

Prévision statistique des concentrations de particules dans l'air : tour d'horizon des principaux outils mathématiques

Airborne pollutant concentration forecast: review of the main mathematical methods

Anda Ionescu*

Résumé

Cet article propose un état de l'art sur la modélisation inverse basée sur l'utilisation des méthodes statistiques pour la prévision des concentrations des particules dans l'air. Ces outils mathématiques sont applicables aux séries temporelles de mesures fournies par les capteurs des réseaux de surveillance de la qualité de l'air pour faire des prévisions et envisager des éventuelles mesures d'urgence ; dans le domaine de l'air intérieur, leur utilisation sera d'autant plus précieuse que peu de modèles déterministes sont applicables à l'heure actuelle. Cet état d'art inclut d'abord les modèles linéaires, en mettant l'accent sur les modèles autorégressifs ; parmi les modèles non linéaires, les réseaux de neurones ont été beaucoup utilisés, et on trouve un nombre non négligeable d'applications différentes dans la littérature. On citera également l'utilisation plus originale des systèmes dynamiques basés sur la théorie du chaos ou du filtre de Kalman. Des résultats plus performants ont été obtenus par des modèles hybrides (en combinant plusieurs modèles) ou en appliquant un prétraitement aux données d'entrée.

Mots-clés

particules, modélisation inverse, modèles autorégressifs, réseaux de neurones artificiels, systèmes dynamiques, théorie du chaos, filtre de Kalman, modèles hybrides

Abstract

This paper is a review of the main mathematical methods applied for an inverse modelling of particulate matter concentrations. These mathematical tools can be applied to time series data provided by the air pollution monitoring networks in order to forecast the air pollutant concentration and help to take urgent decisions; indoor air quality forecast based on statistical modelling is a very important topic because there are very few applications in this field. The following models are reviewed: linear ones, and in particular autoregressive models, but also non linear models, especially the artificial neural networks, extensively applied in the literature. Some models originally applied for particulate matter concentrations forecast are the dynamic systems based on the chaos theory and the Kalman filter. Many authors concluded that best performance is obtained by hybrid modelling (combining several models) or applying a specific pre-processing to input data.

Keywords

particulate matter, inverse modelling, autoregressive models, artificial neural networks, dynamic systems, chaos theory, Kalman filter, hybrid models

(*) Centre d'Études et de Recherche en Thermique, Environnement et Systèmes (CERTES) - Université Paris-Est Créteil, avenue du Général de Gaulle, 94010 Créteil Cedex.

Introduction

En novembre 2012, la revue *Pollution Atmosphérique* a consacré un numéro spécial à la pollution par les particules. L'effet néfaste de ce polluant sur la santé a mérité une place privilégiée, le nombre important de communications en témoigne (cf. Marano, 2012 ; Pascal *et al.*, 2012 ; Médina, 2012 ; Chanel, 2012). Un deuxième effet, la participation des particules au changement climatique, a été également débattu (cf. Coll, 2012 ; Momas et Bergametti, 2012) ; ce sujet a été abordé de manière plus exhaustive dans le numéro spécial *Pollution atmosphérique et changement climatique. Comment traiter les deux enjeux simultanément ?* publié par la même revue en avril 2009. Un autre effet important est l'impact des particules sur le patrimoine bâti, développé dans le numéro spécial *Effets de la pollution atmosphérique sur les matériaux*, édité en octobre 2007.

Devant les effets néfastes importants des particules sur la santé, le climat, les écosystèmes et le patrimoine bâti, ce polluant atmosphérique a été rajouté parmi les polluants réglementés. La législation européenne sur la qualité de l'air impose la surveillance depuis dix ans des PM_{10} et depuis 2008 pour les $PM_{2,5}$ (Villenave *et al.*, 2012).

La surveillance de la qualité de l'air fait appel à plusieurs moyens, parmi lesquels figurent les stations fixes des réseaux de surveillance de la qualité de l'air, qui mesurent les polluants réglementés de manière automatique, en continu (Bouvier, 2012).

«*Nous vivons dans un monde de particules [...]. L'environnement intérieur pourrait être considéré comme un abri*» (Ramalho *et al.*, 2012). Il s'est avéré que les niveaux de particules dans l'air intérieur peuvent être plus importants que ceux extérieurs. L'intérêt pour la qualité de l'air intérieur est indéniable (un futur numéro spécial de la revue *Pollution Atmosphérique* y sera consacré), mais l'état des connaissances est beaucoup moins avancé que pour la qualité de l'air extérieur. Les environnements intérieurs souffrent d'un manque de données concernant les particules (Ramalho *et al.*, 2012).

La surveillance fait également de plus en plus appel à des outils mathématiques et à des modélisations destinés au calcul de la distribution spatiale de la pollution et à la prévision de la qualité de l'air (Bouvier, 2012) : (i) prévision à court terme, pour faciliter la communication vers le public et envisager des mesures d'urgence et (ii) prévision prospective de l'impact des scénarios de réduction des émissions (Rouil, 2012).

«*Les modèles de qualité de l'air ont fortement progressé ces dernières années [...]. Pour les particules,*

des améliorations doivent être encore apportées aux modèles numériques, notamment pour mieux prendre en compte certaines sources et mieux reproduire la formation des aérosols organiques secondaires» (Momas et Bergametti, 2012). Les modèles de chimie transport (CTM) donnent des résultats assez fiables pour la plupart des polluants, mais ils sont utilisables à une échelle régionale : grille de résolution allant de 25 km à quelques km (Rouil, 2012).

L'assimilation de données relève d'un processus mathématique par lequel l'évolution d'un modèle numérique est contrainte à rester proche d'une réalité partiellement observée. Cette technique vise à prendre la meilleure partie de deux informations dont on dispose : d'une part, un modèle numérique décrivant par l'intermédiaire des équations différentielles la physique et la chimie du système (CTM), et d'autre part, les observations *in situ* afin de corriger les prédictions du modèle, d'estimer l'état du système et d'évaluer la précision de cette estimation (Zhang *et al.*, 2012). La prévision des concentrations de particules en assimilation des données dans les modèles de chimie transport 3D progresse en ce moment en utilisant une assez large variété d'algorithmes et différents types d'observations, comme les concentrations de polluants mesurées par plusieurs stations au sol ou bien des mesures optiques obtenues grâce à la télédétection (Tombette *et al.*, 2009 ; Pagowski et Grell, 2012 ; Wang *et al.*, 2013).

