Comparaison de techniques de « Data Mining » pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE

A comparison of Data Mining techniques for the statistical adaptation of ozone forecasts of chemistry-transport MOCAGE model

Philippe BESSE*, Hélène MILHEM*, Olivier MESTRE*,**, Anne DUFOUR***, Vincent-Henri PEUCH***

Résumé

Le modèle MOCAGE développé par Météo-France est capable de simuler les interactions entre les phénomènes dynamiques, physiques et chimiques depuis l'échelle de la planète jusqu'à l'échelle régionale. Il permet d'effectuer des prévisions de qualité de l'air et fait partie du dispositif d'alerte des pouvoirs publics. Les prévisions de ce modèle déterministe sont entachées de biais, que l'on cherche à corriger par « adaptation statistique ». Dans cette étude, on s'intéresse aux prévisions de concentrations d'ozone sur cinq sites pour lesquels les prévisions déterministes sont particulièrement mises en défaut. On compare les résultats de méthodes linéaires basées sur la régression linéaire (avec ou sans interactions) avec des méthodes non linéaires : réseaux de neurones, arbres de segmentation, modèles d'agrégation (bagging, forêt aléatoire) et séparateur à vaste marge. Les meilleures méthodes sont une méthode linéaire, l'analyse de covariance lorsque l'on introduit les interactions entre prédicteurs et la méthode de la forêt aléatoire. On passe alors d'un écart-type d'erreur de prévision d'environ 36 µg/m³ pour MOCAGE à un écart-type d'erreur de prévision de l'ordre de 26 µg/m³ sur fichiers tests après adaptation statistique.

Mots clés

Adaptation statistique. Prévision. Ozone. Data mining.

Abstract

The MOCAGE model, developed by Météo-France is able to simulate interactions between dynamic, physical and chemical processes of the atmosphere, from global to regional scales. This model allows air quality forecasts, and is part of the pollution warning system. The forecasts of this physical model being biased, statistical models are calibrated between forecasts and observations, in order to realize a statistical forecast based on MOCAGE outputs. In this study, we focussed on ozone forecasts over five peculiar sites for which the deterministic approach alone exhibits large errors. Standard linear regression techniques are compared with non-linear methods, such as neural networks, CART regression trees and aggregated models (bagging and random forest), Support Vector Machines. Standard covariance analysis performs well when interactions between predictors are added to the model, as well as random forests. MOCAGE mean prediction error estimated by cross-validation is then reduced from 36 µg/m³ to 26 µg/m³.

Keywords

Model Output Statistics. Forecast. Ozone. Data mining.

** Laboratoire de statistiques et probabilités – Université Paul Sabatier – 118 route de Narbonne – 31062 Toulouse Cedex.*

*** Météo-France – École nationale de la météorologie – 42 avenue Coriolis – 31057 Toulouse Cedex.*

**** Météo-France – CNRM/GMGEC – Toulouse – 42 avenue Coriolis – 31057 Toulouse Cedex.*

1. Introduction

En ce début de XXIe siècle, l'homme continue de rejeter des quantités considérables de polluants dans l'atmosphère. Malgré les progrès enregistrés, l'air que nous respirons reste encore trop souvent une menace pour la santé humaine et pour l'environnement.

La concentration moyenne de l'ozone près du sol a quadruplé depuis un siècle du fait des activités humaines. Outre l'augmentation de la pollution de fond, durable et globale, le risque survient lorsque cette concentration augmente fortement de manière locale. C'est par exemple le cas lorsqu'un temps ensoleillé et calme persiste pendant plusieurs jours. Le rayonnement solaire favorise la production d'ozone à partir du dioxyde d'azote. On parle alors de pics de pollution à l'ozone. Ce gaz, très oxydant, altère alors les fonctions respiratoires et une exposition prolongée devient dangereuse, particulièrement pour les personnes âgées, les asthmatiques et les jeunes enfants.

La loi du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie reconnaît à chaque citoyen le droit de respirer un air de qualité. Cette loi impose un seuil d'information lorsque la concentration en ozone dépasse 180 µg/m3 et un seuil d'alerte pour une concentration de 240 µg/m3.

Les mesures à prendre sont de plusieurs types : prévenir les personnes sensibles pour qu'elles évitent de sortir, limiter les activités scolaires de plein air, agir sur les émissions de polluants en limitant, par exemple, la circulation automobile. Pour établir ces mesures, il est nécessaire de prédire dès 16 h (locales) soit 14 h TU (temps universel) la pollution du lendemain.

Le modèle MOCAGE (modèle de chimie atmosphérique à grande échelle) développé par le CNRM est capable de simuler les interactions entre les phénomènes dynamiques, physiques et chimiques depuis l'échelle de la planète jusqu'à l'échelle régionale [1, 2]. Il permet d'effectuer des simulations de qualité de l'air et fait partie du dispositif d'alerte des pouvoirs publics.

