Analyse Statistique de la Pollution par les PM10 en Haute-Normandie

François-Xavier Jollois, Jean-Michel Poggi, Bruno Portier

Analyse Statistique de la Pollution par les PM_{10} en Haute-Normandie

François-Xavier Jollois[a] & Jean-Michel Poggi[b] & Bruno Portier[c]

[a] *Laboratoire CRIP5, Université Paris Descartes, France,*
francois-xavier.jollois@parisdescartes.fr

[b] *Laboratoire de Mathématiques, Orsay, France and Université Paris Descartes, France,*
jean-michel.poggi@math.u-psud.fr

[c] *Laboratoire de Mathématiques, INSA de Rouen, France,*
bruno.portier@insa-rouen.fr

Résumé :
Ce travail porte sur l'analyse de la pollution par les particules PM_{10} en région Haute-Normandie entre 2004 et 2006. A l'aide de trois méthodes, les forêts aléatoires, les modèles additifs non linéaires et les mélanges de modèles linéaires, on modélise les effets des variables sur la pollution par les PM_{10} et on dégage les variables importantes en distinguant polluants et variables météorologiques. Dans la deuxième partie, on s'intéresse à une quantification d'une part locale et d'une part globale de la pollution par les PM_{10}, en essayant de donner un sens à ces notions dans ce contexte purement statistique sans aucune information directe sur les sources.

Mots clés : PM10, Pollution, Forêts aléatoires, Régression, Classification, Importance des Variables.

Abstract:
The problem is to analyze PM_{10} pollution during 2004-2006 in Haute-Normandie area using six different monitoring sites and to quantify the effects of variables of different types, mainly meteorological versus other pollutants measurements. Three modern non-parametric statistical methods, namely random forests, mixture of linear models and nonlinear additive models are first used to investigate it. Then, a second part focuses on an attempt of quantification of what we call in a broad sense a local part and a regional part of PM_{10} pollution.

Keywords: PM10, Pollution, Random Forests, Regression, Classification, Variable Importance.

1 Introduction

Let us briefly sketch the context of the work[1]. 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, has a network of a dozen of stations measuring every quarter of an hour, sometimes from 10 years, the concentrations of PM_{10} particles whose diameter is less than 10 μm, and expressed in way in a short time interval. The european regulation sets that PM_{10} daily average cannot exceeds $50\mu g/m^3$ more than 35 days per year. The objectives of the work 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, to distinguish situations in which the origin of particles is mainly local or rather the contrary distant or natural. The analysis is based on the PM_{10} concentrations from 2004 to 2006, and the associated weather data.

The bibliography about statistical analysis of PM_{10} contains hundreds of references. So we only mention a few typical ones, differing by their objectives and by the statistical tools used to investigate it: Salvador *et al.* (2004), Chavent *et al.* (2007), Karaca *et al.* (2005), Smith *et al.* (2001).

The talk focus on two aspects: pollution modeling and quantification of a local part and a regional part of PM_{10} pollution. We will introduce and motivate the three main methods used to handle the problem:

- random forests focusing on relative importance of variables and variable selection issues as well as marginal effects of variables;
- partially nonlinear additive model using two original climatic variables to partition data and model each cluster;
- cluster wise linear modeling.

Next, we will focus on an attempt of quantification of what we call in a broad sense a local part and a regional part of PM_{10} pollution. Finally, let us mention that the statistical study has been made using the R software.

2 Data

Among twelve monitoring stations for PM_{10} localized in Haute-Normandie, 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

[1]This work takes place in a scientific collaboration between Air Normand (see the website http://www.airnormand.fr/) from the applied side and Paris-Descartes University and INSA of Rouen from the academic side (see Jollois et al (2008)).

station JUS, the traffic station GUI, the second most polluted in the region, and GCM which is an industrial one 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 seaside. Lastly, we focus on the station AIL near Dieppe, because it is rural and coastal, and *a priori* not influenced by the social and industrial activity.

