Mixture of experts for sequential PM10 forecasting in Normandy (France)

B. Auder [1], J. M. Poggi [2,*] and B. Portier [3]

[1] Univ. Paris-Sud Orsay, benjamin.auder@math.u-psud.fr
[2] Univ. Paris-Sud Orsay et Univ. Paris Descartes, jean-michel.poggi@parisdescartes.fr
[3] Normandie Université, INSA Rouen, bruno.portier@insa-rouen.fr

*Corresponding author

Abstract. *Within the framework of air quality monitoring in Normandy, we experiment the methods of sequential aggregation for forecasting concentrations of PM10 of the next day. Besides the field of application and the adaptation to the special context of the work of the forecaster, the main originality of this contribution is that the set of experts contains at the same time statistical models built by means of various methods and different sets of predictors, as well as experts which are deterministic chemical models of prediction modeling pollution, weather and atmosphere.*

Numerical results on recent data from April 2013 until March 2014, on three monitoring stations, illustrate and compare various methods of aggregation. The obtained results show that such a strategy improves clearly the performances of the best expert both in errors and in alerts and reaches an unbiased observed-forecasted scatterplot, difficult to obtain by usual methods.

Keywords. *Air quality; Forecasting; Mixture of experts; PM10; Sequential prediction.*

1 Introduction

In Normandy (France), Air Normand together with Air COM monitor air quality. During a recent research project between academia and air quality agency (see Poggi, Portier 2011), the statistical forecasting of PM10 has been considered, with the aim of improving warning procedures. It led to the development of operational procedures allowing to forecast the daily average of the PM10 for the current day and for the next day on various horizons of forecast integrating the meteorological information and the model outputs statistics.

More generally, Air Normand has various operational tools for the analysis of episodes and for the interpretation of measures, in view of decisions. However these complementary tools, statistical or deterministic models, local or global, often supplying different forecasts especially because of the various space and time resolutions considered.

In this paper, we evaluate the interest of using sequential aggregation or mixing of experts to develop decision-making tools for the forecasters of Air Normand.

In the context of sequential prediction, experts make predictions at each time instance, and the forecaster determines, step by step, the future values of an observed time series. To build his prediction, it/he has to combine before each instant the forecasts of a finite set of experts. To do so, adopting the deterministic and robust view of the literature of the prediction of individual sequences, three basic references can be highlighted: the survey paper by Clemen (1989), the book of Cesa-Bianchi and Lugosi (2006) and the paper by Stoltz (2010) in French.

In the application framework at hand, empirical studies are particularly valuable and we can mention some studies. In the area of climate Monteleoni et al. (2011), in the field of the air quality Mallet (2010), Mallet et al. (2009), the use of the quantile prediction of the number of daily calls in a call center Biau et al. (2011) and finally the prediction of electricity consumption Devaine et al. (2013). These studies focus on the rules of aggregation of a set of experts and examine how to weight and combine these experts.

Besides the field of application and the adaptation to the special context of the work of the forecaster, the main originality of this work is that the set of experts contains together:
- experts coming from statistical models constructed using different methods and different sets of predictors;
- experts defined by deterministic models of physicochemical prediction modeling pollution, weather and atmosphere. The models are of similar nature but of different spatial and time resolutions with or without statistical adaptation;
- and finally references such as persistence, as usual.

The aforementioned studies combine "homogeneous" methods: only statistical methods or only deterministic ones. Sequential prediction allows mixing several models built on very different assumptions in a unified approach that does not require any prior knowledge about the internal way to use for each expert to generate predictions. It is therefore particularly suitable for our application.
Note that the recent reference Gaillard et al. (2014) is also interested in the design of new experts to be included in the combination, suggesting potential to enrich the basic core of the initial experts, using resampling methods.

2 Data, basic models and aggregation methods

The study period extends from April 3, 2013 to March 18, 2014 (351 days). We have measures of the daily average concentration of PM10 (including the volatile fraction) and in addition the forecasts of the day for the day after of the daily average concentration of PM10 coming from 9 different prediction models. We consider two urban background stations in the Air Normand network HRI (Le Havre) and PQV (Rouen) and an urban background station in the Air COM network LIS (Lisieux).

Numerous forecasting models of different nature are available:
- four statistical models (see Poggi, Portier (2011)): a mixture of regression models with two classes, two linear models (one fitted on slightly polluted days and the other one, on polluted days), and a non-linear additive model;
- three numerical models (Esmeralda and two PREV'AIR models at different spatial resolutions (see `http://www2.prevair.org/` and `http://www.esmeralda-web.fr/`);
- two deterministic models with statistical adaptation (Esmeralda and PREV'AIR);
- the persistence model.

Among the methods used for sequential aggregation strategy, we can distinguish two subsets of different nature.
First, the methods, starting from an initial weighting between the experts and initial performance, are changing the weights adaptively updating the weights at each step. In this category, we will focus on the exponential weight method, called **EWA (Exponential Weighted Average)**.
The second category consists of methods that optimize at every step a global criterion on the history of measurements and expert predictions. Of course, the past can be restricted to a window or the observations can be weighted to emphasize recent ones without omitting those of the past however

distant. In this category, we will focus on the minimization of a quadratic criterion with a quadratic penalty on the coefficients of the mixture, which regularize them (it is the **ridge regression (RR)** framework) or with a l_1-penalty which tend to cancel small coefficients (it is the lasso regression framework).

3 The results

The three main conclusions about the value of the strategy of the mixture of experts are as follows.

