Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France)

Michel Misiti [a], Yves Misiti [a], Jean-Michel Poggi [a,c,d], Bruno Portier [b]

[a] Laboratoire de Mathématiques, Université Paris-Sud, bat. 425, 91405 Orsay, France

[b] Laboratoire de Mathématiques, INSA de Rouen, France

[c] Université Paris Descartes, France

[d] INRIA Saclay, Ile de France, France

Abstract:

Mixture of linear regression models is used for the short-term statistical forecasting of the daily mean PM_{10} concentration. Hourly concentrations of PM_{10} have been measured in three cities in Haute-Normandie (France): Rouen, Le Havre and Dieppe. The Haute-Normandie region is located at northwest of Paris, near the south side of Manche sea and is heavily industrialized. We consider six monitoring stations reflecting the diversity of situations: urban background, traffic, rural and industrial stations. We have focused our attention on recent data from 2007 to 2011.

We forecast the daily mean PM_{10} concentration by modeling it as a mixture of linear regression models involving meteorological predictors and the average concentration measured on the previous day. The values of observed meteorological variables are used for fitting the models but the corresponding predictions are considered for the test data, leading to realistic evaluations of forecasting performances, which are calculated through a leave-one-out scheme on the four years.

We discuss in this paper several methodological issues including estimation schemes, introduction of the deterministic predictions of meteorological models and how to handle the forecasting at various horizons from some hours to one day ahead.

Keywords:

Particulate matter, PM10, forecasting, mixture of linear models, air quality

1. Introduction

We consider in this work carried out with Air Normand, the official association for air quality monitoring in Haute-Normandie (France), the problem of forecasting the daily average PM_{10} concentration using a statistical model.

The context from the pollution viewpoint is summarized by the objective given by the European regulation which prescribes that the daily average concentration of particulate matter PM_{10} cannot exceeds 50 $\mu g\ m^{-3}$ more than 35 days per year, in addition to some criterion based on mean values over large time intervals. Pollution comes from various origins, natural or due to human activity and the objective is, through the joint statistical analysis of PM_{10} concentrations and meteorological parameters, the short term forecasting of PM_{10}. Four different forecasting horizons are considered, from some hours to one day ahead.

The context from the statistical viewpoint can be sketched by citing some papers illustrating the diversity of the statistical modeling techniques used for such purposes. Let us mention first, Grivas and Chaloulakou (2006) giving a detailed introduction and Dong et al. (2009) highlighting methods and models. Most of the papers uses neural networks (NN) based forecasting, see Paschalidou et al. (2011), or multiple linear modeling (LM), see Stadlober et al. (2008). Nevertheless, various other methods are used, let us cite the paper of Zolghadri and Cazaurang (2006) using extended Kalman filter for one station in Bordeaux (France), or the paper by Slini et al. (2006) comparing CART trees, NN, LM and regression on principal components, or the paper by Corani (2005) using local polynomials for nonlinear regression, and the paper by Sfetsos and Vlachogiannis (2010) using clusterwise linear models, in two stages: a supervised clustering followed by local linear modeling. With respect to this last approach, we propose to merge the two stages by using mixture of linear regressions allowing to estimate in the same time the clusters and the local models.

In this paper, mixtures of linear regression models are used for the statistical forecasting of the daily mean PM_{10} concentration. Hourly concentrations of PM_{10} have been measured in three cities in Haute-Normandie (France): Rouen, Le Havre and Dieppe. The Haute-Normandie region is located at northwest of Paris, near the south side of Manche sea and is heavily industrialized. We consider monitoring stations reflecting the diversity of situations: urban background, traffic, rural and industrial stations. We have focused our attention on recent data from 2007 to 2011.

We discuss several methodological issues including various estimation schemes, the introduction of the deterministic predictions of meteorological models and the way to handle the forecasting at various horizons from some hours to one day ahead. In a preliminary work, Poggi and Portier (2011) examined in a restricted context the interest of the mixture of linear models for forecasting purposes. In this paper, this preliminary work has been extended in three directions. First, the forecasting problem is fully addressed in the time domain: short-term forecasting and 4 horizons are considered, from some hours to one day ahead. Second, the performance evaluation includes also the performance evaluation in a real context using meteorological predictions instead of observations at various horizons. Third, the study is carried out on a network of monitoring which reflects the diversity of situations as well as meaningful part of the crucial difficulties for the forecasting problem encountered in

Haute-Normandie region.

Our paper is organized as follows. Section 2 presents the data, PM_{10} and meteorological, and describes the mixture of linear models framework which is the statistical method examined in this study. Section 3 states the specific forecasting problem addressed here and develops various methodological insights including estimation schemes, introduction of the deterministic predictions of meteorological models and forecasting at different horizons. Section 4 presents the forecasting results obtained at four different horizons from some hours to one day ahead. Finally, Section 5 contains a short conclusion.

2. Materials and methods

The Haute-Normandie region (over 1,8 million people in 2008) is located in north-western France, near the south side of Manche sea and at northwest of Paris. The region Haute-Normandie is heavily industrialized and has two large agglomerations Rouen (more than 490,000 inhabitants, the 36th city in France) and Le Havre (more than 250,000) which are also two major harbors.

We have considered pollution data from 1 April 2007 to 31 March 2011, collected by Air Normand. The data set consists of PM_{10} daily mean concentrations and many meteorological data as well as weather and pollution forecasts coming from numerical models. The statistical tools used in this work are not classical in the context of PM_{10} forecasting but as already mentioned in the introduction, have been recently experimented by Poggi and Portier (2011).