Les prévisions de ce modèle physique sont entachées de biais, que l'on cherche à corriger par « adaptation statistique ». Cela consiste à calibrer un modèle statistique entre prédicteurs issus de MOCAGE pour une échéance et un endroit donnés et les observations de concentration d'ozone correspondantes. Dans cette étude, on compare, pour cinq sites, les résultats de méthodes linéaires basées sur la régression [3] (avec ou sans interactions) avec des méthodes non linéaires : réseaux de neurones [4], arbres de régression CART [5], modèles d'agrégation (bagging [6], forêt aléatoire [7]) et séparateur à vaste marge [8]. Des techniques d'agrégation de modèles ont déjà été appliquées avec succès sur des problèmes de pollution par ozone [9] mais sans y intégrer la prévision déterministe (MOCAGE).

2. Données

Le modèle MOCAGE est utilisé en opérationnel pour la prévision de la qualité de l'air depuis 2001. Il fait notamment partie de la plate-forme nationale de prévisions PREV'AIR* pour laquelle il délivre quotidiennement des prévisions jusqu'à trois jours d'échéance. Les prévisions utilisées ici sont celles du premier jour de prévisions MOCAGE (de 0 à 24 heures d'échéance) pour la version du modèle utilisée en opérationnel lors de l'été 2005 ; elles sont disponibles pour l'été 2002, août 2003 et mai-septembre 2005.

Cinq sites sont étudiés : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache et Plan-de-Cuques. Ils sont choisis pour deux raisons : la prévision déterministe est particulièrement médiocre sur ces sites et ils présentent un nombre important de pics de pollution durant la période d'étude. Les observations sont effectuées par les Associations agréées pour la surveillance de la qualité de l'air (AASQA) locales et proviennent des bases de données gérées par l'ADEME : la BDQA (base de données de la qualité de l'air) pour 2002 et 2003 et BASTER (base de données en temps réel) pour 2005.

L'échéance considérée est le lendemain à 15 h TU (17 h locales, heure habituelle du maximum quotidien de pollution à l'ozone).

On cherche à prévoir l'ozone observé O_3 à partir de différents prédicteurs prévus par MOCAGE à 15 h TU : O_{3M} concentration en ozone prévu par MOCAGE, T température, FF force du vent, NO et NO_2 logarithmes des concentrations en monoxyde et dioxyde d'azote, H_2O racine carrée du rapport de mélange de la vapeur d'eau. Les transformations log et racine carrée permettent de symétriser la distribution des prédicteurs NO, NO_2 et H_2O. Pour tenir compte également des variations possibles des sources de pollution, on introduit une variable qualitative jour à deux modalités « ouvrable » et « férié/fin de semaine ». De plus, les estimations des modèles se faisant sur l'ensemble des données, on introduit donc également un facteur « station » à cinq modalités « Aix-en-Provence », « Rambouillet », etc.

3. Méthodes

Certaines des méthodes de modélisation et prévision utilisées dans ce travail sont bien connues et leur usage est largement répandu : régression linéaire ou quadratique, réseaux de neurones ; elles ne sont pas rappelées ici. D'autres, algorithmiques, sont plus récentes et issues de la communauté informatique (« machine learning ») ou encore résultent de l'interface entre statistique et théorie de l'apprentissage. Elles méritent quelques mots d'introduction.

** http://www.prevair.org/*

3.1. Arbres de régression

Cette technique n'est pas très récente [5] mais connaît un grand engouement, notamment dans les applications au marketing. En effet, elle conduit à la construction d'arbres binaires de décision très simples à interpréter. D'autre part, elle est souvent à la base des algorithmes d'agrégation de modèles. Un arbre est construit de façon récursive, chaque nœud étant défini par une variable explicative et une valeur seuil si la variable est quantitative, un partage des modalités si celle-ci est qualitative. Ce choix est fait par optimisation d'un critère qui vise à construire des feuilles les plus homogènes possibles au sens de la variable à prédire : variance inter pour une variable à prévoir quantitative, entropie ou concentration de Gini pour le cas qualitatif. Comme pour tout modèle, une bonne prévision nécessite un « réglage » de la complexité, c'est-à-dire du nombre de paramètres estimés. En effet, un modèle qui ajuste bien n'est pas nécessairement un modèle qui prévoit bien, car il est sujet à une forte variabilité. Dans le cas d'un arbre, l'optimisation (ou élagage) est obtenue par minimisation de l'erreur estimée par validation croisée.

3.2. Agrégation de modèles *(bagging)*

Un arbre construit selon le procédé précédent peut être un modèle très instable, c'est-à-dire très dépendant de l'échantillon d'apprentissage sur lequel il a été estimé. L'idée simple sur laquelle repose l'agrégation de modèles consiste à moyenner plusieurs prévisions afin d'en réduire la variance. Dans le cas de la prévision d'une variable qualitative, la moyenne est remplacée par un vote : la modalité prédite est celle qui est la plus fréquemment obtenue par l'ensemble des prédicteurs. Idéalement, si l'on dispose de m échantillons indépendants, la loi forte des grands nombres nous indique que la variance est divisée par racine de m. En pratique, cela nécessite trop d'observations. Breiman [6] a proposé d'estimer un grand nombre de modèles (d'arbres), avant de faire la moyenne des prédictions, sur des échantillons *« bootstrap »* de l'échantillon initial. Un échantillon *bootstrap* est obtenu par n tirages aléatoires avec remise dans l'échantillon initial de taille n. Les échantillons ainsi obtenus ne sont bien évidemment pas indépendants mais l'instabilité des arbres peut rendre l'ensemble tout à fait performant : chaque arbre est de faible biais tandis que leur moyenne est de faible variance.