First, the comparative **performance** is stable across the three considered stations and the contribution of the combination is clear. Qualitatively, the RR method of mixture of experts provides unbiased predicted-observed scatterplots to be compared with the one of best expert and the EWA mixture (see Figure 1). For quantitative performance: RR (Ridge Regression) is dominant for alerts (see the Threat Score TS) and EWA (Exponential Weighted Average) is the best for the RMSE (Root Mean Square Error). In addition, note that the two methods outperform the best expert.

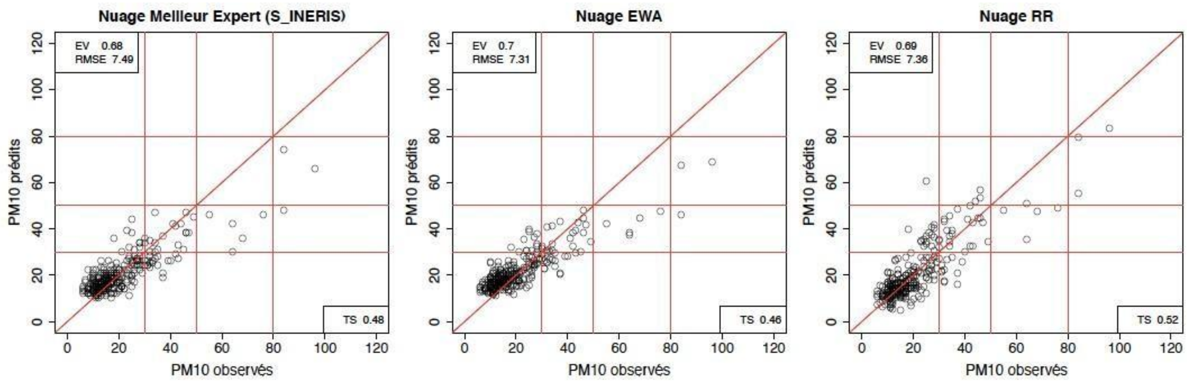

Figure 1: Scatterplots predicted-observed of the mixtures of experts for station HRI (Rouen). Best expert on the left, EWA in the middle and RR on the right.

Even if EWA is not the best one, it deserves to be considered in the future, especially because history will increase and we will more suitably chose the time window. So, they have to be considered as RR methods for deeper investigations in the next step towards operational implementation.

Second, **cooperation between statistical and deterministic** models is helpful, and if two models are often associated with higher weights, all of the models are involved. Indeed, for all the stations, 3 families of methods (deterministic, statistical Air Normand, other statistics) are useful. The sums of the absolute values of weight per family range between 0.2 and 0.4 for EWA and between 0.5 and 3 (and 2 fast enough) to RR. Finally note that the persistence has a small but positive and useful contribution.

Third, if the most useful **methods** are few, no basic method is to depart since an interpretation of the unbiased forecasts of RR and good performance balanced between false alarms and missed alarms, is twofold: first, for each day some methods overestimate and others underestimate (and of course these subsets change with time) and secondly RR (as well as lasso) is best able (with respect to EWA type methods) to take advantage of this situation because it is not constrained by the convexity of weights.

Acknowledgments. This work is the result of a research collaboration between, on the one hand, for the applied side, Air Normand (`http://www.airnormand.fr`) and secondly, for the academic side, the University of Orsay and INSA Rouen. We thank Véronique Delmas, Head of Air Normand, for the problem, the data and for supporting the statistical study. We also thank Pierre Gaillard, Yannig Goude and Michel Bobbia for fruitful discussions.

References

[1] Biau, G., Patra, B. (2010), Sequential quantile prediction of time series. IEEE Transactions on information Theory 57(3), 1664–1674.

[2] Cesa-Bianchi, N., Lugosi, G. (2006), Prediction, Learning, and Games, Cambridge University Press.

[3] 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.

[4] Clemen, R.T. (1989), Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5(4), 559–583.

[5] Devaine, M., Gaillard, P., Goude, Y., Stoltz, G. (2013), Forecasting electricity consumption by aggregating specialized experts, Machine Learning 90(2), 231–260.

[6] Dietterich T.G. (2000), Ensemble methods in machine learning. In Multiple classifier systems, 1-15, Springer Berlin Heidelberg.

[7] Freund, Y., Schapire, R.E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

[8] Gaillard, P., Stoltz, G., van Erven, T. (2014), A second-order bound with excess losses. ArXiv:1402.2044.

[9] Gaillard, P., Goude, Y. (2015), Forecasting electricity consumption by aggregating experts; how to design a good set of experts. Book chapter to appear in Lecture Notes in Statistics, Springer.

[10] Mallet, V., Stoltz, G., Mauricette, B. (2009), Ozone ensemble forecast with machine learning algorithms. Journal of Geophysical Research 114(D05307).

[11] Mallet, V. (2010), Ensemble forecast of analyses. Coupling data assimilation and sequential aggregation. Journal of Geophysical Research 115(D24303).

[12] Monteleoni, C., Schmidt, G.A., Saroha, S., Asplund, E. (2011), Tracking climate models. Statistical Analysis and Data Mining 4(4), 372–392.

[13] Poggi J.-M., Portier B. (2011), PM10 forecasting using clusterwise regression, Atmospheric Environment, 45(38), 7005-7014.

[14] Siwek K., Osowski S., Garanty K., and Osowski S. (2009), Ensemble of predictors for forecasting the PM10 pollution, Proceedings of VXV International Symposium on Theoretical Electrical Engineering (ISTET), pp. 318-322.

[15] Stoltz G. (2010), Agrégation séquentielle de prédicteurs: méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique, Journal de la Société Française de Statistique, 151(2):66-106.