Lorsqu'on passe du niveau régional au niveau local, d'autres outils sont utilisés par les AASQA françaises : les modèles gaussiens urbains, les modèles de rue (Rouil, 2012)...

Une recherche bibliographique sur la prévision à court terme des niveaux de particules montre que dans d'autres pays on utilise beaucoup les modèles statistiques de prévision, basés sur les mesures effectuées dans le passé sur les mêmes polluants et éventuellement sur d'autres paramètres influents. Ce sont des modèles empiriques et qui ne prennent pas en compte les mécanismes de formation et dispersion de ces polluants à partir de leurs sources, comme les modèles déterministes le font.

Cet article propose un état de l'art sur la modélisation inverse basée sur l'utilisation des modèles statistiques pour la prévision des concentrations des particules dans l'air. Ces outils sont applicables aux séries temporelles de mesures fournies par les capteurs des réseaux de surveillance de la qualité de l'air pour envisager des mesures d'urgence. La plupart des travaux qui seront cités dans la suite concernent la modélisation d'une série temporelle de concentrations de polluants fournie par un capteur d'un réseau de surveillance de la qualité de l'air, et la prévision

se fait localement, à la station de mesure. On pourrait envisager de faire cette prévision pour plusieurs stations de mesure, puis procéder à une interpolation spatiale par une méthode géostatistique (Chauvet, 1999), par exemple ; nous n'avons pas trouvé d'étude basée sur cette démarche dans la littérature de spécialité consultée.

Il est intéressant à noter que les modèles présentés sont applicables également dans le domaine de l'air intérieur, où leur utilisation sera d'autant plus précieuse que peu de modèles déterministes sont développés à l'heure actuelle pour l'environnement intérieur.

On commencera par les modèles linéaires, en mettant l'accent sur les modèles autorégressifs ; parmi les modèles non linéaires, les réseaux de neurones ont été beaucoup utilisés, et on trouve un nombre non négligeable d'applications différentes dans la littérature. On citera également l'utilisation plus originale des systèmes dynamiques basés sur la théorie du chaos ou du filtre de Kalman. Des résultats plus performants ont été obtenus par des modèles hybrides (en combinant plusieurs modèles) ou en appliquant un prétraitement aux données d'entrée.

1. Modèles linéaires

1.1. Régression simple et multiple (RLM)

On peut citer tout d'abord les modèles classiques linéaires : la régression linéaire simple (on exprime la concentration d'un polluant en fonction d'une seule variable qui peut être le temps, un facteur météorologique, etc.) ou la régression linéaire multiple (la concentration de polluant est exprimée en fonction de plusieurs variables explicatives). La RLM a été utilisée dans plusieurs études, mais dans la plupart des cas à titre de comparaison. Goyal *et al.* (2006) ont étudié la prévision des niveaux totaux de particules respirables (moyennes journalières) dans deux métropoles : Delhi et Hong Kong. Les auteurs ont développé un premier modèle basé sur une RLM, les variables explicatives étant des paramètres météorologiques : vitesse du vent, rayonnement solaire, humidité relative et température de surface. La RLM a montré que les contributions de certaines variables étaient significatives (vitesse du vent, humidité relative), mais le rayonnement solaire n'était pas une variable influente. Le modèle explique 58 % de la variance des concentrations journalières de particules, avec une erreur quadratique moyenne de $76 \mu\text{g}\cdot\text{m}^{-3}$, ce qui n'est pas très satisfaisant. Le deuxième modèle utilisé est de type ARIMA, et le troisième, une combinaison linéaire des deux premiers ; ils seront brièvement présentés dans la section 4. La RLM a été utilisée à titre de

comparaison dans d'autres travaux, comme ceux de Perez et Salini (2008), Slini *et al.* (2006) (*cf.* section 4), ou bien Diaz-Robles *et al.* (2008), (*cf.* section 5). Généralement, les différents auteurs ont retenu la RLM comme une alternative intéressante en raison de sa simplicité de mise en œuvre, et menant à des résultats plutôt convenables, mais avec des performances plus faibles que celles des modèles plus élaborés. Une façon particulière d'appliquer la RLM est d'utiliser les composantes principales à la place des variables explicatives, lorsque ces dernières sont corrélées entre elles (ce qui arrive souvent), mais on ne peut pas donner d'exemple d'application pour la prévision des particules.

1.2. Modèles autorégressifs

Tout en restant dans la classe des modèles linéaires, si on se réfère particulièrement aux séries temporelles, la théorie développée principalement par Box et Jenkins est adaptée à ce type de données (Box *et al.*, 2008 ; Brockwell et Davis, 2002, 2009). On recherche dans ce cas une structure temporelle dans la série de données, qui servira comme base pour la prévision. La différence par rapport aux modèles de régression évoqués précédemment, où on utilisait principalement des variables autres que la variable prédite pour expliquer sa variance, est que dans le cas des séries temporelles on cherche une autorégression, c'est-à-dire qu'on cherche à expliquer et/ou prédire la variable par rapport à sa propre structure temporelle. Il y a donc la grande classe des modèles appelés autorégressifs. À toute série temporelle ne correspond pas forcément un modèle autorégressif (AR), car elle doit vérifier certaines propriétés statistiques. Souvent, on est amené à appliquer certaines transformations (comme celles de Box-Jenkins) pour qu'elle vérifie les propriétés statistiques nécessaires (e.g. la stationnarité). Autour des simples modèles AR, en fonction des propriétés des séries temporelles, plusieurs autres modèles ont été développés, comme les ARMA (qui incluent une moyenne mobile : Moving Average), ARIMA (avec Intégration)... Tous ces modèles sont de type stochastique.

Certains auteurs soutiennent que, d'un point de vue conceptuel, c'est ce type de modèle statistique qui serait le plus adapté à la prévision des concentrations des polluants. Les modèles AR ont été développés principalement dans un but de prévision, et ils ont donné de bons résultats, à condition que la variable (la série temporelle) soit autocorrélée. Parfois, les résultats peuvent être améliorés si on utilise, en plus, l'information fournie par une autre variable, appelée exogène. Les séries temporelles de polluants sont souvent autocorrélées et on peut obtenir des résultats plus ou moins satisfaisants en utilisant ces modèles. Souvent, rajouter une variable exogène, par exemple

un paramètre météorologique, conduit à une amélioration des résultats.