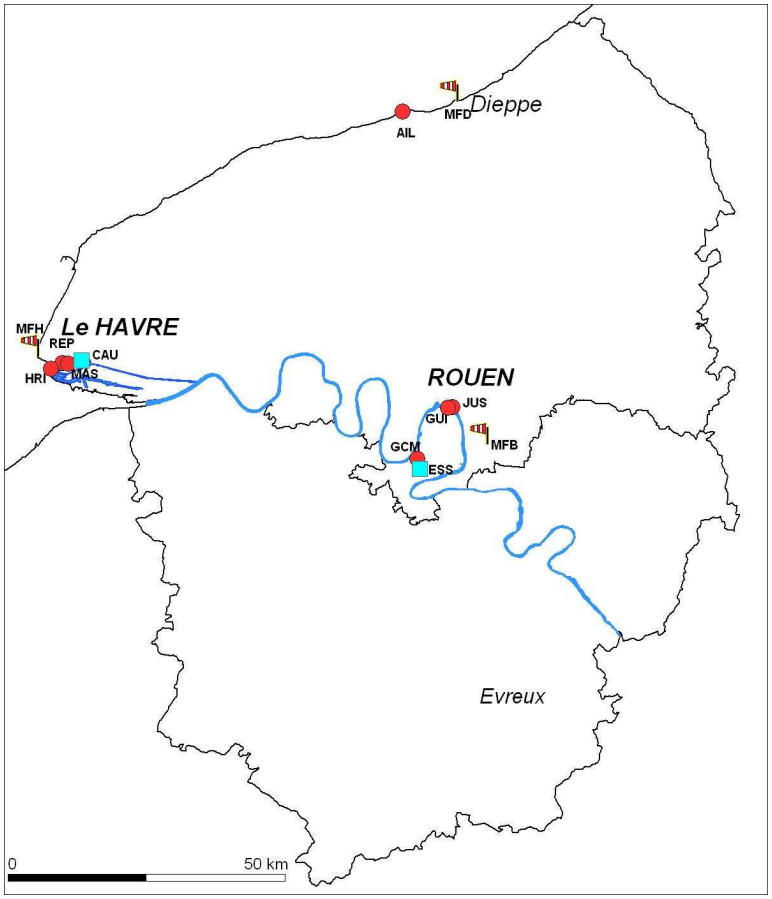

Figure 1: Map of the Haute-Normandie area locating the different monitoring sites of Air Normand and Météo France.

The pollution data analyzed are the TEOM PM_{10} daily mean concentrations and concern the period 2004-2006 (1096 days) coming from the six chosen monitoring sites of Air Normand.

To analyze the PM_{10} concentrations, we have daily meteorological data coming from three monitoring sites of MétéoFrance. The different meteorological parameters, which are calculated from hourly measurements on the period 0h-24h GMT, are the following ones: the daily temperature (min, max and mean), the maximum and mean daily wind speed, the daily total rain, the daily mean atmospheric pressure, the daily relative humidity (min, max and mean), the most frequently observed wind direction and the wind direction

associated with the maximum daily wind speed. We also have the temperature gradients measured by two monitoring sites of Air Normand at Rouen and Le Havre

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 the social activity and especially related to traffic while sulfur dioxide SO_2 captures the consequences of industrial activity.

3 Three nonlinear methods for PM_{10} modeling

Let us shortly present the three nonlinear statistical methods used to analyze and model PM_{10} pollution. Random forests are a very powerful method for prediction and variable importance quantification, introduced by Breiman (2001). 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). Some methodological remarks can be found in Genuer *et al.* (2008). By computing the marginal effects of each variable on the PM_{10} pollution, we get a rough idea of the shape of the influence of each, distinguishing pollutants and climatic variables. In addition, variable importance score allows 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 models are then considered. They are regression models by classes built according to different principles.

The first one is based on generalized additive models widely used (see the pioneer works of Buja *et al.* (1989), Hastie, Tibshirani (1990)) and particularly attractive since they represent an interesting compromise between the linear regression model and the fully nonparametric one. The associated R package is `mgcv` developed by Wood (2006) where the nonlinear functions are estimated using penalized regression splines.

We propose to fit 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, since they appear to be easy to understand and of highly nonlinear effect on PM_{10}.

The second one is based on mixture of linear models and builds class dependent linear models but the building strategy mixes more closely classification 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 allows more flexible classification as well as simpler models within a class but of course the classes are less directly interpretable. 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 mixture of linear regression models. The principle is given by Gruen and Leisch (2007) and the corresponding R implementation in Leisch (2004).

