

# Mélange de prédicteurs pour la prévision séquentielle de la pollution par les PM10 en Haute Normandie

Benjamin Auder <sup>1</sup>, Jean-Michel Poggi <sup>2</sup> & Bruno Portier <sup>3</sup>

<sup>1</sup> Univ. Paris-Sud Orsay, benjamin.auder@math.u-psud.fr

<sup>2</sup> Univ. Paris-Sud Orsay et Univ. Paris Descartes, jean-michel.poggi@parisdescartes.fr

<sup>3</sup> Normandie Université, INSA Rouen, bruno.portier@insa-rouen.fr

**Résumé.** Dans le cadre de la surveillance de la qualité de l'air en Normandie, on expérimente les méthodes d'agrégation séquentielle pour la prévision le jour pour le lendemain, des concentrations de PM10.

Outre le domaine d'application et l'adaptation au contexte concret du travail du prévisionniste, la principale originalité de ce travail est que l'ensemble initial d'experts contient à la fois des modèles statistiques bâtis à l'aide de différentes méthodes et d'ensemble de prédicteurs divers, ainsi que des experts qui sont des modèles déterministes de prédiction physico-chimiques modélisant pollution, météo et atmosphère.

Des résultats numériques sur des données récentes d'avril 2013 à mars 2014, sur trois sites de mesures, illustrent et comparent diverses méthodes d'agrégation. Les résultats obtenus montrent qu'une telle stratégie améliore nettement les performances du meilleur expert tant en erreurs qu'en alertes et assure un « débiaisage » du nuage réalisés-prévus, difficile à obtenir d'ordinaire.

**Mots-clés.** Mélange d'experts, PM10, Prévision, Prédiction séquentielle, Qualité de l'air

**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 work is that initial set of experts contains at the same time statistical models built by means of various methods and groups 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 till 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 the “unbiasedness” of observed-forecasted scatterplot, difficult to obtain by usual methods.

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

## 1. Introduction

En France, la surveillance de la qualité de l'air est prise en charge dans chaque région par une association officielle, agréée par le Ministère de l'Ecologie. En Normandie, composée de deux régions administratives, Air Normand, basée à Rouen (Haute-Normandie), et Air COM, basée à Caen (Basse-Normandie), surveillent la qualité de l'air. En plus de leurs fonctions de contrôle, leur rôle est d'informer le public sur l'air qu'il respire. Ainsi, pour accomplir leurs missions, Air Normand et Air COM mesurent la qualité de l'air avec des analyseurs disséminés dans la région, et publient ces mesures, en particulier par l'utilisation du web pour informer le public sur son exposition à la pollution de l'air.

Dans le cadre d'une collaboration de recherche récente (cf. Poggi, Portier 2011) entre Air Normand, l'INSA de Rouen et l'Université d'Orsay, la prévision statistique des PM10, dans le but d'améliorer les procédures d'alerte a été examinée. Le travail a conduit à la mise au point de procédures opérationnelles permettant de prévoir la moyenne journalière des PM10 pour le jour en cours et pour le lendemain à divers horizons de prévision en intégrant aussi les informations météorologiques et les sorties de modèles numériques. Plus largement, Air Normand dispose d'outils opérationnels riches et variés pour l'analyse d'épisodes et l'interprétation des informations mesurées, en vue de décisions. Cependant ces outils complémentaires, statistiques ou déterministes, fournissent souvent des prévisions différentes en particulier, à cause des différentes échelles en temps et en espace considérées.

Il s'agit dans ce travail d'une première étude visant à évaluer l'intérêt d'élaborer des outils d'aide à la décision des prévisionnistes d'Air Normand basés sur l'utilisation de techniques d'agrégation séquentielle combinant ces diverses sources de prévision.

## 2. Prévision séquentielle

Dans le cadre de la prédiction séquentielle, des experts font des prédictions au fur et à mesure du temps, et le prévisionniste doit déterminer pas à pas, les valeurs futures d'une série temporelle observée. Pour construire sa prédiction, il dispose et combine avant chaque échéance, l'avis d'un ensemble fini d'experts. Pour ce faire, il adopte le point de vue déterministe et robuste de la littérature de la prédiction des séquences individuelles dont trois références de base sont le papier de revue par Clemen (1989), le livre de Cesa-Bianchi et Lugosi et l'excellent papier en français de Stoltz (2010).