On peut citer un exemple d'application des modèles AR dans le domaine de l'environnement intérieur. Issarayangyun et Greaves (2007) ont analysé le comportement temporel de la concentration de $PM_{2.5}$ à l'intérieur d'un habitacle automobile, au pas de temps de la minute. Deux modélisations ont été employées : l'une purement statistique (ARIMA), l'autre faisant intervenir des variables explicatives (vitesse, conditions de circulation, jour et heure...). L'analyse purement statistique de la seule variable $PM_{2.5}$ a montré que le comportement dynamique est bien représenté par une autocorrélation sur trois à quatre minutes. Les résultats de la deuxième modélisation ont mis en évidence l'influence importante de quelques paramètres : ventilation de l'habitable, vitesse, à l'exclusion des autres ; il reste pourtant incertain, en l'absence de mesures sur l'atmosphère extérieure, si ceci est lié aux conditions extérieures ou à un phénomène de diffusion propre à l'habitable.

Beaucoup d'études ont été basées sur ce type de modèles de prévision, mais la plupart des applications publiées dans la littérature dans le domaine de l'environnement extérieur concerne d'autres polluants que les particules. Les résultats obtenus étaient corrects, sans pour autant pouvoir dire que c'est la panacée. L'autocorrélation n'est pas suffisante pour expliquer toute la variance de la variable à prédire. De plus, même si la partie linéaire est bien modélisée, il reste aussi une partie non linéaire dans les séries temporelles de concentrations des polluants. Récemment, des chercheurs ont développé des modèles autorégressifs non linéaires : NARX, NARIMA, etc. (Pisoni *et al.*, 2009). Jusqu'à présent, ils ont été trop peu utilisés dans le domaine de la pollution atmosphérique pour pouvoir juger de manière globale s'ils sont adaptés ou non à cette problématique et s'ils apportent vraiment quelque chose de plus.

2. Modèles non linéaires

2.1. Réseaux de neurones artificiels (RNA)

Pendant assez longtemps, les résultats des modèles linéaires pour la prévision des polluants ont été jugés comme satisfaisants, sans pouvoir dépasser un certain seuil de performance, car il y avait aussi de la non-linéarité. Les chercheurs se sont penchés alors sur les modèles non linéaires. Il y a eu une vraie explosion d'études basées sur les réseaux de neurones artificiels. Maier et Dandy (2000) ont remarqué «*la tendance parmi les chercheurs d'appliquer les RNA à des problèmes pour lesquels les autres méthodes n'ont pas été efficaces*», c'était devenu

presque la panacée. Parmi les différentes architectures des RNA, c'est le perceptron multicouche (PMC) qui a été le plus utilisé.

Un état de l'art des applications des PMC dans les sciences atmosphériques a été rédigé par Gardner et Dorling (1998). Les auteurs ont conclu que l'apprentissage supervisé d'un PMC peut donner une approximation correcte pour n'importe quelle fonction lisse, sans faire aucune hypothèse sur la distribution des données.

a) Afin d'améliorer les performances obtenues avec des RNA, plusieurs **algorithmes d'apprentissage** ont été testés ; comme ces derniers font appel finalement à la minimisation d'une fonction de coût, les performances des différentes méthodes d'optimisation ont été comparées. D'une manière générale, ce sont les méthodes d'optimisation locale qui ont été préférées, même si elles ne fournissaient qu'un minimum local ; en général, la solution était considérée comme acceptable et la simplicité de mise en œuvre, considérée comme un atout. On peut remarquer une préférence pour les méthodes de second ordre (e.g. quasi-Newton) et pour l'algorithme de Levenberg-Marquardt. Malgré leur complexité, les méthodes d'optimisation globale (e.g. l'optimisation par essaim des particules) n'ont pas été épargnées, et elles ont été implémentées dans le cadre des modèles RNA. Les résultats obtenus étaient supérieurs, mais jusqu'à présent il y a trop peu d'études pour pouvoir dégager une conclusion sur l'opportunité d'utiliser cette méthodologie beaucoup plus complexe.

Lu *et al.* (2003) ont étudié l'amélioration des performances (prévision et vitesse de convergence) d'un PMC en utilisant un algorithme d'optimisation globale (optimisation par essaim de particules – *Particle Swarm Optimization* (PSO) en anglais) par rapport à un apprentissage basé sur la rétropropagation. L'étude a porté sur plusieurs polluants : particules, oxydes d'azote, sur différentes échelles de temps (en travaillant sur des valeurs journalières ou mensuelles), au centre de Hong Kong ; les auteurs se sont basés uniquement sur l'historique des polluants, sans variables exogènes.

b) En général, l'**architecture** du PMC était obtenue par des essais, en faisant varier le nombre de neurones de chaque couche et le nombre de couches. L'obtention des architectures optimales peut être réalisée par élagage (*pruning*).

Corani (2005) a étudié la prévision à 9 h du matin des valeurs de PM_{10} et d'ozone pour la journée qui débute, à Milan en proposant : (i) une architecture parcimonieuse du réseau neuronal par élagage (*pruning* PNN) : suppression des connexions correspondantes

aux paramètres redondants ; (ii) un nouveau type d'apprentissage, appelé « paresseux » (*lazy learning* LL), qui est un algorithme local linéaire de prévision effectuant un apprentissage local chaque fois qu'une prévision est demandée. Les meilleures performances globales ont été obtenues en utilisant l'algorithme paresseux, qui, en outre, est très facile à implémenter ; la meilleure détection des dépassements des différents seuils a été obtenue grâce à l'architecture élaguée.

Bien que ce soit dans un but de prévision, ce sont des architectures « statiques » qui ont été utilisées. On trouve aussi des architectures neuronales « dynamiques », comme les réseaux d'Elman ou une cascade de perceptrons. Étonnamment, les architectures « dynamiques » ne sont que légèrement plus performantes que celles « statiques », sur l'ensemble des études comparatives (peu nombreuses dans le cas de la pollution atmosphérique).

Brunelli *et al.* (2007) ont conçu un réseau neural récurrent d'Elman pour la prévision à 48 heures de la concentration maximale journalière de SO_2 , O_3 , PM_{10} , NO_2 , CO , dans la ville de Palerme (Italie), utilisant comme prédictors météorologiques : la vitesse et la direction du vent, la pression atmosphérique et la température ambiante, moyennées sur les 12 heures précédentes. Le coefficient de corrélation obtenu entre les valeurs prédites et celles enregistrées varie entre 0.72 et 0.97, pour les différents polluants testés, montrant une bonne performance du modèle proposé.

c) Les **fonctions d'activation** constituent une autre caractéristique importante des réseaux neuronaux. Alors que le choix classique des fonctions sigmoïdes était presque figé, certains auteurs ont proposé d'autres types de fonctions, comme les ondelettes ; ce n'est qu'un début, mais qui peut être aussi une piste intéressante à explorer.