2.1. Monitoring stations

We consider in this study a subset of six PM_{10} monitoring stations of the permanent network of Air Normand that contains 13 fixed monitoring sites, mainly spread over the main towns of Haute-Normandie. This subset of stations reflects the diversity of situations that we can find in the region. Indeed, we have chosen three stations in Rouen, namely Palais de Justice (JUS), Guillaume-Le-Conquérant (GUI) and Grand-Couronne (GCM), two stations in Le Havre, namely Cours de la République (REP) and Ecole Herriot (HRI), and one station near Dieppe, namely Phare d'Ailly (AIL). The stations JUS and HRI are urban background ones, GUI and REP are roadside stations, GCM is an industrial station near the cereal harbor of Rouen, which is one of the most important in Europe. Lastly, AIL is a rural and coastal station, and a priori faraway from any local important source of pollution.
In addition to the air pollution monitoring sites, three meteorological monitoring stations of Meteo-France (the French national meteorological service) have been considered: MFB at Rouen, MFD at Dieppe and MFH at Le Havre (see Fig. 1).
Lastly, we use data coming from two meteorological monitoring sites of Air Normand, respectively denoted by ESS at Rouen and CAU at Le Havre in Fig. 1, and we associate to each PM_{10} monitoring site the nearest meteorological station (e.g., JUS station of Rouen is associated with MFB and ESS stations, REP of Le Havre with MFH and CAU, and so on).

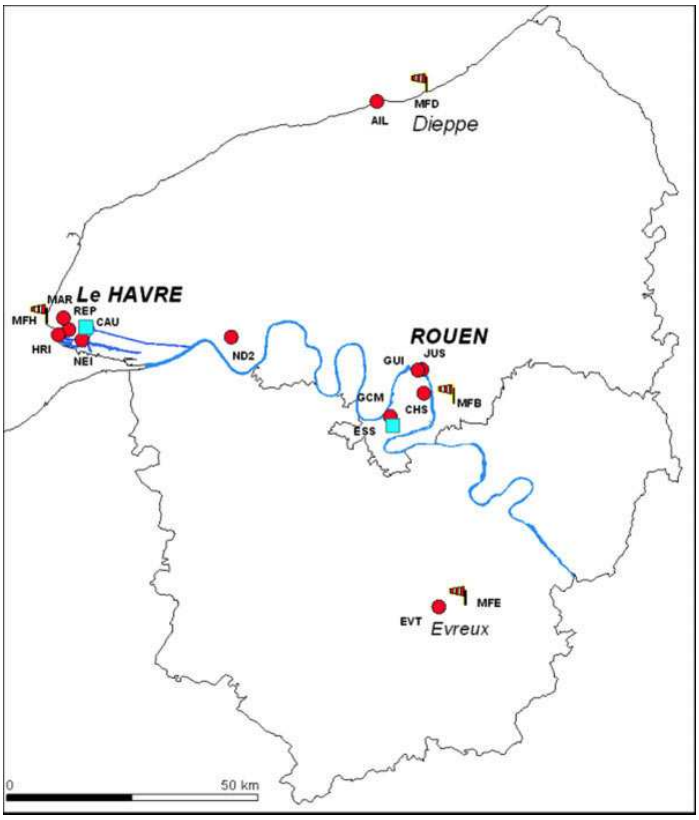

Fig. 1. Map of the Haute-Normandie: monitoring sites of Air Normand and Météo-France.

2.2. Data

2.2.1. Pollution data

We consider TEOM PM_{10} daily mean concentrations coming from six monitoring sites. One can find in Table 1 some basic statistics about the PM_{10} concentrations.

Table 1: Basic statistics of daily mean PM_{10} concentrations (in µg m^{-3}) for the six stations

	JUS	GUI	GCM	HRI	REP	AIL
Minimum	6	7	4	5	8	4
1st Quartile	14	18	13	13	19	12
Median	18	23	17	17	23	15
Mean	19.5	24.3	18.8	19.0	25.3	16.1
3rd Quartile	23	28	22	23	29	19
Maximum	68	90	80	86	88	58
SD	7.8	9.2	8.6	8.5	9.6	6.2
Missing values	16	6	21	30	19	27
No. of values	1461	1461	1461	1461	1461	1461

As it can be seen in Table 2 and Fig. 2, the distribution of the daily mean PM_{10} concentration differs depending on the location of the monitoring stations. Indeed, unsurprisingly the traffic stations GUI and REP measure PM_{10} concentrations higher

than other stations, and the station AIL measures lesser concentrations. The background stations JUS and HRI have almost the same distribution.

In addition, we emphasize that, from our previous study, see Jollois et al. (2009), these monitoring stations are also different in terms of intrinsic difficulty to model and so forecast PM_{10} concentrations with the same predictors, namely JUS and HRI are easy, GCM and AIL are hard and GUI and REP are of medium difficulty. So we capture in this subset of PM_{10} monitoring stations of Air Normand network a meaningful part of the crucial difficulties for the forecasting problem.

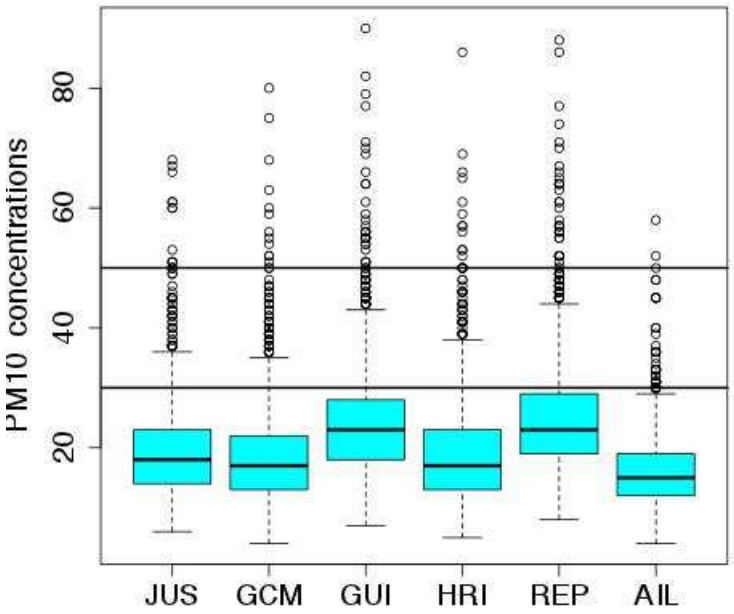

Fig. 2. Boxplots of daily mean PM_{10} concentrations (in µg m^{-3}) for the six stations.

