the explained variance given by the variance of estimated PM_{10} divided by the variance of observed PM_{10}, and the explained deviance equal to 1 minus the variance of residuals divided by the variance of observed PM_{10}. Let us note that, for the linear regression model, both quantities coincide. In addition, as usual the plot of (observed PM_{10}, estimated PM_{10}) allows to assess visually model quality and in particular the estimation quality for episodes, *i.e.* daily PM_{10} concentrations exceeding 50 $\mu g/m^3$.

We find in Figure 15 the plots of (observed PM_{10}, estimated PM_{10}) associated with the three submodels and the global one. These plots are of good quality. We can also mention that many episodes are well estimated, and only two episodes are badly estimated, with predicted values less than 30 $\mu g/m^3$.

As mentioned by a reviewer, the top left diagram seems to show not only more scatter than with the other variables, but also some bias and it seems to indicate that perhaps a three state rainfall variable, no rain, light rain and heavy rain might be more useful. In fact, it is not the case and this probably comes from the fact that such short-scale statistical models necessarily omit some variables, which are better taken into account using dynamical chemical models requiring much more data and a more extensive modeling effort.

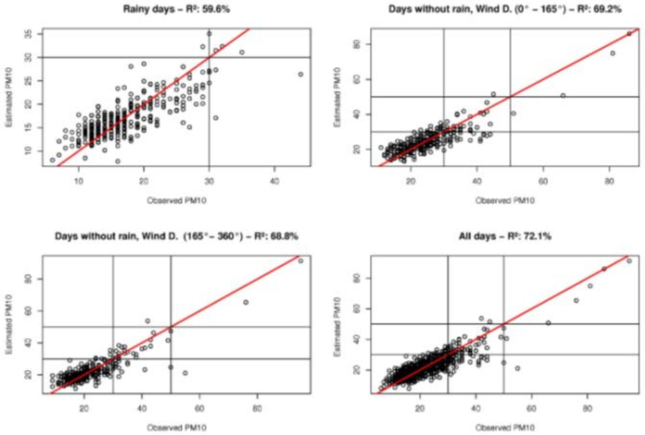

Figure 15. Station JUS, plots (observed PM10, estimated PM10).

Let us describe (without details) this short suggested additional investigation in order to improve the model associated with the rainy days. Two GAM models have been fitted with a variable PL.bin, which is a factor based on PL.som:

- the first one with PL.bin = 1 if PL.som ≤ 9, 2 if 9 < PL.som ≤ 19 and 3 otherwise;
- the second one with PL.bin = 1 if PL.som ≤ 6 and PL.bin = 2 otherwise, where the value of the threshold 6 is obtained using CART.

For each model, the performance is not improved, the percentage of explained variance is the same and the bias remains.

5.2 Conditional GAM across stations

Table 10 gives some characteristics to compare conditional GAM models across stations.

It should be noted that, except for stations HRI and AIL located on the seafront, the most discriminant variable, defining the first regression tree split, is always rain PL.som. The obtained weather types are close to each other: we distinguish rainy days and days without rain, and for these last ones, we separate wind direction from east and from west. The percentage of explained variance is good for stations JUS and REP. It becomes good for GUI and HRI when the SO_2 measured at JUS and the NO_x measured at MAS respectively are added to the model. They are less good for stations GCM and AIL. Probably, for these stations one important piece of information is missing: loads at the cereal grain port located at Rouen north and near the station GCM and, for AIL, one piece of information about air mass movements.

Table 10. Conditional GAM models across stations: weather types and performance.