Viotti *et al.* (2002) ont appliqué les RNA pour la prévision de plusieurs polluants : particules, dioxyde de soufre, oxydes d'azote, ozone, monoxyde de carbone et benzène (prédiction à court et moyen long termes). La particularité de leur étude réside en l'utilisation d'une forme généralisée de la fonction logistique d'activation, avec trois paramètres ajustables, en utilisant l'algorithme classique de rétropropagation. Les entrées sélectionnées ont été reliées aux conditions météorologiques et aux niveaux de trafic.

d) La **sélection des entrées** est un point crucial pour tout modèle ; ceci peut être fait par élagage ou par d'autres techniques, comme les algorithmes génétiques, qui est une technique d'optimisation globale. Bien que peu courante dans le domaine de la qualité de l'air, la technique des forêts aléatoires, basée

sur plusieurs arbres de décision, permet de calculer les scores individuels d'importance des régresseurs et donc une sélection plus judicieuse des entrées du modèle. Cette démarche, couplée avec une régression « *clusterwise* » a donné de très bons résultats en prévision des niveaux de particules.

e) Une amélioration des performances peut être obtenue en appliquant un **prétraitement des données d'entrée** d'un RNA. Celui-ci peut être une classification des données (par exemple l'établissement des typologies météorologiques) à la place des données brutes (on peut même parler dans ce cas d'une modélisation hybride) ou bien l'optimisation de la période sur laquelle on effectue la moyenne temporelle des données d'entrée. Un autre exemple de prétraitement de données concerne l'application des RNA sur les composantes principales. Comme les RNA sont à la base une technique de régression (non linéaire), on retrouve les mêmes contraintes concernant les propriétés statistiques des données d'entrée (inter-corrélations, indépendance).

Hrust *et al.* (2009) ont étudié la prévision des concentrations de NO_2 , O_3 , CO , PM_{10} à Zagreb (Croatie). Les auteurs ont utilisé une nouvelle approche basée sur des familles de modèles de régression univariée afin de sélectionner l'intervalle de moyennage (entre 1 h et 97 h) pour les variables d'entrée d'un PMC ; ils ont calculé les moyennes temporelles mobiles des variables d'entrée et ils ont construit un modèle pour expliquer la sortie avec une seule entrée, moyennée différemment. Les résultats optimaux ont été sélectionnés comme entrées pour un PMC, et la prévision a été faite à 6 h du matin pour la journée (jusqu'à 23 h). Dans certains cas, ce moyennage optimal sur les valeurs passées a apporté une amélioration des performances par rapport à l'utilisation des données moyennes horaires, et l'indice de concordance obtenu a été entre 0.91 et 0.97 (meilleures performances pour l'ozone et les moins bonnes pour les PM_{10}).

f) Il est difficile de trouver un RNA « optimal », car il y a beaucoup de caractéristiques qui interviennent et dont il faut choisir le type, et beaucoup de paramètres à ajuster. Des chercheurs proposent d'utiliser les **plans d'expériences** pour concevoir un RNA optimal, mais cette procédure semble assez lourde et pas très convaincante.

La méthode statistique des plans d'expériences a été appliquée par Balestarssi *et al.* (2009) afin d'améliorer la détermination des paramètres d'un RNA dans un problème de prévision des séries temporelles non linéaires. Cette approche constitue une alternative systématique à l'approche classique qui consiste en des essais.

g) En général, on arrive à bien approximer une fonction par un modèle RNA. Le piège est que le RNA peut s'ajuster également au bruit présent dans les données ; alors on dit que le réseau «a appris par cœur» les exemples qui lui ont été présentés et dans ce cas il n'a pas une bonne capacité de prévision. Pour éviter ce problème de sur-apprentissage, on utilise des techniques de **régularisation bayésienne**, **l'arrêt prématuré** de l'apprentissage ou bien une combinaison des deux. Même avec une très bonne performance en apprentissage, un RNA peut donner des résultats peu satisfaisants en prévision et même en approximation d'une fonction, sur l'ensemble du test.

2.2. Systèmes dynamiques basés sur la théorie du chaos

Dans la majorité des cas étudiés, les RNA se sont avérés plus performants que les modèles linéaires de régression ou autorégressifs, grâce à leur propriété de non-linéarité. On peut alors envisager l'utilisation d'autres modèles non linéaires, comme les systèmes dynamiques dont les propriétés sont issues de la théorie du chaos (Kantz et Schreiber, 2003). Il y a très peu d'applications dans le cas des données de pollution atmosphérique (on peut en citer trois, dont deux concernent la prévision de l'ozone et aucune les particules), ce qui ne permet pas d'avoir une idée plus claire sur ses possibles performances. Les résultats (peu nombreux) sont encourageants, ce qui signifierait que la dynamique de certains polluants peut être considérée comme chaotique à certaines échelles temporelles.

3. Modèles probabilistes

3.1. Chaînes de Markov

Une autre classe de modèles sont les modèles probabilistes, comme les chaînes de Markov, mais ils ont été assez peu appliqués dans le domaine de la pollution de l'air. Ces modèles ont un grand potentiel et méritent une bonne réflexion quant à leur application dans la prévision des concentrations de polluants.

Dong *et al.* (2009) présentent une application des HSM (Hidden semi-Markov models) à la prédiction à 24 h des $PM_{2.5}$ dans l'atmosphère ; à la différence des modèles markoviens classiques, ces modèles introduisent directement une probabilité de rester dans chaque état, ce qui est donc d'une plus grande applicabilité mais nécessite une estimation supplémentaire. Les modèles sont estimés par des techniques de type apprentissage. L'application à des données réelles (prédiction à deux niveaux : faible/important) ne présente cependant pas de comparaison par exemple à des HMM, ni d'analyse structurelle des modèles obtenus.

3.2. Filtre de Kalman étendu

Lorsqu'on souhaite disposer d'un système dont le comportement peut s'adapter à des changements sur le court terme sans nécessiter des ajustements fréquents, le filtre de Kalman peut s'avérer très efficace. Des chercheurs ont utilisé le Filtre de Kalman Étendu (*Extended Kalman Filter-ExKF*) comme système de modélisation adaptatif non linéaire de l'espace des états pour la prévision des niveaux maximaux de pollution. La structure du modèle non linéaire est basée sur les relations empiriques entre les concentrations mesurées et les autres polluants primaires, ainsi que les facteurs météorologiques.