3.3. Random forest

Breiman [7] a par la suite proposé une amélioration de l'algorithme de *bagging* en introduisant un aléa supplémentaire afin de rendre les estimations de chaque modèle plus « variables » et donc, d'une certaine façon, plus « indépendantes entre elles ». À chaque étape de construction d'un arbre, la variable et le seuil optimaux ne sont pas cherchés sur l'ensemble des variables mais sur un sous-ensemble de taille réduite tiré aléatoirement. Ainsi, chaque arbre obtenu est sous-optimal mais il se trouve que l'agrégation de ces modèles conduit, en pratique, à de meilleurs résultats en prévision.

Une autre approche, le *boosting,* n'a pas été utilisée ici. Comme les précédentes, elle construit un ensemble de modèles mais sur une base adaptative. Chaque nouveau modèle donne plus de poids aux observations mal prédites à l'itération précédente. Quelle que soit la méthode utilisée, il est important de noter que celle-ci conduit à l'estimation d'un nombre considérable de paramètres (pour chaque modèle) mais sans pour autant conduire à une situation de sur-apprentissage, d'où la pertinence des prévisions.

3.4. Séparateurs à vaste marge (SVM)

Les SVM [8] sont des outils très récents directement issus des travaux de Vapnik en théorie de l'apprentissage machine. Leur principe est plus délicat à expliciter en quelques lignes. La première version de SVM visait à la séparation de deux classes d'unités statistiques observées sur p variables quantitatives. Le principe de base est la recherche, lorsqu'il existe, d'un hyperplan linéaire séparateur des deux classes. Celui-ci est optimal au sens où il maximise un critère (la marge) de sorte qu'il soit le plus éloigné possible des deux sous-ensembles qu'il discrimine. Ceci conduit à la résolution du problème de maximisation sous les contraintes que les observations soient du bon côté de cet hyperplan. Les contraintes actives de ce problème correspondent alors aux observations à la frontière de leur classe, elles déterminent la position de l'hyperplan et sont appelées vecteurs support. Souvent, en pratique, cette séparation n'est pas possible, aussi, le problème est transformé en ajoutant une pénalisation (paramètre à régler) autorisant des observations à se trouver mal classées, avec plus ou moins de facilité selon la valeur du paramètre. De plus, la recherche d'un séparateur non linéaire est rendue linéaire en plongeant le problème dans un espace de plus grande dimension H muni d'un produit scalaire défini par une fonction bilinéaire positive appelée noyau :

$$< x, y >_H = < F(x), F(y) > = k(x, y)$$

L'« astuce » principale de cette démarche vient du fait que le problème d'optimisation et sa solution s'expriment uniquement par l'intermédiaire du produit scalaire sans qu'il soit nécessaire d'expliciter la fonction non linéaire F. Seule la connaissance de la fonction noyau est nécessaire ; des noyaux polynomiaux ou gaussiens sont souvent utilisés en pratique. Enfin, cette démarche a été étendue du problème de discrimination à deux classes au problème de régression. L'intérêt principal de cette approche est un meilleur contrôle du sur-apprentissage dans la mesure où la complexité du modèle ne dépend pas du nombre de variables mais du nombre de vecteurs supports.

4. Procédure de comparaison des modèles

4.1. Validation des modèles

Le graphique des valeurs observées en fonction des valeurs prédites, ainsi que le graphique des résidus, toujours en fonction des valeurs prédites, permettent de vérifier graphiquement les propriétés souhaitables des modèles d'adaptation statistique : linéarité de la réponse, variance constante des erreurs de prévision en fonction de la valeur prédite (homoscédasticité).

4.2. Erreur quadratique moyenne et écart-type d'erreur de prévision

L'erreur quadratique moyenne (EQM) est la moyenne des carrés des écarts entre valeurs prédites et observées. L'EQM est généralement surévaluée lorsqu'elle est calculée sur les données ayant servi à l'apprentissage du modèle. Pour éviter ce biais, les données sont divisées aléatoirement en deux sous-échantillons, d'apprentissage et de test. Les paramètres des modèles statistiques sont estimés sur l'échantillon d'apprentissage, l'EQM étant calculée sur l'échantillon test. Cette procédure est répétée 50 fois, afin de pouvoir estimer la distribution de l'EQM. L'écart-type d'erreur de prévision est la racine carrée de l'EQM ; il est directement exprimé en µg/m³.