A final remark from this descriptive point of view: as it can be seen, using the threshold of 50 µg m^{-3} mentioned in the introduction as the level of reference is high and somewhat unrealistic because these measures comes from TEOM measurements only and are optimistic measures of pollution. So, in order to take into account the fact that TEOM measurements do not integrate the volatile fraction, we use a smaller value for the threshold of 30 µg m^{-3} to quantify the forecasting performance in terms of exceedances in a more useful way. Of course this threshold value has no impact on the designed model.

2.2.2. Meteorological data

We have retained the meteorological indicators selected in Poggi and Portier (2011): the daily mean temperature (Tmoy, in °C), the daily mean atmospheric pressure (PAmoy, in hPa), the daily mean wind speed (VVmoy, in m/s) and the daily maximum gradient of temperature (GTmax, in °C). The different variables are calculated from hourly measurements during the period 0h-24h GMT. The gradient of temperature is defined as the daily maximum of the hourly differences between the temperature at 2 meters altitude and the temperature at 100 meters. This indicator gives us an idea of the mixing height.

Table 2: Summary statistics for meteorological variables at MFB and ESS stations, data associated to PM_{10} concentrations measured in Rouen.

	Tmoy (in °C)	VVmoy (in m/s)	PAmoy (in hPa)	GTmax (in °C)
Minimum	-6.8	1	974.4	-1.8
1st Quartile	5.9	2.9	1011	-0.7
Median	10.7	3.7	1017	0.4
Mean	10.3	4.1	1016	1.01
3rd Quartile	15.2	4.9	1023	2.3
Maximum	23.9	12	1041	14.6
SD	6.1	1.6	9.7	2.2
Missing values	2	1	2	0
No. of values	1461	1461	1461	1461

Table 2 presents a summary of the basic statistics for the meteorological variables used at Rouen and coming from MFB station for Tmoy, VVmoy and PAmoy, and ESS station for GTmax. Let us note that the values observed in other stations are of the same order of magnitude.

2.2.3. Meteorological predictions

Numerical weather predictions come from the Arpege-France system. This system is used by Météo-France to deliver weather predictions up to three days ahead. Arpege covers the entire planet with a mesh size of side 15 km over France. In our study, we associate to each of the three meteorological stations the nearest location in the Arpege's grid. In addition, two forecast horizons are considered: the morning for the next day and the afternoon for the next day.

Of course, it is not of interest in this paper to evaluate the quality of these forecasts but since these model outputs will be used instead of the corresponding measures, let us briefly comment the accuracy of the predictions.
The considered meteorological predictions are accurate, for example for the station of Rouen, the scatterplots of observed meteorological variables and Arpege predictions made the yesterday morning for the current day lead to high percentages of explained variance EV. Namely EV=0.97, 0.99 or 0.8 for Tmoy, PAmoy and VVmoy respectively. Nevertheless, since some spatial corrections are introduced, for some days the temperature gradient can be imperfectly predicted and the dispersion is higher leading to an explained variance about 0.52.

2.3. Mixture of linear models

The framework of mixture models, see McLachlan and Peel (2000), is widely used for model based clustering. It offers, in the regression context, a global and unified way to handle clusterwise linear models by simultaneously optimize the clusters and the local models. Let us briefly recall it.

The observations are supposed to be a random sample drawn from a mixture of standard linear models. More precisely, if y is dependent variable and x the vector of

explanatory variables, the conditional density h of y is a finite mixture of K components of the following form:

$$h(y|x,\psi) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k)$$

where π_k is the prior probability of component k. In each component, $\theta_k = (\beta_k^T, \sigma_k^2)$ is the parameter vector of the Gaussian density function f of mean $\beta_k^T x$ and variance σ_k^2. Parameter ψ denotes simply the vector of all local parameters put together.

So, starting from a learning set, the estimation procedure allows identifying a mixture model defined by the number of components, the prior probability distribution and the local linear models.

The posterior probability that observation (x, y) belongs to cluster j is given by

$$P(j|x,y,\psi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_{k=1}^{K} \pi_k f(y|x,\theta_k)}$$

leading to posterior probability distribution of the cluster labels. It allows to assign each observation to a cluster by taking the maximum a posteriori probability.

Such mixture models are usually estimated, using the maximum-likelihood framework, by the expectation-maximization (EM) algorithm, which iteratively repeats two steps until convergence, see Dempster et al. (1977). 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. The estimation of models for various fixed number of components are then compared using the BIC criterion (Schwarz 1978) for example, giving a strategy to select the number of components.

From a computational viewpoint, the estimation of mixture models can easily be handled thanks to the `flexmix` R package described in Gruen and Leisch (2007, 08), Leisch (2004). The criterion used for selecting the number of clusters is the BIC model selection criterion, based on asymptotic analysis, which is convenient in this situation because it is parsimonious and the number of observations is important.

Let us give two definitions of the prediction delivered by a given model: the fuzzy one and the hard one. More precisely, let us denote

$$p_k(y) = P(k|x,y,\psi)$$

then the fuzzy prediction for a given x and a given estimated model can be set to the properly weighted combination of clusterwise predictions:

$$\widehat{y^{(w)}} = \sum_{k=1}^{K} p_k\ (y)\ \widehat{y_k}$$

where $\widehat{y_k}$=$\widehat{\beta_k^T} x$ is the prediction given by the model of component k.

The second choice, the hard one is obtained by using a single model corresponding to the most probable cluster:

$$\widehat{y^{(p)}} = \widehat{y_{kmax}} \text{ where } k_{max} = \text{argmax } p_k(y)$$

In the forecasting context for a given x and a given estimated model, the $p_k(y)$ are generally unknown, since the y value is unknown. Even the cluster from which the observation belongs to is no longer accessible. Then the two previous predictions can be estimated thanks to the estimation of these posterior probabilities. This point will be addressed in the next section.

3. Some methodological insights

3.1. The forecasting problems

Let us begin this section by introducing the different forecasting problems to be addressed in this work. A simple way is to comment **Fig. 3**. The daily mean PM_{10} concentration of day *J* is to be predicted, mainly one day before and the prediction should be refined on an intraday term. More precisely, four different horizons have been considered (see the top of the time axis of **Fig. 3**): one-day ahead at 10 a.m. and 5 p.m. of day *J-1* and on intraday basis at 7 a.m. and 1 p.m. of day *J*. Consequently the more recent meteorological model outputs can be considered and are introduced in the prediction model and four different instants of introduction are available for Air Normand (see the bottom of the time axis of **Fig. 3**): one-day ahead at 9 a.m. and 11 a.m. of day *J-1* and on intraday basis at 3 a.m. and 11 a.m. of day *J*.