Zolghadri et Cazaurang (2006) ont développé un système de modélisation adaptatif non linéaire de l'espace des états pour la prévision un jour à l'avance de la concentration journalière maximale de PM_{10} à Bordeaux. La structure du modèle non linéaire est basée sur les relations empiriques entre les concentrations mesurées de PM_{10} et les autres polluants primaires, ainsi que les facteurs météorologiques. Les auteurs ont utilisé l'algorithme du Filtre de Kalman Étendu (*Extended Kalman Filter-ExKF*) pour estimer la prédiction de l'état étendu 1 jour plus tard, contenant les paramètres du modèle et la valeur moyenne journalière de PM_{10} . Une caractéristique clé de ce système est le fait que son comportement peut s'adapter à des changements sur le court terme de la pollution atmosphérique et, par conséquent, le modèle gère bien la nature de ces phénomènes sans nécessiter des ajustements fréquents. La performance globale du modèle dépend du choix de l'état initial (qui se fait *offline*), ensuite les paramètres du modèle sont mis à jour chaque jour (*online*), grâce au mécanisme adaptatif. Dans cette étude, l'initialisation a été faite en utilisant un ensemble restreint de données, par l'intermédiaire d'un algorithme d'optimisation non linéaire de type moindres carrés basé sur la méthode de Levenberg-Marquardt ou de Gauss-Newton. L'application de l'ExKF explique 70 % de la variance de la concentration moyenne journalière de PM_{10} avec une erreur absolue moyenne inférieure à $4.5 \mu g/m^3$, et la valeur de l'indice de concordance vaut environ 0.9. On peut conclure que le modèle est performant.

D'autres chercheurs ont utilisé le filtre de Kalman pour la prévision des dépassements, son rôle étant celui d'un filtre passe-bas. Les filtres utilisés pour la modélisation des systèmes dynamiques, issue du domaine du traitement du signal et de l'automatique représentent une voie très peu explorée dans le domaine de la pollution atmosphérique, mais très intéressante et prometteuse.

4. Comparaison de plusieurs modèles sur les mêmes données

Dans la littérature, plusieurs études ont été consacrées à la comparaison des différents modèles sur les mêmes jeux de données. Les conclusions ne permettent pas d'arriver à un consensus. De plus, la comparaison ne s'est toujours pas faite sur les mêmes critères ; parfois c'était une performance globale, d'autres fois un indice de succès de dépassement du seuil d'alerte qui étaient utilisés.

Kukkonen *et al.* (2003) ont effectué une comparaison entre cinq modèles neuronaux, un modèle linéaire statistique et un autre déterministe pour prévoir les concentrations de NO_2 et de PM_{10} enregistrées par deux stations situées au centre de Helsinki. Mise à part les concentrations de polluants mentionnées, les entrées utilisées ont été les données de trafic et celles météorologiques. L'horizon de prédiction choisi est de 24 heures. Pour éviter le sur-apprentissage, les auteurs ont utilisé une technique de régularisation Bayésienne. Les résultats obtenus ont mis en évidence que les RNA qui sont des modèles non linéaires sont légèrement plus performants que le modèle déterministe ou celui linéaire statistique. Les auteurs ont calculé plusieurs statistiques, dont on peut citer la valeur moyenne annuelle de l'indice de concordance qui a été de 0.91 pour les concentrations de particules.

Perez et Salini (2008) ont testé trois méthodes statistiques pour la prévision des concentrations de $\text{PM}_{2.5}$ pour le lendemain à Santiago (Chili) : PMC, RLM et *clustering*. Les données d'entrée sont les concentrations précédentes de $\text{PM}_{2.5}$ mesurées par quatre stations et les prévisions météorologiques ; les sorties sont les valeurs maximales obtenues par une moyenne mobile sur 24 heures à chaque station. La base de données s'étend sur quatre ans. L'erreur relative en valeur absolue se situe autour de 20 % pour toutes les méthodes. Bien que les trois méthodes testées peuvent être utilisées comme des outils opérationnels, c'est le *clustering* qui détecte le mieux les situations de fortes concentrations (indice de concordance de 0.9).

Slini *et al.* (2006) ont utilisé une base de données sur 7 ans constituée des données journalières de PM_{10} et des informations sur les variables météorologiques au centre de la ville, comme le maximum, la moyenne et le minimum journaliers de la température de l'air, la température du point de rosée et les valeurs moyenne et maximale de la vitesse du vent. Les auteurs ont utilisé les six premières années pour ajuster le modèle et la septième pour le tester. Les méthodes abordées : la RLM, l'analyse CART, l'ana-

lyse en composantes principales (ACP) et les RNA (le PMC) ont montré que celles qui restituaient le mieux les tendances étaient l'ACP et le PMC (erreur quadratique moyenne de $7.8 \mu\text{g}\cdot\text{m}^{-3}$), et la méthode CART pour l'indice de concordance (0.58). Les auteurs ont conclu que des variables supplémentaires (trafic, profil du vent, classe météo...) seraient nécessaires pour améliorer les performances de la prévision.

La méthode de régression linéaire multiple a été comparée à un modèle ARIMA pour la prévision des moyennes journalières de concentrations de particules dans deux métropoles : Delhi et Hong Kong avec un horizon de prévision journalier (Goyal *et al.*, 2006). Les auteurs ont constaté la supériorité du modèle ARIMA par rapport à la RLM ; en combinant les deux modèles, ils ont augmenté encore les performances ; dans un ordre décroissant, ils ont obtenu pour les trois modèles une erreur quadratique moyenne de $42 \mu\text{g}\cdot\text{m}^{-3}$, $20 \mu\text{g}\cdot\text{m}^{-3}$ et $18 \mu\text{g}\cdot\text{m}^{-3}$ à Hong Kong et des valeurs plus élevées à Delhi : $76 \mu\text{g}\cdot\text{m}^{-3}$, $63 \mu\text{g}\cdot\text{m}^{-3}$ et $59 \mu\text{g}\cdot\text{m}^{-3}$.