Station	First split variable	Partition	Explained variance

JUS	PL.som	PL.som > 0
		PL.som=0, DV.maxvv < 165°
		PL.som = 0, DV.maxvv ≥ 165°
			72.1 %

GUI	PL.som	PL.som > 0
		PL.som=0, DV.maxvv < 180°
		PL.som = 0, DV.maxvv ≥ 180°
			66.6 %

		SO_2 measured at JUS significant	70.2 %

GCM	PL.som	PL.som > 0
		PL.som=0, DV.maxvv < 135°
		PL.som = 0, DV.maxvv ≥ 135°
			55.1 %

REP	PL.som	PL.som > 0
		PL.som=0, DV.maxvv < 155°
		PL.som = 0, DV.maxvv ≥ 155°
			70.7 %

		SO_2 measured at MAS weakly significant	72.8 %

HRI	DV.maxvv	PL.som > 0
		PL.som=0, DV.maxvv < 155°
		PL.som = 0, DV.maxvv ≥ 155°
			66.8 %

		NO_x measured at MAS significant especially for the estimation of episodes	71.6 %

AIL	DV.maxvv	PL.som > 0
		PL.som=0, DV.maxvv < 165°
		PL.som = 0, DV.maxvv ≥ 165°
			37.9 %

One can find in Table 11 a summary of the structure of models for each of the six stations. Three weather types are considered: denoted by 1 for rainy days, 2 for days without rain and wind from the east and 3 for days without rain with westerly wind. A cell of the table is empty if the corresponding effect is non-significant and the variable is not involved in the final model, and it contains l or n if the effect is significant and linear (l) or non linear (n).

It should be noted that when a pollutant is present, it is present for all weather types. In addition, for stations GUI, REP and HRI for which the three pollutants are not systematically measured, if the data are augmented using pollutants measured in nearby stations, these pollutants are significant and therefore present in the models. The minor changes caused by introduction of new data are summarized in the Table 12 and appear in capital letters.

Finally, concerning the general shape of the individual effects, except for those which are linearized through weather classification, non-linear estimated effects are similar to those previously obtained using random forests.

Table 11. Structure of conditional GAM models across stations.

	JUS			GUI			GCM			REP			HRI			AIL		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
NO	l	l	n	n	n	l	-	-	-	n	n	n	-	-	-	-	-	-
NO_2	l	n	l	n	l	l	-	-	-	l	n	l	-	-	-	-	-	-
SO_2	l	n	n	-	-	-	n	n	n	-	-	-	n	n	l	-	-	-
T.max		n	n			l	n					n	n		n	n	n	N
T.min	n			n	n			n	n	n				n			n	
VV.moy					n	l		n	n		n		l	n			n	N
VV.max	n	l		n			n			n		n			n	n		
PA.moy	l	n		l	n	l	l	n	l	l	n		l		l	l		N
GTlehavre			n	n				l	n	n	l	n	n			n		
GTrouen	n					l	n		l					n	n			N
HR.max							l										n	
HR.moy	n	n		n		l		n		l						l		N
HR.min					l						n			l			n	
DV.maxvv	n	n	n			l	n	l				n				n		
DV.dom										n			n		n	n	n	
PL.som																		

Table 12. Structure of conditional GAM models across stations. Data augmentation for GUI: SO_2 at JUS, for REP, SO_2 at MAS and for HRI, NO_x and SO_2 at MAS.

	JUS			GUI			GCM			REP			HRI			AIL		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
NO	l	l	n	L	N	l	-	-	-	n	n	L	N	N	N	-	-	-
NO2	l	n	l	L	l	l	-	-	-	l	n	l	N	N	L	-	-	-
SO2	l	n	n	N	N	N	n	n	n	L	L	N	L	n	l	-	-	-
T.max		n	n		N	N	n					L	n		n	n	n	n
T.min	n			n				n	n	n				n			n	
VV.moy						N		n	n		n		l	n			n	n
VV.max	n	l		n	L		n			n		n			n	n		
PA.moy	l	n		l	n	N	l	n	l	l	n		N	L	l	l		n
GTlehavre	N		n	n				l	n	n	l	n	n	N		n		
GTrouen						N	n		l						n			n
HR.max							l										n	
HR.moy	n	n		n	L	l		n		l						l		n
HR.min											n			N			n	
DV.maxvv	n	n	n		N	l	n	l		N		n				n		
DV.dom													n		n	n	n	
PL.som										N								

6. Mixture of linear models

6.1 Mixture of linear models for JUS

In this section, we build a model consisting of the construction of several classes with a linear model for each class. The classes (and the linear models) are obtained to better adjust the global model to data. The optimal number of classes is also automatically selected using a penalized criterion making a tradeoff between model fitting and model complexity. The method is based on a mixture of linear regression models. The principle is given by Gruen and Leisch (2007) and the corresponding R implementation in 2004.