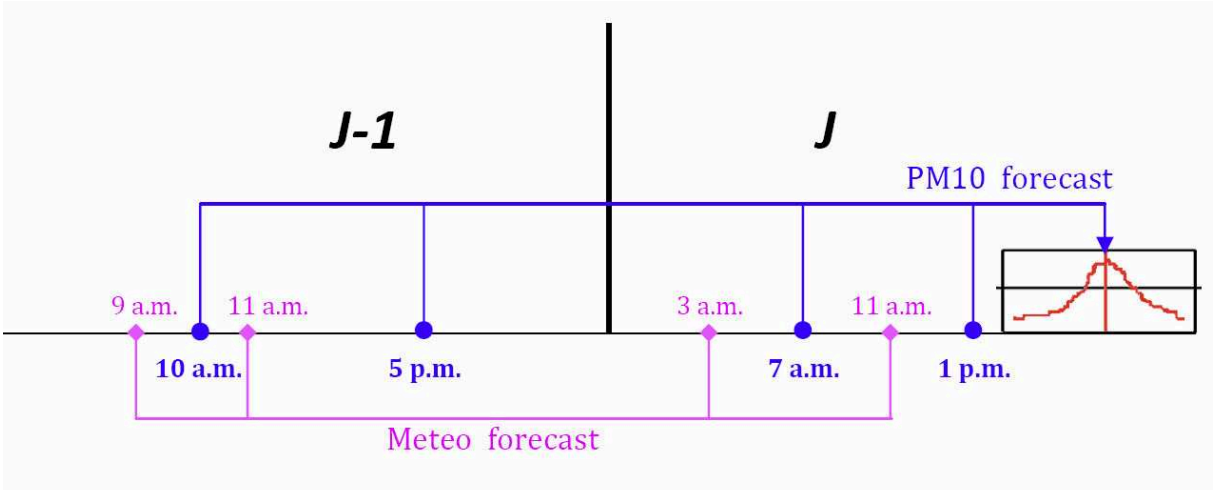

Fig. 3. The short-term forecasting problems: four horizons from some hours to one day ahead.

3.2. From models for analysis to models for forecasting

3.2.1. Variables

In a previous paper dedicated to analysis of PM_{10} concentrations, we have used of course meteorological variables as explanatory variables but also other pollutants: NO, NO_2 and SO_2 which are interesting markers of social and economic activities. The model explaining the daily mean PM_{10} concentration of day *J* naturally involves

explanatory variables observed during the same day. From this point of view, a first idea is to start from this model suitable for analysis to derive a forecasting model by trying to forecast each predictor for which the value is not available.

The situation is quite different for pollutants and for meteorological parameters. Since the pollutants are very difficult to forecast, the resulting model is of poor quality. An alternative widely used strategy is to introduce the PM_{10} of day *J-1* which exhibits hugely larger importance on predictions. On the contrary, the meteorological variables involved in the models for analysis are often very well predicted. Nevertheless some of them like rain or wind direction are difficult to predict and then are omitted in the forecasting models.

3.2.2. Estimating a posteriori probabilities

In the previous section, we have introduced two different predictions, called “fuzzy” and “hard” which cannot be directly used in the forecasting context since they involve posterior probabilities. It should be noted that an oracle telling us in which cluster the day belongs to suffices to calculate the prediction. So these quantities are to be estimated thanks to the estimation of these posterior probabilities.

The most natural idea is based on the use of PM_{10} forecasting models provided by the different clusters and on the replacement of observed PM_{10} by these estimates in the formula of posterior probabilities.
Denoting, for a given x, by $\widehat{y_k}$ the prediction of y given by the model of cluster k, the corresponding predictions are

$$\hat{y} = \sum_{k=1}^{K} p_k\ (\widehat{y_k})\ \widehat{y_k} \Big/ \sum_{k=1}^{K} p_k\ (\widehat{y_k})$$

and, for the hard variant:

$$\hat{y} = \widehat{y_{k*}} \text{ where } k^* = \text{ argmax } p_k(\widehat{y_k})$$

and mixing the two proposals:

$$\hat{y} = \sum_{k=1}^{K} p_k\ (\widehat{y_{k*}})\ \widehat{y_k}$$

since $\sum_{k=1}^{K} p_k\ (\widehat{y_{k*}}) = 1$. The main drawback of these natural estimates is that the $p_k(\widehat{y_k})$ are all together high, of the same order of magnitude, and often lead to wrong cluster assignation. So, these interesting solutions are substantially improved by estimating posterior probabilities directly by replacing y by a forecast $\widehat{y_{GAM}}$ coming from a weighted generalized additive model GAM. Even if GAM is not the best model used solely in that case, it is competitive for the posterior probability estimation thanks to its simplicity and its ability to handle different weights for observations.

Avoiding details that are out of the scope of this paper, let us recall that nonlinear additive modeling was introduced by Breiman and Friedman (1985) and disseminated thanks to the work of Hastie and Tibshirani (1990). It generalizes multiple linear regression allowing the effect of each explanatory variable to be nonlinear but preserving the additivity of the different effects on the dependent

variable. These models are more flexible than linear models but they preserve the ease of interpretation by the additivity of effects of the individual regressors. To estimate nonlinear additive models, we use the R package *mgcv* developed by Wood (2006).

The mean square error (MSE) is the criterion usually minimized in GAM estimation. The weighted GAM used here is obtained simply by multiplying each term of the MSE of the form $(y - \hat{y})^2$ by the observed value leading to a weighted term $y(y - \hat{y})^2$ of the mean.

Finally, the estimator we use is of the following form:

$$\hat{y} = \sum_{k=1}^{K} p_k \ (\widehat{y_{GAM}}) \ \widehat{y_k}$$

This variant leads to the better forecasting performance in terms of errors and threshold exceedances.