5. Résultats

5.1. Valeurs observées en fonction des valeurs prévues

On compare les résultats des sorties brutes de MOCAGE aux résultats donnés après adaptation statistique par les diverses méthodes sur un fichier d'apprentissage (Figure 1).

On constate dans la figure 1 que toutes les adaptations statistiques améliorent fortement les sorties brutes du modèle MOCAGE. Les meilleurs résultats semblent être fournis par deux méthodes non-linéaires : celle de la forêt aléatoire et celle du séparateur à vaste marge. Le réseau de neurones n'apporte pas d'améliorations sensibles par rapport aux autres méthodes.

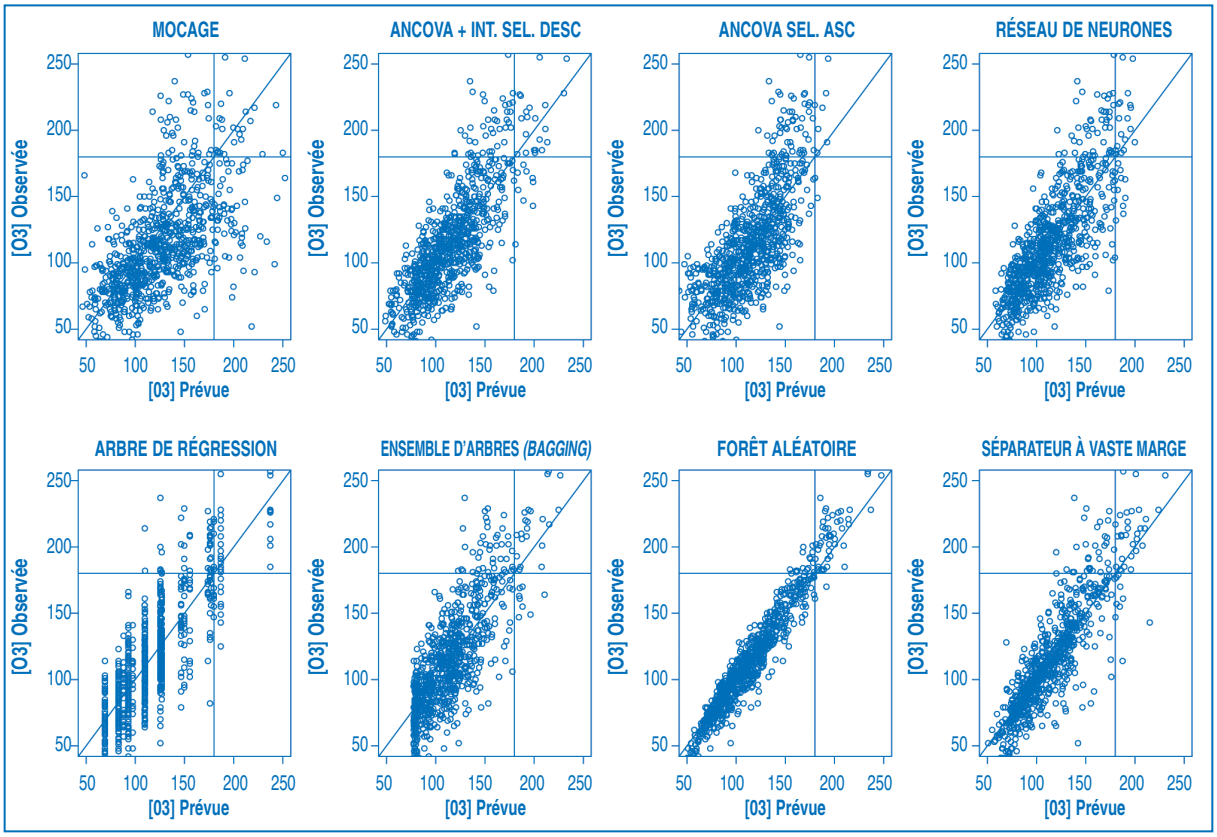

Figure 1.
Valeurs de concentration d'ozone observées en fonction des valeurs prévues (en µg/m³) par le modèle MOCAGE brut et corrigé par adaptation statistique : modèles linéaires d'analyse de covariance (ANCOVA) avec interactions (+ INT, sélection descendante) puis sans interactions, réseau de neurones, arbre de régression (CART), ensemble d'arbres *(bagging),* forêt aléatoire et séparateur à vaste marge. La première bissectrice ainsi que deux droites correspondant au seuil d'information ont été tracées.

Observed ozone concentrations (g/m³) versus raw MOCAGE forecasts and MOS models based on MOCAGE: linear model (ANCOVA) with (+INT, backward selection) and without interactions, neural network, regression tree (CART), bagging, random forest and SVM. The first bisecting line and information levels are also drawn.