Dans le cadre applicatif qui nous occupe, les travaux empiriques sont particulièrement précieux et on peut mentionner quelques études. Dans le domaine du climat Monteleoni et al. (2011), de la qualité de l'air Mallet (2010), Mallet et al. (2009), la prédiction quantile du nombre d'appels quotidiens dans un call center Biau et al. (2011) et enfin la prévision de la consommation électrique Devaine et al. (2013). Ces études se concentrent sur les règles d'agrégation d'un ensemble d'experts et examine la manière de pondérer ces experts.

L'apport de notre étude est multiple, tout d'abord le domaine d'application, ou encore l'adaptation au contexte concret du travail du prévisionniste d'astreinte, mais la principale originalité est que l'ensemble initial d'experts contient à la fois :

- des experts qui proviennent des modèles statistiques bâtis à l'aide de méthodes différentes et d'ensemble de prédicteurs différents ;
- des experts qui sont des modèles déterministes de prédiction physico-chimiques modélisant pollution, météo et atmosphère. Les modèles sont de même nature mais de résolution spatiale différente avec adaptation statistique ;
- et enfin d'autres qui sont des références comme par exemple la persistance, comme cela est classique.

Les études mentionnées précédemment ne mélangent que des méthodes « homogènes », soit uniquement des méthodes statistiques, soit déterministes. Le cadre de la prédiction séquentielle permet de mélanger plusieurs modèles construits selon des hypothèses très différentes en une approche unifiée qui ne requiert a priori aucun prérequis. Il est donc particulièrement adapté à notre application.

Il faut noter que la référence récente Gaillard et al. (2014) s'intéresse aussi à la conception de l'ensemble des experts à inclure dans la combinaison, en proposant des pistes pour enrichir par rééchantillonnage le noyau de base des experts initiaux.

### 3. Les données et les modèles de base

La période d'étude s'étend du 3 avril 2013 au 18 mars 2014 (351 jours). On dispose des mesures de la concentration moyenne journalière en PM10 (incluant la fraction volatile) ainsi que des prévisions du jour pour le lendemain de la concentration moyenne journalière en PM10 issues de 8 modèles de prévision. On considère deux stations de fond urbain du réseau d'Air Normand, HRI (Le Havre) et PQV (Rouen) et une station de fond urbain du réseau d'Air C.O.M, LIS (Lisieux).

Neuf modèles de prévision de base (sans compter la persistance) sont disponibles :

- CLM : modèle de mélange de régression linéaire à deux classes ;
- CLM1 : modèle linéaire de la classe 1 (des jours peu pollués) du modèle CLM ;
- CLM2 : modèle linéaire de la classe 2 (des jours pollués) du modèle CLM ;
- GAM : modèle additif non linéaire ;
- S\_AIRPARIF : modèle Esmeralda (AirParif) avec adaptation statistique ;
- S\_INERIS : modèle Prev'air (INERIS) avec adaptation statistique ;
- D\_ESMERALDA : modèle brut Esmeralda ;
- D\_PREVAIR et D\_PREVAIR2 : deux modèles Prev'air à des résolutions spatiales différentes ;

auxquels on rajoutera

- PERSIS : modèle de persistance.

On trouvera des détails sur les quatre premiers modèles dans Poggi, Portier (2011), et dans <http://www2.prevaair.org/> et <http://www.esmeralda-web.fr/> pour les derniers.

Une première étude préliminaire sur les prédicteurs de base permet de noter que les modèles les plus performants du point de vue de l'erreur de prévision sont différents de ceux permettant de mieux prévoir les fortes concentrations en PM10. Ainsi, les modèles considérés sont

complémentaires et suffisamment riches pour envisager une approche de type « mélange de prévisions ».

#### 4. Les méthodes d'agrégation comparées

Parmi les méthodes qui sont utilisées, on peut en distinguer deux types de nature différente.

Tout d'abord, les méthodes qui, partant d'une pondération initiale entre les experts et de performances initiales, font évoluer les pondérations de façon adaptative en mettant à jour les pondérations à chaque pas. Dans cette catégorie, on s'attachera à la méthode des poids exponentiels, dite **EWA (Exponential Weighted Average)** et consistant à mettre à jour les poids selon la politique suivante :

$$p_{k,t} = c \cdot \exp(-\eta \sum_{s=1,t-1} L_s(x_{k,s}))$$

où la constante  $c$  assure simplement que  $\sum_k p_{k,t} = 1$ .