Note that, if this scheme is not entirely satisfactory because we would prefer to stay within the framework of mixture of linear models this scheme still has a theoretical justification since the additive assumption is reasonable and the use of such global GAM provide estimations of the posterior probabilities.

3.3. Perfect prognosis, MOS and OBS

Another classical issue, see for example Wilks (1995), is to examine how to deduce from a model identified and fitted using the measures of meteorological variables a model for forecasting or at least a realistic prediction equation involving available quantities (measures or model output statistics).
Classically, three ways, abbreviated by MOS (Model Output Statistics), OBS (for observations only) and PP (for perfect prognosis), are considered and compared.
The idea of MOS is to estimate directly the realistic prediction equation and then to fit a model using the past values of the meteorological forecasts, instead of the measures. This strategy suffers from two main drawbacks which are the non-stationarity of these past values and the fact that this strategy implies to handle multiple models depending on future and present instants (the actual time instant and the horizon).
The strategy OBS leads to use only observed meteorological predictors at the actual time instant. This strategy, which is classical in the statistical approach of time series, leads to poor results. The explanation is that we implicitly forecast meteorological using only observations without any deterministic model.
The last strategy PP consists in fitting the model using the observed values of the meteorological variables one day ahead and replacing it by meteorological predictions coming from deterministic models in the forecasting equation. This strategy has three decisive pros: it does not suffer from frequent updates of deterministic model, only one single model is to be fitted for one day ahead forecasting (only the meteorological predictions differ) and it leads to parsimonious models since the most powerful meteorological predictors can be considered.

3.4. Validation scheme

For the validation step, there are a lot of classical solutions based on cross-validation (CV) ideas that are efficient and fair to estimate the true error, without optimistic bias. Nevertheless, in such problems, it is clearly better to use some kind of block-CV in order to preserve some internal homogeneity in each block in order to measure conveniently the estimation error avoiding the confusion with the impact of non-stationarity. For obvious reasons, such phenomenon naturally arises due to seasonal differences for blocks of length less than one year. So, we propose to perform a CV on years, that is a leave-one year out scheme.

The whole data set is composed of four years: 2007 to 2010. More precisely 2007 stands for the time period from April 1, 2007 to March 31, 2008 and so on for the other years. Then, to evaluate the quality of models we based the analysis on four models constructed according to the following scheme: 1 year to predict (test year in red on Fig. 4), 3 years to estimate the model (learn year in green on the Fig. 4).

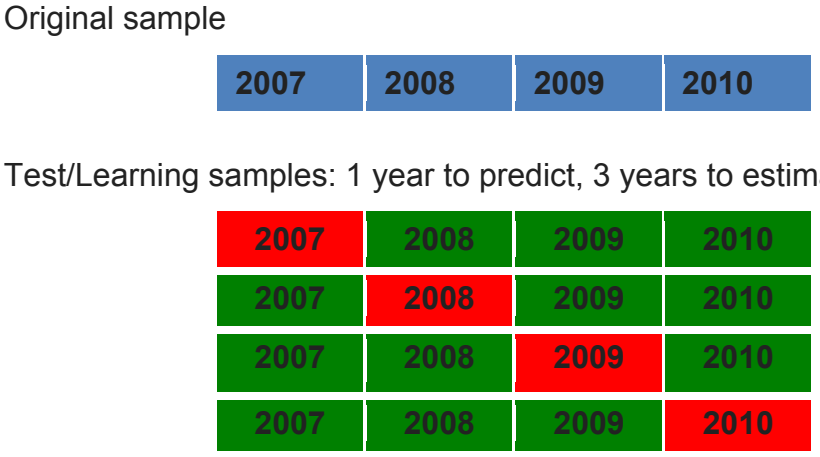

Fig. 4. Validation scheme: a leave-one year out.

Of course, the forecasting of 2007 data using 2008, 2009 and 2010 or more generally using future data for modeling data of the past can appear artificial since in general air quality should be improving in time as appropriate environmental policy is applied. We emphasize that this validation scheme is mainly used to select the model and not to assess actual performance.

4. Results

4.1. Forecasting one-day ahead: models and results

The results analyzed in this section come from a forecasting model defined by a mixture of linear regressions with 2 clusters and where a GAM model is used for estimating a posteriori probabilities. The local regressions involve the basic meteorological variables (Tmoy, VVmoy, PAmoy, GTmax), the PM_{10} concentration of the day before (PM10_jm1), the prediction delivered by the deterministic model Prev'air (Pvr_j) and finally a calendar variable about the day to predict distinguishing weekdays from weekends (TypJ). The scores are evaluated using the previously introduced cross validation on years (leave one year out).

Four scores are given to evaluate performance (see Chaloulakou et al., 2003 for a detailed description of these indicators). The percentage of explained variance EV and the root mean square error RMSE for the error performance, the skill score SS and the threat-score TS (range between [0, 1] with a best value of 1). for the threshold exceedances performance. This last one is computed for a threshold of 30 µg m^{-3} and measures the relative improvement with respect to persistence forecast (range between [-1, 1] and a value of 0.5 or more indicates a significant improvement in skill).

The results are illustrated on the urban station JUS and, if necessary on the traffic station REP, but the tables refer to the performances reached on the six considered stations.

4.1.1. From learning to test performance

Of course, the training set performances are not relevant since it is well known that measures on training set are optimistically biased by definition but it can be interesting to consider it as a kind of reference for the testing performance.

Then the first question is to evaluate the loss from learning to testing performance. Fig. 5 illustrates this for the urban station JUS giving predicted versus observed PM_{10} plots, from Learn (on the left) to Test (on the right). The loss is small as it can be seen visually: the scatterplots are quite similar. In addition, the performance in terms of error measured by the root mean square error RMSE is only slightly increased (from 4.97 to 5.24) and the performance in terms of threshold exceedances measured by the threat-score TS is only slightly decreased (from 0.41 to 0.37).
Let us remark that the explained variance is high, close to 0.55, and that the skill score is positive and even close to 0.5 indicating that the persistence method is clearly outperformed.