Pour le modèle linéaire sans interactions, les prédicteurs sélectionnés (et le signe du coefficient correspondant) sont : O_{3M} (+), T (+), FF (–), NO (–), NO_2 (+), H_2O (+), ainsi que le facteur station. L'ozone prévu par MOCAGE est bien évidemment conservé. On retrouve l'influence positive de la température sur la concentration d'ozone, la vitesse du vent étant un élément de dispersion du polluant. Le facteur station est également retenu, les effets locaux devant être pris en compte dans l'adaptation statistique. Les mêmes prédicteurs sont retenus par sélection ascendante ou descendante. Lorsque l'on introduit les interactions, on constate que la plupart des prédicteurs retenus interagissent significativement avec le facteur station, ce qui signifie que chaque site réagit différemment aux différentes conditions de température, vent, etc.

Par construction, le modèle CART (arbre de régression) ne prévoit qu'un ensemble fini de valeurs. L'ensemble d'arbres obtenu par « *bagging* » apporte une première amélioration, encore plus nette avec la forêt aléatoire. Dans ce dernier modèle, les variables les plus importantes sont l'ozone prévu par MOCAGE et la température.

5.2. Résidus en fonction des valeurs prévues

Lorsque l'on trace les résidus des différents modèles (c'est-à-dire la différence entre l'ozone prévu et l'ozone réellement mesuré) en fonction des valeurs prévues (Figure 2), on constate la forte hétéroscédasticité de MOCAGE : la variance des erreurs augmente avec la concentration prévue. Cet effet est bien corrigé par les adaptations statistiques. Les plus faibles résidus sont observés pour la forêt aléatoire. Il est à noter que, pour ce dernier modèle, une correction supplémentaire est appliquée, les résidus initiaux étant en effet biaisés : leur moyenne varie en fonction de la valeur prévue. Une régression linéaire simple entre sorties brutes du modèle de forêt aléatoire et observations est donc estimée, puis appliquée en sortie des prévisions par forêt aléatoire.