The first result is that, whatever the station, the number of classes is always two. In addition, the two models can then easily be interpreted since the data have been standardized before any segmentation. Indeed, the value and the sign of the intercept allow to easily qualify the classes: positive (or negative) means more (or less) polluted than the average station and the absolute value gives the intensity.

So, one class can be interpreted as the most polluted one and the analysis of the associated model allows to characterize these days while the comparison between the two models captures the differences between the two situations.

For each station, the results are collected in three figures. The first one presents the values of each criterion for choosing the number of clusters (for JUS, see Figure 16), the second one contains a bar-charts representation of the two linear models (for JUS, see Figure 17), and the last one allows to assess the whole model giving the estimated *vs.* observed diagram (for JUS, see Figure 18). For the station JUS, we proceed as follows:

1. First, scale data:
   ```
   x = as.data.frame (scale(jus_comp))
   ```
2. Compute models with k = 1,…,7 clusters:
   ```
   res = stepFlexmix(formula(x), data = x, k =
   1:7, nrep = 10, control = list(iter.max = 500))
   ```
3. Select the best model according to the BIC criterion:
   ```
   resBIC = getModel (res , 'BIC ')
   ```
4. Finally, compute the parameter confidence bounds:
   ```
   resBICfit = refit ( resBIC )
   ```

6.2 Mixture of linear models across stations

Let us first characterize what we call the most polluted class. Table 13 contains the most significant pollutants and weather variables in the corresponding model. We can first notice that the number of days belonging to this class is relatively small (between 57 and 226). For the stations where the three pollutants are measured, NO appears each time with either NO_2 (for REP), or SO_2 (for JUS, GUI and HRI). The rain is important for stations of Le Havre (REP and HRI) and for GCM while the wind directions are important for the stations located in town, Rouen (JUS and GUI) and Le Havre (REP and HRI) as well as for AIL.

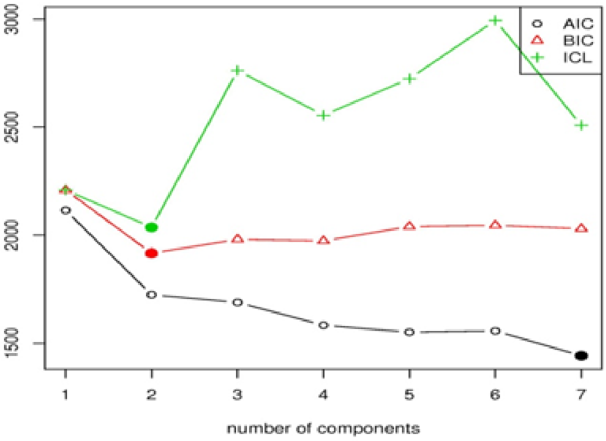

Figure 16. Criteria for the choice of the number of clusters. For JUS, the number 2 is selected.

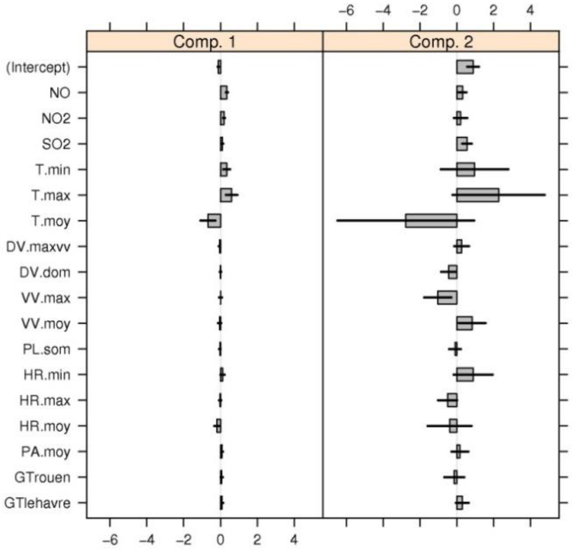

Figure 17. Coefficients for both models, for JUS.