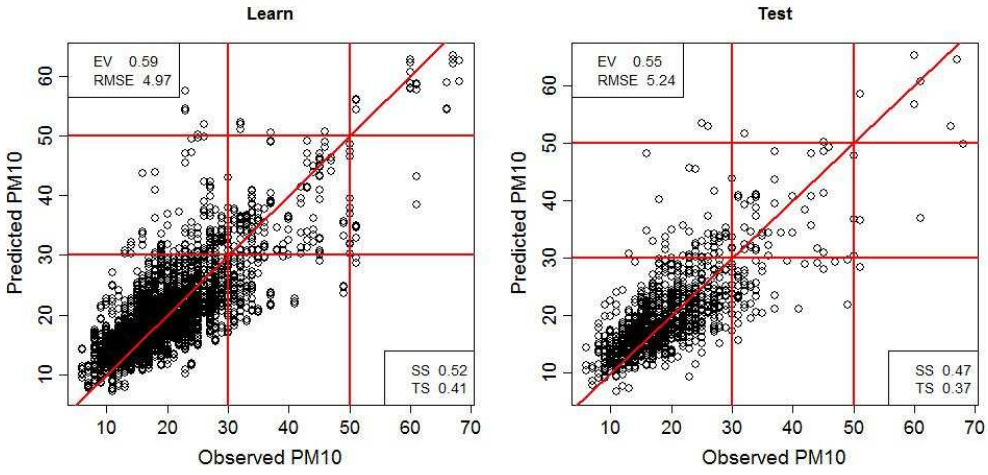

Fig. 5. Performance: predicted versus observed PM_{10} plots, from Learn (on the left) to Test (on the right) on JUS Station.

As it can be seen in Fig. 6, the same phenomenon occurs for the traffic station REP. The RMSE is only slightly modified (from 6.63 to 6.93) as well as the TS.

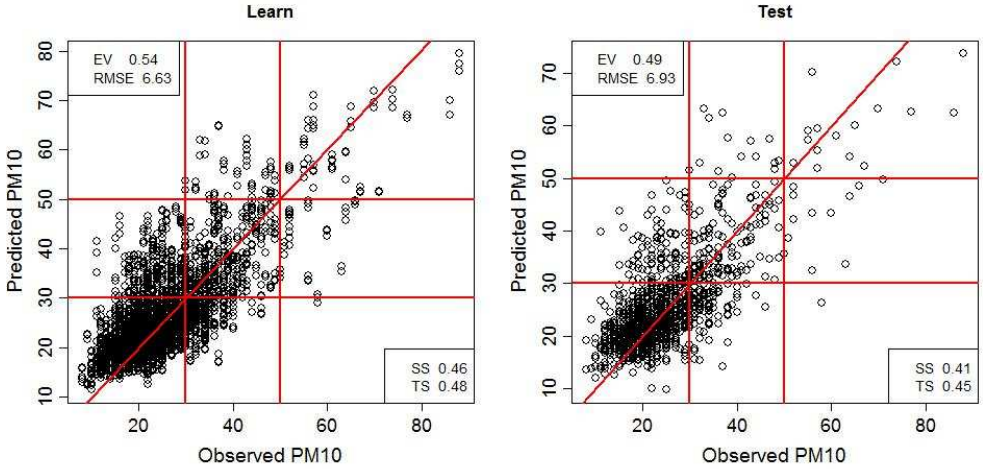

Fig. 6. Performance: predicted versus observed PM_{10} plots, from Learn (on the left) to Test (on the right) on REP Station.

4.1.2. From observed meteo to predicted meteo

The second question is to evaluate the loss on the learning performance observed by replacing measured meteorological variables by the meteorological predictions available the day before, in the morning and in the afternoon respectively. The performance is stable for the RMSE (from 5.24 to 5.8) as well as the TS (from 0.37 to 0.35) and the situation is globally stable as it can be seen in Fig. 7.

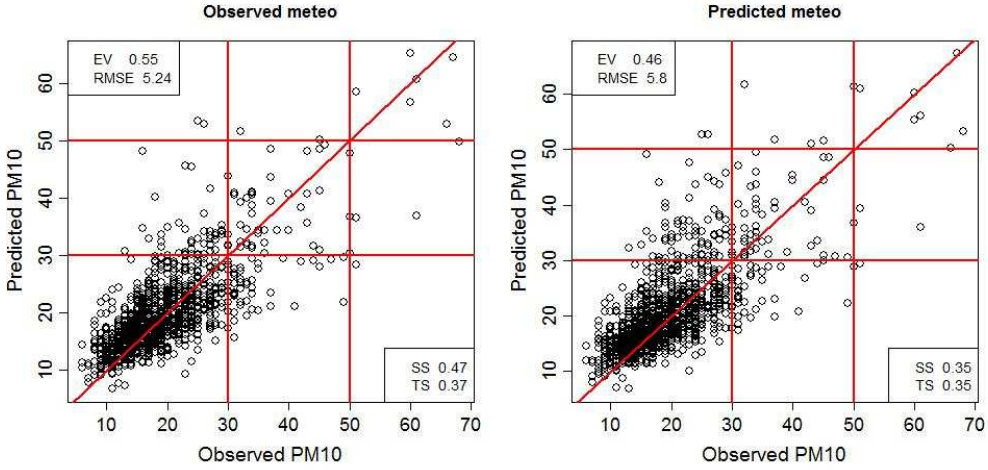

Fig. 7. Performance: predicted versus observed PM_{10} plots, from observed meteo (on the left) to predicted meteo (on the right) on JUS Station.

For each of the two horizons, a table allows to appreciate the forecasting performance evolution from observed meteo to predicted meteo.

Table 3: One day ahead forecasting performance using predicted meteo the day before in the morning.

	EV	RMSE	SS	TS
AIL	0.31	5.17	0.22	0.15
GCM	0.40	6.72	0.37	0.27
GUI	0.39	7.29	0.35	0.40
HRI	0.31	7.02	0.18	0.33
JUS	0.46	5.80	0.35	0.35
REP	0.23	9.44	-0.10	0.45

As it can be seen the performance is satisfactory for the typical stations but let us give some additional comments.
The best results in terms of exceedances are obtained for the traffic station REP (with a TS about 0.45) but the corresponding RMSE is the higher one across the stations (about 9.5). Results of medium quality are obtained for urban stations GUI, HRI and JUS with a TS about 0.3 to 0.4 and a RMSE about 6 to 7. The industrial station GCM near the cereal harbor of Rouen is a little bit hard to predict leading to a smaller TS about 0.27. Finally, the rural station AIL is not polluted, leading to a poor TS but a very satisfactory RMSE near 5.