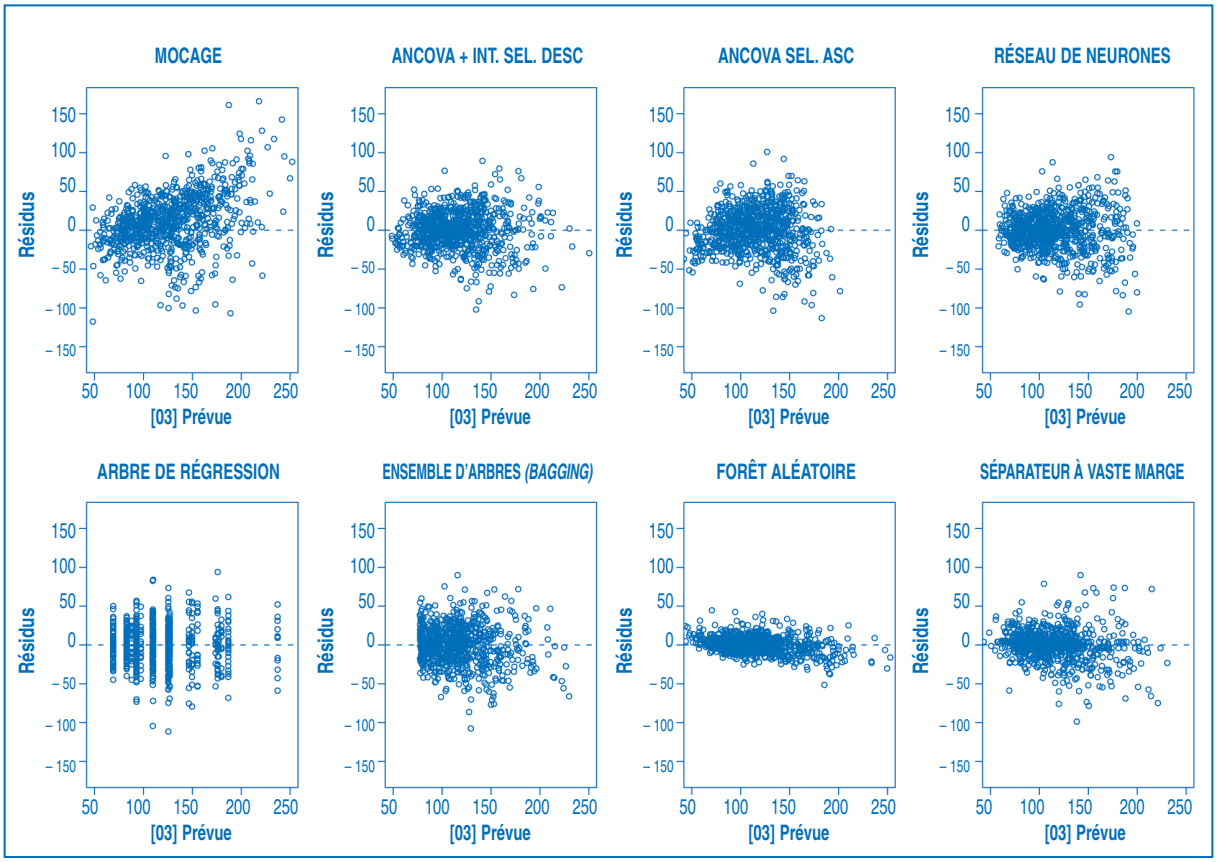

Figure 2.
Résidus estimés en fonction des valeurs prévues (en µg/m3) par le modèle MOCAGE brut et corrigé par adaptation statistique : modèles linéaires d'analyse de covariance (ANCOVA) avec (+ INT, sélection descendante) puis sans interactions, réseau de neurones, arbre de régression (CART), ensemble d'arbres *(bagging),* forêt aléatoire et séparateur à vaste marge.
Estimated residuals versus forecast concentrations (g/m3) given by MOCAGE and MOS models based on MOCAGE: linear model (ANCOVA) with (+INT, backward selection) and without interactions, neural network, regression tree (CART), bagging, random forest and SVM.

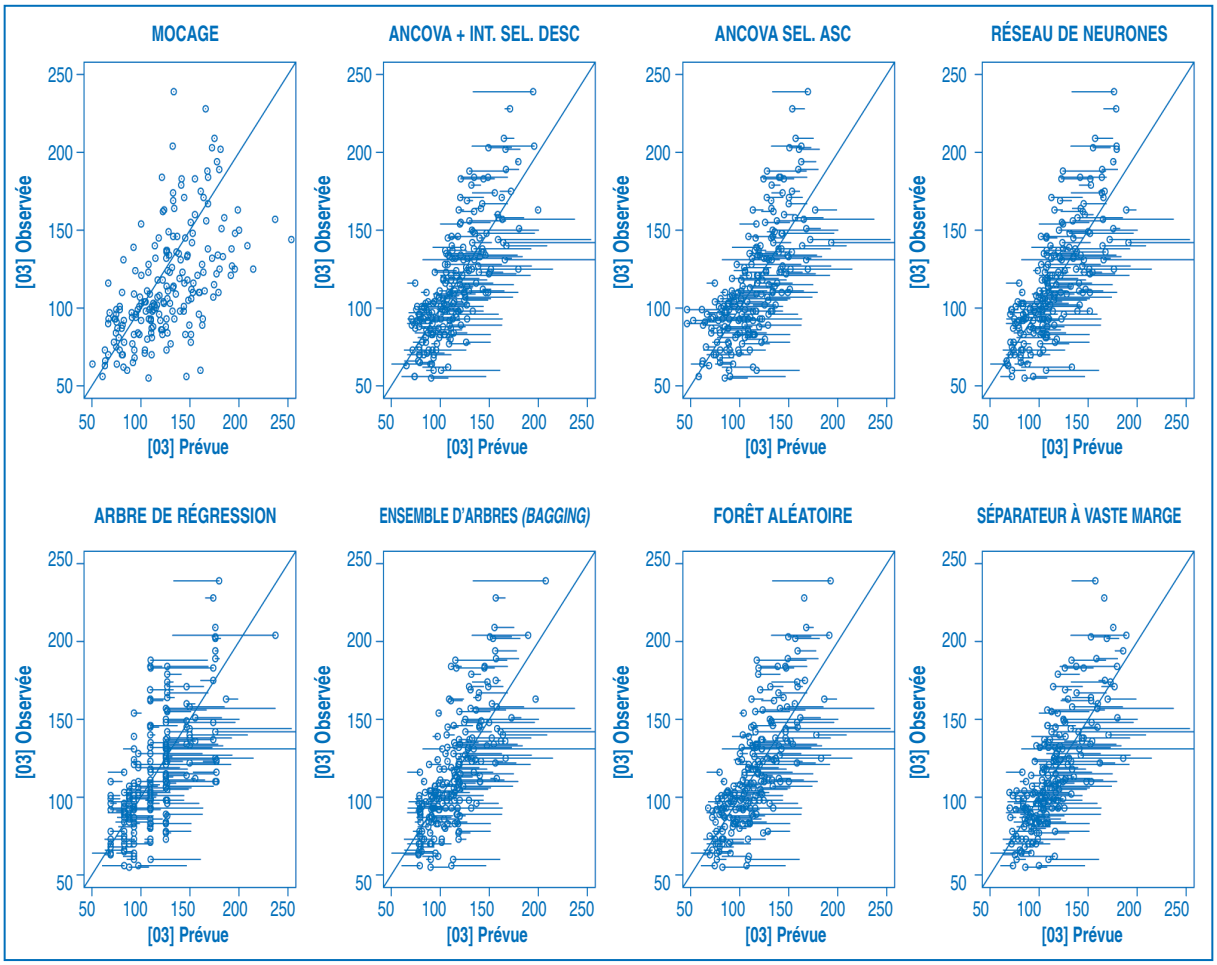

Figure 3.