Let us now characterize the differences between the two classes focusing on the comparison between the two models. The magnitude of the pollutant coefficients allows to quantify the importance of each pollutant. Thus, we collected in Table 14 these values for each station (except of course for the rural station AIL). We can note that for Rouen stations (GCM, JUS and GUI), the coefficients of SO_2 are much larger in the most polluted class. On the contrary, in Le Havre (REP and HRI) the NO_x coefficients are the higher.

7. Local part and regional part

In this section, we focus on a quantification of what we call in a broad sense a local part and a regional part of PM_{10} pollution, trying to give meaning to these concepts in a purely statistical context without neither direct information nor measurements about sources.

The first key point is to start from the distinction between the different groups of explanatory variables: the pollutants and three groups of meteorological variables. The second key idea is the spatial nature of the network of six stations and to take advantage of the specificity of the rural station AIL for which there is *a priori* no local pollution sources.

Starting from the results of Table 8 (see section 4.3), the main idea is to use PM_{10} pollution measured at AIL (denoted by PM_{10}^{AIL}) as an indicator of the spreading pollution at the regional scale. It is supposed to capture the pollution phenomenon at greater or lesser extent (regional or more) and to not be affected specifically by a major local production. This is supported by the distributions of concentrations for the six stations: for the 33 days for which PM_{10}^{AIL} exceeds 30 μg/m^3, the median is around 40 μg/m^3 and the first quartile around 33 μg/m^3.

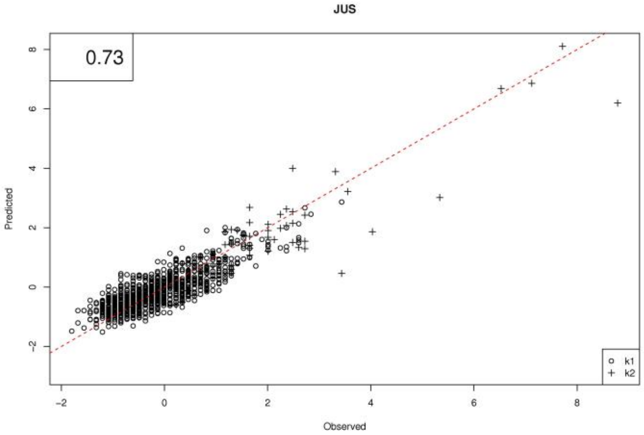

Figure 18. Observed values vs. predicted values, JUS

Table 13. Characterization of the most polluted class across the stations.

Station	Number of days	Explained Variance	Pollutants	Meteo. Variables
GCM	226	0.64	SO_2 (+)	PL.som (-)
JUS	57	0.73	SO_2 (+), NO(+)	VV.max (-), VV.moy (+), DV.dom (-)
GUI	73	0.66	SO_2 (+), NO(+)	GTlehavre (+), DV.dom (-)
AIL	154	0.51		DV.maxvv (-), HR.min (+), VV.moy (+), GTlehavre (+)
REP	83	0.75	NO_2 (+), NO (+)	GTlehavre (+), DV.maxvv (-), PA.moy (+), PL.som (-)
HRI	120	0.79	NO (+), SO_2 (+)	PL.som (-), VV.max (+), GTlehavre (+), GTrouen (-), DV.maxvv (-)

Table 14. Characterization of the difference between the two classes across the stations.