Poggi et Portier (2011) ont comparé plusieurs méthodes de prévision statistique des concentrations horaires de PM_{10} à Rouen pendant l'hiver ; quatre saisons d'hiver ont été utilisées pour développer le modèle, et une saison pour le tester, les données utilisées étant les prédictors météorologiques et la concentration moyenne de PM_{10} mesurés la veille. Les auteurs ont testé trois méthodes : la persistance, les modèles additifs non linéaires généralisés et la régression *clusterwise*, qui a donné de loin les meilleurs résultats. Pour la sélection des prédictors, les auteurs ont utilisé la méthode des forêts aléatoires (*random forest*) basée sur plusieurs arbres de décision. Cette méthode permet de calculer les scores individuels d'importance des régresseurs. La méthode appelée ici «régression *clusterwise*» est légèrement différente de celle connue dans la littérature sous cette appellation. La principale hypothèse est que les observations proviennent d'un mélange de composantes dans des proportions inconnues et, dans chaque composante, les observations sont modélisées par un modèle de régression linéaire. Le but est de déterminer les paramètres de chaque modèle linéaire et les paramètres qui définissent les composantes. Dans le contexte de *clustering*, chaque objet est supposé être généré par l'une des composantes du modèle de mélange ajusté. La partition et les modèles de mélange sont finalement obtenus par le principe du maximum de vraisemblance. Toutes les procédures utilisées sont disponibles comme bibliothèques dans le logiciel gratuit R.

5. Modèles hybrides ou avec prétraitement des données d'entrée

Depuis quelque temps, on remarque une tendance à utiliser des modèles hybrides ou à appliquer un prétraitement des données d'entrée. Ceci serait justifiable par le fait que les dynamiques des polluants ne correspondent pas à des phénomènes purement linéaires ou purement non linéaires, mais plutôt à un mélange des deux. On pourrait penser qu'un modèle non linéaire peut très bien être utilisé pour modéliser un phénomène linéaire, mais en fait il est moins facilement adaptable à la fois à la partie linéaire qu'à celle non linéaire. C'est ainsi que l'on a pu trouver des études avec des modèles hybrides.

La combinaison ARIMAX+PMC a été conçue pour que chacun des modèles prenne en compte une partie de la structure de la série temporelle. Dans une première étape, on développe un modèle ARIMAX qui prend en compte les variables météorologiques comme exogènes. En une deuxième étape, les auteurs appliquent le PMC sur les résidus du modèle précédent. Les résultats du modèle hybride sont meilleurs que ceux obtenus avec chaque modèle individuellement ou par RLM, pour la prévision des niveaux de particules.

Diaz-Roblès *et al.* (2008) ont développé un modèle hybride linéaire-non linéaire pour la prévision des niveaux de PM à Temuco (Chili). La combinaison ARIMAX-RNA a été conçue pour que chacun des modèles prenne en compte une partie de la structure de la série temporelle. Le modèle ARIMA prend en compte les variables météorologiques comme exogènes (ARIMAX). Une fois le modèle ARIMAX développé, les auteurs appliquent, dans un second temps, le PMC sur les résidus du modèle précédent. Les résultats du modèle hybride sont meilleurs que ceux obtenus avec chaque modèle individuellement, ainsi que par RLM. L'indice de succès obtenu pour la prévision de la valeur maximale sur les 24 heures suivantes a été de 87 %.

Voukantsis *et al.* (2011) ont d'abord comparé les typologies (*patterns*) de pollution de l'air dans les villes de Thessalonique et de Helsinki à l'aide d'une analyse en composantes principales (corrélations entre les variables et les composantes principales pour interpréter les composantes). Des similarités ont été trouvées entre les deux villes. Pour une sélection optimale des entrées, les auteurs ont utilisé une approche hybride, en combinant une régression linéaire avec un PMC. Bien que les villes soient assez différentes du point de vue qualité de l'air, les performances de l'approche hybride ont été très similaires et jugées comme satisfaisantes.

Hoi *et al.* (2009) ont proposé un modèle autorégressif avec entrées exogènes (TVAREX : *time-varying autoregressive model with exogenous input*) pour la prévision un jour à l'avance de la concentration journalière maximale de PM₁₀ à Macau. L'estimation de l'état a été réalisée en utilisant l'algorithme du filtre de Kalman. Ce modèle a été comparé à un PMC utilisant l'algorithme d'apprentissage basé sur l'optimisation de Levenberg-Marquardt, avec les mêmes données d'entrée. Comme performances globales, les deux modèles se sont avérés comparables, mais le modèle TVAREX a été plus performant pour reproduire les pics de pollution, grâce à sa nature adaptative inhérente.

L'objet de l'étude de Liu (2009) a été de simuler les concentrations journalières de PM₁₀ à Ta-Liao, Taiwan, à partir d'une base de données de quelques années, en utilisant le modèle Box-Jenkins ARIMA en version à plusieurs variables, combiné avec l'analyse factorielle et la décomposition en composantes principales. Les auteurs ont utilisé comme variables explicatives les facteurs météorologiques (direction et vitesse du vent, température du point de rosée) et les autres polluants (ozone O₃ et les oxydes d'azote, NO_x) pour examiner leur influence potentielle sur les variations journalières de PM₁₀. L'analyse factorielle a été appliquée afin de déterminer les contributions des plus importantes sources de PM₁₀, qui ont été prises en compte. Par l'intermédiaire des composantes principales, l'environnement a été pris en compte de manière globale, sur une région surveillée par quatre capteurs. Ceci a permis d'améliorer la prévision des valeurs extrêmes.

Chelani et Devotta (2006) affirment que les séries temporelles de concentrations de polluants ne sont ni purement linéaires, ni purement non linéaires et, par conséquent, appliquer uniquement une technique linéaire ou uniquement non linéaire peut conduire à des résultats non adéquats. Les auteurs ont développé un modèle hybride basé sur les modèles de type ARIMA pour la partie linéaire couplée avec un modèle dynamique non linéaire basé sur la théorie du chaos. Les résultats du modèle hybride sont meilleurs que ceux de chaque modèle appliqué individuellement. Les auteurs proposent cette approche hybride surtout lorsque les variables explicatives des concentrations de polluants ne sont pas disponibles.

6. Outils

Taylor *et al.* (2007) présentent un *toolbox* MatLab qui implémente les techniques dites DBM (*database mechanistics*) développées sur les vingt dernières années, et son application à des thématiques environnementales. Grossièrement, ces méthodes sont consacrées à l'identification de modèles, sous

forme d'état ou de fonction de transfert, à paramètres constants, variables dans le temps ou dépendant de l'état, l'accent étant mis sur la robustesse et la parcimonie de l'estimation, dans un but affiché d'interprétation physique des modèles. Cette boîte à outils constitue un ensemble versatile de méthodes, mais son intérêt pour les problèmes spécifiques de pollution atmosphérique n'est pas démontré.