Effet de l'adaptation statistique sur les prévisions de concentration d'ozone (en µg/m³) : modèle MOCAGE brut et corrigé par adaptation statistique : pour chaque observation, un trait horizontal relie la prévision par le modèle MOCAGE brut à la prévision après adaptation statistique, matérialisée par le symbole « o ». L'effet des différentes techniques est ainsi matérialisé : modèles linéaires d'analyse de covariance (ANCOVA) avec (+ INT, sélection descendante) puis sans interactions, réseau de neurones, arbre de régression (CART), ensemble d'arbres (*bagging),* forêt aléatoire et séparateur à vaste marge. La première bissectrice est également tracée.

Influence of MOS on ozone concentration forecasts (g/m³) : raw and MOS MOCAGE. For each observation, an horizontal line links the raw MOCAGE forecast to the MOS modified forecast ("o" symbol). The effect of each statistical technique is then shown: linear model (ANCOVA) with (+INT, backward selection) and without interactions, neural network, regression tree (CART), bagging, random forest and SVM. The first bisecting line is also drawn.

5.3. Effet de l'adaptation statistique

On montre dans la figure 3 l'effet de l'adaptation statistique sur les prévisions de concentration d'ozone pour un fichier test.

On constate bien que les prévisions MOCAGE, fortement dispersées, sont ramenées vers la première bissectrice.

D'autres résultats, non développés ici, montrent que si l'on s'intéresse à la prévision des dépassements du seuil d'information de 180 µg/m³, l'adaptation statistique permet de réduire les nombreuses fausses alertes générées par MOCAGE. En revanche, le taux de détection n'est pas amélioré. Lorsque MOCAGE ne prévoit pas un pic réellement observé, l'amélioration apportée par les adaptations statistiques est moins flagrante.

5.4. Écarts-types d'erreur de prévision estimés sur fichiers tests

Pour estimer plus précisément la qualité des adaptations statistiques, l'écart-type d'erreur de prévision est estimé sur 50 échantillons tests tirés au hasard, à partir des modèles statistiques estimés sur 50 échantillons d'apprentissage indépendants des données test. Les boîtes à moustaches des 50 écarts-types ainsi calculés pour chacune des méthodes sont

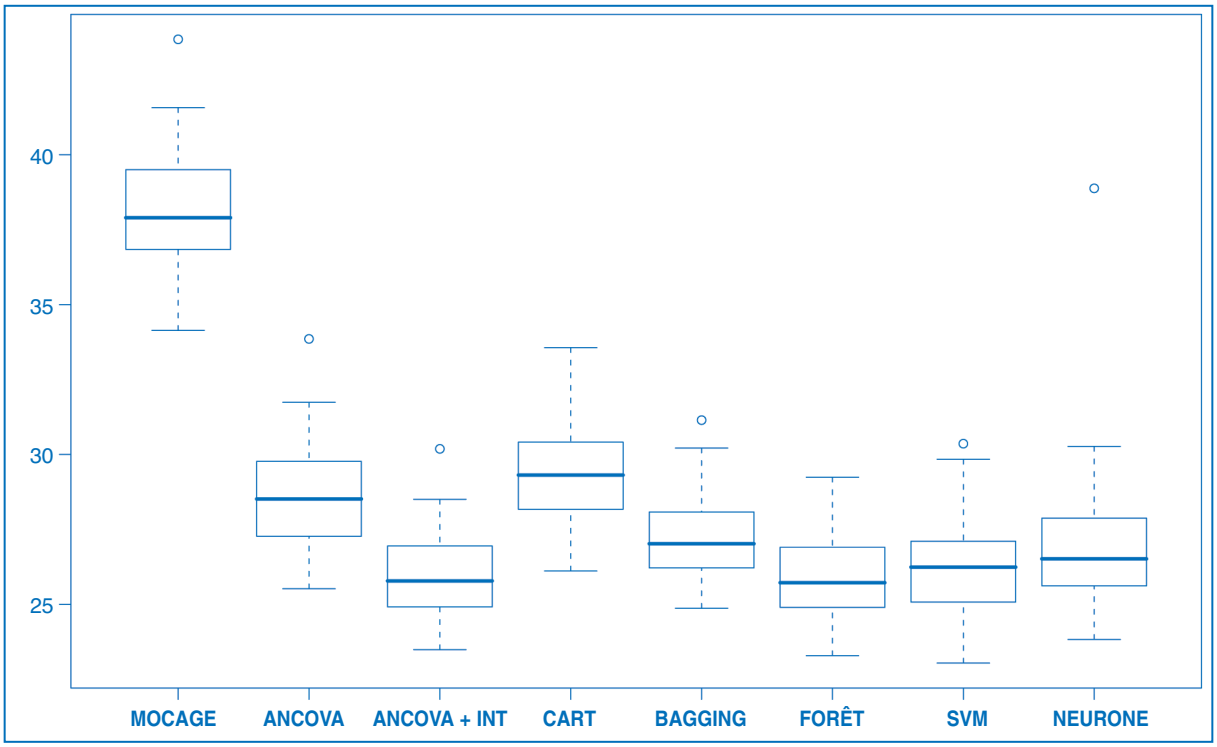

Figure 4.
Boîtes à moustaches des écarts-types d'erreur des prévisions de concentration d'ozone (en µg/m3) : modèle MOCAGE brut et corrigé par adaptation statistique : modèles linéaires d'analyse de covariance (ANCOVA) sans puis avec (+ INT) interactions, arbre de régression (CART), ensemble d'arbres *(bagging),* forêt aléatoire, séparateur à vaste marge et réseau de neurones.
Boxplots of standard deviations of ozone forecasts errors (g/m3): MOCAGE and MOS models based on MOCAGE : linear model (ANCOVA) with (+INT, backward selection) and without interactions, regression tree (CART), bagging, random forest, SVM and neural network.