Remark that if the objective of the work is to compare the accuracy achieved across different locations and different forecasting approaches, some additional work could be necessary to provide for the various indicators of performance a kind of "confidence" interval to take into account the additional variability due, for example, by the replacement of measures by predictions. But this is out of the scope of this paper.

Table 4: One day ahead forecasting performance using predicted meteo the day before in the afternoon.

	EV	RMSE	SS	TS
AIL	0.29	5.14	0.22	0.15
GCM	0.41	6.62	0.38	0.29
GUI	0.39	7.17	0.37	0.40
HRI	0.31	7.02	0.18	0.33
JUS	0.46	5.69	0.37	0.34
REP	0.32	8.04	0.20	0.40

Comparing the two previous Table 3 and Table 4, it can be seen that the forecasting performance is stable from the morning to the afternoon of the day before.

4.2. Intraday Forecasting: models and results

Let us now switch to the intraday forecasting problem. We consider two different horizons. Of course these forecasting tasks are simpler problems and for which we can expect both simpler models and better forecasting performance, with respect to one-day ahead forecasting problems.

The first discussion is to choose the variables considered in the model. Of course the basic meteorological variables (Tmoy, VVmoy, PAmoy, GTmax) are needed as well as the calendar variable (TypJ) but of course the endogenous variables needs to be adapted since we have observed PM_{10} concentrations of the beginning of the day: PM10_7h or PM10_13h respectively. In fact in the final models, it appears that the deterministic prediction delivered by Prev'air is not useful, so the currently available mean of the first hours of the day is the only endogenous variable.

The second idea is to imagine that a single linear model can suffice instead of a clusterwise regression model. Let us illustrate this point by a simple unweighted fit using GAM of JUS station, on the whole sample ignoring the calendar variable. As it can be seen the individual effects are essentially linear, especially those of large amplitude. In addition, the only nonlinear identified parts are related to a very small number of observations.

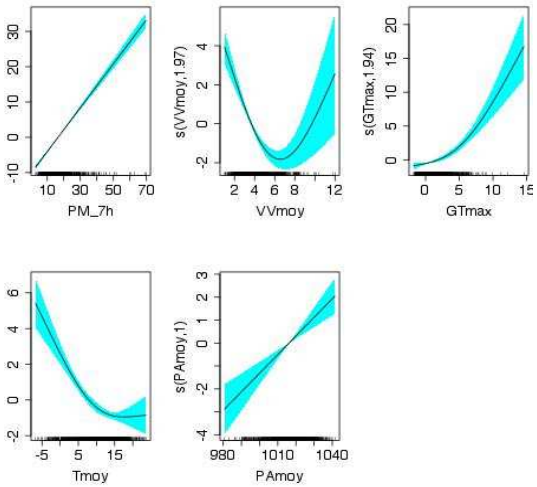

Fig. 8. A simple unweighted fit using GAM for JUS station: estimated effects of explanatory variables.

So we decide to consider a single linear multiple regression model (LM) but of course to benefit from the possibility to use different weights for the days in the least squares criterion, we use linearly weighted LM. Each day of the learning set is weighted by its daily mean PM_{10} concentration. Again, the structure of the forecasting model used is the same for the two horizons and the six stations.

In the sequel, the performance evaluation is calculated using cross validation on years, leave one year out.
For each of the two horizons, a table allows to appreciate the forecasting performance evolution. As in the previous sections, four scores are given: EV, RMSE for the error performance, SS and TS for the threshold exceedances performance.

Table 5: Intraday forecasting performance using predicted meteo in the morning.

	EV	RMSE	SS	TS
AIL	0.62	4.27	0.48	0.37
GCM	0.67	5.19	0.63	0.43
GUI	0.65	5.56	0.63	0.53
HRI	0.65	5.68	0.46	0.44
JUS	0.68	4.55	0.60	0.51
REP	0.65	6.33	0.50	0.55

As it can be seen the performance is satisfactory and very homogeneous despite the extreme diversity of the selected stations. Indeed, the TS is about 0.4 to 0.55 and the RMSE varies from 4.2 to 6.3. The explained variance is about 0.65.

Table 6: Intraday forecasting performance using predicted meteo in the afternoon.

	EV	RMSE	SS	TS
AIL	0.80	2.99	0.74	0.49
GCM	0.84	3.50	0.83	0.63
GUI	0.85	3.66	0.84	0.66
HRI	0.87	3.19	0.83	0.68
JUS	0.86	3.00	0.82	0.64
REP	0.85	3.85	0.82	0.73

Comparing Table 5 and Table 6, it can be seen that, from 7 a.m. to 1 p.m., a huge gain is reached both for errors and alarms, as expected.

Let us illustrate this conclusion by two plots of Fig. 9 giving the weighted LM forecasting performance at 7 a.m. on the left and at 1 p.m. on the right for the urban station JUS.

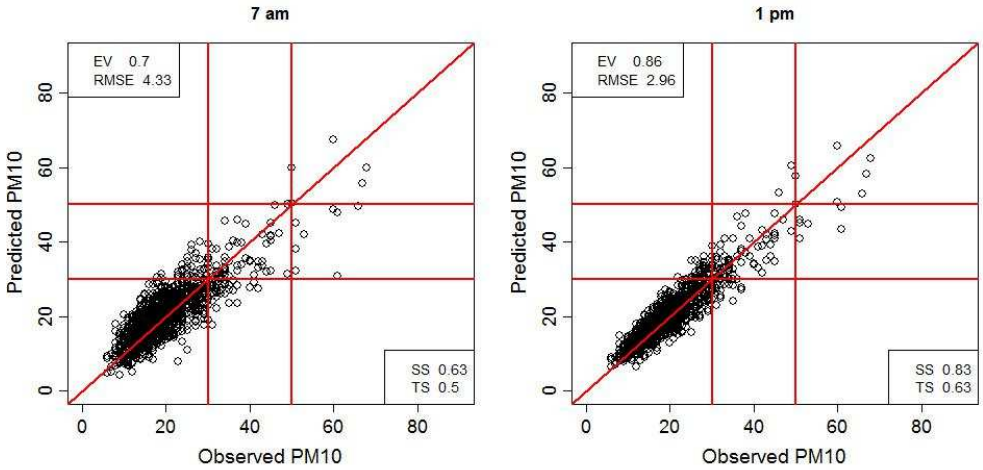

Fig. 9. Forecasting performance at 7 a.m. on the left and at 1 p.m. on the right for the urban station JUS.