4 Local part and regional part

We then 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 make profit of the specificity of the rural station `AIL` for which there is *a priori* no local pollution sources.

The main idea is to use PM_{10} pollution measured at `AIL` (denoted by PM_{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 be not affected specifically by a major local production.

The importance of the variable PM_{AIL} in previous models when they are complemented by the introduction of this new variable leads to the following behavior : its importance is considerable, while the importance of meteorological variables significantly decrease. At the contrary, importance of pollutants remain stable, for all the stations.

So, these elements are compatible with the idea that PM_{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.

In addition, the effects obtained by fitting additive models are weakly increasing and weakly nonlinear. So the conclusion is that by introducing this new variable and canceling the meteorological variables, the model is linearized. Then concentrating on models involving pollutants locally measured and PM_{10} from `AIL`, we quantify more directly the respective parts of these two factors by fitting a simple linear model and computing the standardized coefficients.

Acknowledgements

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 Haute-Normandie land for supporting this work. Lastly, we thank Météo-France for authorizing the aggregated meteorological data to be publicly available.

Bibliographie

[1] L. Breiman, J.H. Friedman, R.A. Ohlsen, C.J. Stone (1984), *Classification and Regression Trees*, Belmont.

[2] Breiman, L. and Friedman, J. H. (1985), *Estimating optimal transformations for multiple regressionand correlation*, J. Am. Stat. Assoc., 80, 580-619.

[3] L. Breiman (2001), *Random Forests*, Machine Learning 45 (1), 5-32.
[4] L. Breiman, A. Cutler (2005), *Random Forests*, Berkeley. http://www.stat.berkeley.edu/users/breima
[5] Buja A., Hastie T. J. and Tibshirani, R. J. (1989), *Linear smoothers and additive models*,Ann. of Stat., 17, 453-510.
[6] Chavent M. , Guégan H., Kuentz V., Patouille B., Saracco J. (2007), *Apportionment of air pollution by source at a French urban site*, Case Studies in Business, Industry and Government Statistics (CSBIGS), Vol 1-Issue 2, 119-129.
[7] R. Genuer, J-M. Poggi, C. Tuleau (2008), *Random Forests: some methodological insights*, PreprintINRIA-Select, 1-38.
[8] B. Gruen and F. Leisch (2007). *Fitting finite mixtures of generalized linear regressions in R*. Computational Stat. and Data Analysis, 51(11), 5247-5252.
[9] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*,Chapman & Hall.
[10] T. Hastie, R. Tibshirani, J.H. Friedman (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*,Springer.
[11] R.J. Hathaway, J.C. Bezdek (1993), *Switching regression models and fuzzy clustering*, IEEE Trans. Fuzzy Systems 1 (3), 195-203.
[12] F-X. Jollois, J-M. Poggi, B. Portier (2008), *Three non-linear statistical methods to analyze* PM_{10} *pollution in Rouen area*, 41 pages, submitted for publication.
[13] F.X. Jollois, J.M. Poggi, B. Portier (2008), *Analyse statistique de la pollution par les particules en Haute-Normandie*, Rapport de contrat de recherche, Air Normand.
[14] F. Karaca, O. Alagha, F. Erturk (2005), *Statistical characterization of atmospheric PM10 and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey*, Chemosphere 59, 1183-1190.
[15] F. Leisch (2004). *FlexMix: A general framework for finite mixture models and latent class regression in R*. Journal of Statistical Software, 11(8). http://www.jstatsoft.org/v11/i08/
[16] A. Liaw, M. Wiener (2002), *Classification and Regression by randomForest*, R News,2(3), 18-22.
[17] G. McLachlan, D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics
[18] P. Salvador, B. Artinano, D. G. Alonso, X. Querol, A. Alastuey (2004), *Identification and characterisation of sources of PM10 in Madrid (Spain) by statistical methods*, Atmospheric Environment 38, 435-447.
[19] S. Smith, F. T. Stribley, P. Milligan, B. Barratt (2001), *Factors influencing measurements of PM10 during 1995-1997 in London*, Atmospheric Environment 35, 4651-4662.
[20] S.N. Wood (2006), *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.