Tableau 1.
Moyenne, médiane et écart-type des écarts-types d'erreur des prévisions de concentration d'ozone (en µg/m3) : modèle MOCAGE brut et corrigé par adaptation statistique : modèles linéaires d'analyse de covariance sans, puis, avec interactions, arbre de régression (CART), ensemble d'arbres *(bagging),* forêt aléatoire, séparateur à vaste marge et réseau de neurones. Statistiques calculées sur 50 fichiers tests.
Mean, median and standard deviation of standard deviations of ozone forecasts errors (g/m3) : MOCAGE and MOS models based on MOCAGE : linear model (ANCOVA) with (+INT, backward selection) and without interactions, regression tree (CART), bagging, random forest, SVM and neural network. Those statistics are computed on 50 randomly generated test samples.

Méthode	Moyenne (µg/m3)	Médiane (µg/m3)	Écart-type (µg/m3)
MOCAGE brut	38,1	37,9	2,1
Analyse de covariance sans interactions	28,6	28,5	1,6
Analyse de covariance avec interactions	26,1	25,8	1,4
Arbre de régression	29,3	29,3	1,5
Ensemble d'arbres *(bagging)*	27,2	27,0	1,5
Forêt aléatoire	25,9	25,7	1,4
Séparateur à vaste marge	26,2	26,2	1,6
Réseau de neurones	27,0	26,5	2,3

données dans la figure 4. Le tableau 1 renferme les moyennes, médianes et écarts-types de ces scores.

On note l'amélioration apportée par toutes les méthodes par rapport à MOCAGE, ainsi que le bon comportement du modèle linéaire d'analyse de covariance lorsque les interactions sont présentes.

L'arbre de régression (CART) est grandement amélioré par les techniques d'agrégation de modèles : *bagging* (ensemble d'arbres) mais surtout forêt aléatoire.

Au vu des résultats, la complexité des séparateurs à vaste marge ou des réseaux de neurones ne se justifie pas.

Les deux meilleures méthodes sont donc l'analyse de covariance avec interactions et les forêts aléatoires. La nette supériorité de la forêt aléatoire, telle qu'elle apparaît sur fichier d'apprentissage, est beaucoup moins évidente sur les fichiers tests. Sur cet exemple, la forêt aléatoire semble présenter une tendance au surajustement.

6. Conclusion

Cette étude, réalisée sur cinq sites, illustre l'intérêt des techniques d'adaptation statistique qui permettent d'améliorer à moindre coût les prévisions déterministes du modèle MOCAGE. Dans notre étude de cas, on réduit l'écart-type d'erreur de prévision de 36 µg/m^3 à 26 µg/m^3 pour les meilleures méthodes. L'analyse de covariance avec interactions et la méthode des forêts aléatoires donnant d'excellents résultats, elles devront être privilégiées pour réaliser l'adaptation statistique sur l'ensemble des postes.

Remerciements

Tous nos remerciements à Marie Paya, Lionel Fugon et Yuqing Mei (étudiants INSA Toulouse) pour avoir codé une partie des programmes ayant servi à cette étude.

Références

1. Peuch VH, Amodei M, Barthet T, Cathala ML, Josse B, Michou M, Simon P. MOCAGE : modèle de chimie atmosphérique à grande échelle. Proc. of Météo-France Workshop on Atmospheric Modelling. Toulouse, décembre 1999 : 33-6.
2. Dufour A. Simulation et prévision de la qualité de l'air aux échelles continentale et régionale. Thèse de l'Université Paul Sabatier, décembre 2006 : 245 p.
3. Azaïs JM, Bardet JM. Le modèle linéaire par l'exemple. Dunod, Paris 2005 : 326 p.
4 Haykin T. Neural network, a comprehensive foundation. 2nd edition. Prentice Hall 1998 : 842 p.
5. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New edition. Chapman & Hall/CRC 1984 : 368 p.
6. Breiman L. Bagging predictors. *Machine Learning* 1996 ; 26 (2) : 123-140.
7. Breiman L. Random forests. *Machine Learning* 2001 ; 45 (2) : 5-32.
8. Schölkopf B, Smola A. Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond. MIT Press 2001 : 644 p.
9. Ghattas B. Prévisions des pics d'ozone par arbres de régression, simples et agrégés par Bootstrap. *Revue de statistique appliquée* 1999 ; 47(2) : 61-80.