Au départ de l'algorithme, en général, les poids sont tous égaux à  $1/K$  et lorsque le paramètre  $\eta$  est convenablement choisi, le regret est en  $O(1/\sqrt{T})$ . On pourra utiliser la méthode décrite dans Stoltz (2010) pour le choisir.

Mentionnons enfin, dans cette même famille, la méthode des poids polynomiaux (dite ML-poly).

La seconde catégorie est constituée des méthodes qui, à chaque pas optimisent un critère global portant sur l'historique des mesures et des prédictions des experts. Naturellement, le passé peut être restreint à une fenêtre ou les observations peuvent être pondérées de manière à privilégier les observations récentes sans omettre celles du passé même lointain. Dans cette catégorie, on s'attachera à la minimisation d'un critère quadratique :

- avec une pénalisation quadratique sur les coefficients du mélange, qui les régularise, c'est le cadre de la **régression ridge (RR)** ;  
Typiquement, on part de poids tous égaux à  $1/K$  et ensuite, pour chaque  $t$  donné, on choisit les poids par minimisation sous contraintes  

$$p_{.,t} = \arg \min \{ \sum_{s=h,t-1} (y_s - \mathbf{u} \cdot \mathbf{x}_s)^2 + \lambda \|\mathbf{u} - \mathbf{p}_0\|_2^2 ; \mathbf{u} \in \mathbb{R}^K \}$$
où  $\lambda$  est la pénalité ridge à choisir convenablement et  $h$  est une fenêtre ;  
temporelle conduisant à n'optimiser le critère que les  $h$  prédictions les plus récentes.
- avec une pénalisation  $l_1$  :  

$$p_{.,t} = \arg \min \{ \sum_{s=h,t-1} (y_s - \mathbf{u} \cdot \mathbf{x}_s)^2 + \lambda \|\mathbf{u} - \mathbf{p}_0\|_1 ; \mathbf{u} \in \mathbb{R}^K \}$$
qui a tendance à annuler les coefficients, c'est le cadre de la régression lasso.

#### 5. Les résultats

Les principales conclusions quant à l'intérêt du mélange pour les trois stations retenues sont les suivantes.

Tout d'abord, les *performances* comparées sont stables et l'apport de la combinaison est manifeste : à la fois du côté du débiaisage, il suffit pour cela de regarder l'allure des graphes réalisés-prévus des mélanges et les comparer à celui du meilleur expert. Il en est de même du

côté des performances : RR (Régression Ridge) est dominant pour les alertes et EWA (Exponential Weighted Average) l'est pour le RMSE.

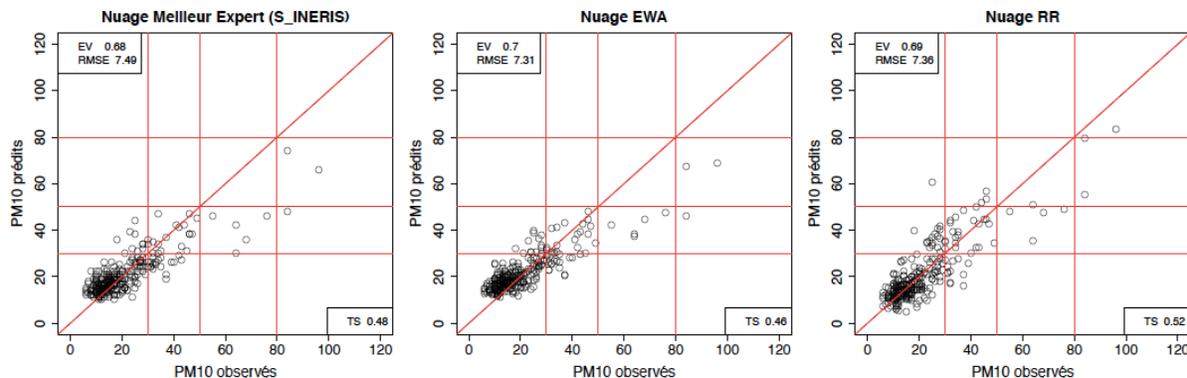


Figure 1: graphes des PM10 réalisés-prévus sur la station HRI. Meilleur expert à gauche, EWA au centre et RR (Régression Ridge) à droite.