	Polluted class			Non-polluted class		
Station	NO	NO_2	SO_2	NO	NO_2	SO_2
GCM			*0.70*			*0.18*
JUS	0.31	0.20	0.56	0.35	0.19	0.10
GUI	0.32	0.26	0.60	0.25	0.21	0.08
REP	0.36	0.57	0.01	0.34	0.19	0.30
HRI	0.55	0.13	0.31	0.08	0.03	0.43

The importance of the variable PM_{10}^{AIL} in previous models when they are complemented by the introduction of this new variable leads to the Table 15. Its importance is around 64 to 86, which is considerable, while the importance of meteorological variables significantly decreases. On the contrary, the importance of pollutants remains stable, for all the stations.

These elements are compatible with the idea that PM_{10}^{AIL} reflects diffuse pollution in the sense that it does not significantly change the importance of local markers while it hugely affects weather variables ones.

Fitting a random forest for each of the five other stations, considering only pollutants and PM_{10} concentration at AIL, discarding all the meteorological variables, then leads to Table 16.

In addition, the effects obtained by fitting additive models (not reported here) are weakly increasing and weakly nonlinear (except sometimes for extreme levels of SO_2) at least for the default window choice (leading to slight oversmoothing). So the conclusion is that by introducing this new variable and canceling the meteorological variables, the model is linearized. Concentrating then on models involving locally measured pollutants and PM_{10} from AIL, we propose to quantify more directly the respective parts of these two factors by fitting a simple linear model and computing the standardized coefficients (see Table 17).

Then, by summing the coefficients associated with pollutants of similar behavior, we obtain the quantification given by Table 18. This leads to the

Table 15. RF variable importance across stations, including PM_{10}^{AIL}.

		Meteo			
	Pollutants	Negative	Positive	Mixed	PM_{10}^{AIL}
GCM Rouen	33 (SO_2)	23 (PL)	17 (GT)	16 (DV)	68
JUS Rouen	31 (NO) 18 (SO_2)	19 (PL)	16 (GT)	13 (DV)	78
GUI Rouen	40 (NO_2) 17 (SO_2)	18 (PL)	26 (PA)	20 (T)	64
REP Le Havre	44 (NO_2) 28 (SO_2)	18 (VV)	19 (GT)	12 (T)	81
HRI Le Havre	41 (SO_2)	12 (VV)	11 (GT)	13 (T)	86

Table 16. Variable importance when replacing meteo by PM_{10}^{AIL}.

	Pollutant now more important	Other pollutants	PM_{10}^{AIL}
GCM Rouen, industrial	73 (SO_2)		98
JUS Rouen, urban	35 (NO)	26, 22 (NO_2, SO_2)	89
GUI Rouen, traffic	42 (NO_2)	31, 20 (NO, SO_2)	77
REP Le Havre, traffic	42 (SO_2)	41, 36 (NO_2, NO)	78
HRI Le Havre, urban	50 (SO_2)	21, 19 (NO, NO_2)	81

Table 17. Linear regression for local and regional parts: standardized coefficients.

	Adjusted R^2	Pollutant now more important	Other pollutants	PM_{10}^{AIL}
GCM Rouen	0.54	0.39 (SO_2)		0.56
JUS Rouen	0.68	0.47 (NO)	0.06, 0.11 (NO_2, SO_2)	0.55
GUI Rouen	0.56	0.38 (NO)	0.16, 0.12 (NO_2, SO_2)	0.49
REP Le Havre	0.74	0.32 (SO_2)	0.28, 0.26 (NO, NO_2)	0.47
HRI Le Havre	0.75	0.49 (SO_2)	0.23, -0.04 (NO, NO_2)	0.50

Table 18. Local and regional parts: sum of standardized coefficients.

	NO, NO_2	SO_2	PM_{10}^{AIL}
GCM Rouen	0.39		0.56
JUS Rouen	0.53	0.11	0.55
GUI Rouen	0.54	0.12	0.49
REP Le Havre	0.54	0.32	0.47
HRI Le Havre	0.19	0.49	0.50

Table 19. Quantification of local and regional parts.