A similar situation can be observed for the traffic station REP inspecting Fig. 10.

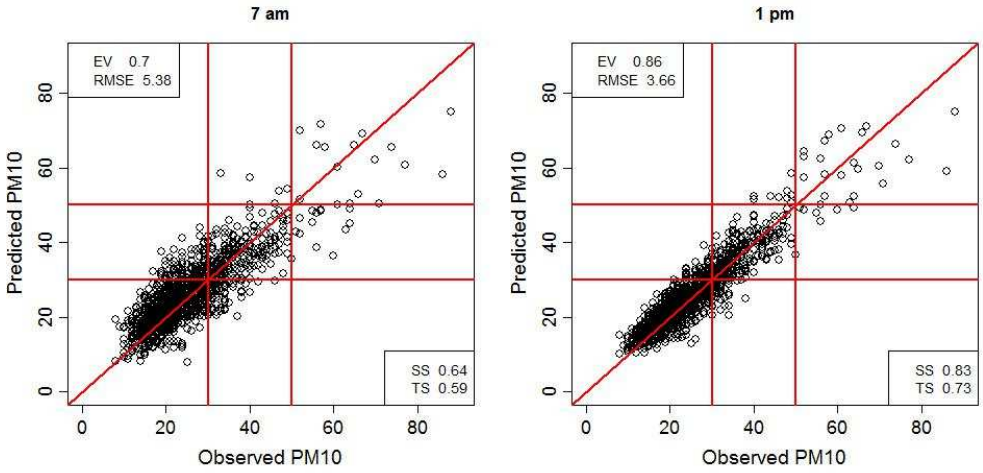

Fig. 10. Forecasting performance at 7 a.m. on the left and at 1 p.m. on the right for the urban station REP.

5. Conclusion

In this paper, we forecast the daily mean PM_{10} concentration by modeling it as a mixture of linear regression models involving meteorological predictors and the average concentration measured on the previous day. The values of observed meteorological variables are used for fitting the models but the corresponding predictions are considered for the test data, leading to realistic evaluations of forecasting performances, which are calculated through a leave-one-out scheme on the four years. We discuss several methodological issues including estimation schemes, introduction of the deterministic predictions of meteorological or numerical models and how to handle the forecasting at various horizons from some hours to one day ahead.

These models have been implemented very recently on the entire network of monitoring stations of the Normandie region (see the Air Normand website `http://www.airnormand.fr/`) following the methodology outlined in this paper. More precisely the different models, for one-day ahead and intraday horizons, are fitted and used in a complete prediction scheme by replacing, for the prediction of the next day, the daily mean PM_{10} concentration of day *J-1* which by its forecast conveniently chosen among the intraday predictions according to the time instant of calculation of the prediction for the next day. Of course, it is still too early to make a comprehensive assessment but early results are promising.

Acknowledgements

This work comes from a scientific collaboration between Air Normand, see the website `http://www.airnormand.fr/`, from the applied side and Orsay University and INSA Rouen from the academic side. We would like to thank Véronique Delmas and Michel Bobbia, from Air Normand, for providing the problem as well for supporting the statistical study. In addition, we want to thank Air Normand and Météo-France for providing PM_{10} data and meteorological data as well as predictions respectively.

References

Bobbia, M., Jollois, F.X., Poggi, J.M., Portier, B., 2011. Quantifying local and background contributions to PM10 concentrations in Haute-Normandie using random forests. Environmetrics 22, 758-768.

Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. The Science of the Total Environment 313, 1-13.

Corani, G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling 185, 513-529.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood for incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1-38.

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D., 2009. PM2.5 concentration prediction using hidden semi-markov model-based times series data mining. Expert Systems with Applications 36, 9046-9055.

Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of PM10 hourly concentrations, in the greater area of Athens, Greece. Atmospheric Environment 40, 1216-1229.

Grün, B., Leisch, F., 2007. Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis 51, 5247-5252.

Grün, B., Leisch, F., 2008. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. Journal of Statistical Software, 28, 1-35.

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

Jollois, F.X., Poggi, J.M., Portier, B., 2009. Three nonlinear statistical methods to analyze PM10 pollution in Rouen area, Case Stud. Bus. Ind. Gov. Stat. 3, 1-17.

Leisch, F., 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. Journal of Statistical Software 11, 1-18.

McLachlan, G., Peel, D., 2000. Finite mixture models, Wiley series in probability and statistics.

Paschalidou, A.K., Karakitsios, S., Kleanthous, S., Kassomenos, P.A., 2011. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Environmental Science and Pollution Research 18, 316-327.

Poggi, J.-M., Portier, B., 2011. PM10 forecasting using clusterwise regression, Atmospheric Environment 45, 7005-7014

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

Sfetsos, A., Vlachogiannis, D., 2010. Time series forecasting of hourly PM10 using localized linear models. Journal of Software Engineering and Applications 3, 374-383.

Slini, T., Kaprara, A., Karatzas, K., Moussiopoulos, N., 2006. PM10 forecasting for Thessaloniki, Greece. Environmental Modelling & Software 21, 559-565.

Stadlober, E., Hörmann, S., Pfeiler, B., 2008. Quality and performance of a PM10 daily forecasting model. Atmospheric Environment 42, 1098-1109.

Wilks, D.R., 1995. Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press.

Wood, S.N., 2006. Generalized additive models: an introduction with R. Chapman & Hall/CRC.

Zolghadri, A., Cazaurang, F., 2006. Adaptive nonlinear state-space modelling for the prediction of daily mean PM10 concentrations, Environmental Modelling & Software 21, 885-894.