Ensuite, les *méthodes* EW, si elles sont devancées, restent intéressantes surtout si l'historique s'accroît et que l'on choisit convenablement la fenêtre temporelle. Elles sont donc à conserver au même titre que les méthodes RR.

Enfin, la *coopération entre modèles statistiques et déterministes* est utile, et si S\_INERIS et ESMEALDA ressortent du lot, l'ensemble des méthodes est utile. En effet, pour l'ensemble des stations, les 3 familles de méthodes (déterministes, statistiques d'Air Normand, statistiques autres) sont utiles. Les sommes des valeurs absolues des poids par famille oscillent entre 0.2 et 0.4 pour EWA et entre 0.5 et 3 (puis 2 assez rapidement) pour RR. Enfin notons que la persistance qui a la contribution la plus faible puisqu'elle est seule dans la famille, a une contribution modeste mais positive.

Ainsi, si les méthodes les plus utiles sont peu nombreuses, aucune méthode n'est à écarter car une interprétation du débiaisage de RR et des bonnes performances équilibrées entre fausses alertes et alertes non faites, est double : d'une part, chaque jour voit des méthodes surestimer tandis que d'autres sous-estiment et d'autre part RR est le plus à même à tirer parti de cette situation puisqu'il n'est pas contraint par la convexité.

## Remerciements

Ce travail est issu d'une collaboration de recherche entre, d'une part, du côté appliqué, Air Normand (<http://www.airnormand.fr>) et d'autre part, pour le côté académique, l'Université d'Orsay et l'INSA Rouen.

Nous tenons à remercier Véronique Delmas, d'Air Normand, pour le problème, les données ainsi que pour avoir soutenu l'étude statistique. Nous remercions également Pierre Gaillard, Yannig Goude et Michel Bobbia pour les discussions fructueuses.

## Bibliographie

- [1] Biau, G., Fischer A., Guedj B., Malley B., COBRA: A Nonlinear Aggregation Strategy, arXiv e-print (arXiv:1303.2236) (2013)
- [2] Biau, G., Patra, B., Sequential quantile prediction of time series. *IEEE Transactions on Information Theory* 57(3), 1664–1674 (2011)
- [3] Cesa-Bianchi, N., Lugosi, G., *Prediction, Learning, and Games*, Cambridge University Press (2006)
- [4] Chaloulakou A., Saisana M., Spyrellis N., Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens, *The Science of the Total Environment* 313, 1-13 (2003).
- [5] Clemen, R.T., Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5(4), 559–583 (1989)
- [6] Devaine, M., Gaillard, P., Goude, Y., Stoltz, G., Forecasting electricity consumption by aggregating specialized experts, *Machine Learning* 90(2), 231–260 (2013)
- [7] Dietterich T.G. Ensemble methods in machine learning. In *Multiple classifier systems*, 1-15, Springer Berlin Heidelberg (2000)
- [8] Freund, Y., Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119–139 (1997)
- [9] Gaillard, P., Stoltz, G., van Erven, T., A second-order bound with excess losses. *ArXiv:1402.2044* (2014).
- [10] Gaillard, P., Goude, Y., Forecasting electricity consumption by aggregating experts; how to design a good set of experts. Book chapter to appear in *Lecture Notes in Statistics*, Springer (2015)
- [11] Gerchinovitz S., Mallet V., Stoltz G., A further look at sequential aggregation rules for ozone ensemble forecasting, *Technical report, INRIA and École Normale Supérieure de Paris* (2008)
- [12] Mallet, V., Stoltz, G., Mauricette, B. Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research* 114(D05307) (2009)
- [13] Mallet, V., Ensemble forecast of analyses. Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research* 115(D24303) (2010)
- [14] Monteleoni, C., Schmidt, G.A., Saroha, S., Asplund, E., Tracking climate models. *Statistical Analysis and Data Mining* 4(4), 372–392 (2011)
- [15] Poggi J.-M., Portier B., PM10 forecasting using clusterwise regression, *Atmospheric Environment*, 45(38), 7005-7014 (2011)
- [16] Siwek K., Osowski S., Garanty K., and Osowski S., Ensemble of predictors for forecasting the PM10 pollution, *Proceedings of VXXV International Symposium on Theoretical Electrical Engineering (ISTET)*, pp. 318-322 (2009)
- [17] Stoltz G., 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 (2010).