	Local part (pollutants), %	Regional part %
GCM Rouen, industrial	41	59
JUS Rouen, urban	49	51
GUI Rouen, traffic	52	48
REP Le Havre, traffic	53	47
HRI Le Havre, urban	49	51

breakdown given by Table 19, based on the sum of the coefficients of the most important pollutants on the one hand and the coefficient of PM_{10}^{AIL} on the other hand.

The main conclusion is that the respective parts appear to be balanced (around 50%) except for GCM where the regional part exceeds the local one (41%-59%). This exception must be considered with caution because the model does not take into account the loads at grain port of Rouen, for which the data are not yet available.

8 Conclusion

We exhibited in this case study a methodology for variable selection, non-linear modeling and importance variable quantification using three modern nonparametric statistical methods (random forests, mixtures of linear models and nonlinear additive models) to investigate the problem of the statistical analysis of air pollution in a French area.

To conclude about the relative merits of the three approaches, both to this case study, and perhaps in general, the key idea is that these three different tools offer different views of the data, focusing on different tasks and are to be used simultaneously. Indeed, random forests are very powerful for prediction and variable importance quantification. However a random forest does not define an explicit model since it builds a prediction model which is an aggregation of regression trees. So we consider regression models by classes of two kinds. The first one is based on generalized additive models and proposes weather type dependent nonlinear additive models, leading to explicit and easy to understand classes. The second one is based on mixture of linear models and also builds clusterwise regression models but the building strategy combines more closely clustering and regression fitting allowing more flexible classification as well as simpler models within a class but of course yielding less directly interpretable classes.

From the PM_{10} modeling viewpoint, the results allow us to analyze PM_{10} concentrations and confirm the environmental knowledge of the phenomenon of air pollution by fine particles. Of course, this kind of methodology can be useful when applied to any dataset and any problem involving nonlinear data analysis and modeling in the context of environmental data. In addition, the last section focusing on an attempt of quantification of a local part and a regional part of PM_{10} pollution illustrates how different approaches are merged to construct such a tricky evaluation.

Let us finally mention that, in addition to some simple examples included in the paper, the appendix and related online material provide full R code as well as the complete data set.

Acknowledgements

This work results from a scientific collaboration between Air Normand (see the website http://www.airnormand.fr/) from the applied side and Paris Descartes University and INSA in Rouen from the academic side (see Jollois *et al.* 2008). We would like to thank Véronique Delmas and Michel Bobbia, from Air Normand, for providing the problem and the pollution data as well for supporting the statistical study. In addition, we want to thank the regional council of the Haute-Normandie area for supporting this work and Météo-France for authorizing the aggregated meteorological data to be publicly available. Last, we thank the anonymous referees for their useful suggestions and comments.

Appendix: Associated Data and R Code

A.1 Data

The data used in this paper are available in the associated archive, containing 14 files. These data are in two formats:

- Specific R format, in the JolloisPoggiPortier.Rdata, readable with the following command:
  ```
  load('JolloisPoggiPortier.Rdata')
  ```
- Semicolon-separated text format, in the JolloisPoggiPortier_XXX.txt and JolloisPoggiPortier_XXX_comp.txt files. The XXX represents the desired station (AIL, GCM, HRI, REP, JUS, GUI). The second file, ended by _comp.txt, contains data without missing values.

Table 20 describes the 18 variables. For GCM station, only the pollutant SO_2 is available, and there is no pollutant for AIL station.

Table 20. Data dictionary. For wind direction, 0° corresponds to north.

Name	Description	Units
PM10	Concentration of PM_{10}	μg/m^3
NO, NO2, SO2	Concentration of NO, NO_2, SO_2	μg/m^3
T.min, T.max, T.moy	Minimum, maximum and mean temperature	°C
DV.maxvv, DV.dom	Maximum speed and dominant wind direction	°
VV.max, VV.moy	Maximum and mean wind speed	m/s
PL.som	Daily rainfall	Mm
HR.min, HR.max, HR.moy	Minimum, maximum and mean relative humidity	%
PA.moy	Mean air pressure	hPa
GTrouen, GTlehavre	Temperature gradient	°C

A.2 R code

Three R scripts are associated with this paper, one for each of the three considered methods:

`ModNL_GAM.R`	Non-linear additive models
`VarImp_Effect_RF.R`	Random forest
`ClusReg_CR.R`	Mixture of linear regressions

They allow one to apply the methods for each station, and save the results in a specific directory for each method and for each station.