Conclusion

Après ce passage en revue des différents modèles appliqués pour la prévision des niveaux de particules dans l'air, on peut conclure qu'il y a une large variété de modèles qui ont été utilisés, dont on peut rappeler : les régressions linéaires (simples ou multiples), modèles linéaires généralisés, analyses CART, régression «*clusterwise*» ; les modèles autorégressifs (AR, ARMA, ARIMA, ARMAX, ARIMAX, NARX...) ; les réseaux de neurones (en particulier le PMC) ; la dynamique du chaos ; les modèles probabilistes ; les filtres (Kalman). Dans ce papier, on n'a pas donné d'exemple d'utilisation de la logique floue, mais quelques applications dans ce domaine se présentent dans la littérature. L'utilisation des réseaux bayésiens n'a pas été rencontrée, mais elle pourrait être envisagée.

On ne peut pas conclure qu'il existe une classe de méthodes qui soit la plus adaptée, car chaque cas d'étude a ses particularités, l'information disponible est très variable en quantité et en qualité, et chaque méthode a ses avantages et ses inconvénients.

Étant donné que les séries temporelles de concentrations de polluants sont souvent autocorrélées, l'utilisation d'un modèle autorégressif semble justifiée. Il faudrait prendre en compte que dans la plupart des cas, la dynamique est non linéaire ; plus précisément,

il semble plus juste de considérer une partie linéaire et une partie non linéaire, et d'utiliser un modèle hybride. Un modèle linéaire AR peut être combiné avec un modèle non linéaire de type RNA ou un système dynamique issu de la théorie du chaos.

L'utilisation d'une analyse CART ou des réseaux bayésiens pourrait être intéressante du point de vue interprétation physique des résultats.

Enfin, la théorie du traitement du signal et de l'automatique pourrait s'avérer adéquate pour modéliser une série temporelle en faisant des ajustements fréquents. Les résultats dépendront aussi bien du modèle choisi, que de la pertinence des données d'entrée.

Le nombre important d'applications présentées montre l'intérêt de la communauté scientifique internationale pour les méthodes de prévision statistique basées sur la série temporelle de la concentration d'un polluant ; les principales applications envisageables sont dans le cadre de l'étude de la qualité de l'air intérieur, mais aussi comme outil de prévision à court terme pour faciliter la communication vers le public et envisager des mesures d'urgence au niveau des AASQA en France.

L'auteur tient à remercier le Centre Scientifique et Technique du Bâtiment (CSTB Champ-sur-Marne) pour le soutien financier d'un contrat de recherche sur lequel cette synthèse bibliographique est basée, ainsi que les différents collègues ayant participé à l'élaboration du rapport : Viorel Petreus, de l'Université Technique de Constructions de Bucarest, Olivier Ramalho (CSTB Champ-sur-Marne) et Yves Candau du CERTES.

Références bibliographiques

Anh V, Lunney K, Peiris S. Stochastic models for characterisation and prediction of time series with long-range dependence and fractality. *Environmental Modelling & Software* 1997 ; 12 (1) : 67-73.

Balestrassi PP, Popova E, Paiva AP *et al.* Design of experiments on neural network's training for nonlinear time series forecasting. *Neurocomputing* 2009 ; 72 : 1160-78.

Bouvier F. Le dispositif français de surveillance de la qualité de l'air. *Pollution atmosphérique. Climat, santé, société* 2012. N° spécial particules, novembre : 35-6.

Box GE, Jenkins GM, Reinsel GC. Time Series Analysis : Forecasting and Control, 4th Revised Edition. Wiley 2008.

Brockwell PJ, Davis RA. Introduction to Time Series and Forecasting. Springer 2002.

Brockwell PJ, Davis RA. Time Series: Theory and Methods, 2nd Edition. Springer 2009.

Brunelli U, Piazza V, Pignato L *et al.* Two-days ahead prediction of daily maximum concentrations of SO₂, O₃, PM₁₀, NO₂, CO in the urban area of Palermo, Italy. *Atmospheric Environment* 2007 ; 41 : 2967-95.

Chanel O. L'évaluation monétaire des effets sanitaires associés à la pollution atmosphérique. *Pollution atmosphérique. Climat, santé, société* 2012. N° spécial particules, novembre : 80-3.

Chauvet, P. Aide-mémoire de la géostatistique linéaire. *Cahiers de Géostatistique* 1999, fascicule 2, École Nationale Supérieure des Mines de Paris, Centre de Géostatistique, Fontainebleau.

Chelani AB, Devotta S. Air quality forecasting using a hybrid autoregressive and nonlinear model. *Atmospheric Environment* 2006 ; 40 : 1774-80.

Coll I. Connaissances sur les enjeux environnementaux et climatiques des particules. *Pollution atmosphérique. Climat, santé, société* 2012 ; n° spécial particules, novembre : 26-30.

Corani G. Air quality prediction in Milan : feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling* 2005 ; 185, 2-4, July : 513-29.

Diaz-Robles LA, Ortega JC, Fu JS *et al.* A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas : The case of Temuco, Chile. *Atmospheric Environment* 2008 ; 42 : 8331-40.

Dong M, Yang D, Kuang Y *et al.* PM_{2.5} concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications* 2009 ; 36 : 9046-55.

Dorling SR, Foxall RJ, Mandic DP, Cawley GC. Maximum likelihood cost functions for neural networks models of air quality data. *Atmospheric Environment* 2003 ; 37 : 3435-43.

El Raey M, Shalaby EA, Ghatass ZF, Marey HS. Time Series Analysis of Ambient Air Concentrations in Alexandria and Nile Delta Region, Egypt. *Proceedings of the 2nd Environmental Physics Conference*, 18-22 Feb. 2006, Alexandria, Egypt.

Foxall RJ, Cawley GC, Dorling SR, Mandic DP. Error functions for prediction of episodes of poor air quality. *Proceedings of the International Conference on Artificial Neural Networks (ICANN-2002)*, Springer Lecture Notes on Computer Science, 2415, Madrid, Spain, August 2002 : 1031-36.

Gardner MW, Dorling SR. Artificial neural network (the multilayer perceptron) – a review of applications in the atmospheric sciences. *Atmospheric Environment* 1998 ; 32 : 2627-36.

Gardner MW, Dorling SR. Statistical surface ozone models : an improved methodology to account for non-linear behaviour. *Atmospheric Environment* 2000 ; 34 : 21-34.

Goyal P, Chan AT, Jaiswal N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment* 2006 ; 40 : 2068-77.