Caution: when a GAM model is fitted on a weather defined class of small size, an error may occur when a numerical variable is confused with a factor. So, to still use the automatic version in such situations, it is necessary to delete the corresponding variable (to be search among VV.max, HR.max, DV.maxvv or DV.dom).

REFERENCES

Breiman, L., J.H. Friedman, R.A. Ohlsen, and C.J. Stone. 1884. *Classification and Regression Trees*, Belmont.

Breiman, L. and J. H. Friedman. 1985. "Estimating optimal transformations for multiple regression and correlation", *J. Am. Stat. Assoc*., 80: 580-619.

Breiman, L. 2001. *Random Forests*, Machine Learning 45(1): 5-32.

Breiman, L. and A. Cutler. 2005. *Random Forests*, Berkeley CA.

Buja, A., T. J. Hastie and R. J. Tibshirani. 1989. "Linear smoothers and additive models", *Ann. of Stat*., 17: 453-510.

Chavent, M., H. Guégan, V. Kuentz, B. Patouille, and J. Saracco. 2007. "Apportionment of air pollution by source at a French urban site", *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, 1(2): 119-129.

Dempster, A., N. Laird, and D. Rubin. 1977. "Maximum likelihood for incomplete data via the EM algorithm", *Journal of the Royal Statistical Society* 39(B): 1-38.

Genuer, R., J-M. Poggi, and C. Tuleau. 2008. Random Forests: some methodological insights, Research report INRIA 6729, 1-32. http://hal.inria.fr/inria-00340725/fr/

Jollois, F.X., J.M. Poggi, and B. Portier. 2008. Analyse statistique de la pollution par les particules en Haute-Normandie, Technical Report, Air Normand.

Gruen, B. and F. Leisch. 2007. "Fitting finite mixtures of generalized linear regressions", *R. Computational Statistics and Data Analysis*, 51(11): 5247-5252.

Hastie, T. and R. Tibshirani. 1990. *Generalized Additive Models*, Chapman & Hall.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Hathaway, R.J. and J.C. Bezdek. 1993. "Switching regression models and fuzzy clustering", *IEEE Trans. Fuzzy Systems*, 1(3): 195-203.

Karaca, F., O. Alagha, and F. Erturk. 2005. "Statistical characterization of atmospheric PM$_{10}$ and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey", *Chemosphere*, 59: 1183-1190.

Leisch, F. 2004. FlexMix: A general framework for finite mixture models and latent class regression", *R. Journal of Statistical Software,* 11(8): 1-35.

Liaw, A. and M. Wiener. 2002. Classification and Regression by randomForest, *R News*, 2(3):18-22.

McLachlan, G. and D. Peel. 2000. *Finite Mixture Models*, Wiley Series in Probability and Statistics.

Salvador, P., B. Artinano, D. G. Alonso, X. Querol and A. Alastuey. 2004. Identification and characterization of sources of PM$_{10}$ in Madrid (Spain) by statistical methods, 38: 435-447.

Smith, S., F. T. Stribley, P. Milligan and B. Barratt. 2001. "Factors influencing measurements of PM$_{10}$ during 1995-1997 in London", *Atmospheric Environment*, 35: 4651-4662.

Schwarz, G. 1978. Estimating the Dimension of a Model, *Annals of Statistics,* 6: 461-464.

Stone, C.J. 1986. The dimensionality reduction principle for generalized additive models, *Ann. of Stat.,* 14: 592-606.

R Development Core Team. 2007. *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Wood, S.N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Correspondence: Jean-Michel.Poggi@math.u-psud.fr