Hies T, Treffeisen R, Sebald L, Reimer E. Spectral analysis of air pollutants. Part 1 : elemental carbon time series. *Atmospheric Environment* 2000 ; 34 : 3495-502.

Hoi KI, Yuen KV, Mok KM. Prediction of daily averaged PM₁₀ concentrations by statistical time-varying model. *Atmospheric Environment* 2009 ; 43 : 2579-81.

Hrust L, Bencetić Klaić Z, Križan J *et al.* Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment* 2009 ; 43 : 5588-96.

Issarayangyun T, Greaves S. Analysis of minute-by-minute exposure to fine particulates inside a car – A time-series modelling approach. *Transportation Research Part D* 2007 ; 12 : 347-57.

Kantz H, Schreiber T. Nonlinear Time Series Analysis, Cambridge University Press, 2003.

Karatzas KD, Kaltsatos S. Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. *Simulation Modelling Practice and Theory* 2007 ; 15 : 1310-9.

Kukkonen J, Partanen L, Karppinen A *et al.* Extensive evaluation of neural network models for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modeling system and measurements in central Helsinki. *Atmospheric Environment* 2003 ; 37 : 4539-50.

Liu PW. Simulation of the daily average PM₁₀ concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. *Atmospheric Environment* 2009 ; 43 : 2104-13.

Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables : a review of modelling issues and applications. *Environmental Modelling & Software* 2000 ; 15 : 101-24.

Marano F. Mécanismes d'action des particules atmosphériques fines et ultrafines. *Pollution atmosphérique. Climat, santé, société* 2012, n° spécial particules, novembre : 20-3.

Médina S. Évaluation de l'impact de la pollution urbaine dans 25 villes européennes. *Pollution atmosphérique. Climat, santé, société* 2012, n° spécial particules, novembre : 70-1.

Momas I, Bergametti G. Processus et enjeux sanitaires des particules. *Pollution atmosphérique. Climat, santé, société* 2012, n° spécial particules, novembre : 7.

Niska H, Hiltunen T, Karppinen A *et al.* Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence* 2004 ; 17 : 159-67.

Nunnari G, Nucifora AF, Randieri C. The application of neural techniques to the modelling of time-series of atmospheric pollution data. *Ecological Modelling* 1998 ; 111 : 187-205.

Pagowski M, Grell GA. Experiments with the assimilation of fine aerosols using an ensemble Kalman filter. *J. Geophys. Res.* 2012 ; 117 : D21302.

Pascal M, Ung A, Médina S, Declercq C. Évaluation de l'impact sanitaire de la pollution atmosphérique : méthodes et résultats. *Pollution atmosphérique. Climat, santé, société* 2012 ; n° spécial particules, novembre : 24-5.

Perez P, Salini G. PM_{2.5} forecasting in a large city : Comparison of three methods. *Atmospheric Environment* 2008 ; 42 : 8219-24.

Pisoni E, Farina M, Carnevale C, Piroddi L. Forecasting peak air pollution levels using NARX models. *Engineering Applications of Artificial Intelligence* 2009 ; 22 : 593-602.

Poggi JM, Portier B. PM₁₀ forecasting using clusterwise regression. *Atmospheric Environment* 2011 ; 45 : 7005-14.

Ramalho O, Lucas JP, Mandin C, *et al.* Niveaux de particules dans les environnements intérieurs en France. *Pollution atmosphérique. Climat, santé, société* 2012 ; n° spécial particules, novembre, 37-44.

Rouïl L. Les modèles régionaux et les modèles urbains : nouveaux outils d'aide à la décision. *Pollution atmosphérique. Climat, santé, société* 2012 ; n° spécial particules, novembre : 119-21.

Saporta G. Probabilités, analyses des données et statistiques. Editions Technip, 1990.

Siwek K, Osowski S. Improving the accuracy of prediction of PM₁₀ pollution by the wavelet transformation and an ensemble of neural predictors. *Engineering Applications of Artificial Intelligence* 2012 ; 25 (6), September : 1246-58.

Slini Th, Karatzas K, Moussiopoulos N. Statistical analysis of environmental data as the basis of forecasting : an air quality application. *The Science of the Total Environment* 2002 ; 288 : 227-37.

Slini T, Kaprara A, Karatzas K, Moussiopoulos N. PM₁₀ forecasting for Thessaloniki, Greece. *Environmental Modelling & Software* 2006 ; 21 : 559-65.

Taylor CJ, Pedregal DJ, Young PC, Tych W. Environmental time series analysis and forecasting with the Captain toolbox. *Environmental Modelling & Software* 2007 ; 22 : 797-814.

Tombette M., Mallet V., Sportisse B. PM₁₀ data assimilation over Europe with the optimal interpolation method. *Atmos. Chem. Phys.* 9, 2009: 57-70.

Viotti P., Liuti G., Di Genova P. Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. *Ecological Modelling* 148, 2002 : 27-46.

Villeneuve E., Aymoz G., Beekmann M., Baeza-Squiban A., Colosio J. La pollution de l'air par les particules : des connaissances scientifiques au service des acteurs. *Revue Pollution atmosphérique. Climat, santé, société. N° spécial particules*, Novembre 2012 : 10-16.

Vlachogianni D., Kassomenos P., Karppinen A., Karakitsios S., Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM₁₀ in Athens and Helsinki. *The Science of the Total Environment* 409, 2011 : 1559–1571.

Voukantsis D., Karatzas K., Kukkonen J., Räsänen T., Karppinen A., Kolehmainen M. Intercomparison of air quality data using principal component analysis, and forecasting of PM₁₀ and PM_{2.5} concentrations using artificial neural networks, in Thessaloniki and Helsinki. *The Science of the Total Environment* 409, 2011 : 1266–1276.

Wang X.K., Lu W.Z. Seasonal variation of air pollution index : Hong Kong case study. *Chemosphere* 63, 2006 : 1261–1272.

Wang Y., Sartelet K., Bocquet M., Chazette P. Assimilation of ground versus lidar observations for PM₁₀ forecasting. *Atmos. Chem. Phys.* 13: 2013, 269-283.

Zhang Y, Bocquet M., Mallet V., Seigneur C., Baklanove A. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment* 60: 2012, 656–676

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

Zainuddin Z., Pauline O. Modified wavelet neural network in function approximation and its application in prediction of time-series pollution data. *Applied Soft Computing* 11, 2011 : 4866–4874.

Zemouri R., Gouriveau R., Zerhouni N. Defining and applying prediction performance metrics on a recurrent NARX time series model. *Neurocomputing* 73, 2010 : 2506–2521.